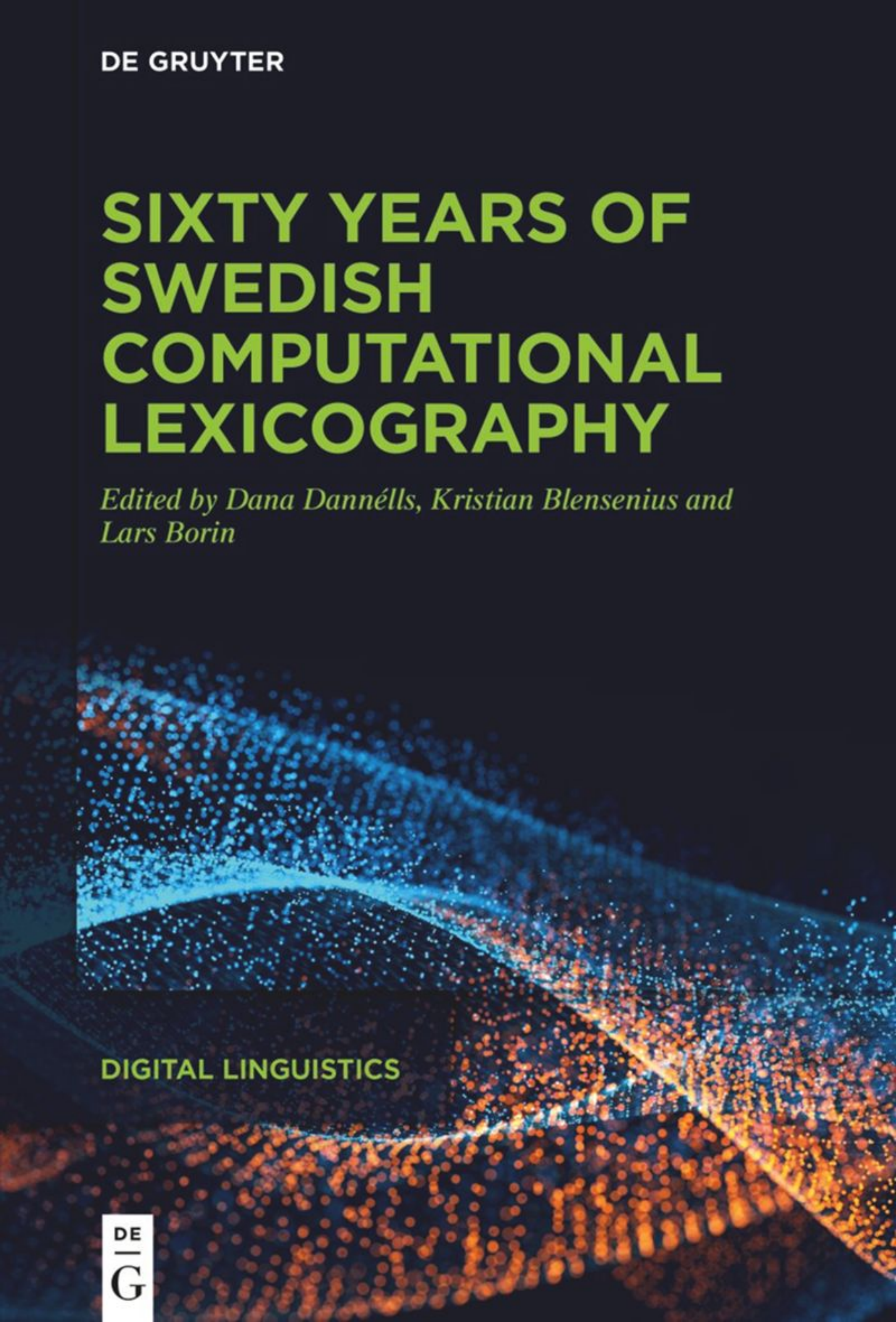
DE GRUYTER

SIXTY YEARS OF SWEDISH COMPUTATIONAL LEXICOGRAPHY

Edited by Dana Dannélls, Kristian Blensenius and Lars Borin



Sixty years of Swedish computational lexicography

Digital Linguistics

Edited by
Andreas Witt

Volume 3

Sixty years of Swedish computational lexicography

Edited by
Dana Dannélls, Kristian Blensenius, and Lars Borin

DE GRUYTER

The work on this volume as well as its open-access publication have been supported by funding from the Swedish Research Council for the national research infrastructure *Språkbanken* (grants 2017-00626 and 2023-00161) and by a publication grant from the Åke Wiberg Foundation (grant H24-0218).

ISBN 978-3-11-157713-5
e-ISBN (PDF) 978-3-11-157723-4
e-ISBN (EPUB) 978-3-11-157809-5
ISSN 2751-1278
DOI https://doi.org/10.1515/9783111577234

This work is licensed under the Creative Commons Attribution 4.0 International License. For details go to https://creativecommons.org/licenses/by/4.0/.

Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

Library of Congress Control Number: 2025936147

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

© 2025 with the author(s), editing © 2025 Dana Dannélls, Kristian Blensenius, and Lars Borin, published by Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin
This book is published with open access at www.degruyterbrill.com.
Cover image: piranka/E+/Getty Images
Printing and binding: CPI books GmbH, Leck

www.degruyterbrill.com
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com

Contents

Acknowledgments

Over the last six decades, numerous talented people have devoted many of their most productive years to the work whose results are described in this book, leading a development that turned Språkbanken and its host institution the University of Gothenburg to an outstanding center for research and development activities that combine lexicology, lexicography, and language technology. Without these individuals, this book could not have been written, which we as editors and authors of this volume hereby gratefully acknowledge.

We would also like to express our gratitude to the series editor Andreas Witt for not only considering this book for the *Digital Linguistics* series, but actively encouraging us to make an effort to complete it earlier rather than later. The process of going from brainchild to finished book was considerably helped along by our friendly and always helpful contacts at De Gruyter, first Svetoslava Antonova-Baumann and later her successor as acquisition editor, Hana Ikenaga. Finally, the nitty-gritty of practical manuscript preparation was made so much easier by the quick and insightful responses to all our questions by Teodor Borsa, the production manager for our camera-ready book manuscript, and our content editor at De Gruyter, Albina Töws. Thank you all!

The work on this volume was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*. The open-access publication fee has been covered in part by a publication grant from the Åke Wiberg Foundation (grant H24-0218). Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Dana Dannélls, Kristian Blensenius, and Lars Borin

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-203

Part I: **Introduction and background**

Lars Borin, Dana Dannélls, and Kristian Blensenius

1 Introduction

Abstract: This volume provides an account of the pioneering computational lexicographic research and development work conducted at the University of Gothenburg over the last 60 years. Starting from Sture Allén's efforts to reorient Swedish lexicography using digital corpus data in the 1960s – a novel notion at the time – the University of Gothenburg has developed into a leading center for computational lexicography in Sweden, synergistically combining lexically based computational linguistic text processing with computational methodology development for lexicography aiming at producing dictionaries for humans.

Keywords: computational lexicography, language technology, lexicography, lexicology, research infrastructure

1 Introduction: Swedish computational lexicography

This volume bears the at least four ways ambiguous title *Sixty years of Swedish computational lexicography*.

Computational lexicography has been used in the literature to refer to two fairly different kinds of activity. It can refer to work in computational linguistics aiming to build automatic systems for various kinds of analysis of words in text (corpora), for instance lemmatization and morphological analysis. This kind of work goes back

Acknowledgments: The work on this volume was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*. Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Dana Dannélls, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: dana.dannells@svenska.gu.se
Kristian Blensenius, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: kristian.blensenius@gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-001

almost to the beginning of the discipline, with much groundwork laid already in the 1950s, including early machine learning attempts; for a survey of the latter, see Hammarström & Borin (2011). It was very actively pursued in the 1980s, following the introduction of finite-state morphology to the field; see Sproat (1992: Ch. 3) for an overview.

It can also mean the use of computers as an aid in compilation of dictionaries, which is a more recent development. The dictionaries are not, of course: they have been around for about as long as we as a species have had writing. The history of dictionaries starts with monolingual (Sumerian) and bilingual (Akkadian and Sumerian) lexical lists on Mesopotamian cuneiform clay tablets from the fourth and third millennia BCE (Civil 1990). Dictionaries are known from most ancient and modern civilizations (Durkin 2015).

Compared to this, the computer is a latecomer, but one which has made a real difference to dictionary-making. The computer is an artificial symbol manipulator, extremely well suited to handle human language, the symbolic system par excellence, as amply evidenced these days by the large language models behind the chatbots that rapidly have made *artificial intelligence* (or AI) a household word. The computer has radically changed the ways in which dictionaries are compiled and used.

The other source of ambiguity in the title of this volume is the word *Swedish*, which could be paraphrased as 'in Sweden' or 'for Swedish'.

In fact, the story to be told in this book contains something of all four possible meanings of *Swedish computational lexicography*, much like the famous rabbit–duck illusion lets you see two animals at the same time (Figure 1).

Computational linguistics was born not long after the first modern computers were built in the 1940s, fueled by the dream of fully automatic translation between languages, leading among other things to the developments outlined above, where computational linguists endeavored to design and program (rule-driven) systems for automatic lexical and syntactic analysis of unrestricted text, building computational lexical resources and grammars to this end. This intersected – at least in some places, and notably in Gothenburg in Sweden – with another development, viz. in lexicographical practice. With the advent of electronic corpora in the 1960s, some pioneers realized how access to these text masses could serve to "turbo-charge" the empirical aspect of lexicography by putting unprecedented volumes of authentic language data at the disposal of dictionary compilers. The exponential growth of computing power that we have seen since Gordon Moore formulated his eponymous law, also sixty years ago, in 1965 (Moore 2006), ensured that lexicographers also could draw upon a wide variety of increasingly powerful computational tools in aid of their work. One such pioneer was Sture Allén, in the 1960s a PhD candidate in the Department of Scandinavian Languages at the University of Gothenburg, and eventually a professor of computational linguistics at the same university. His vision

Figure 1: The rabbit–duck illusion (image in public domain; source: https://en.wikipedia.org/wiki/File:Kaninchen_und_Ente.svg)

of how corpora and computers could modernize Swedish lexicography ultimately has led to the multifaceted combination of computational linguistic and lexicographical research and development activities that we describe in this book.

These activities address mostly Swedish (the language), but it is also true that the Språkdata/Språkbanken group at the University of Gothenburg has been quite unique in Sweden in its methodical and consistent pursuit of research and development in computational lexicography in both senses of this term.

This book is first and foremost an exercise in *historiography*. The initial idea for putting together this volume came from the realization that the research unit at the University of Gothenburg where all the authors in this volume are active has been pursuing its unique and pioneering combination of computationally informed lexicography and lexicographically informed computational linguistics for close to 60 years, and that the research e-infrastructure Språkbanken Text which was established in support of this research will celebrate its 50th anniversary in 2025, but that this development has been only fragmentarily documented, almost exclusively in Swedish. Hence, we felt that this history deserved to be written down for an international audience.

2 This volume

This volume is organized into five thematic parts. This and the following chapter (Chapter 2) make up the first, introductory part. Chapter 2 provides a brief account of the history of computational lexicography at the University of Gothenburg over the

last sixty years, as a background to the following 13 chapters, which address various aspects of this history and also provide some glimpses of the present-day state of computational lexicography in Gothenburg.

Part II: Dictionaries for humans

In the second part of the volume, we give an account of two of the products of the lexicographical projects at the University of Gothenburg: the two main dictionaries of contemporary Swedish, (*Svenska Akademiens ordlista*, 'The Swedish Academy Glossary'; SAOL in short) and *Svensk ordbok utgiven av Svenska Akademien* ('The Contemporary Dictionary of the Swedish Academy'; SO in short). Each dictionary is dedicated its own chapter.

SAOL (Chapter 3) is a general-language, monolingual dictionary of Swedish that has served as a reference for spelling and inflection since its first edition in 1874 (SAOL 1 1874). It is produced at the University of Gothenburg since the 1980s, and it is incorporated into the Språkbanken Text research unit since 2021. In its 15th edition, planned for 2026, SAOL has evolved from a printed book to also become a widely accessible digital resource available through the dictionary portal Svenska.se alongside SO, the contemporary definition dictionary, and the historical Swedish Academy Dictionary SAOB.

SO (Chapter 4) is a corpus-based monolingual dictionary constituting the most complete description of modern Swedish vocabulary of today. It provides detailed semantic descriptions, pronunciation via audio files, language examples, idioms, historical information, and more. SO has evolved through a number of editions since it was originally developed from the Lexical Database project at the University of Gothenburg in the 1970s, with the latest edition (SO 2021) published digitally in 2021.

Part III: Lexical resources for machines

The notion of "case" (or semantic roles) as a fundamental linguistic concept for understanding sentence structure and the relationships between words (Fillmore 1968) has been a source of inspiration for the computational lexicographic work at the University of Gothenburg over the years (see, e.g., Toporowska Gronostaj 1996). This work resulted in the development of several lexical-semantic databases (Järborg 1996) primarily intended for dictionary production and lexicological investigations, such as the Semantic Database (SDB) – a lexical database with formalized semantic roles and valency frames.

The trend of emphasizing lexical semantics was continued in a number of lexical resources for natural language processing developed by Språkbanken starting in the early years of this millennium. This part of the volume provides an overview of these computational lexical resources, their uses in automatic language processing, and their relationship to the dictionaries for humans. It consists of four contributions, which are oriented towards large-scale, structured collections of lexical data that can be used to improve automatic text processing.

Chapter 5 describes the history of SweFN++, the rich lexical macroresource of Språkbanken Text, from the early stages of its creation. As the name indicates with the plus-plus suffix, it is an enhanced resource that gathers a variety of freely available modern and historical lexicons, including both Swedish and multilingual lexicons, for the purpose of enabling automatic text analysis.

Chapter 6 describes the origin of Saldo, the "pivot" lexical-semantic resource of what started off as SweFN++, and that has been central to the development of what is today Språkbanken's Lexical Research Infrastructure. The chapter describes the internal organization of the resource, illustrated with sample entries.

In Chapter 7 an updated version of the Swedish FrameNet (SweFN) is presented. The chapter looks at the harmonization process of the resource with respect to Global FrameNet, an initiative aiming to merge framenets from different languages into a unified multilingual lexical-semantic network (Torrent et al. 2020).

This part of the book is concluded by Chapter 8, an account of a number of the semantic lexical resources currently accessible through Språkbanken Text, including a Roget-style thesaurus and wordnet-like resources, each of which has been developed in a different framework and instantiates a different theory. They have, under the umbrella of SweFN++, been partly interlinked with Saldo.

Part IV: A computational infrastructure for dictionary making and lexical research

From a computational linguistic perspective, the idea of harmonizing multiple hierarchical structures to convey syntactic and semantic depth, while simultaneously enriching this computational system with corpus examples, has been a source of inspiration for building an infrastructure for Swedish language technology.

The chapters in the *infrastructure* part of the volume describe the central parts of a language technology based research infrastructure for computational lexicography.

Språkbanken Text at the University of Gothenburg is a division of the national digital research infrastructure Språkbanken. It has been conducting research in natural language processing and served as a research infrastructure for preparing and publishing Swedish language resources for 50 years. In Chapter 9 the history

of Språkbanken Text is briefly presented, and an overview is given of its resource collections and publicly available research platforms. Two of these are described in the following chapters.

Chapter 10 presents Korp, Språkbanken's word research platform, its origins and the various methodological and other decisions that have informed its design, with an emphasis on functionality that is important in the context of lexicology and dictionary compilation.

The development of Karp, Språkbanken's data editing platform, is introduced in Chapter 11. It takes the lexicographer's point of view in its presentation of the features of the editing interface, highlighting the importance of close collaboration between the lexicon's editorial team and the interface developers.

Part V: Case studies

For the *case studies* part of the volume, we have simply asked the authors – researchers in or close to Språkbanken Text – to write about some aspect of their ongoing research with relevance for the topic of this volume.

The first case study offered in Chapter 12 discusses ongoing efforts to create a new digital version of A.F. Dalin's Dictionary of the Swedish language (*Ordbok öfver svenska språket*, 1850–1853). The authors take both a corpus based approach and the lexicographer perspective to explore the potential for analyzing the dictionary. Additionally, it highlights challenges that arise when moving from printed to digital version.

The second case study presented in Chapter 13 investigates lexical change with diachronic lexical resources and corpora. The authors specifically explore the opportunities for diachronic lexical research provided by the inclusion of SAOLhist Plus, a dataset that combines digitized versions of successive editions of the SAOL dictionary (1874–2015).

The third case study concerns the Swedish ConstructiCon (SweCcn; see Chapter 14), a digital repository of Swedish construction descriptions incorporated into Språkbanken's Karp infrastructure. SweCcn is a practical implementation of construction grammar (e.g., Fillmore 1988), containing over 400 construction entries. The chapter discusses the development of a network model for SweCcn, distinguishing relations in three domains: relations between constructions, relations within constructions, and external relations. The relations between and within constructions form a multidimensional construction network, both organizing the constructiCon and providing the user with versatile search options. The external relations consist of cross-linguistic connections via MoCCA and FrameNet, as well as lexical links to other resources within Språkbanken Text.

Last, but not least is the case study, presented in Chapter 15, on the lexicographic benefits of integrating the contemporary dictionary SO into Språkbanken's text research platform Strix, highlighting how document vectors in Strix can serve as a methodological tool to support the development of a dictionary.

References

Civil, Miguel. 1990. Sumerian and Akkadian lexicography. In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand & Ladislav Zgusta (eds.), *Dictionaries: An international encyclopedia of lexicography. Second volume*, 1682–1686. Berlin: De Gruyter.

Durkin, Philip. 2015. A chronology of major events in the history of lexicography. In Philip Durkin (ed.), *The Oxford handbook of lexicography*, 605–615. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199691630.002.0008.

Fillmore, Charles J. 1968. The case for case. In Emmon Bach & Robert Harms (eds.), *Universals in linguistic theory*, 1–88. New York: Holt, Rinehart, & Winston.

Fillmore, Charles J. 1988. The mechanisms of "construction grammar". *Annual Meeting of the Berkeley Linguistics Society 1988*. 35–55.

Hammarström, Harald & Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2): 309–350.

Järborg, Jerker. 1996. *Formaliserad lexikologi: Rapport från ett långtidsprojekt (Preliminär version)* [Formalized lexicology: Report from a long-term project (Preliminary version)]. (Research Reports from the Department of Swedish No. GU-ISS-96-3) Gothenburg: Department of Swedish, University of Gothenburg.

Moore, Gordon. 2006. Moore's law at 40. In David C. Brock (ed.), *Understanding Moore's law: Four decades of innovation*, 67–84. Philadelphia: Chemical Heritage Foundation.

SAOL 1. 1874. *Ordlista öfver svenska språket utgifven af Svenska Akademien* [Glossary of the Swedish language published by the Swedish Academy]. 1st edn. Stockholm: P.A. Norstedt & söner.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

Sproat, Richard (ed.). 1992. *Morphology and computation*. Cambridge: MIT Press.

Toporowska Gronostaj, Maria. 1996. *Integrated valensbeskrivning: mot ett formaliserat verbvalenslexikon* [Integrated valence description: Towards a formalized verb valence lexicon]. Gothenburg: University of Gothenburg. (PhD thesis).

Torrent, Tiago T., Collin F. Baker, Oliver Czulo, Kyoko Ohara & Miriam R. L. Petruck (eds.). 2020. *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*. Marseille: European Language Resources Association.

Lars Borin and Louise Holmer

2 Background: a brief history of computational lexicography in Gothenburg

Abstract: Swedish computational lexicography and computer-supported lexicology have a long history at the University of Gothenburg. Starting sixty years ago, the Språkdata research group pioneered corpus-supported lexicography for Swedish, forming the basis for successive editions of two main dictionaries of contemporary Swedish, SAOL and SO. Language technological lexical resources for Swedish have been developed by the research unit/research infrastructure Språkbanken Text since the turn of the millennium, most recently within the framework of the *Swedish FrameNet++* initiative. After close to two decades of separation, these two largely mutually independently developed strands of computational lexicography have recently joined forces under the umbrella of *Språkbanken's Lexical Research Infrastructure* to advance the field technologically, methodologically, and scientifically.

Keywords: computational lexicography, corpus linguistics, dictionary, language technology, lexical resource

1 Introduction

Today, the two universities in Gothenburg – University of Gothenburg and Chalmers University of Technology – are the home of one of the strongest and most well-known academic language technology research environments in Sweden, with sev-

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*. Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Louise Holmer, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: louise.holmer@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-002

eral groups pursuing cutting-edge research on various aspects of computers and language.

The long history that has produced the present state of affairs started because of a scientific interest in lexicographical method development on the part of Sture Allén, who initiated this development in the 1960s while still pursuing his PhD studies in Scandinavian languages at the University of Gothenburg. Allén realized that the fledgling computer technology, together with increasing availability of born-digital text due to advances in printing technology, could provide invaluable support both in lexically oriented linguistic studies and in dictionary production. As a result of this, in addition to being leading in Swedish language technology, the University of Gothenburg also boasts a top-notch lexicographical research group responsible for the compilation of the two main reference dictionaries of modern-day Swedish, SAOL and SO (see below in this chapter and Chapters 3 and 4 in this volume). This lexicographical research and dictionary production form part of the activities of *Språkbanken Text*, a language technological research and development (R&D) unit and research infrastructure continuing the computational lexicographical tradition established sixty years ago at the University of Gothenburg by Sture Allén.

The chronology of nomenclature used for the organizational entity that is today represented by Språkbanken Text is not completely straightforward. The timeline of the R&D activities that today fall under the aegis of Språkbanken Text looks approximately as in Figure 1, in terms of its shifting organizational structure, which will be explained in more detail in the following sections of this chapter.

For convenience, we will refer to this entity as "Språkdata" when we talk about the period from 1965 to about 2003,[1] and as "Språkbanken" or "Center for Lexicology and Lexicography", as appropriate, when discussing the period 2003–2021, and finally as "Språkbanken (Text)" when describing the present-day situation, since 2021. Note that the last-mentioned (Språkbanken Text) is the topic of a separate chapter (Chapter 9 in this volume), while the present chapter will mainly describe the period up until about 2021.[2] In the remainder of this chapter we provide a brief historical overview where we trace this development from the 1960s onwards, in order to provide a background to the later, more detailed chapters covering various aspects of this development and its present-day manifestations.

1 Thus Språkdata lives as a name both before and after its formal existence as an administrative unit inside the University of Gothenburg, a usage that agrees well with actual practice.

2 This chapter contains a considerably reworked and extended version of a conference paper presented at the first Swedish Huminfra conference, in January 2024 (Borin & Holmer 2024).

1965 Sture Allén defends his PhD thesis in Scandinavian languages.
1965 The compilation of the first major Swedish text corpus, *Press 65*, is initiated, to support vocabulary research and dictionary writing.
1972 Allén becomes the first professor of computational linguistics in Sweden and the *Språkdata* computational linguistics research unit is established as a separate administrative entity in the Department of Scandinavian Languages.
1975 *Språkbanken* is established with government funding as a "service organ" for text corpus and lexical data (initially named the *Logotheque*).
1977 The Språkdata research unit becomes a separate department at the University of Gothenburg.
1991 Språkdata is merged with the Department of Scandinavian Languages; the name of the new department becomes *Department of Swedish*.
2003 The *Center for Lexicology and Lexicography* (CLL; *Lexikaliska institutet* in Swedish) is established, in official recognition of the fact that the language technological and lexicographic activities had gone their separate ways already some time ago. The language technological activities are continued separately from the CLL under the Språkbanken name.
2014 Språkbanken receives funding for the period 2014–2018 for establishing and coordinating the Swedish node of the European research infrastructure CLARIN ERIC.
2018 The national research infrastructure Språkbanken receives funding from the Swedish Research Council, initially for the period 2018–2024, and subsequently to the end of 2028. In connection with this *Språkbanken Text* is adopted as the official name of the Gothenburg division.
2021 The Center for Lexicology and Lexicography is merged into Språkbanken Text, and is discontinued as an organizational unit.

Figure 1: A brief history of computational lexicography in Gothenburg over the last 60 years. See the text of this chapter for a more detailed account

2 The Språkdata era: 1965–2003

The topic of the present volume – the combination of computers and lexicography – has a long and distinguished history at the University of Gothenburg. Six decades ago, in 1965, not long after the compilation of the well-known *Brown Corpus* of American English (Francis & Kučera 1967), Sture Allén initiated the collection of digital texts for the first major Swedish text corpus – the one-million word *Press 65* – in order to address lexicographical research questions and aims such as "In a broad sense, what are the lexical units of Swedish as represented by a large corpus? How common are they, and how are they distributed over different text types?" (Allén 2014: 61, our translation).

Not long after defending his PhD thesis (Allén 1965a,b), Allén founded a research group that soon became a departmental division and eventually a separate depart-

ment of computational linguistics, commonly referred to as *Språkdata*.[3] Språkdata pursued corpus-supported Lexicology and Lexicography for many years, and also activities aimed at developing computational linguistics as a discipline in Gothenburg, in Sweden, and in the Nordic countries.

An important historical milestone was the establishment in 1975 of *Språkbanken* ('the Language Bank') as a dedicated research infrastructure – referred to at the time as a "service organ" (Allén [1980] 1999: 303–304) – operated by Språkdata in support of Swedish linguistic research in general and the local lexicographical activities in particular,[4] with a remit to be "nationally responsible for the collecting, storing, processing, and providing of linguistic material in machine-readable form" (Allén [1980] 1999: 304).

In 1977 Språkdata was turned into an independent department – the Department of Computational Linguistics – a status that it kept until 1991, when it was merged with its former mother institution, the Department of Scandinavian Languages, into the Department of Swedish. During this time, computational linguistics underwent a period of strong expansion at the University of Gothenburg, a development very much promoted by the newly formed department, in collaboration with the Department of Linguistics. The two departments jointly launched a four-year undergraduate study program in computational linguistics with the first intake of students in 1984.

Already the same year as the new department was formed, Språkdata arranged the first meeting in what was to become the series of NoDaLiDa conferences, now the Nordic-Baltic computational linguistics conference, held biennially in a Nordic or Baltic country on odd-numbered years.[5]

The lexicographical projects in Gothenburg, aiming at publishing dictionaries, produced two large Swedish print dictionaries in the 1980s, namely the first edition of the Swedish monolingual *Svensk ordbok* ('Swedish dictionary'; SOB 1986; in short SOB) (Malmgren & Sköldberg 2013) and the 11th edition of *The Swedish Academy Glossary* (SAOL 11 1986; henceforth SAOL). The two lexical databases underlying the dictionaries[6] (SAOL and SOB, as well as their revised, later editions, respectively:

3 The name *Språkdata* was formed by telescoping word formation from the name of the original departmental division, *Avdelningen för språklig databehandling*, 'The Division of Computational Linguistics' (literally 'The Division of Linguistic Data Processing').

4 When established with government funding in 1975, this infrastructure was initially named *Logoteket* 'the Logotheque', but within a few years the name Språkbanken caught on.

5 In 2017, NoDaLiDa returned to Gothenburg for the first time since the conference series was initiated in order to mark its 40th anniversary.

6 In the terminology used in this chapter, a *dictionary* and its underlying *lexical database* are intended for human consumption, while a *lexical resource* is intended for use in NLP systems. We

SAOL 12 1998; SAOL 13 2006; SAOL 14 2015; and SO 2009; 2021, referred to as SO) have so far been treated as separate entities during the years of development. They have been revised and refined, one dictionary at a time, often by more or less the same lexicographers. SAOL and SO are financed by the Swedish Academy (Sköldberg et al. 2019), and the editorial staff is employed by the University of Gothenburg. SAOL and SO are ongoing projects, with new editions appearing every few years, and there are separate chapters devoted to both dictionaries, Chapters 3 (SAOL) and 4 (SO) in this volume. The work on SAOL is also featured in Chapter 10 and that on SO in Chapter 15 in this volume.

The lexicographical projects continued through the last decade of the 20th century, resulting in the successor of SOB (1986), *Nationalencyklopedins ordbok* ('The Dictionary of the National Encyclopedia'; NEO 1995) and the Swedish extended core vocabulary making up the source language part of the series of LEXIN (*Lexikon för invandrare* 'Lexicon for immigrants') dictionaries (Gellerstam 1999).

Språkdata's researchers also devoted their attention to lexicology and studies of lexical change. In the late 1980s and the 1990s, the history of Swedish lexicography before the 20th century was investigated in an ambitious research project (Hannesdóttir & Ralph 1988a,b; Ralph 2001), followed at the beginning of the 21st century by an extensive project investigating the development of the Swedish vocabulary since 1800 (Malmgren 2000).

Already in the 1970s, ambitious plans had been drawn up for a large and multifaceted lexical database – the *Gothenburg Lexical Database* (GLDB) – for dictionary production (see Ralph, Järborg & Allén 1977 and Chapter 5 in this volume). Although what actually materialized was a series of databases, one for each dictionary (see Chapter 5 in this volume), it far surpassed anything achieved internationally in computational linguistics at that time. The Språkdata work on GLDB started considerably earlier – by about a decade (Ralph 1977) – than the awakening of a more general interest in matters lexical in computational linguistics, as evidenced in i.a. work by Boguraev & Briscoe (1989) and Wilks, Slator & Guthrie (1996). We would do well to recall what the state of the art was like at the time. For example, in a workshop on "Linguistic Theory and Computer Applications" held in 1985 at the University of Manchester (Whitelock et al. 1987), one of the discussants (Stuart Shieber) is cited as noting that the average number of lexical entries in the systems presented at the workshop is around 1,500, and if you do not count the outlier, the machine translation system Rosetta, this is reduced to about 25 entries (Ritchie 1987: 234).

---

further use the term *lexicon* to include all the three mentioned, i.e., dictionaries, lexical databases, and lexical resources.

Even if the primary and most explicitly elaborated purpose of the GLDB was to compile a modern Swedish reference dictionary based on empirical language data in the form of text corpora, the relevance of this work to language technology was clearly recognized and acknowledged from the very start of the GLDB endeavor (Ralph, Järborg & Allén 1977; Ralph 1979). The version of GLDB underlying NEO (1995) was used as the point of departure for a lexicological project on formalized lexical semantics, the *Semantic Database* (SDB; e.g., Järborg 1999; 2001; 2003), that later served as the basis for the creation of some lexical-semantic resources for NLP (see Chapter 8 in this volume).

Språkdata also launched an initiative to build a lexical databased referred to as the *Swedish Morphological Database*, basing it on the inflectional information in the 12th edition of SAOL (see Chapter 3 in this volume). This database came to good use in Språkdata's lexicographical work, and the plan was to also employ it for morphological annotation of Språkbanken's corpora (Gellerstam, Cederholm & Rasmark 2000), but this plan was ultimately not realized, because of later developments described below.

3 The fork in the road: 2003–2021

In the early years of this millennium, the lexicographical and language technological strands parted ways. In late 2002, a new director was appointed to lead Språkbanken, namely one of the authors of this chapter (Borin), whose background was in general computational linguistics rather than in Swedish lexicography, and as a consequence the thrust of Språkbanken's activities started to shift from traditional corpus linguistics with a lexicographical and lexicological focus in the direction of mainstream language technology. The two strands developed largely separately over the following two decades, being somewhat non-communicating vessels with regard to researchers as well as databases and research output.

3.1 The Center for Lexicology and Lexicography: 2003–2021

The Center for Lexicology and Lexicography (CLL) was officially established in 2003, and for the following two decades, it pursued corpus-based lexicography, with fairly little interaction with the language technological research going on in Språkbanken.

The Center was also closely connected to one of the Department's five profile areas (the others being grammar, text and interaction, multilingualism, and language technology). The production of dictionaries as well as research on lexicography

and lexicology have been main activities for the CLL. These standard reference dictionaries of contemporary Swedish constitute the most important results during the period mentioned above:

- *Svenska Akademiens ordlista* 'The Swedish Academy Glossary', SAOL (13th ed. 2006), the standard reference on spelling and inflection of Swedish words. SAOL 13 was also further developed into electronic versions like SAOLPlus (2007), a CD featuring advanced search options, and later apps for iOS and Android. See Chapter 3 in this volume.
- *Svensk ordbok utgiven av Svenska Akademien* 'The Contemporary Dictionary of the Swedish Academy' (SO, 2009). The print dictionary of 2009 was later developed into e-versions and adapted for the web and for apps. See Chapter 4 in this volume.
- Lexin, 4th ed. 2011, a digital-only dictionary. The Lexin dictionaries are based on a Swedish monolingual dictionary translated into the most spoken immigrant languages. The material of Lexin has since been transferred to the Institute for Language and Folklore (Gellerstam 1999; Lexin 2025).
- SAOLhist, a web resource dedicated to the previous editions of SAOL, where the user can search for headwords in the different editions and also compare the headwords between editions, etc. See Chapter 13 in this volume.
- SAOL, 14th ed. (2015), print and e-versions. SAOL 14 was developed with the print dictionary in mind, and e-versions were developed after the release of the book.
- SO 2021, digital-only. This new edition of SO is the first of the dictionaries financed by the Swedish Academy to be a digital-only product. SO 2021 has also been released in different e-versions.

In addition to the initiatives listed above, the CLL was also involved in an Icelandic dictionary project, ISLEX. The ISLEX project was carried out in a cooperation between Denmark, the Faroe Islands, Finland, Iceland, Norway, and Sweden (ISLEX 2025).

The CLL was also involved in the testing and evaluation of the web resource Svenska.se, launched by the Swedish Academy in 2017. Svenska.se includes SAOL, SO, and the historical dictionary *Svenska Akademiens ordbok* (SAOB), presented together.

After a thorough external evaluation of the department's different research groups and their relations, it was evident that the CLL was too small and too dependent on language technology to continue its existence on its own, separately from Språkbanken. In 2021, the CLL was formally merged with Språkbanken Text, and at the time of writing this, in early 2025, the Swedish Academy contemporary dictionaries project involves all in all ten lexicographers and language technologists.

3.2 The language-technological turn in Språkbanken

As already mentioned above, during the Språkdata period, Språkbanken served as a repository of text corpora and lexical data: a "service organ" (Allén [1980] 1999), made up by a combination of a "text bank" and a "word bank" (Gellerstam & Sjögreen 1994). Plans had been made for adding automatic linguistic annotations to the corpora, for instance by using the Swedish Morphological Database, but that aspect of Språkdata's activities had progressed very slowly compared to the advances made in lexicography and lexicology. At the beginning of the 2000s Språkbanken could offer access to only one automatically annotated corpus, the part-of-speech tagged Parole corpus (about 20 million words of mixed-genre texts).[7] Other corpora in Språkbanken allowed only for various forms of text-word and string searches.

Beginning in 2003, Språkbanken went through a slow process of re-orientation of its main activities, with the goal to develop computational tools for automatic linguistic annotation of the considerable amounts of text collected in Språkbanken's corpora, thereby turning the corpora into resources that could be used for language technology R&D. Following general practice in the field, the software and language resources used should be open and freely available, ideally for all purposes, not least to ensure reproducibility of research.

Keeping to the topic of the present volume, we focus our attention here on annotation of lexical and inflectional properties of text words. At the time, there was no freely available full-sized Swedish digital lexical resource that could be used as the basis of morphological analysis and lemmatization of our corpora. Nor could the content of the Swedish Academy dictionary databases developed in-house be made openly available because of commercial commitments.

Instead, Språkbanken initiated a parallel computational lexicographic project, that took its point of departure in SAL (*Svenskt associationslexikon* 'Swedish Associative Thesaurus'), a semantic Swedish dictionary developed by Lennart Lönngren at Uppsala University between 1987 and 1992 (Lönngren 1988; 1998). SAL could be turned into a full-sized – containing slightly over 72,000 entries in its first release – semantic dictionary of Swedish with complete morphological specifications (inflectional paradigms plus compounding forms) provided for all entries (Borin, Forsberg & Lönngren 2008), released under an open (CC-BY) license allowing all kinds of use, including for commercial purposes (Borin, Forsberg & Lönngren 2013; Borin et al. 2021). The latest official release (Saldo 3.3, from 2025) holds about 148,000 entries, i.e., Saldo is approximately comparable in size to SAOL (with upwards of 126,000 headwords in the latest edition, SAOL 14). See Chapter 6 in this volume.

7 https://spraakbanken.gu.se/en/resources/parole (last accessed: April 4, 2025)

Even if there was no free Swedish computational lexicon available before Saldo that was both large and general enough for the intended purposes, the long history of lexicographical activity in Språkdata and Språkbanken had left behind a number of smaller and more specialized computational lexicons, resulting from various projects carried out through the years, to which could be added initiatives started elsewhere, such as the (partial) *Swedish WordNet* compiled at Lund University (Viberg et al. 2002), or the crowdsourced *People's Synonym Lexicon* created and maintained at the Royal Institute of Technology in Stockholm (Kann & Rosell 2006). The *Swedish FrameNet++* (SweFN++) project was initiated around 2008, with two complementary and interlocking aims. One aim was to combine the rich, painstakingly compiled linguistic information hidden in these formally as well as content-wise quite heterogeneous resources into one unified lexical macroresource, SweFN++. The other aim was to create a computational infrastructure facilitating development of the resources themselves as well as research based on their content (Borin et al. 2021; Dannélls, Borin & Heppin 2021). See further Chapters 5, 7, 8, and 11 in this volume.

In 2014, Språkbanken was awarded substantial funding from the Swedish Research Council (SRC) together with 8 other partners nationwide (universities and cultural heritage institutions) to set up and coordinate a Swedish node of the European research infrastructure CLARIN ERIC, *Swe-CLARIN*.[8] Språkbanken's lexical resources played a key role in the project proposal, and part of this funding could be used to increase activity in the SweFN++ project. Further research infrastructure funding was forthcoming in 2018, in the form of a seven-year SRC grant for a national research infrastructure initiative led by Språkbanken under the name *Språkbanken & Swe-CLARIN* (2018–2024). Again, work on lexical resources under the SweFN++ umbrella figured prominently in the infrastructure activities.

4 A new synthesis: 2021–

The two strands of research and development described above were again brought together into one unit in 2021, with an expressed synergistic aim. The lexicographical projects formerly organized under the Center for Lexicology and Lexicography were officially made a part of Språkbanken Text, and merged with the SweFN++ activities under a new umbrella designation: *Språkbanken's Lexical Research Infrastructure*. In a way, this move signaled a return to the pre-2003 organization, but at a considerably higher level of technical and methodological maturity, the former originating pri-

8 https://www.clarin.eu/ (last accessed: April 4, 2025); https://sprakbanken-clarin.lingfil.uu.se/index_en.html (last accessed: April 4, 2025)

marily in Språkbanken Text and the latter contributed in equal and complementary parts by the two strands of lexicographical R&D that have now joined forces.

Furthermore, the underlying databases of SAOL and SO have been migrated to the Karp data editing platform, that has been under active development for over a decade as a tool for working with formally structured language data (Borin, Forsberg & Roxendal 2012), notably the lexical resources that make up SweFN++, in particular the Swedish FrameNet (Dannélls et al. 2021) and the Swedish ConstructiCon (Lyngfelt et al. 2018; see also Chapter 14 in this volume). The migration has also resulted in a long sought-after union, and to some extent harmonization, of the two sibling dictionary database structures (SAOL and SO), into a combined rich lexical database, *Salex*. See Chapter 11 in this volume.

5 Looking ahead

As outlined above, Språkbanken – now carrying forward the legacy of the Gothenburg brand of computational lexicographical R&D – has grown considerably over the years from its beginnings in the 1960s. The main focus of its present incarnation – *Språkbanken Text* – is on language technological research rather than corpus linguistics as at the beginning of its existence. The lexicographical element has been very much present throughout its history, even during the years of separation from the dictionary-compilation work, as described above (see also Chapter 9 in this volume).

We see a bright future for computational lexicography in Gothenburg. With the recent developments described in Section 4, the strengths of the two strands that were pursued separately for two decades are synergistically combined. The result is a vibrant and multifaceted research environment intertwined with and supported by a closely integrated cutting-edge computational infrastructure for working with lexical data. This will advance Swedish computational lexicography technically, methodologically, and scientifically, and serve a broad range of R&D purposes, in particular in the humanities and social sciences.

We will now be able to draw both on highly information-rich Swedish lexical databases compiled and enriched over several decades by highly trained lexicographers and on the most recent language technologies built on deep learning and AI. Some promising directions for the short and medium term future are development of new or improved sophisticated computational tools for mining very large text corpora to: find evidence for new words and word usages, as well as obsolescing word usages (Forsberg, Sikora & Sköldberg 2023; Holmer et al. 2024); investigate phraseology and multiword expressions (Borin 2021; Sköldberg 2022); track the historical

development of the Swedish lexicon (Viklund & Borin 2016; Sköldberg & Holmer 2017; Petersson & Sköldberg 2021; Adesam et al. 2021; Chapter 13 in this volume); contribute to the state of the art of lexical typology (Borin 2012; Borin, Comrie & Saxena 2013); and, of course make better dictionaries (for human consumption) and lexical resources (for computer processing of Swedish text).

References

Adesam, Yvonne, Peter Andersson, Lars Borin & Gerlof Bouma. 2021. A lexical resource for computational historical linguistics. In *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 98–121. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.04ade.

Allén, Sture. 1965a. *Grafematisk analys som grundval för textedering med särskild hänsyn till Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655* [Graphemic analysis as a basis for text editing with special reference to Johan Ekeblad's letters to his brother Claes Ekeblad 1639–1655]. Gothenburg: Department of Scandinavian Languages, University of Gothenburg.

Allén, Sture. 1965b. *Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655: Utgivna med inledning, kommentar och register* [Johan Ekeblad's letters to his brother Claes Ekeblad 1639–1655: Edited with an introduction, commentary, and index.] Gothenburg: Department of Scandinavian Languages, University of Gothenburg.

Allén, Sture. [1980] 1999. The language bank concept. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 302–310. (Originally published in Joseph Raben & Gregory Marks (eds.), *Data bases in the humanities and social sciences*, 171–176. Amsterdam: North-Holland). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allén, Sture. 2014. Språkvetenskaplig databehandling [Computational linguistics]. In *Personliga tillbakablickar över ämnesområden vid Göteborgs universitet: Seniorakademien Dokumentationsserie del 2*, 61–66. http://hdl.handle.net/2077/51552. Gothenburg: Seniorakademien vid Göteborgs universitet.

Boguraev, Bran & Ted Briscoe (eds.). 1989. *Computational lexicography for natural language processing*. London: Longman.

Borin, Lars. 2012. Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Diana Santos, Krister Lindén & Wanjiku Ng'ang'a (eds.), *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, 53–65. Berlin: Springer.

Borin, Lars. 2021. Multiword expressions: A tough typological nut for Swedish FrameNet++. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 221–259. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The Intercontinental Dictionary Series: A rich and principled database for language comparison. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 285–302. Berlin: De Gruyter Mouton.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2008. The hunting of the BLARK: SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf & Beáta Megyesi (eds.), *Resourceful language technology: Festschrift in honor of Anna Sågvall Hein*, 21–32. Uppsala: Department of Linguistics & Philology, Uppsala University.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. DOI: 10.1007/s10579-013-9233-4.

Borin, Lars, Markus Forsberg, Lennart Lönngren & Niklas Zechner. 2021. Swedish FrameNet++: Lexical samsara. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 69–95. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. *International Conference on Language Resources and Evaluation (LREC) 2012*. 474–478.

Borin, Lars & Louise Holmer. 2024. Tradita innovare, innovata tradere: The Gothenburg approach to computational lexicography. *Proceedings of the Huminfra Conference (HiC 2024)*. 41–50.

Dannélls, Dana, Lars Borin, Markus Forsberg, Karin Friberg Heppin & Maria Toporowska Gronostaj. 2021. Swedish FrameNet. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 37–65. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (eds.). 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Forsberg, Markus, Justyna Sikora & Emma Sköldberg. 2023. *Words unboxed: Discovering new words with Kubord*. Stockholm: National Library of Sweden. KBLab blog post: https://kb-labb.github.io/posts/2023-08-29-kubord/.

Francis, W. Nelson & Henry Kučera. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

Gellerstam, Martin. 1999. LEXIN: Lexikon för invandrare [LEXIN: Lexicons for immigrants]. *LexicoNordica* 6: 3–17.

Gellerstam, Martin, Yvonne Cederholm & Torgny Rasmark. 2000. The Bank of Swedish. *International Conference on Language Resources and Evaluation (LREC) 2000*. np.

Gellerstam, Martin & Christian Sjögreen. 1994. *Språkbanken: En språklig referensdatabas* [Språkbanken: A linguistic reference database]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Hannesdóttir, Anna Helga & Bo Ralph. 1988a. Early dictionaries in Sweden: Traditions and influences. *Symposium on Lexicography IV. Proceedings of the Fourth International Symposium on Lexicography April 20–22, 1988 at the University of Copenhagen*. 265–279.

Hannesdóttir, Anna Helga & Bo Ralph. 1988b. Projektet "Lexikografisk tradition i Sverige" [The project "Lexicographic tradition in Sweden"]. In Gertrud Pettersson (ed.), *Studier i svensk språkhistoria*, 74–85. Lund: Lund University Press.

Holmer, Louise, Ann Lilliestöm, Emma Sköldberg & Jonatan Uppström. 2024. Time to say goodbye revisited: On the exclusion of headwords from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 443–452.

ISLEX. 2025. *ISLEX* [The ISLEX Project]. [Online resource] https://islex.arnastofnun.is/se/about/. Accessed on 2025-03-09.

Järborg, Jerker. 1999. *Lexikon i konfrontation* [Lexicons in confrontation]. (Research Reports from the Department of Swedish No. GU-ISS-99-6) Gothenburg: Department of Swedish, University of Gothenburg.

Järborg, Jerker. 2001. *Roller i Semantisk databas* [Roles in the Semantic database]. (Research Reports from the Department of Swedish No. GU-ISS-01-3) Gothenburg: Department of Swedish, University of Gothenburg.

Järborg, Jerker. 2003. *Formaliserade semantiska samband mellan enheter i GLDB* [Formalized semantic relations among items in GLDB]. (Research Reports from the Department of Swedish No. GU-ISS-03-1) Gothenburg: Department of Swedish, University of Gothenburg.

Kann, Viggo & Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 105–110.

Lexin. 2025. *Lexin ordböcker* [Lexin Dictionaries]. [Online resource] https://lexin.se/. Accessed on 2025-03-22.

Lönngren, Lennart. 1988. Lexika, baserade på semantiska relationer [Lexicons based on semantic relations]. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 229–236.

Lönngren, Lennart. 1998. A Swedish associative thesaurus. *Proceedings of the European Association for Lexicography (EURALEX) 1998: Vol. 2*. 467–474.

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 41–106. Amsterdam: John Benjamins.

Malmgren, Sven-Göran. 2000. *Projektet Det svenska ordförrådets utveckling 1800–2000: Utgångspunkter* [The project The Evolution of the Swedish Vocabulary 1800–2000: Points of departure]. (ORDAT No. 1) Gothenburg: Department of Swedish, University of Gothenburg.

Malmgren, Sven-Göran & Emma Sköldberg. 2013. The lexicography of Swedish and other Scandinavian languages. *International Journal of Lexicography* 26(2): 117–134. DOI: 10.1093/ijl/ect008.

NEO. 1995. *Nationalencyklopedins ordbok* [The Dictionary of the National Encyclopedia]. Höganäs: Bra böcker.

Petersson, Stellan & Emma Sköldberg. 2021. Semantic change in Swedish: From a lexicographic perspective. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 149–167. Berlin: Language Science Press.

Ralph, Bo. 1977. Projektet lexikalisk databas [The Lexical Database project]. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 79–81.

Ralph, Bo. 1979. Lexikologi som datalingvistik [Lexicology as computational linguistics]. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 161–170.

Ralph, Bo. 2001. Orden i ordning: Den historiska framväxten av en lexikografisk tradition i Sverige [Words in order: The historical development of a lexicographic tradition in Sweden]. *Nordiska studier i lexikografi 5*. 282–321.

Ralph, Bo, Jerker Järborg & Sture Allén. 1977. *Svensk ordbok och lexikalisk databas: Förstudierapport* [The dictionary *Svensk ordbok* and the lexical database: A pilot study report]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Ritchie, Graeme. 1987. The lexicon. In Pete Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson & Paul Bennett (eds.), *Linguistic theory and computer applications*, 225–256. London: Academic Press.

SAOL 11. 1986. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 11th edn. Stockholm: Norstedts.

SAOL 12. 1998. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 12th edn. Stockholm: Norstedts.

SAOL 13. 2006. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 13th edn. Stockholm: Norstedts.

SAOL 14. 2015. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 14th edn. Stockholm: Norstedts.

Sköldberg, Emma. 2022. Phraseological theory, evidence in corpora and lexicographical practice: On collocations in a monolingual dictionary of Swedish. In Kristian Blensenius (ed.), *Valency and constructions: Perspectives on combining words*, 155–182. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Sköldberg, Emma & Louise Holmer. 2017. Ordböcker som språkhistoriska källor [Dictionaries as historical linguistic sources]. *Svensklläraren: Tidskrift för svenskundervisning* 61(3): 20–21.

Sköldberg, Emma, Louise Holmer, Elena Volodina & Ildikó Pilán. 2019. State-of-the-art of monolingual lexicography for Sweden. *Slovenščina 2.0* 7(1): 13–24. DOI: 10.4312/slo2.0.2019.1.13-24.

SO. 2009. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. Stockholm: Svenska Akademien.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Viberg, Åke, Kerstin Lindmark, Ann Lindvall & Ingmarie Mellenius. 2002. The Swedish WordNet project. *Proceedings of the European Association for Lexicography (EURALEX) 2002*. 407–412.

Viklund, Jon & Lars Borin. 2016. How can big data help us study rhetorical history? *Selected Papers from the CLARIN Annual Conference 2015*. 79–93.

Whitelock, Pete, Mary McGee Wood, Harold L. Somers, Rod Johnson & Paul Bennett (eds.). 1987. *Linguistic theory and computer applications*. London: Academic Press. 225–256.

Wilks, Yorick, Brian Slator & Louise Guthrie. 1996. *Electric words: Dictionaries, computers, and meanings*. Cambridge: MIT Press.

Part II: **Dictionaries for humans**

Louise Holmer and Kristian Blensenius

3 SAOL: a Swedish dictionary for all times

Abstract: *Svenska Akademiens ordlista*, SAOL 'The Swedish Academy Glossary', is a Swedish monolingual contemporary dictionary. It comprises about 126,000 headwords, providing information on part of speech, orthography, and inflection for Swedish words. The first edition was published in 1874. Hence, SAOL celebrates 150 years of setting an orthographical norm for everyday written Swedish. The latest print edition was published in 2015, and the forthcoming edition is planned for early 2026 in print and e-versions. SAOL is accessible through the dictionary web portal Svenska.se together with the comprehensive definition dictionary SO 'The Contemporary Dictionary of The Swedish Academy' and the historical SAOB 'The Swedish Academy Dictionary'.

Since the 1980s, the editorial staff of SAOL is housed at the University of Gothenburg, and from 2021, the researchers and editors of SAOL are incorporated in the Språkbanken Text research unit.

This chapter gives an overview of SAOL and its history, the different editions, and focus for the future.

Keywords: contemporary dictionaries, lexicography, SAOL, Svenska.se, Swedish Academy

1 Introduction

Svenska Akademiens ordlista, 'The Swedish Academy Glossary', often abbreviated as SAOL, is one of the most renowned dictionaries in Sweden. It boasts the longest tradition of all Swedish monolingual dictionaries, whether measured by number of editions, numbers of reprints, or sales figures. It functions as an authoritative

Acknowledgments: The work on the dictionaries SAOL and SO is financed by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*, and the work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161).

Louise Holmer, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: louise.holmer@svenska.gu.se
Kristian Blensenius, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: kristian.blensenius@gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-003

standard for orthography and inflection of contemporary Swedish vocabulary. This chapter provides an account of SAOL, with special reference to its history, lexicographical features, and usage. The dictionary's full name in Swedish is *Svenska Akademiens ordlista över svenska språket*, 'The Swedish Academy Glossary of the Swedish Language', but for the sake of convenience the abbreviated form SAOL is being used throughout this chapter.

As noted, SAOL serves as an informal norm for the orthography and inflection of Swedish words, in the absence of an *official* norm for Swedish. SAOL is often compared to the similar Danish dictionary *Retskrivningsordbogen* 'The orthographical dictionary', which is the official norm for spelling and inflection in Denmark. The difference between these two Nordic dictionaries lies in their different roles: in Denmark, official documents and texts written in the educational system, etc. must adhere to the norm established in *Retskrivningsordbogen* (Dansk Sprognævn 2025). In Sweden, there is no such official orthographical regulation anymore, although this was the case in the beginning of the 20th century (see Sections 3.1 and 3.2; see Buchmann 2015 for German orthography; see also Chapter 13 in this volume).

In this chapter, the focus is on the following features and main characteristics of SAOL:

– SAOL and its more than 150 years of existence;
– its 15 editions, most of them fully revised;
– SAOL's specific niche in contrast to similar dictionaries; and
– SAOL's development over time, and users' opinions.

SAOL has previously been the subject of extensive research, including works such as Sigurd (1986b), Gellerstam (2009b), Holm (1951), Ralph (2009), Malmgren (2009; 2014), Holmer (2016), and others. However, most of these studies are written in Swedish for a Swedish-speaking audience. We therefore aim to summarize the key issues focusing on the broader context, acknowledging that it is almost impossible to do justice to a dictionary with a long tradition and many editions like SAOL, as well as to the previous editors' detailed accounts of several of the editions.

The Swedish Academy has historically funded SAOL, and today they also fund two other dictionaries: the historical and comprehensive *Swedish Academy Dictionary* (SAOB 1898–2023), and the large definition dictionary of present-day Swedish, *The Contemporary Dictionary of the Swedish Academy* (SO 2021), the latter further described in Chapter 4 in this volume. SAOL and SO are today compiled at the University of Gothenburg, whereas the editors of SAOB are situated in Lund, employed directly by the Swedish Academy.

2 Swedish reference dictionaries – a historical overview

When the Swedish Academy was founded in 1786, by His Majesty King Gustav III, one of its aims was linguistically oriented: to compile a Swedish dictionary and a reference grammar (Ralph 2009: 34). Gustav III was highly inspired by the French Academy, and he selected the thirteen original members of the Swedish Academy himself (today their number is eighteen). The motto of the Academy was, and still is, *Snille och smak* 'Talent and taste'.

A pronounced purpose for the Academy is to actively promote "the purity, vigor and majesty" of the Swedish language, i.e., its clarity, expressiveness and prestige (Svenska Akademien 2025). SAOL, SO, and SAOB of today all contribute to this purpose, as do several other initiatives by the Academy. The Academy has also provided partial financial support to several other lexicographical projects, including Olof Östergren's comprehensive *Nusvensk ordbok* ('Dictionary of Contemporary Swedish', see Östergren 1919–1972), Erik Wellander's antibarbarus *Riktig svenska* ('Proper Swedish' 1939; see Wellander 1939 and Sigurd 1986a), and many others.

2.1 Early orthographical initiatives

The first edition of SAOL (SAOL 1 1874) was not the first attempt to set down authoritative guiding principles regarding Swedish spelling. In 1801, a member of the Swedish Academy, Carl Gustaf Leopold, published *Afhandling om svenska stafsättet* 'Thesis on the Swedish way of spelling' in the Academy publication series (Leopold 1801). In his thesis, Leopold suggested spellings like *absolut* and *choklad*, which are the preferred spellings in Swedish today as well. He also proposed the use of double consonants in short function words such as *att*, *till*, and *upp* 'to/that', 'to' and 'up', respectively; addressing inconsistencies in their earlier spelling.[1] Generally, for foreign words, he was an advocate of a spelling more aligned with Swedish (Leopold 1801; Santesson 2001).

In 1869, Sweden hosted a Nordic orthographical conference. The aim was to discuss, and hopefully also rein in, the orthographical chaos that reigned in Sweden. Another aim was to try to bring the spelling norms in Danish, Norwegian, and Swedish closer to each other. The mottos of the conference were *samnordiskhet* 'common

1 See for example these entries in SAOB (1898–2023) for spelling variations from 1520 to the present day.

Nordicness' and *ljudlikhet* 'phonological similarity', i.e., spelling in compliance with phonology (see Ståhle 1970).

The Swedish Academy was not present at the conference, since none of its representatives were invited. After the meeting, a progressive linguistic society was founded, arguing for a more sound-based orthography of Swedish. Its members published a journal called *Nystavaren* 'The New-speller', and managed to initiate an organized Swedish spelling reform. Although not all of their suggestions were adopted, there was a major Swedish spelling reform in 1906, disseminated to the public through a royal decree. The spelling reform and its history and later implications for SAOL are covered by Malmgren (2009) and Ralph (2009). To the present day, Swedish still uses the orthography established by that spelling reform. The most prominent changes were the following:

– *f*, *fv*, and *hv* ⇒ *v* (in most cases): *golfvet* – *golvet* 'the floor', *hvila* – *vila* '(to) rest';
– *dt* ⇒ *tt*: *godt* – *gott* 'good, delicious'

SAOL registered most of these changes listed above, albeit not directly.

Apart from the orthographical discussions, it had become evident that the completion of the historical dictionary SAOB would be severely delayed. When the progress of the work on SAOB turned out to be less rapid than foreseen, the Swedish Academy started to plan for a handbook of orthography which could provide a partial descriptive framework for the considerably more ambitious SAOB project. Hence, SAOL was seen as a short and light version of SAOB. The editors of that time were based in Stockholm and closely associated with the Swedish Academy.

In the second half of the 19th century, several Swedish dictionaries and orthographical reference works were available on the Swedish market. Most of them are now more or less forgotten by the public, and most likely also by lexicographers and other linguists. The most notable dictionary from that time is the Swedish dictionary compiled by A.F. Dalin who completed the first Swedish monolingual dictionary (Dalin 1850–1853; Hannesdóttir 1998; Malmgren 1988; see Chapter 12 in this volume).

2.2 The literature on SAOL

Several authors and publications investigate various aspects of SAOL and its different editions. Many publications are written by the editor in chief of a specific edition, whereas others are contributed by editors or linguistically oriented members of the Swedish Academy. The authors of this chapter form no exception in this regard – we are since April 2023 the editor in chief (Louise Holmer) and assistant editor in chief (Kristian Blensenius) of the 15th edition of SAOL. As of 2025, a total of ten lexicographers and language technologists work with SAOL 15.

The main works dedicated to SAOL are *Nytt och gammalt i Svenska Akademiens ordlista* 'New and old in The Swedish Academy Glossary' (Holm 1951), Johannisson (1974), Mattsson (1974), *Svenska Akademien och svenska språket* 'The Swedish Academy and the Swedish language' (Allén, Sigurd & Loman 1986), and *SAOL och tidens flykt* 'SAOL and the flight of time' (Gellerstam 2009b). Together with Malmgren (2014), editor in chief for SAOL 14 (2015), they all offer historical overviews of SAOL and highlight novelties and major revisions in the different editions.

The later editions of SAOL, as well as the dictionary portal Svenska.se, have all been the subject of scientific reviews in especially the *LexicoNordica* journal.[2] Other than that, there are relatively few extensive scholarly works where SAOL is the scientific object of study (Holmer 2016; 2022). The same goes for the other two dictionaries financed by the Swedish Academy, SO and SAOB, although Rosqvist (2014) and Nilsson (2023) have written doctoral theses about the latter.

The introductory text in a dictionary constitutes a specialized genre, where editors, publishers, or funders can explain the functions of the dictionary, the intended user, the linguistic orientation and the theoretical and methodological, and hence, lexicographical, decisions (Svensén 2009; Holmer 2016). However, these texts in general tend to be short and merely outline in a descriptive way what users can expect from the dictionary. As for the different editions of SAOL, each of these includes at least some kind of introduction, although they vary from a few pages to more than 30 pages in length (for an overview, see Holmer 2016: 59–60). These introductions taken together make up a kind of long-term documentation of the guiding principles behind SAOL, both in general and for every separate edition.

3 Main features of the different editions

SAOL's main features are information on orthography, inflection, and part of speech (POS). In the earlier editions, information on POS might be given only for ambiguous words, or implicitly with the addition of inflectional suffixes, whereas the latest editions from SAOL 12 (1998) and onwards provide explicit POS information for every headword. Although somewhat slimmed-down in its microstructure (the entry and its text; see Svensén 2009 for terminology), SAOL also provides information on word segmentation in morphemes or syllables (less so in the early editions and somewhat more in the later editions), usage, and register.

2 *LexicoNordica* is the journal of lexicography in the Nordic countries published by *Nordisk förening för lexikografi* 'Nordic association for lexicography'.

Many of the editions, although quite faithful to the main objectives, possess several special features, or, in some cases, peculiarities. The subsections of the present section are devoted to description of and comments on the editions, with special regard to their similarities and differences. It is, of course, impossible to describe everything in detail. The intention is instead to give a thorough and fair illustration of the main features.

It should be noted that the linguistic terminology used in SAOL varies between editions. It stands to reason that norms for abbreviations will vary in a language over a period of 150 years, and the conventions for abbreviations in SAOL have followed suit. It should also be noted that the ways of categorizing the vocabulary and of assigning parts of speech to the words also vary, in linguistics in general, as well as in SAOL. This may not always be stated nor explained in the introduction to every new edition.

3.1 SAOL 1–6, 1874–1889: Orthographical guidance with royal blessings

The first edition of SAOL was published in 1874 and printed in 2,000 copies, which were quickly sold out (Johannisson 1974). From the first proposal in the Academy 1869, to the realization of SAOL, only five years passed. The proposal came from the quite conservative and historically oriented linguist Johan Erik Rydqvist, a member of the Swedish Academy. The editor in chief of SAOL 1, however, was F.A. Dahlgren, a lexicographer and author. The editors of the first SAOL had access to A.F. Dalin's *Ordbok öfver svenska språket* 'Dictionary of the Swedish language' (Dalin 1850–1853; see Chapter 12 in this volume) as well as Rydqvist's historical linguistic treatise *Svenska språkets lagar* 'The laws of the Swedish Language' (Rydqvist 1850).

The first edition comprises about 34,000 headwords. Its introduction repeats almost verbatim Rydqvist's earlier text on SAOL as a descriptive framework of sorts for the SAOB, and that SAOL provides information on spelling, inflection, and gender (for nouns). As Sigurd (1986b: 197) points out, the introduction refers (surprisingly) frequently to linguistics as a science, and especially to historical linguistics.

SAOL 1 includes "främmande ord" 'foreign words', but only to a limited extent, something that might be considered a shortcoming. Compared to Dalin's dictionary, published 20 years earlier, SAOL 1 lacks fairly common headwords like *amiral* 'admiral', *faktura* 'invoice', *jaguar* 'jaguar', *jalusi* 'shutter', *madrigal* 'madrigal', and many more. One explanation for this is that Rydqvist took a generally conservative stance. Another one has to do with gender information for nouns. SAOL 1 provides such information (m. for masculine nouns, f. for feminine nouns, and n. for neuter nouns). This can be seen in Figure 1, where the headword *morot* 'carrot' is labelled "s. f.",

Mogande, part. pres. (till mogen ålder kommen). Brukas stundom i lagspråket.

Mogen (*moget; mogne, -a; mognare, mognast*), adj. **-het** (utan pl.), s. f. **Moget,** adv. **Mogna** (*-ar, -ade, -at, -ad*), v. intr. **-ande,** s. n. **-ing,** s. f. **Mognad** (utan pl.), s. m.

Moja sig (*-ar, -ade, -at*), v. refl.

Morla (*-ar, -ade, -at*), v. impers. **-ande,** s. n.

Morot (pl. *-rötter*), s. f.

Morra (*-ar, -ade, -at*), v. intr. **-ande,** s. n.

Morsk (samdr. af *mordisk*), adj. **-het** (utan pl.), s. f.

Mortel (pl. *mortlar*), s. m. **-stöt** (pl. *-ar*), s. m.

Figure 1: The first edition of SAOL, 1874, some of the words in M

meaning 'noun, feminine'. The last headword in the figure, *mortel* 'mortar', is labelled "s. m.", 'noun, masculine'. To provide information on masculine and feminine for nouns was considered a bit old-fashioned at that time, since nouns often were referred to as *den* 'it' instead of *han/hon* 'he/she' or *honom/henne* 'him/her'. Even in the Dalin dictionary, the gender marker *reale* was in use for masculine and feminine nouns, but Rydqvist, being an advocate of the older (in retrospect even obsolete) three-gender system in Swedish, insisted on keeping this information about gender for nouns in SAOL 1, along with neuter, n. This led to considerably fewer loan words in the dictionary. Such words (often French and English ones) would have to have been treated according to some kind of principle – either they are assigned the same gender as in their source language, or their gender is assigned in analogy with other words in the target language, e.g., because of their phonology. The lexicographers would have had to make a major effort to assign m. or f. to all borrowed nouns, and this circumstance may have effected the number of foreign words (see Teleman 2003: 138–144; Chapter 13 in this volume; Malmgren 2009).

In Figure 1, an example of the macro- and microstructure of SAOL 1 is shown. The first headword, *mogande* 'maturing', has a POS label indicating the present participle, a short semantic description "till mogen ålder kommen" 'having come to a mature age' and a usage note, "brukas stundom i lagspråket" 'occasionally used in the language of law'. In the first edition, there are occasional notes on etymology, in contradiction to the statements made in the introduction. In some entries, digressions on correct and incorrect use are also made.

The other entries shown in the figure display less information than *mogande*. The mix of headwords with descriptions of various sort, and headwords with only POS-information and maybe inflectional suffixes, can be said to represent the majority of the entries in SAOL 1.

The editions 2–5 might best be considered reprints rather than new editions, due to very few added entries (Johannisson 1974). These editions were published between 1874 or 1875 and onwards, and few changes were made between editions.

The 6th edition (SAOL 6 1889) is the first one to show considerable differences compared to its predecessors. Rydqvist had passed away in 1878 and the energetic new-speller Esaias Tegnér Jr (1843–1928) had joined the editorial staff. Tegnér was a member of the Swedish Academy, and he was also the editor in chief of SAOB in Lund.

The 6th edition comprises about 41,000 headwords. New features include, according to its introduction, more compounds, more derivatives, and more loanwords (SAOL 6 1889: XXX; Sigurd 1986b: 201). The main development, however, is the application of new orthographical principles:

– gt => kt: *bugt – bukt* 'bay', *pligt –plikt* 'duty';
– q => variant with k: *qvinna – kvinna* 'woman';
– e/ä variants: *elg – älg* 'moose, elk', *prest – präst* 'priest'.

New (and nowadays non-controversial) spellings like *röd – rött* (the non-neuter and neuter forms of *röd* 'red') and *vid – vitt* (the non-neuter and neuter forms of *vid* 'wide') were not included in SAOL 6, as they were seen as too newfangled ("de torde i de flestes ögon vara alltför dristiga nyheter" 'they ought to, in the eyes of most people, be too drastic novelties', Sigurd 1986b: 286). The Swedish Academy's point of view was that the spelling with a 'v' for the older 'fv' would cloud the visual impression, and the Academy did not wish to disturb the established order. The 6th edition was also made the official orthographical canon for Swedish schools (Gellerstam 2009c: 57).

As for further reading, Johannisson (1974), Sigurd (1986b), Ralph (2009), Gellerstam (2009a), and Malmgren (2009; 2014) all provide interesting information about the early editions and the editorial work in relation to them.

3.2 SAOL 7, 7.5 and 8, 1900–1923: Conservativists vs. new-spellers

Among the main features of the 7th edition (SAOL 7 1900) are its increased number of headwords, in total ca. 71,000, and its complex compromises between a more conservative orthography and a modern approach to spelling and gender in nouns. One particular decision that had impact on the number of lemmas in SAOL 7 was that regularly formed (deverbal) nomina actionis ending in *-ande* were excluded as headwords. This led to about 3,000 excluded deverbal nouns, making room for neologisms and other derivatives (see SAOL 7 1900: 1 and Holmer 2022).

The 7th edition was compiled in Lund instead of Stockholm, with Otto Hoppe (1857–1919) as the editor in chief (Johannisson 1974). Like his predecessor, he was

also the main editor for SAOB at the same time. This transfer to Lund and SAOB led to benefits for SAOL, as the editors were able to make use of the SAOB material.

As noted, in 1906, a decree was issued by His Majesty the King, establishing a new way of Swedish spelling. The Swedish Academy was quite conservative at the time, and reacted by publishing a golden mean, a radically scaled down version of SAOL, called *Ordförteckning över svenska språket* 'Inventory of the Swedish language' (Ordförteckning 1916). The somewhat odd publication was proffered in an attempt to moderate the criticism against the Academy of being (too) conservative regarding the simplified spelling, and at the same time work in compliance with His Majesty's decree and its directives. In the *Ordförteckning*, all the orthographical novelties are applied, with one major exception. The Academy chose to keep the dt-spelling in neuter forms of non-participial adjectives ending in *-d* in their non-neuter form (*god – godt*, *röd – rödt*, where the official Swedish and the public would prefer (and use) *god – gott*, *röd – rött* etc.). The *Ordförteckning över svenska språket* presents an even more reduced list of words than the other editions of SAOL, and today it remains a mere curiosity, and has, for example, not been added in the historical reference work SAOLhist (see Chapter 13 in this volume).

In the following list, the macro- and microstructure of the *Ordförteckning* are shown, illustrated by lemmas beginning in *k-*. Headwords are in boldface, no semantics or POS information is provided, and there is a dagger (†) marking that the old and first listed spelling *rödt* is preferred over the new spelling *rött* (but the more modern spelling *rött* is the only valid spelling in present-day Swedish).

krank. -het. krans. -formig.
krans|a. -ning.
krapp. -färg. röd. (*-rödt* l. † *rött*)
kras. krasa.
krasch, äv. **krach.**

The lemmas above, starting with *krank* 'ill' and ending in *krach* 'crash', form a very short excerpt of the glossary. Nevertheless, these few examples are representative of the overall structure of the *Ordförteckning*.

A new edition – the 8th – appeared only in 1923 (SAOL 8 1923). It is mainly known for its Swedish equivalents accompanying foreign loanwords, especially those from French and classical languages, but not loanwords from Germanic languages (Gellerstam 2009a) and for its conservative principle of keeping the old-fashioned dt-spelling, although marked in the microstructure with the text "SvAk", to signal that the Swedish Academy was opposed to the new spelling. The lemma list in SAOL 8 was somewhat expanded into approximately 77,000 headwords. There was also a wish from the editor in chief Ebbe Tuneld to provide pronunciation and a separate

antibarbarus chapter on "language correctness" (*språkriktighet*), but these ideas were never realized (Malmgren 2009).[3]

Apart from (contrived) Swedish equivalents to loanwords, the 8th edition is also known for its attempt to introduce Swedish spellings for some loanwords, like *burgogne*, *klaun*, and *skejlejt* for *bourgogne*, *clown*, and *skylight*. However, these suggestions for Swedish spellings were not picked up by the public and they were removed by the editors in the 9th edition.

3.3 SAOL 9–11, 1950–1986: Growth spurt, everyday language and semantic austerity

The ninth edition is by far the most exhaustive one, with about 150,000 headwords (SAOL 9 1950). When it was published in 1950, 27 years had passed since the previous edition of SAOL. The editor in chief, Pelle Holm, was also the editor in chief of SAOB in Lund, and he was able to make use of the manuscripts forming the basis for SAOB. He describes the large increase of words as being due to new societal phenomena, sports and leisure, technology, etc. (Holm 1951).

The ninth edition is first and foremost known for its many compounds, often consisting of three nouns (or a combination of nouns plus words belonging to other POS), like *mansdagsverke* 'man's day's work', *råttfällsfabrik* 'rat trap factory', and *stridsvagnsformation* 'tank formation'.

An example from the 9th edition shows the structure of the etymologically based nests with the simplex first, followed by a large set of compounds. In the following example, the entry *kärlek* 'love', and about half of its compounds registered in SAOL 9, are included. The layout of the example is to a large extent in line with the original layout (but not fully). In this nest of compounds, no POS label has been provided, although *kärleksdrama* 'love drama' is a noun, whereas the following headword *kärleksdrucken* 'love struck, strongly infatuated' is an adjective.

kärlek *-en* s. **kärleks|affär. -betygelse. -brev. -bud** relig. **-dikt. -diktning. -drama. -drucken. -dryck. -full. -fullhet. -förbindelse. -förhållande. -förklaring. -gnabb. -griller. -gud. -gudinna. -gåva. -handel. -hat.**

In SAOL 10 (SAOL 10 1973), the editor in chief Gösta Mattsson revised the list of headwords and excluded about 20,000 words, mostly compounds. The editorial

3 Today, there are public institutions specifically charged with dealing with matters concerning the Swedish language, providing advice about spelling, language use, etc., such as *The Institute for Language and Folklore*, and in particular its section *The Language Council*.

staff of SAOB was also involved in the work on SAOL (Mattsson 1974: 57), but the 10th edition was the last one to be completed in Lund. In comparison with the previous edition, the 10th one was a bit stricter with regard to the description of plural forms and the orthography of English loanwords. Plurals in *-s*, such as *gangster – gangsters* 'gangster – gangsters', were gone and replaced with the Swedish plural form *gangstrar*. English loans like *jet*, *pop*, and *pub* were listed with the variant forms *jett*, *popp* and *pubb*, which better represent traditional Swedish spelling conventions, especially in the definite forms *jett – jetten* 'the jet', etc. However, users did not seem to care too much about these recommended forms, perhaps since English was now more common in Sweden in general.

For the next edition (SAOL 11 1986), the editor in chief, situated in Lund and engaged in SAOB, Sven Ekbo, wanted to delete about 50,000 headwords from SAOL 10 to make room for considerably more developed semantic descriptions. This would have turned SAOL into more of a general dictionary. The Swedish Academy opposed the idea, and decided to relocate the editorial office to the University of Gothenburg (see, e.g., Allén 1986; Gellerstam 2009a).

SAOL 11 was published in 1986, and the editor in chief in Gothenburg, Martin Gellerstam, kept the original idea of SAOL mainly presenting a list of words. In the same year, an exhaustive and corpus based definition dictionary was published in Gothenburg as well, SOB (1986). The publication of these two dictionaries also marked the 200th anniversary of the Swedish Academy.

The transfer of the SAOL manuscripts from the SAOB editorial team and the facilities in Lund to the research group at the University of Gothenburg also strengthened the scientific basis and approach, drawing on computer-aided lexicography. Also, to transplant the work on a quite special product like a dictionary commissioned by the Swedish Academy, a congregation *sui generis*, into the university instead, is quite special. Swedish universities are public authorities governed by explicit formal regulations about transparency, etc., which may sometimes come into conflict with the interests of external stakeholders, for instance regarding IPR (intellectual property rights).

In SAOL 11, computational linguistic support was used for the first time. The previous edition, SAOL 10, was available on punched tape. The department of *Språkdata* at the University of Gothenburg, had, among other research areas, developed text corpora and performed exhaustive studies on Swedish word frequencies, morphology and phraseology (see Malmgren & Sköldberg 2013 and Borin & Holmer 2024 for research projects and their output; see Chapter 2 in this volume for a brief history of Språkdata and Språkbanken).

The 11th edition was one of the best selling of all the SAOL editions, with about half a million copies sold. It held a larger vocabulary from the domain of colloquial speech, being the first edition to contain the iconic expletive *jävlar* lit. 'devils', among

other words in the more casual register. SAOL 11 was also the first of the editions to appear in a publicly available electronic version (on floppy disk).

3.4 SAOL 12–14, 1998–2015, from strict alphabetical order to *hen*

In SAOL 12 (1998), a strict alphabetical order in the macrostructure was introduced. The classical etymologically sorted nests were abandoned, and instead every headword was set in boldface at the beginning of a new line. Earlier editions that were practicing the nesting principle still implicitly provided semantic and etymological information, since related words were listed and described close to one another.

The most prominent advantage of the new order was that the user had less difficulties finding the searched-for word. The disadvantage was that some of the implicit information on semantics was lost at times, when closely related words might appear many lines apart.

The following example shows the word *handled* 'wrist' in SAOL 12 and the four other headwords following alphabetically. In this example, the first headword – *handled* 'wrist' – and the last one – *handledsväska* 'wrist bag, wristlet' – belong together, but are interrupted by headwords 2–4, which also belong together, but whose senses have nothing to do with wrists.

hand|led s. *-en -er*
hand|leda v., till *[1]leda*
hand|led|are s. *-n*; pl. =, best. pl. *handledarna*
hand|led|ning s. *-en -ar*
hand|leds|väska s.

For the 12th edition, the editors also systematized the morphosyntactic tags for all inflectional forms, creating a morphological database, SMDB (Swedish Morphological Database). SMDB contained the headwords of SAOL 12 and all their inflectional forms. It formed the basis of the extended inflectional information in the different e-versions of SAOL, such as the CD-ROM version SAOL Plus (2007), smart phone apps (2011 and 2017) and the web version on Svenska.se (2017). Thanks to the structure of the database and its combination with the larger corpora of Språkbanken, more scientifically based decisions were made on what headwords to include and exclude. It was also evident which inflectional forms that were in use in text (Berg & Cederholm 2001; Berg 2009; Borin & Holmer 2024).

A new formal model was introduced, the *lemma–lexeme model* (see Allén 1981 for an introduction, and Svensén 2009 for comparison). With the new macrostructure, a

new headword is always given on a new line, instead of being treated in the same nest as its etymologically related word family.

Another principle in relation to this macrostructural change is the idea that even semantically distinct but formally identical headwords that share the same inflectional pattern should form a single entry (one lemma). Also, if two identical word forms have *different* inflectional patterns, they belong to different lemmas. For example, earlier editions listed the noun *ljus* 'light, candle' and the adjective *ljus* 'fair, bright' under the same entry *ljus*, but in accordance with the lemma–lexeme model, they belong to separate entries and are treated as different lexemes.

Other significant qualities of the 12th edition are the dedicated lemma stock with selected Finland-Swedish headwords (Gellerstam 2009a: 76–77), the increased use of corpus-based methods for extraction of candidates for inclusion and exclusion, and the typographically emphasized indications of word segmentation. The word *handledare* 'supervisor' (see above) has a vertical line (|) marking the division between *hand* 'hand' and *ledare* 'leader', whereas the second part *ledare* has a short vertical line (ˈ) marking the derivational suffix *-are* '-er'.

SAOL 13 (2006) shares many similarities with SAOL 12 (1998). Both editions had Martin Gellerstam as editor in chief and Sture Berg as assistant editor in chief. The microstructure of SAOL 13 is the same as that of SAOL 12. The novelties in connection with SAOL 13 are above all related to its appearance in digital versions:

- the development of a CD version of SAOL 13 with full inflectional paradigms for each headword and generous search possibilities (SAOL Plus 2007);
- the publication of the book *SAOL och tidens flykt* 2009 'SAOL and the flight of time';
- SAOL 13 in a free facsimile version online;
- app versions of SAOL 13 for iOS and Android; and
- an electronic service with all headwords from almost all previous editions freely accessible (SAOLhist 2013), see Chapter 13 in this volume.

The printed book sold in considerably fewer copies than the previous edition. Hopes were probably high for the CD version to sell better, but the users had already turned to freely accessible dictionaries online, SAOL unfortunately not being one of these. Instead, the publishing company Norstedts, which has so far published all editions of SAOL and also SO 2009, developed app versions in a joint effort with app developers (first at Isolve AB, later Wang.se). The search functions in SAOL Plus 2007 were provided by Oribi, a company developing technical aids for people with dyslexia. Especially the fuzzy search in SAOL Plus, but also the other search facilities, maintained quite a high standard and had few counterparts in its field (Berg, Holmer & Hult 2008; Berg 2009).

The app versions built upon the work with the CD version, and the apps display inflectional forms together with grammatical information about the forms for all

lexical items that carry inflection. In the print version, inflectional information such as suffixes etc. are mostly given in abbreviated or else condensed forms. The apps were also the object of the first major user study of SAOL (Holmer, Hult & Sköldberg 2015).

From a lexicographic point of view, the introduction of the letter W as an independent letter in the macrostructure was one of the more prominent new features. In earlier editions, words with the initial letter W were sorted under V. The W section of SAOL is however one of the shortest sections in the dictionary; 2 pages in SAOL 14, compared to the section with S as the initial letter, which covers about 230 pages out of a total of 1,596 pages.

The 14th edition was published in 2015, with Sven-Göran Malmgren as editor in chief (SAOL 14 2015). This edition is so far the latest. It comprises about 126,000 headwords. Compared to SAOL 13, about 13,000 new entries were added and 9,000 excluded. Among the new headwords are a considerable number of compounds, semi-automatically extracted from mainly newspaper texts. Also, a POS previously not present was introduced in SAOL when about 400 names were added as entries, mainly geographical names and personal names.

A major change was also the introduction of sense disambiguation with lexeme numbers and more semantic descriptions as a consequence (SAOL 14 2015: XXX–XXXVI, see also Malmgren 2014). All compounds were provided with inflectional information, which was also a novelty. In previous editions, the user was referred to the simplex word for information on inflection. The following example shows the macro- and microstructure of SAOL 14, starting with *kärleksbetygelse* 'love declaration':

kär|leks|be|tyg|else s. ~*n* ~*r*
kär|leks|be|vis s. ~*et*; pl. ~
kär|leks|brev s. ~*et*; pl. ~
kär|leks|bud s. ~*et*; pl. ~ <relig.> till *bud 3*

All compounds have information about POS ("s." for *substantiv* 'noun'), inflectional suffixes, and long and short vertical lines for marking of word segmentation. The forthcoming edition, SAOL 15, will adhere to the same structure.

When it became publicly known that SAOL 14 had included the new, gender-neutral pronoun *hen* (Milles 2013; Haugen & Borin 2018), the headlines even made it into the international press. The inclusion of *hen* also led to reactions from the public, both positive and negative. The discussion of *hen* is actually too large a topic to be fully covered in this chapter, but from a lexicographical point of view, it can be noted that the headword *hen* and its short entry text has received far more attention than most other newly added headwords.

In 2017, the dictionary portal Svenska.se was released. It is probably safe to say that in 2025, more users access SAOL 14 through Svenska.se and the app versions than through the printed book, although the book also has a small but dedicated user group. Apart from Svenska.se, the most popular dictionary site for Swedish is Synonymer.se, a dynamic site that presents a synonym dictionary, a monolingual Swedish dictionary from 2010 from the publishing company Bonniers, example sentences from the internet, inflection tables, synonyms and antonyms suggested by the users, and much more (Holmer & Sköldberg 2016). Other than Synonymer.se, at present there are only two major Swedish online dictionaries, Lexin and Swedish Wiktionary.

3.5 SAOL 15

The focus areas of the ongoing work on SAOL 15 mainly include semi-automated detection of new lemma candidates, exclusion of obsolete headwords, revision of dated entry texts, and targeted search for particularly neologisms in selected semantic fields.

One major task has been to transfer the data from the older and in-house developed systems to Språkbanken Text and creating a functional dictionary writing system interface in alignment with the lexical infrastructure of Karp (see Chapter 11 in this volume). Major efforts have been made to incorporate the databases and the data structure for SAOL and SO into Karp and to continuously validate the dataset. The work is developed in close collaboration between the lexicographers and the language technologists.

Since the beginning of computational lexicography at the University of Gothenburg, the corpora made accessible by Språkbanken Text have grown considerably. The text material available via Korp amounts to several billion words. The methods for extracting candidates for inclusion include comparisons between selected years of newspaper text, such as 2021 with 2020, 2020 with 2019, etc. To discover lemma lacunas in both SO and SAOL, vector analyses with fastText have been performed (Forsberg & Holmer 2024; Forsberg & Sköldberg forthcoming; see Chapter 4 in this volume).

To get a picture of candidates for exclusion, frequency analysis has been performed with an in-house tool comparing all inflectional forms of the SAOL 14 lemmas to selected corpora of 3 billion words (Holmer et al. 2024). This has resulted in a list of 3,000 lemmas with zero occurrences in the modern material, which makes them good candidates for exclusion, although some manual lexicographic curation is needed in addition.

It was an outspoken wish from the Swedish Academy to publish a new edition of SAOL in print, planned for early 2026, along with an update of Svenska.se. There has also been a public demand for a print book, although it is not realistic to expect large sales numbers anymore.

3.6 Overview of editions, lemma counts and editors in chief

To sum up, there are considerable differences in the lemma count and features of the different editions of SAOL. First, it should be noted that the reported headword counts differ somewhat between texts about SAOL, for example the introductions to different editions of SAOL, research articles on SAOL, and statistics from the SAOLhist-project (see Chapter 13 in this volume). The headword counts in Table 1 are taken from an overview by Gellerstam (2009a: 56). It is safe to say that the information on lemma count is more approximate than precise. In Chapter 13 of this volume, the numbers of the different editions are somewhat different, partly because they are based on the lemma count in the database of SAOLhist. However, there is no doubt that the first editions are the smallest and thinnest, whereas the 9th edition holds the largest number of headwords, and that the later editions add up to at least about 120,000 lemmas.

One explanation for the fluctuations in lemma number might be related to the editors and their preferences: in the early editions, foreign words were deliberately neglected, which kept the headword count somewhat low. In the 9th edition, modern society and its many new fields influenced the lemma list, as did linguistic discussions on topics such as word formation and compounding (Mattsson 1974: 81). No later edition has been as generous in including transparent compounds, and the aim has instead been to exclude such non-opaque words, making space for more relevant ones.

In Table 1, every edition's year of publication is shown together with the name of its editor in chief and number of headwords.

4 Relations to other dictionaries and to users

Since the publication of its first edition, numerous official opinions about SAOL have been expressed, for example in press reviews, newspaper articles, comments from various word puzzle players, and many more. Although SAOL was rather quickly recognized and appreciated as an official orthographical aid within the Swedish school system, it has also always faced criticism for its headword selection, principles

Table 1: First year of publication of the SAOL editions, their editors in chief, and approximate number of headwords, from Gellerstam (2009a) and the SAOLhist project (in parentheses)

Year	Edition	Editor in chief	Headwords	
1874	SAOL 1	Fredrik August Dahlgren	35,000	(34,000)
1889	SAOL 6	Fredrik August Dahlgren	40,000	(42,000)
1900	SAOL 7	Otto Hoppe	71,000	(72,000)
1923	SAOL 8	Ebbe Tuneld	85,000	(79,000)
1950	SAOL 9	Pelle Holm	155,000	(149,000)
1973	SAOL 10	Gösta Mattsson	135,000	(140,000)
1986	SAOL 11	Martin Gellerstam	115,000	(117,000)
1998	SAOL 12	Martin Gellerstam	120,000	(119,000)
2006	SAOL 13	Martin Gellerstam	125,000	(123,000)
2015	SAOL 14	Sven-Göran Malmgren	126,000	(127,000)

for spelling variants, comments on usage and register, and so on. When the gender-neutral pronoun *hen* was included in SAOL 14, some users stated that they would never buy or use SAOL again. Others were very pleased, since the inclusion of a word in SAOL tends to provide it with some kind of official status.

In the early years, SAOL gained public attention mainly from its function as an orthographical aid in the educational system. In the late 1800s, newspapers reported from many teachers' conferences that SAOL was adopted as the accepted norm for spelling and inflection, primarily due to the lack of serious competitors.

When describing SAOL, it is important to distinguish it from SO, the contemporary definition dictionary of the Swedish Academy (see Chapter 4 in this volume). It is also of great value for the lexicographers to investigate users' opinions of SAOL. In this section, an overview of the major differences to SO is made, and users' opinions are exemplified.

4.1 SAOL and SO: similarities and differences

SAOL and SO have for a long time been supported by two different computational lexicographical databases. As described above, SAOL has a long history and was relocated to Gothenburg from Lund in the early 1980s. The SO database, however, was originally developed at the University of Gothenburg (see Chapters 2 and 4 in this volume). The two dictionaries have their own niches, respectively, and as mentioned, the main niche for SAOL is orthography and inflection and a more normative approach, whereas SO's main area of expertise is semantics and phraseology, and many other lexical properties of the Swedish lexicon. SO also adopts a more descriptive approach.

However, since the launch of Svenska.se in 2017, SAOL and SO, and also the historical SAOB, are accessible side by side, and results for search queries are displayed in parallel for the three dictionaries. This is generally appreciated by users but can also be confusing, when SAOL and SO at times describe the same word in different ways. The lack of harmony between SAOL and SO arises for example when the same headword is assigned different parts of speech (one example is *rykande*, 'smoking', characterized as an adjective in SAOL and an adverb in SO), when the order of subsenses is inconsistent between the dictionaries (for example in chronological order in SO according to etymology, but in frequency order based on modern usage in SAOL), when the number of senses vary, etc. (Bäckerud, Nilsson & Sköldberg 2020; Blensenius, Holmer & Sköldberg 2021; Sköldberg 2023).

One reason is due to their different purposes, another is that the development of the two dictionaries has so far not been synchronized or coordinated (at least not until 2024), and a third one is that their underlying databases have been developed separately, resulting in different structures.

4.2 User studies and user input

Thus, SAOL has been a resource for orthography and inflection for more than 150 years, but how is it actually used? In order to gain better understanding of users' behavior and lexicographical needs, the editors have conducted two user studies, the first one after the e-versions of SAOL were published (Holmer, Hult & Sköldberg 2015) and in 2024, another one about SAOL in general (Holmer in preparation).

The first study, (Holmer, Hult & Sköldberg 2015), focused on the app version of SAOL 13. SAOL 13 was published in a print version in 2006, was transformed into a CD version in 2007, and was further developed and released as apps. The apps offered somewhat fewer features compared to the CD, but they were available free of charge for the users. The aim of this study was mainly to collect opinions on the app, so as to see what users' opinions were of the first release, and what users wished for future versions. Users in general appreciated SAOL as a whole, valued the price (i.e., no cost at all) and were positive to the fact that the Swedish Academy actually was interested in making SAOL more available. Suggestions for the future included, among other things, to expand abbreviations.

The second study was aimed at linguists and language experts and dealt with SAOL's approach to language policy and language counseling. The majority of respondents' opinions stated, in brief, that SAOL is an important tool for everyone involved in linguistically oriented tasks, and that it is expected that SAOL should be somewhat conservative in current languages debates.

Apart from these two studies, which focus solely on SAOL, Bäckerud, Nilsson & Sköldberg (2020) investigate the use of Svenska.se, where SAOL of course forms a relevant part.

However, the editors of SAOL and SO receive emails and comments from users on a regular basis. Through the dictionary portal Svenska.se, users also post questions, comments, and suggestions. These different ways of collecting user input complement each other, and the editors of today have a fairly good overview of different views on SAOL from both the public and linguistically oriented users.

5 SAOL in the near and distant future

Few lexicographers and other linguists can predict what the future holds for lexicography and dictionaries for humans. As active lexicographers, we naturally hope for our products to reach many users, not only in the present, but also for many years to come. At the time of writing, there are concrete plans for one more print edition of SAOL, planned for early 2026. Alongside the book release, the content on Svenska.se will also undergo partial updates, and a slightly updated version of SO will be published. The 15th edition of SAOL will be incorporated into Svenska.se, replacing the 14th edition. Additionally, the app versions of SAOL and SO will be updated.

In this section, we outline the current workflow and focus areas of SAOL 15, and take a daring look into the crystal ball, to try to speculate on what SAOL 25 will look like in 2125.

5.1 The SAOL workflow

The current influx of new words to SAOL and also SO, both neologisms and older words that might fill lacunas, come from different sources. To guarantee the scientific basis of SAOL 15, the editors rely on language technology and computer aided solutions developed at Språkbanken Text at the University of Gothenburg (see Chapters 2, 4, 9, 10, and 11 in this volume; Forsberg & Holmer 2024).

The suggestions for candidates for inclusion in SAOL are mainly collected from the following sources:

- selected corpora containing contemporary Swedish, available through Korp (see Chapter 10 in this volume);
- recently added headwords, derivatives and compounds in SO (see Chapter 4 in this volume);

– manually extracted examples; and
– suggestions from users and linguistically oriented colleagues.

In order to keep the list of headwords relevant and up to date, and to guarantee that SAOL continues to be a contemporary dictionary, the editors also exclude obsolete headwords, subsenses, and quotations (Berg, Holmer & Sköldberg 2010; Diamond 2015; Holmer et al. 2024). Ever since the 11th edition in 1986, the lexicographical work with SAOL has been corpus based, and it is almost unthinkable to imagine anything else as far as methodology is concerned. The corpus based work is crucial in many stages of the process: extracting neologisms and inflectional patterns from selected texts, excluding obsolete words, investigating constructional patterns and valency, the hunt for first occurrences, etc. (see examples in Chapter 10 in this volume).

The inclusion of the research group of lexicographers in Språkbanken Text (see Chapter 2 in this volume) facilitates and secures the long-term technical stability. In addition to this, hopes are that in the longer term, both SAOL and SO may in some way be integrated in the computational lexical infrastructure of Språkbanken.

5.2 Three key characteristics

In order to summarize the specific traits of SAOL and its relevance as of today, we would like to focus on especially three of its main characteristics. The first one is that SAOL registers the *general* language, not specialized language. The second, that it comprises a *selection* of words, not every valid Swedish word. The third is that SAOL is *one of several options* for a user in search for different kinds of lexical information.

The idea that SAOL deals with the general language is quite old. This means that words that clearly belong mostly to certain regions in Sweden very rarely are registered in SAOL. Instead, there are several dedicated dialect dictionaries and web resources available for this subject, for example published by The Institute for Language and Folklore (Isof.se 2025). Words from domain specific language are included if they can also be considered a part of the general language. See, e.g., Landqvist, Sköldberg & Holmer (2024) for lexicographical considerations regarding words from the medical domain.

SAOL comprises a selection of words. There is an old belief that SAOL registers the valid Swedish vocabulary, and that if a word is not included in SAOL, it is not part of the approved Swedish. This is a misconception (see Gellerstam 2009c; Holmer 2022). It is, for example, simply not possible to include all compounds for every simplex. One example is the word *vatten* ‘water’, which is represented with about 230 compounds in SAOL 14, like *vattendjur* ‘water animal’, *vattenfärg* ‘water color’ and *vattenlöslig* ‘water soluble’. In the selected corpora, the initial compound member *vatten-* occurs

in around 2,000 compounds, of which most are *not* part of the SAOL list of headwords. Good candidates from the corpora, which might very well be included in SAOL 15, are for example *vattenransonering* 'water rationing', *vattenskydd* 'water protection' and *vattenpark* 'water park'.

SAOL is one of several dictionary resources. SAOL focuses on spelling and inflection, and some users expect to find more semantics, synonyms and etymology. Luckily, there are more comprehensive dictionaries, like SO and SAOB for semantics, examples of usages, constructions, etymology, and so on, and on Synonymer.se the user can find a lot of synonyms, antonyms, user generated content, etc. Obsolete words that have been excluded from SAOL can be accessed via SAOLhist (SAOLhist 2013), and learners of Swedish looking for a Swedish dictionary can consult Lexin.[4] All these resources are nowadays available free of charge online. A user who does not find the sought-for word in SAOL, can easily access any of the other dictionaries (and there are also several others, not mentioned here).

5.3 SAOL in the year 2125

The exact extent of the future for SAOL 100 years from now is somewhat unclear, and to make speculations about print dictionaries in the far future will perhaps lead to very little. As with most projects, issues of time and money are key. For a start, the Swedish Academy and the University of Gothenburg have an agreement about mutual cooperation with financing of the dictionaries SAOL and SO by the Academy until the end of 2028. During this period, a print version of SAOL will be published, and the dictionary portal Svenska.se (2025) will be revised and updated in alignment with the new SAOL edition, the revision of SO 2021, and the now completed SAOB.

The publication of SAOL 15 in print was considered a bit old fashioned by some users, while others cherish the book features and are convinced that there will still be books and print dictionaries in 2125. However, the plans for SAOL 16 are still in their infancy. SAOL 15 post print will most likely be revised with regard to correction of obvious errors – always present – and perhaps a reprint will be published. However, sales figures might affect the outcome of future plans for print. The Danish *Retskrivningsordbogen*, published in late 2024 in print, will be updated digitally from now on, and the latest e-version will be the authorized one (Dansk Sprognævn 2025). It is not impossible that a similar suggestion will be made for forthcoming versions of SAOL, although it is also likely that there will be a print SAOL 16.

4 https://www.isof.se/flersprakighet/publikationer/lexin-ordbocker (last accessed: April 4, 2025)

5.4 Closing remarks

When the first edition of SAOL was published more than 150 years ago, the lexicographers probably had little knowledge about what the glossary users might do with the final product. And, as mentioned, when is a dictionary considered a "final product" anyway? SAOL has over the years increased its lemma count from 34,000 to about 128,000 (SAOL 15). The page count has increased from about 300 to 1,600 pages, and the pieces of information accompanying most of the headwords are also more generous nowadays, from close to zero in the editions of the 20th century to at least POS information and inflectional suffixes plus often hints of semantics in SAOL 14 and 15.

SAOL is now mainly accessed through Svenska.se together with its close relatives SO and SAOB. The idea of a print dictionary as a standalone product is perhaps seriously challenged by the change of times and the preference for e-solutions. Nevertheless, there is actually no contradiction in producing a print dictionary and an e-version based on the same dataset.

References

Allén, Sture. 1981. The lemma-lexeme model of the Swedish lexical data base. In Burghard B. Rieger (ed.), *Empirical semantics II*, 376–387. Bochum: Brockmeyer.

Allén, Sture. 1986. Nytt och gammalt: Om arbetet med ordboken, ordlistan, grammatiken, symposierna och belöningarna [New and old: About the dictionary, the glossary, the grammar, the symposia, and the prizes]. In Sture Allén, Bengt Loman & Bengt Sigurd (eds.), *Svenska Akademien och svenska språket*, 240–273. Stockholm: Norstedts.

Allén, Sture, Bengt Sigurd & Bengt Loman. 1986. *Svenska Akademien och svenska språket: Tre studier* [The Swedish Academy and the Swedish language: Three studies]. Stockholm: Norstedts.

Bäckerud, Erik, Pär Nilsson & Emma Sköldberg. 2020. Så används Svenska Akademiens ordböcker på nätet: Implicit och explicit feedback från användarna [How the Swedish Academy's dictionaries are used online: Implicit and explicit feedback from the users]. *Nordiska studier i lexikografi 15*. 91–101.

Berg, Sture. 2009. Om ordböjning och SAOL Plus [On word inflection and SAOL Plus]. In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 139–165. Stockholm: Norstedts.

Berg, Sture & Yvonne Cederholm. 2001. Att hålla på formerna: Om framväxten av Svensk morfologisk databas [To keep to the forms: On the development of the Swedish Morphological Database]. In Sture Allén, Sture Berg, Sven-Göran Malmgren, Kerstin Norén & Bo Ralph (eds.), *Gäller stam, suffix och ord: Festskrift till Martin Gellerstam den 15 oktober 2001*, 58–69. Gothenburg: Meijerbergs arkiv för svensk ordforskning.

Berg, Sture, Louise Holmer & Ann-Kristin Hult. 2008. Saol Plus: A new Swedish electronic dictionary. *Proceedings of the European Association for Lexicography (EURALEX) 2008*. 291–296.

Berg, Sture, Louise Holmer & Emma Sköldberg. 2010. Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2010*. 567–576.

Blensenius, Kristian, Louise Holmer & Emma Sköldberg. 2021. SAOL 14 som rättesnöre: Diskussion om den senaste upplagan [SAOL 14 as a guiding standard: Discussion on the latest edition]. *LexicoNordica* 28: 39–58.

Borin, Lars & Louise Holmer. 2024. Tradita innovare, innovata tradere: The Gothenburg approach to computational lexicography. *Proceedings of the Huminfra Conference (HiC 2024)*. 41–50.

Buchmann, Franziska. 2015. Spelling dictionaries. In Philip Durkin (ed.), *The Oxford handbook of lexicography*, 310–324. Oxford: Oxford University Press.

Dalin, Anders Fredrik. 1850–1853. *Ordbok öfver svenska språket* [Dictionary of the Swedish language]. *Vol. I–II*. Stockholm: Self-published.

Dansk Sprognævn. 2025. *Retskrivningsordbogen* [The Orthographical Dictionary]. [Online resource] https://ro.dsn.dk/. Accessed on 2025-01-27.

Diamond, Graeme. 2015. Making decisions about inclusion and exclusion. In Philip Durkin (ed.), *The Oxford handbook of lexicography*, 532–545. Oxford: Oxford University Press.

Forsberg, Markus & Louise Holmer. 2024. Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text [Data access, methodological development and lexicographical work at Språkbanken Text]. *LexicoNordica* 31: 61–79.

Forsberg, Markus & Emma Sköldberg. Forthcoming. Ord med liknande kontext sökes! Om ordvektorers roll i svensk lexikografi [Words with similar context wanted! On the role of word vectors in Swedish lexicography]. *Den 17. konferansen om leksikografi i Norden*.

Gellerstam, Martin. 2009a. SAOL i många upplagor [SAOL in many editions]. In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 53–83. Stockholm: Norstedts.

Gellerstam, Martin (ed.). 2009b. *SAOL och tidens flykt. Några nedslag i ordlistans historia* [SAOL and the flight of the time. Selected milestones in the history of the dictionary]. Stockholm: Norstedts.

Gellerstam, Martin. 2009c. Vad är Svenska Akademiens ordlista? [What is the Swedish Academy Glossary?] In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 11–30. Stockholm: Norstedts.

Hannesdóttir, Anna Helga. 1998. *Lexikografihistorisk spegel: Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga* [History of lexicography reflected: The development of monolingual Swedish lexicography from the bilingual]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Haugen, Einar & Lars Borin. 2018. Danish, Norwegian and Swedish. In Bernard Comrie (ed.), *The world's major languages*, 3rd edn., 127–150. London: Routledge.

Holm, Pelle. 1951. *Nytt och gammalt i Svenska Akademiens ordlista* [New and old in the Swedish Academy Glossary]. Stockholm: Norstedt.

Holmer, Louise. 2016. *Grammatik i SAOL. En undersökning av grammatisk information i* Svenska Akademiens ordlista *över 130 år* [Grammar in SAOL. An investigation of grammatical information in the Swedish Academy Glossary over 130 years]. Gothenburg: MISS.

Holmer, Louise. 2022. *Neutrala substantiv på* -ande *i text och ordbok* [Deverbal neutral nouns ending in *-ande* in text and dictionary]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Holmer, Louise. in preparation. Så tycker språkvetare om SAOL: En attitydundersökning inför den femtonde upplagan av *Svenska Akademiens ordlista* [What linguists think about SAOL: An attitude survey for the upcoming fifteenth edition of the *Swedish Academy Glossary*]. Språkbanken Text blog post to appear at https://spraakbanken.gu.se/blogg. Gothenburg: University of Gothenburg.

Holmer, Louise, Ann-Kristin Hult & Emma Sköldberg. 2015. Spell-checking on the fly? On the use of a Swedish dictionary app. *Proceedings of the Electronic lexicography in the 21st century (eLex) 2015 conference*. 356–371.

Holmer, Louise, Ann Lilliestrom, Emma Sköldberg & Jonatan Uppström. 2024. Time to say goodbye revisited: On the exclusion of headwords from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 443–452.

Holmer, Louise & Emma Sköldberg. 2016. Synonymer.se i fokus: Om användningen av en svensk ordbokssajt [Synonymer.se in focus. On the use of a Swedish dictionary site]. *Svenskans beskrivning 34* 34. 215–228.

Isof.se. 2025. *Institutet för språk och folkminnen* [The Institute for Language and Folklore]. [Online resource] https://isof.se/. Accessed on 2025-01-26.

Johannisson, Ture. 1974. Ordlistans första nio upplagor – en historik [The first nine editions of the dictionary: A history]. In Ture Johannisson & Gösta Mattsson (eds.), *Svenska Akademiens ordlista under 100 år*, 3–54. Stockholm: Svenska språknämnden.

Landqvist, Hans, Emma Sköldberg & Louise Holmer. 2024. Hur kan *appendicit*, *blodförgiftning* och *hyperaktivitetssyndrom* behandlas? Medicinens fackområde i Svensk ordbok utgiven av Svenska Akademien [How can *appendicitis*, *blood poisoning* and *hyperactivity disorder* be addressed? The medical field in the *Contemporary Dictionary of the Swedish Academy*]. *Svenskans beskrivning 38: Del II*. 89–106.

Leopold, Carl Gustaf af. 1801. *Afhandling om svenska stafsättet* [Thesis on the Swedish way of spelling]. Stockholm: Svenska Akademien.

Malmgren, Sven-Göran. 1988. Almqvist, Dalin och den svenska definitionsordbokens födelse [Almqvist, Dalin and the birth of the Swedish definition dictionary]. In Gertrud Pettersson (ed.), *Studier i svensk språkhistoria*, 195–213. Lund: Lund University Press.

Malmgren, Sven-Göran. 2009. Tre viktiga språkförändringar speglade i SAOL [Three important language changes reflected in SAOL]. In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 119–137. Stockholm: Norstedts.

Malmgren, Sven-Göran. 2014. Svenska Akademiens ordlista genom 140 år: Mot fjortonde upplagan [The Swedish Academy Glossary thorough 140 years: Towards the fourteenth edition]. *LexicoNordica* 21: 81–98.

Malmgren, Sven-Göran & Emma Sköldberg. 2013. The lexicography of Swedish and other Scandinavian languages. *International Journal of Lexicography* 26(2): 117–134. DOI: 10.1093/ijl/ect008.

Mattsson, Gösta. 1974. Ordlistans tionde upplaga – riktlinjer och kommentarer [SAOL's tenth edition – guidelines and comments]. In Ture Johannisson & Gösta Mattsson (eds.), *Svenska Akademiens ordlista under 100 år*, 57–116. Stockholm: Svenska språknämnden.

Milles, Karin. 2013. En öppning i en sluten ordklass? Den nya användningen av pronomenet *hen* [An opening in a closed part of speech? The new use of the pronoun *hen*]. *Språk och stil: Tidskrift för svensk språkforskning* 23(1): 107–140.

Nilsson, Pär. 2023. *Lexikal betydelseutveckling i teori och praktik: En analys av fem definitionsformler i Svenska Akademiens ordbok och de semantiska förändringsmekanismer som de beskriver* [Lexical semantic change in theory and practice: An analysis of five semantic labels in the Swedish Academy dictionary and the meaning-changing mechanisms they describe]. Lund: Lund University. (PhD thesis).

Ordförteckning. 1916. *Ordförteckning över svenska språket utgiven av Svenska Akademien* [Inventory of the Swedish language published by the Swedish Academy]. Stockholm: Norstedts.

Östergren, Olof. 1919–1972. *Nusvensk ordbok* [Dictionary of present-day Swedish]. Stockholm: Wahlström & Widstrand.

Ralph, Bo. 2009. När ordboken blev en ordlista [When the dictionary became a list of words]. In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 31–52. Stockholm: Norstedts.

Rosqvist, Bodil. 2014. Hårt arbete *och* sträng vila*: Svenska kollokationer i lexikografisk och lexikologisk belysning* [*Hard work* and *strict rest*. Lexicographical and lexicological perspectives on Swedish collocations]. Gothenburg: University of Gothenburg. (PhD thesis).

Rydqvist, Johan Erik. 1850. *Svenska språkets lagar: Kritisk afhandling* [The laws of Swedish: A critical investigation]. Stockholm: Norstedts.

Santesson, Lillemor. 2001. Leopolds förteckning över främmande ord 1801: En diakronisk studie av ordurval och stavning [Leopold's inventory of foreign words 1801: A diachronic study of word selection and spelling]. *Språk och stil: Tidskrift för svensk språkforskning* 10: 87–128.

SAOB. 1898–2023. *Svenska Akademiens ordbok* [The Swedish Academy Dictionary]. Lund: Gleerups.

SAOL 1. 1874. *Ordlista öfver svenska språket utgifven af Svenska Akademien* [Glossary of the Swedish language published by the Swedish Academy]. 1st edn. Stockholm: P.A. Norstedt & söner.

SAOL 6. 1889. *Ordlista öfver svenska språket utgifven af Svenska Akademien* [Glossary of the Swedish language published by the Swedish Academy]. 6th edn. Stockholm: P.A. Norstedt & söners förlag.

SAOL 7. 1900. *Ordlista öfver svenska språket utgifven af Svenska Akademien* [Glossary of the Swedish language published by the Swedish Academy]. 7th edn. Stockholm: P.A. Norstedt & söners förlag.

SAOL 8. 1923. *Ordlista över svenska språket utgiven av Svenska Akademien* [Glossary of the Swedish language published by the Swedish Academy]. 8th edn. Stockholm: Svenska Bokförlaget & Norstedts.

SAOL 9. 1950. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 9th edn. Stockholm: Svenska Bokförlaget/Norstedts.

SAOL 10. 1973. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 10th edn. Stockholm: Norstedts.

SAOL 11. 1986. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 11th edn. Stockholm: Norstedts.

SAOL 12. 1998. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 12th edn. Stockholm: Norstedts.

SAOL 13. 2006. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 13th edn. Stockholm: Norstedts.

SAOL 14. 2015. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 14th edn. Stockholm: Norstedts.

SAOL Plus. 2007. *Svenska Akademiens ordlista 13* [The Swedish Academy Glossary]. (CD-rom) https://svenska.se/. Accessed on 2025-01-09.

SAOLhist. 2013. *SAOLhist*. [Online resource] Stockholm: Svenska Akademien and University of Gothenburg. https://spraakbanken.gu.se/saolhist/.

Sigurd, Bengt. 1986a. I väntan på en grammatik: Wellanders "Riktig Svenska" [Waiting for a grammar: Wellander's "Proper Swedish"]. In Sture Allén, Bengt Loman & Bengt Sigurd (eds.), *Svenska Akademien och svenska språket*, 216–222. Stockholm: Norstedts.

Sigurd, Bengt. 1986b. SAOL – rikslikaren [SAOL – the guiding standard]. In Sture Allén, Bengt Loman & Bengt Sigurd (eds.), *Svenska Akademien och svenska språket*, 192–215. Stockholm: Norstedts.

Sköldberg, Emma. 2023. "Varför står det olika i SAOL och i SO?" Om (bearbetning av) skillnader mellan Svenska Akademiens samtidsordböcker ["Why do SAOL and SO say it differently?" On (the revision of) the differences between the contemporary dictionaries of the Swedish Academy]. *Nordiska studier i lexikografi 16*. 349–361.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Ståhle, Carl-Ivar. 1970. Det nordiska rättstavningsmötet 1869 och hundra års svensk rättstavning [The Nordic congregation of orthography in 1869 and a hundred years of Swedish orthography]. In *Sprog i Norden*, 5–36. Copenhagen: Dansk Sprognævn.

Svensén, Bo. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Svenska Akademien. 2025. *About the Academy: Historical overview*. Stockholm: Svenska Akademien. https://www.svenskaakademien.se/en/the-academy/history.

Svenska.se. 2025. *Svenska Akademiens ordböcker* [The Swedish Academy dictionaries]. [Online resource] https://svenska.se/. Accessed on 2025-01-09.

Teleman, Ulf. 2003. *Tradis och funkis: Svensk språkvård och språkpolitik efter 1800* [Trad and mod: Swedish language cultivation and language policy after 1800]. Stockholm: Norstedts.

Wellander, Erik. 1939. *Riktig svenska: En handledning i svenska språkets vård* [Proper Swedish: A guide to the care of the Swedish language]. Stockholm: Norstedt.

Emma Sköldberg, Kristian Blensenius, and Louise Holmer

4 SO: the Swedish contemporary dictionary

Abstract: *Svensk ordbok utgiven av Svenska Akademien* ('The Contemporary Dictionary of the Swedish Academy'; SO in short) comprises approx. 65,000 headwords. It is a definition dictionary covering Swedish general language of today. The dictionary is corpus-based and includes, for example, detailed semantic descriptions, information on different types of word combinations, and pronunciation by way of audio files. It also provides historical information about the headwords. SO is the result of an active and continuous production of several dictionaries (starting in 1986). The second and latest edition of SO was published 2021 at the dictionary portal Svenska.se and as smartphone apps. The target groups of SO are native speakers and advanced learners of Swedish. The dictionary is primarily aimed at supporting the users in situations related to reception but also with production related tasks.

Keywords: contemporary dictionaries, SO, definition dictionary, Svenska.se, lexicography, Swedish Academy

1 Introduction

The definition dictionary *Svensk ordbok utgiven av Svenska Akademien* ('The Contemporary Dictionary of the Swedish Academy'; SO in short) constitutes the most complete description of modern Swedish vocabulary. The latest version is the 2021 edition (SO 2021) (the first version was titled SOB 1986) and work on updating the

Acknowledgments: The work on the dictionaries SO and SAOL is financed by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*, and the work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161).

Emma Sköldberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: emma.skoldberg@svenska.gu.se
Kristian Blensenius, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: kristian.blensenius@gu.se
Louise Holmer, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: louise.holmer@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-004

dictionary is ongoing (see Section 2.2).[1] In this chapter we present the work on this dictionary in the *Swedish Academy contemporary dictionaries* project at the University of Gothenburg (see Chapters 2 and 3 in this volume). We discuss how various corpora and tools, provided mostly by Språkbanken Text and partly by other organizations, are used in connection with the work to: (1) find, select, and describe new headwords; (2) update and further develop the description of already included headwords; and (3) exclude obsolete headwords from the dictionary. We also address the revisional work on different information categories in SO and what kind of data and methods are used in relation to them.

The outline of the chapter is as follows: in Section 2 we give an account for some central considerations in metalexicography. We also briefly present SO, its background, and what characterizes the dictionary today. Section 3 provides an overview of how Språkbanken's word research platform Korp and the National Library's text resources are used in the lexicographic work. The discussion is based on the previously mentioned editorial tasks related to both new and existing headwords. Section 4 is divided into five subsections, each addressing different types of information found in the entries: formal, semantic, syntactic-semantic, pragmatic, and historical properties of the headwords. The chapter concludes with some final remarks in Section 5.

2 Points of departure

To describe SO, we begin with a brief introduction to key characteristics of dictionaries, primarily using Svensén's terminology (Svensén 2009). Further, we provide a brief description of the origin and historical development of SO, leading to SO of today.

2.1 Metalexicographical considerations

Svensén's lexicographic handbook (Svensén 2009: 12–38) includes a detailed presentation of different types of dictionaries. Based on the format of a dictionary, a distinction is usually made, for example, between printed dictionaries and e-dictionaries such as dictionary apps and websites. A dictionary accessed on a computer, or a mobile device, has considerable advantages over its analogue predecessors. One

1 The authors of this chapter are editor in chief (Emma Sköldberg) and senior lexicographers (Kristian Blensenius and Louise Holmer).

obvious benefit is that e-dictionaries can allow more headwords to be included. The e-dictionary is also more dynamic and can be updated more easily (see, e.g., Rundell 2015).

Svensén (2009) also distinguishes between different types of dictionaries based on the number of languages they cover. While some dictionaries are monolingual, others are bilingual or multilingual. A dictionary can be applied to general vocabulary, a specific technical language (such as legal language), a geographically limited variety (such as Scanian, a dialect spoken in the south of Sweden) or words belonging to a certain style (such as slang). In addition, the dictionary may deal with many different aspects of the headwords (e.g., their pronunciation, inflection, and semantics) or apply to a certain aspect, e.g., their etymology. Focus can be on different language stages, e.g., Old Swedish or Modern Swedish and, in addition, the dictionary can be synchronic or diachronic. Dictionaries also adopt different perspectives on the lexical items that are included, as they have more or less normative ambitions. According to Svensén (2009: 24), the aim of normative dictionaries is to "influence usage".

When compiling new dictionary entries or revising existing ones, the lexicographers often focus on the different information categories usually found in a dictionary (cf. Svensén 2009: 6–8). The information categories can be divided in the following way:

- spelling, inflection, word class, pronunciation (formal information)
- meaning description, cross-references (semantic information)
- language examples, constructions, and phraseology (syntactic-semantic information)
- usage labels (pragmatic information)
- establishment, origin, and kinship (historical information).

In this chapter we discuss each of these categories in relation to SO, and particularly with the editors' continuous work on improving the data analysis and corpus-based methods.

A dictionary can largely be regarded as a tool. The lexicographer must consider the primary target audience, as the form and content of the dictionary must be adapted to fit the intended user(s). Furthermore, the lexicographer must regard the dictionary users' needs and how to meet these requirements, as it affects what information the dictionary should include (see Atkins & Rundell 2008: 28–33; Tarp 2008; H. Bergenholtz & I. Bergenholtz 2011; Rundell 2012 among others). For instance, semantic information is important in tasks related to *reception* such as understanding speech and writing. Information about pronunciation is most relevant in language *production*. In the case of documentation, historical information about the headwords plays a central role (see, e.g., Malmgren 2009b).

2.2 The history of SO

SO is a subset of a lexical database of approx. 214,000 entries managed and further developed at the University of Gothenburg (GU). Work related to the database began at GU in the 1970s in connection with the project *Lexikalisk databas* 'Lexical Database'. The project had two main aims: firstly, to "establish a well-structured lexicon stored in the form of a linked network (database)" and, secondly, that "the database is to be used as the basis for a new Swedish monolingual dictionary of definitions, Svensk ordbok" (SOB; Ralph, Järborg & Allén 1977: 3) (our translations). This dictionary contained 58,000 headwords (SOB 1986: VI). The more theoretical aim was thus just as central as the practical one, the creation of a dictionary. This is evident in the extensive and theoretically elaborate documentation of the structuring of the lexical content. One example concerns the definition format in the database: the definition should be "completely interchangeable with the definiendum [i.e., the word to be defined] in all syntactic (and also morphological) contexts, if necessary after application of some natural syntactic transformation" (Järborg 1989: 22) (our translation). This requirement meant that different word classes received different definition formats: a noun was to be described with an indefinite nominal phrase, an adjective with a relative clause or adjective phrase, a verb with an infinitive phrase, etc. (see also Josephson 2022: 51, 254–255 on the Lexical Database and its role within Swedish linguistics).

Moreover, as stated by Malmgren (1992: 486), "[b]asically, every word should be provided with a true (Aristotelian, if possible) definition; only in exceptional cases was it considered legitimate to give mere synonyms".

Based on the Lexical Database, a series of dictionaries have been produced, e.g., *Svensk ordbok* 'Swedish dictionary' (SOB 1986), *Nationalencyklopedins ordbok* 'The Dictionary of the National Encyclopedia' (NEO 1995) and *Svensk ordbok utgiven av Svenska Akademien* 'The Contemporary Dictionary of the Swedish Academy' (SO 2009). Over time, new headwords have been added, along with expanded information about already included headwords such as facts about their establishment, origin and constructional properties. For idioms, language examples of typical usage have been attached, while pronunciation transcription has been supplemented with audio files (see, e.g., Malmgren 1992; 2002; 2009a,b; Hult 2010; Holmer, von Martens & Sköldberg 2015; Borin & Holmer 2024).

2.3 SO today

Today's SO can briefly be characterized as a monolingual, synchronous, general-purpose, and mainly descriptive dictionary. With its approx. 65,000 headwords, SO

is aimed primarily at users with Swedish as their mother tongue, but also at adult learners with good knowledge of Swedish. Primarily, it aims to support the users with reception-oriented needs, while also assisting with production-oriented tasks (see also Malmgren 2009a,b).

Since 2010, the Lexical Database is owned by the Swedish Academy and the editorial work is regulated by an agreement between the Swedish Academy and GU. The editorial work takes place within Karp, Språkbanken's data editing platform (see Chapter 11 in this volume). GU is, among other things, responsible for ensuring that both new headwords and multiword expressions are added to the database, as well as keeping the description of already included headwords up to date. The entries should include information regarding the headword's spelling, pronunciation, inflection, stylistic value, and meaning, supported by language examples. Additionally, the database should provide information on the phraseology and construction of the headwords.

As mentioned, the second and latest edition of SO was published in 2021. Unlike previous editions, it was only published digitally, in the form of free downloadable apps and on the Swedish Academy's dictionary portal Svenska.se.[2] (See, e.g., Sköldberg 2022; 2024 for editorial positions concerning this edition, and Trap-Jensen 2022b for a review of the edition.)

On Svenska.se, the user consults three dictionaries at the same time: the more normative *Svenska Akademiens ordlista* (SAOL 14 2015), SO (2021), and the historical *Svenska Akademiens ordbok* 'The Swedish Academy Dictionary' (SAOB 1898–2023). (See also Lönnroth 2018 and Bäckerud, Nilsson & Sköldberg 2020.) On the portal, there is also a display mode where users have access to the SO entries only. An important advantage of the web version of SO is its easy accessibility while working on a computer, while the app versions provide improved search functionalities and features such as "Word of the day".

Figure 1 shows the interface of Svenska.se and the search results for the word *spray*. The SO entry, published in 2021, is placed in the middle column. The SAOL entry (shown in the left column) originates from the 14th edition from 2015. The SAOB entry, in the right column, was published in 1985.

Figure 1 reveals that the word *spray* is treated in partially different ways in the three dictionaries. Also, the spelling variants of the word are presented in different ways. In both SAOL and SAOB, *sprej* is the first form, and *spray* is the second form. As already mentioned, SAOL is more normative, thus often highlighting variants more in line with Swedish spelling conventions (see Chapter 3 in this volume). As for SAOB, the choice of the first form *sprej* has been influenced by the notation in the then

2 https://svenska.se/ (last accessed: April 4, 2025)

Figure 1: The entry *spray* (and *sprej*) in SO and the corresponding entries in SAOL and SAOB on Svenska.se

current edition of SAOL (see Larsson 2012 on normative aspects of the headwords in the historical dictionary). Today, however, the spelling *spray* is most frequent in Swedish texts and hence it is placed before *sprej* in the more descriptive SO.
In the following, we focus on the content of the SO entry *spray*, as shown in Figure 2. Headwords (along with morphological language examples) that are alphabetically in the vicinity of the word *spray*, are displayed in the drop-down menu to the right.

As shown in Figures 1 and 2, the headword is spelled in two ways: *spray* and *sprej*. The order of the variants signals that *spray* is more common in the texts that the SO lexicographers consult (see Section 3). The abbreviation "el." (Sw. *eller* 'or') between the variants indicates that they are considered almost equal in terms of, e.g., frequency and style (see Malmgren 2009b: 96–97; SO 2009: X). The entry also includes the definite singular form and the indefinite plural form for both variants: *sprayen sprayer* and *sprejen sprejer* 'the spray, sprays', respectively (for details on inflection, see Section 4.1). In addition, the headword's word class is presented. In this case, it is a noun, like the clear majority of the headwords in SO.

The pronunciation of the word is communicated both in writing and with linked audio files.[3] As pointed out by Sköldberg (2017: 126), it can be challenging for the

3 SO uses a sort of *respelling*, a phonemic transcription mainly by means of the characters of the Swedish alphabet (Svensén 2009: 117), and not IPA. For the convenience of the reader, the pronunciation examples in this chapter are rendered using IPA, however.

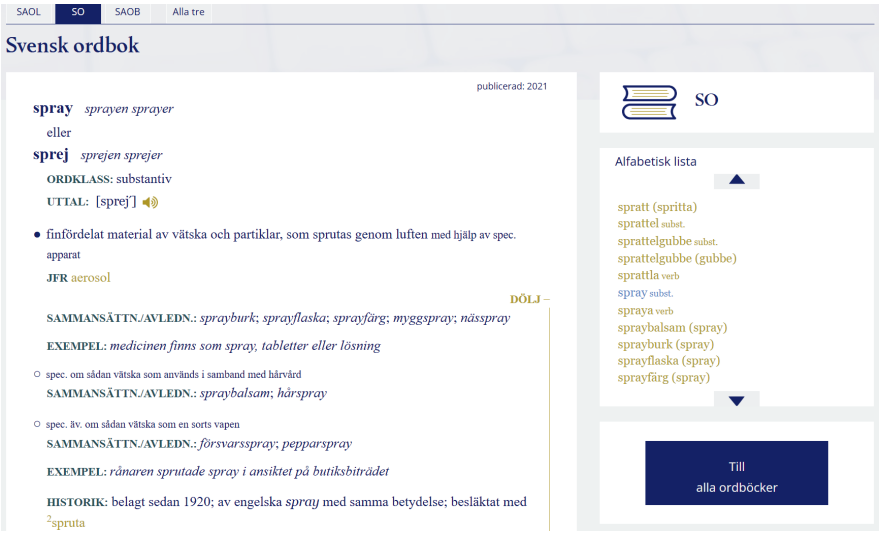


Figure 2: The headword *spray* (and *sprej*) in SO (2021) on Svenska.se

lexicographers to get data on how Swedish words are pronounced. This is, not least, due to the lack of large spoken Swedish corpora. However, the SO lexicographers have made the assessment that the word *spray* is pronounced [sprɛj:] in Swedish. The main sense is indicated by a black circle and conveyed through the definition "finfördelat material av vätska och partiklar, som sprutas genom luften" 'fine matter of liquid and particles, which is dispersed through the air' together with the definition supplement "med hjälp av spec. apparat" 'using special equipment', which is written in a smaller font (more on meaning descriptions in SO in Section 4.2).

The main sense is illustrated by five morphological examples, here compounds,[4] where *spray* forms the first or the last part (e.g., *sprayburk* 'spray can' and *näsспray* 'nasal spray'. There is also a syntactic language example: *medicinen finns som spray, tabletter eller lösning* 'the medicine is available as spray, tablets, or solution'. These examples complement the information provided by the definition. Also, the cross-reference to another headword in the dictionary, in this case the noun *aerosol*, contributes to the meaning description.

The headword has two subsenses, each marked with white circles. Both are stated to be special uses derived from the main sense. The subsenses are also illustrated by language examples. However, no pragmatic information is provided in the

4 According to Swedish orthography, compounds are written as one word, without spaces or hyphens delimiting the components.

entry, which indicates that the word has been assessed as unmarked with respect to style and usage. (See Svensén 2009: 315 on different ways of marking that a headword deviates from the majority of the lexical units described.)

Finally, in terms of establishment, origin, and kinship, it is stated that the headword has been used in written Swedish texts since 1920 and that it is of English origin. The English word *spray* has the same meaning. Etymologically, the headword is also related to the Swedish verb *[2]spruta* 'spray'.

In SO (2021), the lexicographers have utilized many of the advantages that are associated with an e-dictionary. For example, the cross-references in the form of hyperlinks have been developed. The inflectional forms are also displayed in full text (instead of condensed text). The edition of 2021 reaches a wider audience via Svenska.se than its printed predecessors. Today's SO users probably have varying experiences of dictionary use, and more users than before are likely to be learners of Swedish. Thus, they may need more help in understanding the information in the dictionary.

However, the SO tradition is strong and SO of today is still strongly influenced by the printed format. Rundell (2015: 303) states:

> [M]uch of what we take for granted as "natural" features of dictionaries are in reality expedients. They evolved not because they are the best possible way of conveying information to users, but because they satisfy the imperative of shoehorning large amounts of information into a limited space.

In other words, in an e-dictionary, the lexicographers do not have to use the same amount of textual condensation as in printed dictionaries. Many conventions that were developed in connection with the preparation of printed dictionaries persist in the editorial work and can be difficult to be aware of and overcome (see also Trap-Jensen 2022a and Holmer 2022: 76, 85–88).

As the contemporary dictionaries SAOL and SO are displayed side by side on Svenska.se, the respective functions of both dictionaries can be discussed. While SAOL traditionally takes a more normative approach, the objective among today's SO lexicographers is to refine the *descriptive* perspective in SO. SO (2021) is also more descriptive than SO (2009) (see, e.g., Sköldberg 2017).

Nevertheless, it is worth noting that it is hard to find purely descriptive or purely normative dictionaries. A normative dictionary has a descriptive basis and a descriptive dictionary represents the linguistic value judgments of the lexicographers. Thus, descriptive dictionaries also have a normative dimension (cf. Svensén 2009: 24). Trap-Jensen (2002: 64–65) makes the point that regardless of the individual dictionary's descriptive or normative focus, users often consider the content of the dictionary as correct.

3 Using computational lexicographic resources in dictionary compilation

SO has been corpus-based from the start. The lexicographers' access to Swedish corpora and other text resources has increased significantly since the work with the Lexical Database started, which has been crucial for ensuring the dictionary's quality. Currently, the lexicographers primarily use resources and tools provided by Språkbanken Text, the National Library of Sweden (*Kungliga biblioteket*; KB) and Mediearkivet ('The Media Archive')[5] in their work with SO. But the lexicographers also consult other sources such as the Swedish encyclopedia site Ne.se and search engines such as Google. In the following sections, we introduce these resources and tools briefly. We also present how they are used in relation to the three general tasks in the editorial work, i.e., finding and selecting new headwords and compiling new entries, revising existing entries and excluding headwords. (The use of the resources and tools will be further elucidated in connection with the presentation of different information categories in Sections 4.1–4.5.)

3.1 Resources and tools

Access to a large amount of text, as well as a rich variety of genres, is an important prerequisite for the SO lexicographers to be able to reflect present-day Swedish. At the time of writing (early 2025), the dictionary project at GU benefits from access to more than 16 billion tokens of modern texts, thanks to Språkbanken Text. The word research platform Korp includes corpora with text from, e.g., newspapers (approx. 848 million tokens), novels (approx. 20 million tokens), and social media (approx. 12 billion tokens) from the 20th and 21st centuries. Korp also includes the sub-corpus Kubord (see Chapter 10 in this volume about Korp). To manage these extensive amounts of data and find patterns in them, the lexicographers rely on different kinds of language technology methods and tools. Such methods and tools also make the work more efficient and reduce manual labor and subjectivity in the lexicographic process.

In Korp's search interface, there are three display modes: *the keyword-in-context (KWIC) concordance view*, *the statistics view*, and *the word picture view* (Borin, Forsberg & Roxendal 2012; see also Chapter 10 in this volume). The KWIC concordances make it possible for the lexicographers to review the contexts of a word and discern

5 https://www.retrievergroup.com/sv/product-mediearkivet (last accessed: April 4, 2025)

recurring patterns in how the word is used. They are also useful when searching for language examples. The statistics view makes it relatively easy for lexicographers to compare the frequency of different spelling variants and inflectional forms of a word. It also provides an efficient way to study compounds and derivations. Finally, in the word picture view, the searched word is displayed alongside other words in the corpus with which it stands in particular syntactic relations. For example, for a verb, lists of recurring subjects, objects, and adverbials can be found. The word picture view makes it even easier to review the typical context of a word, which is necessary in describing the headwords' syntagmatic properties. Word pictures can also clarify different senses of a word (see Chapter 10 in this volume on how Korp is used in dictionary compilation).

Another valuable resource in lexicographic work is the text material provided by the National Library of Sweden (KB). The web site *Svenska tidningar* ('Swedish newspapers') includes digitized newspapers and magazines from 1645 to today. With the help of a graph, it is relatively easy to get an overview of how the frequency of a word has evolved over the centuries.

Due to copyright reasons, access to the texts that KB manages is limited. For this reason, the *KB Lab* at the National Library and Språkbanken Text have created data collections based on contemporary newspaper texts that can support lexical research without infringing copyright. The collaboration between KB Lab and Språkbanken Text has, among other things, resulted in the data collections Kubord 1 (Språkbanken Text 2025a) and Kubord 2 (Språkbanken Text 2025b), that are available via Korp. In addition, another data collection, Kubord fastText (Språkbanken Text 2024a), is about to be publicly released.

Kubord 1 is based on roughly 80 subcorpora with various press materials, a total of about 3 billion tokens. In this resource, occurrences in the texts cannot be displayed as KWIC concordances, but it is possible to obtain detailed frequency information about the words, and this goes a long way toward fulfilling certain lexicographic requirements (see further Section 3.2). Kubord 2, which broadly includes the same newspaper material as Kubord 1, provides the user with frequency information about the words, but also statistics on word pairs that stand in a syntactic (dependency) relation to each other, such as verb–subject or noun–attribute. This contextual information enables the users to have word pictures for the words that appear in the corpora (see Forsberg & Holmer 2024).

Kubord fastText plays an important role in connection with studies of the role of word vectors in a lexicographical context. The work with word vectors provides the lexicographers with information about words whose linguistic contexts are similar. In this way, the word vectors supplement the information the lexicographer obtains through KWIC concordances and word pictures (see further Forsberg & Sköldberg 2022; Bouma et al. 2024; Forsberg & Sköldberg forthcoming).

Finally, another resource used in connection with lexicographic work is Mediearkivet. The advantage of this search service is that it offers a wide range of recent texts, including both news and specialized magazines, as well as radio and TV programs with transcriptions. Here, the lexicographers can find contemporary language use (words in context) across different types of media. The disadvantage is that Mediearkivet is not designed to be a linguistic tool: for example, the search capabilities are quite limited compared to Korp (for example, the Mediearkivet texts are not annotated for word class and syntactic structure).

3.2 Selection and inclusion of new headwords

Recently published texts, especially news texts, are essential when the lexicographers are searching for newly coined words that have become relatively established in general language. These new words may either become headwords or morphological language examples in the dictionary. To identify new words, as well as older ones that for some reason have not been captured and incorporated into previous editions of SO, the members of the dictionary project at GU use various methods. For example, they study logfiles of unsuccessful lookups in Svenska.se (see, e.g., Bäckerud, Nilsson & Sköldberg 2020: 95; cf. also Hult 2016: 95,155). Another important method involves comparing the vocabulary in an older corpus in Korp with the vocabulary in a more modern one and then focusing on the words that are exclusively found or more frequently used in the corpus with newer texts. These words are then compared with the headwords that are already included in the dictionary. By doing that, the lexicographers obtain a list of potential new headwords. The words on these lists illustrate how most Swedish words are formed. Many words also reflect our times, changes in society, new sciences and technologies, etc. (e.g., Karlsson 2021).

In connection with this kind of comparison, the subcorpora in Kubord 1 are particularly useful. The design of Kubord is well-suited for examining similarities and differences in the vocabulary between two years of the same newspaper, for example. The subcorpora in Kubord can also be compared with other types of corpora in Korp. Several more informal but well-established words not yet included in SO, have been clarified through comparisons between Kubord 1 and Poeter.se (Språkbanken Text 2024b). The latter corpus includes texts comprising about 106 million tokens from a website where writers can upload their own fiction, mainly poetry, but also novels, short stories, drama, etc. In the work with SO, it is important that the usage of more colloquial words is also investigated.

In connection with the selection of new headwords in SO, aspects like frequency in general language and distribution in different texts are taken into account. The lexicographers get this information through Korp. As for compounds, the degree

of semantic transparency also plays an important role. Lexicalized compounds are included as headwords in SO, whereas more transparent compounds may serve as morphological language examples (SO 2009: IX).

The use of the words that are selected by the lexicographers to be included as new headwords is thoroughly analyzed and described in accordance with the agreement between the Swedish Academy and GU (see Section 2.3). Here, Språkbanken Text's corpora and tools play an important role. For example, when specifying spelling and inflection for a new headword, the statistics view in Korp is used (see Section 4.1 and Chapter 10 in this volume for examples). The KWIC concordances and the word picture view are important in the semantic analysis, as a starting point for the selection of language examples and for the information about the headword's constructional and phraseological behavior. As a source of historical information, especially regarding the year of first evidence in Swedish written texts, the web site Svenska tidningar is primarily used. It contains material from nearly 2,000 different newspapers and journals. The entire newspaper and journal material is typically used, and the results are normally sorted by date by the lexicographers. There are quite a few OCR errors in the material, so the lexicographers must sometimes consult facsimiles of individual pages.

3.3 Revision of existing entries

As already stated, the content of SO has developed over decades. A large proportion of the entries in SO were included already in SOB, the predecessor of SO (see Section 2.2), and many of these entries as well as newer ones need to be updated, not least because the language has changed since the entries were compiled. As an example, consider the pronunciation of headwords that are loanwords, e.g., the headword *router* which, according to SO (2009), is pronounced ['ru:tɛr] or ['raʊtɛr]. There may have been some variation in the pronunciation of the noun during the compilation of SO (2009), but now the pronunciation ['raʊtɛr] is considered to be the generally prevailing one. It is also the only variant found in SO (2021). See Section 4.1 for further information on pronunciation.

Furthermore, a word may have acquired a new sense during the last decades. In some cases, this has resulted in the addition of a new main sense in SO, e.g., the second main sense of the noun *mandolin*, which designates a kitchen utensil ('mandoline'). This sense appeared for the first time in SO (2021). In other cases, the semantic change of a word results in a new subsense in the dictionary. This applies for the noun *hatare* 'hater'. According to SAOB (1898–2023), the word has been used since 1541 in Swedish texts. The traditional sense and the new use in social media have many features in common, but the new sense is considered to be special and

therefore it now forms the basis for a subsense in SO (2021). The work of finding new meanings has traditionally been based on manual methods, but lexicographers now also use language technology methods to more systematically identify meanings that are not yet recorded in SO (see Section 4.2).

Thanks to larger and more varied corpora, different kinds of language changes are noticed more easily by today's lexicographers than by lexicographers of earlier times. For example, corpora with texts from social media may include spelling variants that are not found in corpora based almost exclusively on editorial texts. The relationship between the already mentioned spelling variants *spray* and *sprej* can be mentioned here. In previous editions, including SO (2009), the indication "*sprej* eller ['or'] *spray*" was used, but in SO (2021) the order of the variants has been reversed. The indication "*spray* eller *sprej*" reflects more accurately how the word is spelled in various present-day Swedish texts.

Another headword in SO (2009) that has been revised is *kajennpeppar* 'cayenne pepper' with the alternative spelling form *cayennepeppar*. In more descriptive SO (2021), only *cayennepeppar* is presented, due to the fact that form *kajennpeppar* is used very rarely in the corpora which the SO lexicographers consult (0.0 hits per million tokens, compared to *cayennepeppar*, with 0.5 hits per million tokens). In the more normative SAOL 14 (2015), however, the users find both spelling variants.

Another reason why older entries in SO need to be revised is that the dictionary is now exclusively digital. Today's lexicographers aim to take full advantage of the digital format, including creating more connections among entries by incorporating cross-references in the form of links.

To ensure the lexicographic quality of entries already included in SO, it is fruitful to approach the dictionary content from several perspectives. The review of the data can benefit from, e.g., studying dictionary entries from the letters A–Ö or by studying one category of information in the entries at a time (e.g., pronunciation, inflection, and usage labels). As an example, during the revision process in preparation for SO (2021), the lexicographers conducted a comprehensive review of the information regarding the headword's constructional behavior. This information is typically found in verb entries, but is also present in many noun and adjective entries (see further Blensenius 2019) (see also Section 4.3).

Thematic studies of headwords related to different domains may also be fruitful (see, e.g., Landqvist, Sköldberg & Holmer 2024). For instance, an extensive editorial review before the second edition of SO concerned the headwords related to pedagogy. Between different editions of the dictionary, society, including the Swedish educational system, has changed, and these changes are also reflected in, e.g., Språkbanken Text's corpora. Grading systems, perspectives on education, and the relationship between teachers and students have evolved, and these changes must be reflected in the entries. Otherwise, the content of SO will not be perceived as up to date.

As mentioned, the SO lexicographers utilize the KB newspaper material to date new headwords, i.e., to identify their first occurrence in Swedish texts. Over time, it has become clear that the dating in many existing entries can be revised, as access to written materials is now significantly better than when the information about a headword's establishment in Swedish was originally added to the database and published in NEO (1995).

It is not uncommon for dictionary users to inform the lexicographers that a headword has been in use in Swedish texts for a longer time than stated in SO. By examining the texts available through the website Svenska tidningar, the historical information can be updated accordingly. In the next edition of SO, several headwords will include revised information about their establishment in Swedish (see also Section 4.5).

3.4 Exclusion of headwords

The vocabulary is constantly evolving, and it is expected that the use of some of the headwords in different versions of SO has diminished or nearly ceased. These headwords are at risk of being removed from SO. Lists with so-called "exclusion candidates" are prepared by comparing the headwords in SO with the vocabulary in corpora of contemporary texts in Korp, using automatic language technology tools. All in all, the corpora consist of about 3 billion tokens (modern newspaper texts, texts from the Swedish public service television company, texts from Swedish Wikipedia, and texts published on the website Poeter.se).

A comparison of the headwords in SO (2021) with the vocabulary in the corpora reveals that SO includes a large number of words that do not appear in the texts in any of their inflected forms (cf. Holmer et al. 2024 for a similar comparison between the headwords in SAOL and the vocabulary in the same text material). A more qualitative review of SO headwords that are absent in the corpora, shows that many of them consist of designations for objects, professions, etc. that are hardly used or out of date, e.g., *raderkniv* 'eraser knife' and *kakelugnssättare* 'tile stove setter'. There are also technical headwords with very limited use in general language texts, e.g., the medical noun *kardioskleros* 'cardiosclerosis' and the musical adverb *rallentando* 'rallentando'. The use of words like these will be further investigated by the lexicographers before possible exclusion from the dictionary.

However, it is debatable how many SO entries need to be retired since also the next edition of the dictionary will be published only digitally. This entails that most headwords can remain in SO, even if they are less typical of modern language usage. On the other hand and considering the display of the three dictionaries on Svenska.se, it can be questioned if the synchronous SO needs to account for historical

words that are described in more detail in SAOB. Dictionary users who are interested in older Swedish words which are rare in present-day Swedish can, in addition to consulting SAOB, search in the online resource *SAOLhist* (SAOLhist 2013),[6] which includes, e.g., the first edition of SO (SO 2009).

To sum up, the SO lexicographers' access to different kinds of textual material is crucial given that the dictionary is largely corpus-based. The lexicographers need access to extensive corpora in order to find new words, study the distribution of words, investigate unusual words, analyze semantic change over time, to name a few aspects. The text materials are also the basis for language examples in the dictionary and for the dating of the headwords.

4 Information types in SO (a selection)

In the following, we discuss a selection of information categories in SO, their characteristics, and the empirical data and tools used in the lexicographic work related to these categories. The categories, such as spelling and inflection, are grouped based on the type of information they contribute to the entries (see Section 2.1).

4.1 Formal information

The formal information presented in the SO includes spelling, inflection, word-class information and pronunciation.

To investigate usage patterns, Korp is used, particularly its word-form statistics view. Since entries in Saldo (see Chapter 6 in this volume) are provided with morphological information and inflectional full-form generation, a lemgram lookup is usually enough to find sufficient information about a word's inflectional behavior. This method reveals both common forms and potential usage variations. Given SO's descriptive approach, frequency heavily influences how alternative forms are presented, with the most common forms listed first.

Other text databases, such as Mediearkivet, are also consulted for specific forms, including specialized periodicals and transcribed radio programs. Sampling from these sources provides additional insights.

After deciding to include a headword, its spelling must be determined. As mentioned above, Språkbanken Text's corpora and tools are crucial in this process. Primary corpora include news texts, but fiction, academic writing, and social media

6 http://spraakdata.gu.se/saolhist/ (last accessed: April 4, 2025)

are also analyzed. For instance, the statistics view in Korp is utilized to determine spelling of new headwords. Often, there is only one spelling that dominates in terms of frequency (this applies, for example, to the noun *trigger* 'trigger'). In many cases, however, variation occurs. An example is the adjective *tätslutande* 'tight-fitting' with the variant form *tättslutande*. In such cases, the lexicographer's intuition is combined with tools like a Korp CQP query of the type [word = "tä.*slutande"], which, depending on the corpus selection, may yield varying frequency distributions. In this case, however, the forms are fairly evenly distributed in terms of frequency over different corpora, so it was decided that the shorter *tätslutande* would be listed before the longer *tättslutande*, since other entries are listed that way in SO (i.e., with the shorter form first, e.g, *tätsittande* 'tight-fitting' before *tättsittande*). The notation "el." 'or' is placed between the alternative word forms, signaling equivalent alternatives.

It is important to investigate if there are additional words containing the same varying compound member so that the words are treated consistently, e.g., *tätbebyggd*/*tättbebyggd* 'densely built-up'. The number of variant spellings is now essentially limited to one. For example, *rullad* 'roulade' had two variant forms in earlier editions, *rulad* and *roulad*, but the less frequent *rulad* has since been removed.

Inflectional forms are presented as follows: nouns are generally shown in their definite singular and indefinite plural forms (e.g., *hunden*, *hundar* 'dog.DEF, dog.PL'), verbs in their past tense and supine forms (*hoppade*, *hoppat* 'jump.PST, jump.SUP'), and adjectives in their neuter and plural forms (*gult*, *gula* 'yellow.N, yellow.PL'). Exceptions exist, such as uninflected nouns (e.g., *fanders* 'devil', 'hell'), verbs lacking specific forms (*stinga* 'sting', noted with "preteritum undviks" meaning 'the past tense is avoided'), and adjectives that seldom take a neuter form (*rigid* 'rigid'). The specifics of these variations are listed in the inflection field.[7]

Irregular inflections like *gå*, *gick*, *gått* 'go, go.PST, go.SUP' and *bra*, *bättre*, *bäst* 'good, better, best' pose minimal challenges due to their association with older, well-documented words. The likelihood of encountering them with new words is low. However, inflection patterns can shift over time, often favoring weak inflection over strong. For instance, while *simma*, *sam*, *summit* 'swim, swim.PST, swim.SUP' is recognized, the weak *simma*, *simmade*, *simmat* is now more common, with the less frequent forms marked as "äv." 'also' in SO. Lexicographers also face challenges with words that exhibit multiple inflection patterns of different kinds, and a key question is how many variants should be included in a descriptive dictionary like SO compared to the more normative SAOL. For example, the entry for *papper* in SAOL lists forms like *papperet* and *pappret* (both meaning 'the paper') while SO

7 Glosses in this chapter follow the Leipzig Glossing Rules. The one abbreviation used here not found in the Leipzig Glossing Rules is SUP 'supine (verb form)'.

also includes definite plural forms like *papperna* and *papprena* (both meaning 'the papers'). Balancing the breadth of forms with normative standards remains an open issue.

Another category involves words influenced by foreign inflectional patterns, such as the *-s* plural in *enchiladas* and *wraps*. For newer words like *boomer* 'boomer', *piñata* 'piñata', and *Youtuber* 'Youtuber', identifying usage patterns requires tools like frequency searches in modern corpora. In the case of *piñata*, forms like *piñator* 'piñatas' are also used but less frequently.

Certain plural forms stand out in frequency, such as *britter* 'Brits' and *pengar* 'money', which are more common than their singular forms. These cases are noted with "usually plural" in SO. Research by Blensenius & Martens (2020) reveals additional plural forms, such as *räkor* 'shrimps' (far more frequent than *räka* 'shrimp') and *trakter* 'regions' (more frequent than *trakt* 'region'), underscoring the value of word-form statistics for identifying trends.

Some forms, such as the colloquial plural form *avocadosar* 'avocados' (Holmer & Blensenius 2023), may not appear in the Saldo paradigms used for Korp's annotations and hence require targeted searches. For instance, searching specifically for *avocadosar* in Korp confirms its presence, even if it is absent from the Saldo paradigm for the lemma *avocado*. Similarly, regional forms like *lös* 'shone' (past tense of *lysa* 'shine') may not have been included in the Saldo paradigms for these lemmas, posing additional challenges for lexicographers.

The set of word classes in SO is not explicitly outlined in the dictionary's preface. The user is informed that the dictionary provides information about word-class membership for all entries, and it is merely stated that the set of word classes aligns with the "traditional classification", with one exception. As in the Swedish Academy grammar (SAG; Teleman, Hellberg & Andersson 1999), the older word class of conjunctions is divided into two separate classes: subordinating conjunctions and coordinating conjunctions (SO 2009: XII). The word-class categorization indeed follows SAG in several ways, such as often treating adjectives ending in *-t* as adjectives rather than *t*-adverbs (e.g., the adverbial usage of *lodrätt* 'vertically' as derived from the adjective *lodrät* 'vertical'). However, the word class participle from SAG is not adopted. Overall, the handling of word classes in SO presents a challenge for lexicographers.

As already mentioned, SO offers nearly exhaustive pronunciation information in the form of phonetic transcriptions, including inflected forms when needed. The pronunciation information closely mirrors standard spelling, primarily using regular letters. To conserve space, it is assumed that users are familiar with basic pronunciation rules, so some words are considered to not require pronunciation guidance (e.g., monosyllabic words like *väg* 'road'). However, in the digital versions of SO, audio recordings are available for every word, including alternative forms and al-

ternative pronunciations (like *mozzarella* 'mozzarella', which is provided with three alternative pronunciations).

Due to the fact that the corpus tools do not contain much spoken language, the lexicographers can, if necessary, consult services with recorded material. One such service is Mediearkivet, since it also includes up-to-date recordings of both podcast and TV programs, along with speech transcriptions. Here, one can, for example, search for how a particular word is pronounced, but large-scale studies of pronunciation cannot be conducted.

4.2 Semantic information

Before defining the meaning of a new headword or revising the semantic details of an existing one, the lexicographer must analyze its usage across various corpora to determine whether it should be assigned a single sense or divided into two or more senses. This decision is usually based on morphological, syntagmatic, semantic, paradigmatic, and pragmatic criteria (see, e.g., Atkins & Rundell 2008: 263–315 and Svensén 2009: 205–211). In this context, Lew (2013: 287) highlights two editorial strategies, known as *lumping* and *splitting*, for sense division in monolingual dictionaries. The lumping strategy aims to minimize the number of senses, with each sense encompassing as much semantic content as possible. Conversely, the splitting strategy results in a greater number of narrowly defined senses. The choice of strategy often depends on the type of dictionary being produced and its intended audience (Atkins & Rundell 2008: 267–268). Svensén also addresses the different ways in which senses of polysemous words are structured. According to the author (Svensén 2009: 211–212), the senses can be arranged linearly, i.e., as a number of discrete units in a sequence, or hierarchically, as a number of main senses, to which one or a group of subsenses are associated. For example, a word like *fotavtryck* 'footprint' has a literal and a figurative sense. When the polysemy structure is linear, these two senses are listed as sense 1 and sense 2. In a hierarchical structure, the literal sense might be labeled as sense 1, with the figurative sense listed as 1a, indicating that it is a subsense of the main sense.

In SO, senses are ordered hierarchically, following the guidelines in, e.g., Ralph, Järborg & Allén (1977) and Järborg (1989). In addition, the dictionary explicitly specifies the relationship between the main senses and their respective subsenses (in terms of meaning extension, specialization, figurative use, etc.). The distinction between different senses is relatively fine-grained. In other words, the SO lexicographers are splitters rather than lumpers (see also Rydstedt 2012 and Blensenius & Holmer 2022).

As stated, Ralph, Järborg & Allén (1977) and Järborg (1989) include information on definition format in SOB, the predecessor of SO (see Section 2.2), and the guidelines expressed in those publications still apply in many respects. But the guidelines are not exhaustive. For example, they do not address how interjections should be described (see further Sköldberg & Landqvist 2024). The current SO editorial team strives to uphold the tradition and the stated principles regarding meaning description etc., but sometimes the lexicographers choose to depart from the directives, often to simplify content and better accommodate the needs of new SO user groups.

As the semantic information in SO is central, an important part of the revision work between editions is to investigate *if* the sense of a headword has changed, and if so, *how*. Lexicographers often identify the need to revise the semantic details of an entry during various editorial tasks. For example, they may encounter word senses not previously recorded in SO when using the word picture view in search of collocations and other recurring word combinations. An example of this is the headword *bearbeta* 'process, refine'. One of its main senses, related to reshaping raw materials etc., got a new figurative subsense in connection with SO (2021). A word picture of the lemgram *bearbeta* in the corpus Kubord 2 shows that the word is often combined with objects like *sorg* 'sorrow', *trauma* 'trauma', *upplevelse* 'experience', and *känsla* 'feeling'. These words indicate that the verb is also used figuratively, which is supported by searches in other corpora. Words like *sorg* are now also included among the language examples illustrating the new subsense of the verb in SO.

Other types of close reading of selected parts of SO also yield valuable insights. For example, numerous interjection entries in the database have been revised based on a comparison between the existing entries and how interjections are used in social media texts from the Korp corpus (see Sköldberg & Landqvist 2024). In the next edition of SO, many interjection entries will not include definitions. Instead, the interjections will be described in terms of the function(s) they fulfill (see also Svensén 2009: 241).

The SO lexicographers are also made aware of shortcomings among the meaning descriptions by the dictionary users. A user comment concerned the headword *fruga* in SO (2021) with the sense 'wife (of a certain man)'. According to the user, the definition fails to reflect modern use of the word as it does not cover homosexual relationships. Like many other headwords that have been provided with more inclusive meaning descriptions in SO (2021), the meaning description of the noun *fruga* needs to be reformulated (see further Petersson & Sköldberg 2020).

In addition, SO's editorial members are involved in studies related to computer-aided methods for detecting semantic change in a more systematic way. For example, Sköldberg et al. (2024) have conducted different experiments using the annotation tool DURel which relies on language models for automatic semantic analysis of word usages (Sköldberg et al. 2024: 159; see also Schlechtweg, Schulte im Walde & Eckmann

2018 and Schlechtweg et al. 2024 about the tool). In short, DURel automatically annotates usage pairs based on semantic proximity and then generates semantic clusters, each representing a distinct sense. For these experiments, the SVT corpus in Korp with texts published between 2004 and 2021 has been used.

In one of the experiments, Sköldberg et al. (2024) focused on headwords in SO with only one listed sense. They then examined the number of semantic clusters that DURel created for each headword based on uses in the SVT corpus. When the tool identifies two or more semantic clusters for a headword, it may suggest that the headword has more than one sense in the texts, indicating that the semantic description in SO should be revised. Such headwords can be prioritized for manual inspection during revision work for the next edition.

The experiments have so far drawn the SO lexicographers' attention to subsenses (mainly figurative uses of headwords, but also meaning extensions and specializations) that should be incorporated into the dictionary. One example is the noun *slutspurt* 'final sprint' which, according to SO (2021), means "sista del av spurt" 'last part of sprint'. For this word, DURel identified two semantic clusters. The SO lexicographers, who inspected the corpus samples manually, also confirmed that the headword has at least two different senses. It can refer to the final stage of a competition, but it can also be used figuratively, for example about the last days before a national election. The figurative use of the word is not clearly represented in the current edition of SO and will be added in the next update.

The SO editorial team has also experimented with word vectors to see how they can strengthen the description of the headwords in the dictionary. The studies have shown that there are semantically interesting neighbors in the vector spaces of the examined headword. The neighbors can, for example, form the basis for an increased number of cross-references in the SO entries (see further Forsberg & Sköldberg 2022; Bouma et al. 2024; Forsberg & Sköldberg forthcoming). These cross-references consist primarily of synonyms, antonyms, and co-hyponyms to the headword and, along with the definitions, they constitute an important part of the semantic description of the entries. They are also, according to Malmgren (2009b: 98), "extremely important from a production-oriented point of view". Although the quality of SO is generally high, the recent studies mentioned above support the results by, e.g., Blensenius, Sköldberg & Bäckerud (2021), who state that the semantics in SO can be further developed. Obviously, computer-aided methods can streamline and improve lexicographic work in several ways.

4.3 Syntactic-semantic information

As mentioned above, the SO entries include both morphological and syntactic language examples (see the examples illustrating the use of *spray* in Figure 2). These examples are intended to complement the meaning descriptions and provide insights into the phraseology of the headwords (see further Atkins & Rundell 2008: 452–461, Svensén 2009: 281–297 among others). While many of the examples are more or less direct quotations from corpora, SO also features "semi-authentic examples" as well as editorially created ones (Malmgren 2009a: 16).

In connection with the revision of SO, it has become obvious that many syntactic language examples need to be replaced on an ongoing basis for the content of SO to continue to be perceived as modern. This has led to a large number of outdated language examples in SO (2009) being updated or replaced by new ones that better reflect today's society, new technical aids etc. Some older examples, which have come to be replaced, are also characterized by sexism or gender stereotypes (see further Sköldberg 2020). New examples have also been added to support the construction specifications in the entries. In connection with this revision work, the amount and variety of corpora in Korp have been decisive (see further Sköldberg 2022; 2024).

Over time it has also become clear that the syntactic examples in different versions of SO include a considerable number of collocations, like *brinnande intresse* 'burning interest' and *fatta ett beslut* 'make a decision'. The information about this type of word combination is particularly valuable for language learners, especially in relation to production of Swedish. Thanks to the word picture view in Korp, many previously undocumented collocations have been identified (see further Sköldberg 2022). In order to clarify the information about collocations and further emphasize them, the treatment of the word combinations has been further refined in later editions. In SO (2009), collocations were grouped according to semantic principles in the first part of the section with language examples (see Malmgren 2009b: 98–99). In SO (2021), certain dictionary entries containing multiple collocations (e.g., the headword *läkemedel* 'medicine'), are provided with a dedicated collocation section. This addition not only highlights the collocations but also draws greater attention to the free-standing language examples.

Idiomatic expressions like *ta bladet från munnen* 'to say frankly what one means or thinks' have been part of the Lexical Database since it was established (Järborg 1989: 3). In the 2009 edition, special consideration was taken to improve their description. This category, in total about 4,600 multiword expressions, includes a range of more or less fixed phrases like idioms but also similes, some proverbs and multiword technical expressions. During the revision process, the SO lexicographers carefully considered under which headwords these expressions should be placed. Also, most of the expressions were paired with corpus-inspired language examples (Malmgren

2009a: 18; Hult 2010), which probably contributed to them being significantly easier to understand and use than before.

During the development of the lexical database that would serve as the foundation for SO, plans were made to include constructional information. This information was intended to complement, among other things, the syntactic structure of definitions and syntactic examples by specifying features such as valency (Ralph, Järborg & Allén 1977: 32). The syntactic constructions were to be represented using code labels, with a key provided in the introduction of the dictionary (Ralph, Järborg & Allén 1977: 53), similar to the labeling systems in older dictionaries, such as *vb tr.* for 'transitive verb.' However, this type of information was not included in SOB. It was not until the release of NEO that such valency indications were introduced, making it the first L1 dictionary in the Nordic countries to provide this information (Malmgren & Toporowska Gronostaj 2009: 183). The constructional information was further developed in SO (2009) and later in SO (2021), which included subject specifications for several verbs. Constructional annotations were applied not only to the main senses but also to subsenses.

In SO (2021), constructional information specifies, for instance, the types of objects that transitive verbs take and the prepositions associated with various verbs, adjectives, and nouns. An illustrative example can be found in the entry for the verb *regna* 'rain'. The subject varies depending on the meaning level, alternating between an expletive subject (*det* 'it') and a referential subject (NGT, NGRA 'something', 'some things', respectively), and the verb's transitivity also changes across subsenses. See Figure 3, where the constructional information is underlined.

To determine which constructional features should be included, the primary resource used is Korp's word picture search (see Section 4.2 and Chapter 10 in this volume for examples of such a search). In some cases, additional searches are conducted to examine concordances in greater detail.

4.4 Pragmatic information

As for usage, a significant portion of the headwords in SO are restricted in one way or another, and this is expected to be stated in the relevant entries. For instance, lexicographers may need to comment on the style or register of a word, and such information is provided in square brackets in the entries. As an example, the adjective *påtänd* 'high, stoned' is labeled with "vardagligt" 'everyday language'. The noun *mammon* 'mammon' is, on the contrary, provided with the usage label "högtidligt" 'solemn'. Furthermore, the label "något ålderdomligt" 'somewhat old-fashioned' is found in connection with a number of headwords, e.g., *lebeman* 'man who lives lavishly and often extravagantly'. The same applies to the label "ålderdomligt" 'old-fashioned'

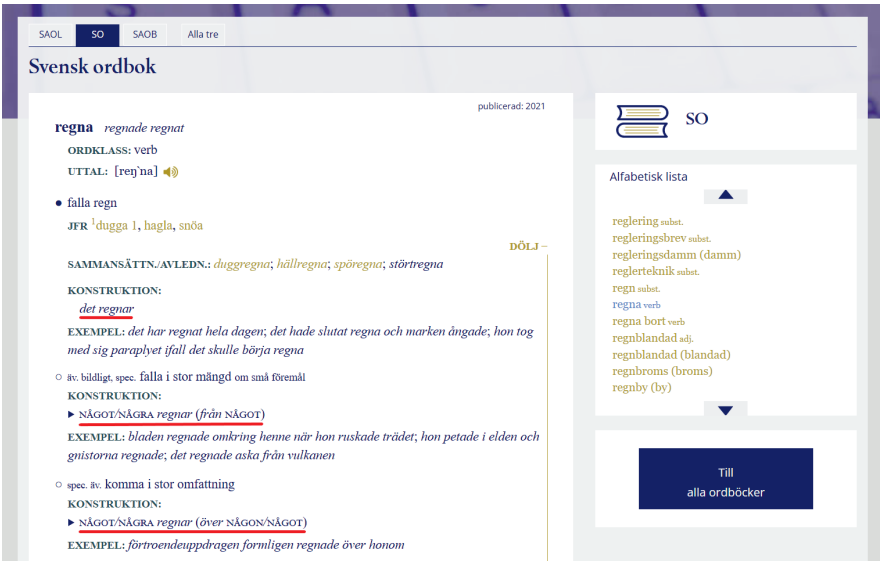


Figure 3: Constructional information underlined for the main sense of *regna* 'to fall as rain': *det regnar* 'it rains', for the subsense 'to fall in large quantities': NÅGOT/NÅGRA *regnar* (*från* NÅGOT) 'something/-some things rain (from something)', and the subsense 'to come in great abundance': NÅGOT/NÅGRA *regnar* (*över* NÅGON/NÅGOT) 'something/some things rain (over someone/something)', respectively

as in the case of *lungsiktig* 'having (predisposition to) pulmonary tuberculosis'. In some cases, lexicographers also deem it important to inform users about a word's negative emotive charge. For instance, the adjective *tvålfager* 'good-looking in a slick way' is combined with the usage label "nedsättande" 'derogatory'. (Labels such as "vardagligt" and "ålderdomligt" are commonly used, but it should be noted that there is no set of predefined labels.)

SO (2021) also states that if a headword, or a specific sense of a headword, is used within a particular field such as psychology or chemistry. The definition of a headword like *dränage* 'drainage' is supplemented by the usage label "medicin 'medicine'. In the same way, the subsense "densitet" 'density' of the noun *täthet* is accompanied by the usage label "fysik" 'physics', specifying its field of application.

A comparison between the three dictionaries funded by the Swedish Academy reveals that the sets of labels are not identical. Moreover, the style and the emotive charge of a headword can change over time. A noun like *tjej* 'girl, woman' has become more stylistically neutral and the noun *indian* 'American Indian, Native American' has come to be associated with more negative connotations in public discourse. For lexicographers, it is somewhat easier to detect a change related to frequency in modern texts or even a stylistic change, based on searches in different corpora. It

is more challenging to establish that a word has acquired negative connotations (Malmgren 2009b: 98). Nonetheless, in the SO team's recent work, both the set and the use of usage labels in the dictionary have been carefully examined and modernized with the goal of making the labels and associated information clearer and more accessible for dictionary users (see Petersson & Sköldberg 2020: 384; see also Trap-Jensen 2022b: 202–205 for a review of this work).

4.5 Historical information

All headwords in SO are provided with information about the time of their first known occurrence in Swedish written texts. For instance, the dictionary notes that relatively modern headwords like *ortorexi* 'orthorexia', *responsiv* 'responsive', and *wow* 'wow' have been used in Swedish texts since 1998, 1919 and 1932, respectively.

Furthermore, SO is (since its first edition 2009) likely the first of its kind among general dictionaries in the Nordic countries to include comprehensive datings from the very earliest sources of evidence in the "ancient language" (Sw. *fornspråket*). The dictionary also indicates the sources of these dates (see Lövfors 2010: preface). For example, the first record of the headword *sol* 'sun', an inscription found on a rock engraved with runes, dates back to the 10th–11th century. Another example is *stövel* 'boot' which has been in use since at least the year 1405. The first record of this word is, according to SO, in a will drawn up by a man named Jösse Johansson.

A large proportion of the headwords are additionally provided with etymological information. According to Malmgren (2009a: 17), 25,000 etymologies were added in NEO (1995). For example, the verb *anskaffa* 'acquire' is stated to have its origin in the German word *anschaffen* with the same meaning, while the word *artificiell* 'artificial' originates from the Latin word *artificialis* and entered Swedish through French.

These etymologies can aid in text comprehension (Lövfors 2010: 7). One example is the etymology of *angelägen* 'keen, anxious, eager', which is explained as follows: "av lågtyska, tyska *angelegen* med samma betydelse, eg. 'som ligger intill, om hjärtat'" ('from Low German, German *angelegen* with the same meaning, actually 'which lies close to, about the heart'). This information about the word's origin can help dictionary users better understand its meaning.

However, as noted by Svensén (2009: 342), the etymologies in SO may be challenging for the users as they tend to be highly condensed. Additionally, the dating in many existing entries may need to be updated due to better access to older texts, especially through the Svenska tidningar site. Despite this, updating the dates and refining etymologies are not currently priorities for the dictionary project, according to the agreement between the Swedish Academy and GU. That said, some entries in the next edition of SO will include updated years of first attestation, for example

barbecue (current year in SO: 1989; updated year: 1868), *espresso* (current year in SO: approx. 1960; updated year: 1924), and *tsunami* (current year in SO: 1973; updated year 1942).

5 Summary and outlook

In this chapter, we have outlined the empirical word research that forms the basis of the ongoing lexicographic work within a project at Språkbanken Text at the University of Gothenburg. The work within the project aims, among other things, to further develop the contemporary dictionary *Svensk ordbok utgiven av Svenska Akademien* (SO). The latest edition of this work was published in 2021 and work on an updated version of the dictionary is ongoing.

We have demonstrated how different types of corpora and tools, provided by Språkbanken Text, as well as other types of text material made available by the National Library, are used in the updating process. The content of various textual materials is central when creating new dictionary entries, revising existing ones, and deciding which entries should be excluded from a more descriptive contemporary dictionary. For instance, examples from different types of corpora are essential when providing information about inflection and semantics.

The resources and tools are applied in various ways and often complement one another. As mentioned earlier, selecting appropriate headwords is a crucial task. Once the words are selected, they must, after careful analysis, be provided with information about spelling, pronunciation, inflection, meaning, combinatorial properties, history, etc. In addition, language examples must be selected, which show how the headwords are used. Editorial work also involves revising existing entries, which applies to multiple categories of information. In other words, lexicographic work draws upon various linguistic research fields. But it also involves interdisciplinary challenges that intersect with digital humanities, particularly those related to the processing, analysis, and presentation of large, complex data sets.

As is well known, practical lexicographic work is bound by tradition, and all the editions of SO follow certain principles, which are presented by, e.g., Ralph, Järborg & Allén (1977) and Järborg (1989). However, these principles were established several decades ago and much has changed since then, especially with the increased access to various types of corpora and language technology-based tools. In addition, SO has transitioned from being a printed dictionary to a fully digital resource, which is displayed alongside SAOL and SAOB. This shift has expanded the dictionary's reach to different types of users with varying needs. Factors such as these naturally influence how the current editorial board approaches the foundational decisions made in the

past. This may lead the lexicographer to deviate from the once established definition formats to make the description of meaning more comprehensible.

Språkbanken Text's corpora, as well as the newspaper materials available via the National Library, have many strengths. The materials serve as excellent support in connection with certain lexicographic tasks, not least when the lexicographer wants to study how a word is spelled or inflected. It is relatively straightforward to compare possible variants and thereby determine which forms should be included in the dictionary and in which order they should be presented. The text materials are also invaluable when lexicographers look for information on how words are combined with one another or when lexicographers want to find examples of how individual words or expressions are used. Nevertheless, it is more challenging to determine what the connotations of a word are based solely on the textual materials that are available today. In addition, the lack of spoken language corpora can create difficulties in obtaining information on how words are pronounced and how distinctly spoken words, such as many interjections, are used.

In any case: thanks to the current corpora at Språkbanken Text, the texts available through the National Library, and modern language technology-based methods and tools, the lexicographic work with SO has become more scientifically grounded and efficient.

References

Atkins, B. T. Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Bäckerud, Erik, Pär Nilsson & Emma Sköldberg. 2020. Så används Svenska Akademiens ordböcker på nätet: Implicit och explicit feedback från användarna [How the Swedish Academy's dictionaries are used online: Implicit and explicit feedback from the users]. *Nordiska studier i lexikografi 15*. 91–101.

Bergenholtz, Henning & Inger Bergenholtz. 2011. A dictionary is a tool, a good dictionary is a monofunctional tool. In Pedro A. Fuertes-Olivera & Henning Bergenholtz (eds.), *e-Lexicography: The internet, digital initiatives and lexicography*, 187–207. London/New York: Continuum.

Blensenius, Kristian. 2019. Revision av konstruktionsuppgifter i Svensk ordbok utgiven av Svenska Akademien [Revision of construction descriptions in The Contemporary Dictionary of the Swedish Academy]. *LexicoNordica* 26: 203–223.

Blensenius, Kristian & Louise Holmer. 2022. How do verbal constructional alternations reflect (sub-)sense distinctions in dictionaries? A case study of a Swedish monolingual dictionary. In Kristian Blensenius (ed.), *Valency and constructions: Perspectives on combining words*, 9–30. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Blensenius, Kristian & Monica Martens. 2020. Analys av relativa ordformsfrekvenser för en bättre ordbok [Analysis of relative word form frequencies for an improved dictionary]. *Svenskans beskrivning 37*. 56–69.

Blensenius, Kristian, Emma Sköldberg & Erik Bäckerud. 2021. Finding gaps in semantic descriptions: Visualisation of the cross-reference network in a Swedish monolingual dictionary. *Proceedings of the Electronic lexicography in the 21st century (eLex) 2021 conference*. 247–258.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. *International Conference on Language Resources and Evaluation (LREC) 2012*. 474–478.

Borin, Lars & Louise Holmer. 2024. Tradita innovare, innovata tradere: The Gothenburg approach to computational lexicography. *Proceedings of the Huminfra Conference (HiC 2024)*. 41–50.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg. 2024. Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten [The art of conducting Swedish lexical research without violating copyright]. *Proceedings of the Huminfra Conference (HiC 2024)*. 161–167.

Forsberg, Markus & Louise Holmer. 2024. Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text [Data access, methodological development and lexicographical work at Språkbanken Text]. *LexicoNordica* 31: 61–79.

Forsberg, Markus & Emma Sköldberg. 2022. Ordvektorer i lexikografiskt arbete [Word vectors in lexicographic work]. In Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (eds.), *Live and learn: Festschrift in honor of Lars Borin*, 37–41. Gothenburg: Department of Swedish, Multilingualism, Language Technology, University of Gothenburg.

Forsberg, Markus & Emma Sköldberg. Forthcoming. Ord med liknande kontext sökes! Om ordvektorers roll i svensk lexikografi [Words with similar context wanted! On the role of word vectors in Swedish lexicography]. *Den 17. konferansen om leksikografi i Norden*.

Holmer, Louise. 2022. *Neutrala substantiv på* -ande *i text och ordbok* [Deverbal neutral nouns ending in *-ande* in text and dictionary]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Holmer, Louise & Kristian Blensenius. 2023. Okynniga pluraler: Normering och bruk av s-plural speglat i SAOL och SO [Insolent plurals: Standardization and the use of s-plural reflected in SAOL and SO]. *Nordiska studier i lexikografi 16*. 153–164.

Holmer, Louise, Ann Lilliestrom, Emma Sköldberg & Jonatan Uppström. 2024. Time to say goodbye revisited: On the exclusion of headwords from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 443–452.

Holmer, Louise, Monica von Martens & Emma Sköldberg. 2015. Making a dictionary app from a lexical database: The case of the Contemporary Dictionary of the Swedish Academy. *Proceedings of the Electronic lexicography in the 21st century (eLex 2015) conference*. 32–50.

Hult, Ann-Kristin. 2010. Kort och gott: Om idiomens språkprov i *Svensk ordbok utgiven av Svenska Akademien* (2009) [In short: On the language sample of idioms in the Contemporary Dictionary of the Swedish Academy (2009)]. *Nordiska studier i lexikografi 10*. 209–222.

Hult, Ann-Kristin. 2016. *Ordboksanvändning på nätet: En undersökning av användningen av Lexins svenska lexikon* [Dictionary use on the internet: Empirical studies of the use and users of the Swedish Lexin dictionary]. Gothenburg: University of Gothenburg. (PhD thesis).

Järborg, Jerker. 1989. *Betydelseanalys och betydelsebeskrivning i Lexikalisk databas: Preliminär version* [Semantic analysis and semantic description in the Lexical Database]. (Research Reports from the Department of Swedish No. GU-ISS-89-01) Gothenburg: Department of Swedish, University of Gothenburg.

Josephson, Olle. 2022. *Språkpolitik* [Language policy]. 2nd edn. Stockholm: Morfem.

Karlsson, Ola. 2021. *Lesserwisser*, *lårskav* och *läppstiftseffekt*: presentation och problematisering av urvalskriterierna för den svenska nyordslistan [*Lesserwisser*, *lårskav* and *läppstiftseffekt*: Presentation and problematization of the selection criteria for the Swedish new words list]. *LexicoNordica* 28: 101–120.

Landqvist, Hans, Emma Sköldberg & Louise Holmer. 2024. Hur kan *appendicit*, *blodförgiftning* och *hyperaktivitetssyndrom* behandlas? Medicinens fackområde i Svensk ordbok utgiven av Svenska Akademien [How can *appendicitis*, *blood poisoning* and *hyperactivity disorder* be addressed? The medical field in the *Contemporary Dictionary of the Swedish Academy*]. *Svenskans beskrivning 38: Del II*. 89–106.

Larsson, Lennart. 2012. Varför inte dub(b)lettformer? Om SAOB som normativ ordbok [Why not doublet forms? On SAOB as a normative dictionary]. *Nordiska studier i lexikografi 11*. 385–395.

Lew, Robert. 2013. Identifying, ordering and defining senses. In Howard Jackson (ed.), *The Bloomsbury companion to lexicography*, 284–302. London: Bloomsbury Publishing.

Lönnroth, Harry. 2018. Portalen svenska.se: en ny digital samlingsplats för språkresurser från Svenska Akademien [The portal svenska.se: A new digital hub for language resources from the Swedish Academy]. *LexicoNordica* 25: 281–292.

Lövfors, Sven. 2010. *Historiska angivelser i allmänna ordböcker: Med särskild hänsyn till* Svensk ordbok utgiven av Svenska Akademien [Historical references in general dictionaries: With special attention to the Contemporary Dictionary of the Swedish Academy]. Gothenburg: Institutionen för svenska språket, Göteborgs universitet.

Malmgren, Sven-Göran. 1992. From Svensk ordbok ('A dictionary of Swedish') to Nationalencyklopediens ordbok ('The Dictionary of the National Encyclopedia'). *Proceedings of the European Association for Lexicography (EURALEX) 1992*. 485–491.

Malmgren, Sven-Göran. 2002. Lexicography in the Nordic countries: Traditions and recent developments. *Proceedings of the European Association for Lexicography (EURALEX) 2002*. 39–54.

Malmgren, Sven-Göran. 2009a. Från Nationalencyklopedins ordbok (1995–96) till Svensk ordbok utgiven av Svenska Akademien (2009) [From The Dictionary of the National Encyclopedia (1995–96) to the Contemporary Dictionary of the Swedish Academy (2009)]. *LEDA-Nyt* (47): 14–20.

Malmgren, Sven-Göran. 2009b. On production-oriented information in Swedish monolingual defining dictionaries. In Sandro Nielsen & Sven Tarp (eds.), *Lexicography in the 21st century: In honour of Henning Bergenholtz*, 93–102. Amsterdam: John Benjamins.

Malmgren, Sven-Göran & Maria Toporowska Gronostaj. 2009. Valensbeskrivning i svenska ordböcker – och några andra [Valency description in Swedish dictionaries – and some others]. *LexicoNordica* 16: 181–196.

NEO. 1995. *Nationalencyklopedins ordbok* [The Dictionary of the National Encyclopedia]. Höganäs: Bra böcker.

Petersson, Stellan & Emma Sköldberg. 2020. To discriminate between discrimination and inclusion: A lexicographer's dilemma. *Proceedings of the European Association for Lexicography (EURALEX) 2021*. 382–386.

Ralph, Bo, Jerker Järborg & Sture Allén. 1977. *Svensk ordbok och lexikalisk databas: Förstudierapport* [The dictionary *Svensk ordbok* and the lexical database: A pilot study report]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Rundell, Michael. 2012. 'It works in practice but will it work in theory?': The uneasy relationship between lexicography and matters theoretical. *Proceedings of the European Association for Lexicography (EURALEX) 2012*. 47–92.

Rundell, Michael. 2015. From print to digital: Implications for dictionary policy and lexicographic conventions. *Lexikos* 25: 301–322.

Rydstedt, Rudolf. 2012. *En matchningsdriven semantisk modell. Mellan ordboken och den interna grammatiken* [A match-driven semantic model: Between the dictionary and the internal grammar]. Gothenburg: University of Gothenburg. (PhD thesis).

SAOB. 1898–2023. *Svenska Akademiens ordbok* [The Swedish Academy Dictionary]. Lund: Gleerups.

SAOL 14. 2015. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 14th edn. Stockholm: Norstedts.

SAOLhist. 2013. *SAOLhist*. [Online resource] Stockholm: Svenska Akademien and University of Gothenburg. https://spraakbanken.gu.se/saolhist/.

Schlechtweg, Dominik, Sabine Schulte im Walde & Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies (NAACL/HLT) 2018: Volume 2 (Short Papers)*. 169–174. DOI: 10.18653/v1/N18-2027.

Schlechtweg, Dominik, Shafqat Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi & Sabine Schulte im Walde. 2024. The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. *Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL) 2024: System Demonstrations*. 137–149.

Sköldberg, Emma. 2017. Innehållet i Svensk ordbok utgiven av Svenska Akademien – eller kampen mellan norm och bruk [The content of the Contemporary Dictionary of the Swedish Academy – or the struggle between norm and usage]. In Saga Bendegard, Ulla Melander Marttala & Maria Westman (eds.), *Språk och norm. Rapport från ASLA:s symposium*, 123–129. Uppsala: ASLA: Svenska föreningen för tillämpad språkvetenskap.

Sköldberg, Emma. 2022. Andra upplagan av Svensk ordbok: Förutsättningar och redaktionella val [Second edition of the Swedish dictionary: Conditions and editorial choices]. *LexicoNordica* 29: 139–152.

Sköldberg, Emma. 2024. Andra upplagan av Svensk ordbok (SO): Förutsättningar, teoretiska överväganden, insatser och mottagande [Second edition of the Swedish dictionary (SO): Conditions, theoretical considerations, efforts, and reception]. *Svenskans beskrivning 38*.

Sköldberg, Emma & Hans Landqvist. 2024. *Sorry*, *shit* and *wow*: A case study of the handling of interjections in three Nordic monolingual dictionaries. *Lexicographica: International Annual for Lexicography* 40(1): 29–57.

Sköldberg, Emma, Shafqat Virk, Pauline Sander, Simon Hengchen & Dominik Schlechtweg. 2024. Revealing semantic variation in Swedish using computational models of semantic proximity: Results from lexicographical experiments. *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 169–182.

SO. 2009. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. Stockholm: Svenska Akademien.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Språkbanken Text. 2024a. *Kubord-fasttext*. [Data set]. DOI: 10.23695/SP99-9H02.

Språkbanken Text. 2024b. *Poeter.se*. [Data set]. DOI: 10.23695/GMCG-MZ48.

Språkbanken Text. 2025a. *Kubord 1*. [Data set]. DOI: 10.23695/VB85-Y587.

Språkbanken Text. 2025b. *Kubord 2*. [Data set]. DOI: 10.23695/CKMP-PV98.

Svensén, Bo. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Tarp, Sven. 2008. *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Tübingen: Max Niemeyer Verlag.

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik* [The Swedish Academy grammar]. Stockholm: Norstedts.

Trap-Jensen, Lars. 2002. Normering og deskription i Den Danske Ordbog – mere eller mindre? [Standardization and description in the Danish Dictionary: More or less?] *LexicoNordica* 9: 63–78.

Trap-Jensen, Lars. 2022a. Researching lexicographical practice. In Howard Jackson (ed.), *The Bloomsbury handbook of lexicography*, 2nd edn., 19–30. London: Bloomsbury Academic.

Trap-Jensen, Lars. 2022b. Svensk ordbok – anden og reviderte udgave [Swedish dictionary: Second and revised edition]. *LexicoNordica* 29: 197–213.

Part III: **Lexical resources for machines**

Lars Borin

5 Swedish FrameNet++: an integrated network of lexical resources

Abstract: *Swedish FrameNet++* was a decade-and-a-half-long research and development initiative with the aim of creating a large and diverse computational lexical macroresource for Swedish (both modern and historical varieties), to be used in a range of automatic computational text processing applications. The resulting resource, also called Swedish FrameNet++ (SweFN++), forms a network of individual lexical resources, interlinked by the use of a set of formally structured persistent identifiers for word senses, lexemes (including multiword expressions), inflectional paradigms, and a set of lexical-semantic relations (synonymy, hyponymy, etc.). In this chapter we describe the history of SweFN++, its design principles, and its chief component resources, as well as its present-day successor, *Språkbanken's Lexical Research Infrastructure*.

Keywords: dictionary, lexical database, lexical infrastructure, lexical resource, Saldo, Swedish FrameNet

1 Background: towards a Swedish computational lexical macroresource

Språkbanken Text's research infrastructure platforms (see Chapters 9 and 10 in this volume) apply language technology (LT; or natural language processing: NLP) tools to carry out various forms of automatic linguistic analysis of textual data. These tools in turn rely on linguistic knowledge in some form in order to accomplish this analysis. This knowledge may be hidden inside the black boxes that are the large language models (LLMs) of recent artificial intelligence (AI) fame, learned

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161). Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded to me in 2024 for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-005

automatically on the basis of huge amounts of raw language data (typically scraped *en masse* from the internet) and often not even straightforwardly expressible in any traditional linguistic conceptualization of language. However, explicitly expressed linguistic knowledge still seems to have a role to play in NLP (Guo, Xu & Ritter 2024; Pedersen et al. 2024), in particular when the requisite massive amounts of language data are unavailable, something that is arguably the case for all but a few of the world's languages. For this reason research is ongoing in the NLP field aiming to find ways of informing LLMs (or complementing them) with more traditional linguistic knowledge.

The richest and most central form of linguistic knowledge about a language in this connection is arguably *lexical knowledge*, that is, knowledge about the *vocabulary* (and lexical morphology) of the language. This is traditionally treated as one kind of linguistic knowledge, in addition to phonology, grammar (inflectional morphology and syntax), semantics, and pragmatics. However, for many intents and purposes the vocabulary *is* the language, where each lexical unit brings with it relevant parts of the other kinds of knowledge, as it were, simply because this is what it means to know a lexical item as a native speaker of a language (Swartz 1992: 223). The task of linguists and lexicographers is then to formulate this multifaceted lexical knowledge in an explicit way using the time-honored conceptual apparatus of their field of inquiry.

Hence it makes sense to draw upon the rich lexical information traditionally investigated and formulated by professional lexicographers compiling dictionaries for humans, and turn it into formally structured lexical resources for use by computer algorithms, i.e., such resources that all the linguistic knowledge in them is not only machine-readable, but also machine-interpretable for use in automatic language analysis systems. This chapter describes our work – still ongoing – that started out with the aim of building a unified Swedish lexical macroresource by recycling and merging existing "orphan" resources, in parallel with creating a new lexical resource, the Swedish FrameNet (see Chapter 7 in this volume). The latter was estimated to require the bulk of the project effort; hence, it determined the name of the whole project.

With time and following some organizational changes in Språkbanken Text, this activity has now grown into a general lexical research infrastructure comprising lexical resources and dictionary databases,[1] as well as computational tools for working with various kinds of lexical data.

1 In the terminology used in this volume, a *dictionary* and its underlying *lexical database* are intended for human consumption, while a *lexical resource* is intended for use in NLP systems. We further use the term *lexicon* to include all the three mentioned, i.e., dictionaries, lexical databases, and lexical resources, as well as some dual-use datasets.

Swedish FrameNet++ (SweFN++) is the umbrella term referring to the first period – roughly 2008–2021 – of this initiative, that during that time constituted the main computational lexicographical activity at Språkbanken. This warrants the inclusion of the present chapter in this volume, despite the fact that a recent anthology (Dannélls, Borin & Heppin 2021) describes the SweFN++ work in detail from several angles. In part, this chapter summarizes the main points of Dannélls, Borin & Heppin (2021), but it also extends it and places it in the context of this volume. In particular, we describe very relevant conceptual links to earlier work at Språkdata on lexical databases, that we did not discuss or even consider in our previous account (other than very indirectly; see Section 2.2), but that have become obvious in connection with the work on the present volume.

2 Computational lexicography in Språkbanken: the compleat word-hoard, and then some

2.1 2008–2021: toward the compleat word-hoard …

2.1.1 Foundations, projects, and outcomes

As described in Chapter 2 in this volume, the former Språkdata research group split into two separate units in the early years of this millennium. While the new Center for Lexicology and Lexicography (CLL) in effect turned away from language technology and worked out different bespoke lexical database solutions for a series of dictionaries, plans for an interlinked computational lexical resource were elaborated in the renewed language-technology oriented Språkbanken. Drawing inspiration from high-profile international initiatives such as Princeton WordNet and Berkeley FrameNet, and using the in-house product Saldo as a basis and pivot resource, a project plan for this was formulated and set into motion. Under the name *Swedish FrameNet++* (SweFN++) and with funding from many different sources this project went on for almost a decade and a half. In a sense it is still ongoing, having now morphed into what we refer to as *Språkbanken's Lexical Research Infrastructure* (SBLRI).

In 2008, the first version of Språkbanken's new lexical resource Saldo was released, and plans were made for its inclusion in our corpus import pipeline. Notably, Saldo was not built from scratch: instead we recycled a lexicon compiled elsewhere – by a linguist at Uppsala University – for different purposes, and complemented it with an inflectional morphological component developed in-house at Språkbanken. Further, the bulk of the entries in the Uppsala lexicon had in fact at one point been

acquired from Språkdata in the form of the lemma list of *Svensk ordbok* (SOB 1986). See Chapter 6 in this volume for more details.

This suggested to us that there could be other, unexplored opportunities for recycling lexical data among datasets resulting from earlier research activities, where the funding had expired. Research funding, whether awarded by external funding agencies or by universities, generally has never (at least not in Sweden) catered for long-term upkeep of resources developed in a research project. This has not been seen as the responsibility of funding agencies. Even if these resources do stay around in some digital archive, they are typically not actively maintained. We conducted a preliminary survey, from which it emerged that a number of such lexical resources had been created for Swedish, either by digitization of preexisting paper dictionaries or through compilation directly into a digital format, in some cases even more than once. Because of the differing requirements and aims of the research projects – both language technological and linguistic – in which these resources had been created, they naturally tended to be heterogeneous as to their content and format. The SweFN++ initiative was devised as a concerted effort to save these resources from the slow spiraling into oblivion resulting from storage formats and software going out of use, and thereby to ensure that the considerable efforts and public funds spent on their development would not have gone to waste.

The initiative also partly arose out of an ambition to seize the opportunity to take advantage of the momentum and experiences gained from the work on Saldo. One of our objectives was to realize two kinds of lexical-semantic resource for Swedish, thereby putting Swedish language technology on a par with that of English and a few other large languages in this regard, namely a Swedish wordnet and a Swedish framenet (see Chapter 8 in this volume). We concluded from our survey that the raw stuff of a wordnet was already present to a considerable extent in existing resources, although in a form requiring a fair amount of format and information-model standardization. The framenet on the other hand would have to be built largely from scratch, although using the English Berkeley FrameNet frame definitions and frame structure as the point of departure (see Chapter 7 in this volume). This became the explicit focus of the planned initiative, even providing the name for it, as a project with the main aim of creating the Swedish FrameNet, to be one of the central lexical resources in an interconnected network of such resources, largely made up of recycled existing datasets, curated and extended to adhere to a standardized information model and data format, viz. those embodied in the system of persistent identifiers of the central resource Saldo (see Chapters 6 and 8 in this volume for details). This then was the "++" aspect of the initiative.

Saldo plays – literally – a pivotal role in SweFN++. It is designed to be the hub in a hub-and-spokes structured lexical resource, and hence used to connect all our other lexical resources into a versatile lexical macroresource providing lexical-

level information on linguistic form (morphology and syntax), linguistic content (word sense definitions and lexical-semantic relations), pragmatics (sentiment and connotation), and text occurrence data (word, lemma, and word sense frequencies). Saldo also connects the SBLRI to the Swedish ConstructiCon (Lyngfelt et al. 2018; Borin & Lyngfelt 2025; see also Chapter 14 in this volume).

Finally, the general enthusiastic spirit in which this work was undertaken was one fostered by Språkdata's long history of working with lexicographical data in a computational setting. This ensured that there was no dearth of competence that could be brought to bear on the lexicographical aspects of the project. The project was also timely, in coinciding with an increasing interest within the language technology field in high-grade lexical information, as evidenced by the use of wordnets, framenets, and thesauruses in text-analysis applications (see Chapter 8 in this volume), as well as the emergence of dependency grammar as the preferred formalism for syntactic analysis.

The SweFN++ work was thus conducted for close to a decade and a half, during the years 2008–2021, involving almost 20 contributing externally funded projects (Borin, Dannélls & Friberg Heppin 2021: 8). This more than anything else stands as a testimony to the central importance of lexical knowledge for almost all aspects of language and linguistic analysis. The main project – also named *Swedish FrameNet++* – received funding by the Swedish Research Council for the years 2011–2015.

The work on SweFN++ had an obvious infrastructural side, whose concrete manifestation was the resulting network of lexical resources that are still available through Språkbanken Text, and the construction of a language-technology informed computational infrastructure for editing lexical resources and for deploying them in automatic language analysis (see Chapters 6, 7, 8, 10, and 11 in this volume). The project also had a scientific side, evidenced by a large number of publications. Borin, Dannélls & Friberg Heppin (2021: 31–35) list 65 publications resulting from the various SweFN++ projects. Together with the 12 chapters in Dannélls, Borin & Heppin (2021) and some publications that were inadvertently omitted from the list, the SweFN++ activities yielded close to 80 publications over their course.

2.1.2 Vive la différence: practical, theoretical, and methodological considerations

The basic idea underlying SweFN++ was at heart a quite simple (and unoriginal) one. It stemmed from the observation that a large amount of high-quality and information-rich lexical datasets lay unused, sometimes due to prohibitive licensing conditions, but even open resources went unutilized, because of missing documentation, incompatible data formats, and other practical problems, but sometimes also because of differing theoretical conceptualizations of some linguistic phenomena. In our initial

survey, we identified about 15 existing lexicons that potentially could be included in SweFN++ (Borin, Dannélls & Friberg Heppin 2021: 10). These included onomasiological and semasiological lexicons of present-day Swedish, lexicons of historical language stages, and multilingual lexicons. Several of these originated in digitized paper dictionaries (see Chapter 12 in this volume for an example), others came in the form of lexical databases or digital dictionaries, and the remainder had been built as lexical resources in language technology projects, with different specific application areas in mind. This great variety naturally presented a challenge for standardization, but also offered an opportunity, allowing missing information to be filled in by assembling pieces of it present in different lexicons, and even allowed for the creation and extension of new component resources, for example the sentiment lexicon SenSaldo, which was partly automatically extended drawing on the lexical-semantic relations present in Saldo (see Chapter 8 in this volume).

SweFN++ gave us numerous opportunities to address three interlinked practical issues: how to harmonize lexical information across disparate resources; which kinds of information to include in the resources; and what kind of software support will most effectively move this work forward.

The SweFN++ initiative also had a strong theoretical dimension, by providing us with rich opportunities – at times even compelling us – to explore the linguistic implications of our work and its results, i.e., the relationship between (traditional Swedish) lexicography on the one hand, and descriptive linguistics, lexical semantics and lexical typology on the other. Two cases in point are the relationship of WordNet synsets to the word senses forming the lexical-semantic backbone of SweFN++ (see Chapter 8 in this volume) and the treatment of *multiword expressions* in SweFN++ (Borin 2021).

2.2 Interlude: the same, only different

Resource recycling is by no means a new idea. As already mentioned, Saldo – the core resource of SweFN++ – is the outcome of two successive rounds of reuse of preexisting lexicons (but a major additional effort – the creation of a computational morphology for Saldo – was required to make it into a useful lexical resource for our purposes; see Chapter 6 in this volume).

It also turned out that SweFN++ in some ways mirrored – or even recycled – a much earlier idea from the history of Språkdata. We were not aware at the time when we were drawing up the plans for our initiative of the grand – and uncannily similar in certain respects in its vision to our aims – earlier plans for a kind of lexical macroresource for dictionary compilation and computational processing of text that had been hatched already at the beginning of the Språkdata era.

Generally, Språkbanken Text has been characterized by a lexical focus throughout its history. Recall that the original impetus for the computational linguistic work at Gothenburg was a desire to compile better dictionaries. See Chapters 2, 3, 4, and 8 in this volume. The design and implementation of a lexical database was high on Språkdata's agenda already in the 1970s. The embryo of this database came out of the pioneering corpus-based work on the Swedish frequency dictionaries (NFO 1 1970; NFO 2 1971; NFO 3 1975; NFO 4 1980) and the *Swedish Academy Glossary* (SAOL; the 11th and 12th editions; see Chapter 3 in this volume), and it was designed to be the basis for compiling *Svensk ordbok* (Ralph, Järborg & Allén 1977; SOB 1986).

This database, later referred to as the *Gothenburg Lexical Database* (GLDB), contained all the lexical information that had gone into SOB (1986) and additional, both formal and semantic, lexical information, in a structure already well-suited for supporting the task of lexicography. Already from its start in the 1960s, the ambitions for the GLDB were considerably greater than simply to compile a corpus-based Swedish dictionary utilizing state-of-the-art computational tools such as database technology and interactive editing of dictionary entries. Even early on a close interconnection between the lexical database and the corpus search system was envisioned.

In several writings appearing throughout the 1970s and 1980s, Sture Allén outlines his vision of a "language bank" (*språkbank* in Swedish) consisting of a "text bank" and a "word bank" (Allén 1970; 1973; [1980] 1999a; [1983] 1999b; [1984] 1999c). The envisioned word bank is an enhanced and extended version of GLDB, which would grow in size and complexity both through traditional lexicographical work and through computational processing of the texts in the text bank. As far as we can see, there is no technical or other documentation of the word bank. It receives brief mention in an information booklet about Språkbanken from the mid-1990s:

> *Word bank construction*
> On the basis of the texts a word bank has been constructed, functioning as a kind of index to all words in texts or processed dictionaries (Svensk ordbok, SAOL, Nusvensk frekvensordbok, etc.). In addition to text words and lexical words (which may comprise ca. 500,000 items in total), there are also graphical words resulting from generation of the inflected forms of SAOL 11 (approximately 900,000 generated graphic words). Through the spell checking system CAPS, developed at Språkdata by Rolf Gavare and at present marketed by the company Omniterm, continuous neologism updates are made available from the typesetters' processing of new texts.[2] (Gellerstam & Sjögreen 1994: 5)

2 Sw. "*Uppbyggnad av ordbank*
På grundval av texterna har en ordbank byggts upp som fungerar som ett slags index till alla ord i texter eller ordboksbearbetningar (Svensk ordbok, SAOL, Nusvensk frekvensordbok osv.). Förutom textord och lexikonord (som tillsammans kan omfatta ca 500 000 enheter) finns också grafiska ord som tagits fram genom generering av böjningsformerna till SAOL 11 (det rör sig om ca 900 000 gener-

After this time, the word bank disappears from both internal documents and external publications. Apparently, it was never realized as envisioned; neither Gellerstam, Cederholm & Rasmark (2000) nor Allén (2002) mentions a "word bank", indicating that work on this resource had not progressed further. The connection to the word bank vision had most likely been lost when there was a generational shift among the lexicographers that coincided in time with the parting of ways of the lexicographical and language technological activities of former Språkdata mentioned above in Section 2.1.1. Sture Allén left his day-to-day engagement in the activities of Språkdata already in the late 1980s and retired officially in 1993. The lexicographers in CLL turned their attention to designing and populating successive lexical databases for the various dictionaries compiled by them. They had a background in Swedish linguistics and had adopted corpus linguistic methods while working on the first dictionaries produced at Språkdata. Their background did not incline them to embrace state-of-the-art language technology methodology.

Further, the work on computational semantic lexicons at the University of Gothenburg had followed a path that kept it apart from the mainstream international computational linguistic community. Because of the strong lexicographical tradition at Språkdata, the impulse to look to the Princeton WordNet for inspiration and guidance – so obviously present in many non-English computational linguistic semantic lexicon projects (e.g., Pedersen et al. 2019) – simply did not arise at Gothenburg. The onomasiological database *Semlex* (or SDB) was developed as a refinement of GLDB as a strictly in-house product, in response to linguistic rather than NLP research questions, and with its prohibitive intellectual property rights inherited from GLDB (see Chapter 8 in this volume).

SweFN++ was thus in effect conceived largely independently of these earlier Språkdata initiatives, because of the organizational and staffing discontinuity prevailing around the turn of the millennium. Språkdata fell apart in two units at a time that coincided with a radical turnover of staff because of a spate of more or less simultaneous retirements. For a considerable period, there was not much professional interaction between the lexicographical activities in CLL, having dictionary compilation as their principal aim, and the new Språkbanken, now primarily a language technology research unit with a notable infrastructure component.

erade graford). Via korrekturläsningssystemet CAPS, utvecklat vid Språkdata av Rolf Gavare och för närvarande marknadsfört av företaget Omniterm, kommer dessutom kontinuerliga uppdateringar av de nyord som framkommer vid sätteriernas körning mot nya texter." – my translation

2.3 … and then some: from SweFN++ to SBLRI

In 2021 the former Center for Lexicology and Lexicography became organizationally included in Språkbanken Text. Hence, the editorial staff of two main dictionaries of present-day Swedish – SAOL and SO (see Chapters 3 and 4 in this volume) – are now part of Språkbanken Text, and their work forms a considerable and important part of its activities.

A natural consequence of this organizational merger is that the two computational lexicography environments – developed separately with very little interaction during the two decades preceding the merger (see Section 2.2 above and Chapter 2 in this volume) – should be combined whenever feasible in order to avoid unnecessary repetition of effort. This is a process that requires some care, mainly because the SweFN++ work had only one purpose: automated linguistic analysis of text. This allowed us to ignore some lexicographically central aspects of the included dictionaries as being irrelevant or at least less relevant for this purpose (e.g., free-text definitions of word senses or pronunciation information).[3]

With two kinds of origin of lexical information – lexical databases vs. lexical resources – and two kinds of target purpose – compilation of dictionaries of present-day Swedish vs. automatic linguistic analysis of modern and historical Swedish texts – the lexical research infrastructure necessarily will become more complex. For example, some of the pieces of lexical information ignored earlier while building lexical resources should now be standardized and made editable by lexicographers. Also, the dictionaries compiled in Språkbanken Text are revised at regular intervals, primarily by adding new items and removing obsolete and obsolescing vocabulary. The lexical resources on the other hand simply grow in number of lexical items; normally, items are not removed from the lexical resources, only added.[4] The criteria for the addition of new items are also different. Our lexicographers rely to a large extent on corpus frequency data in order to determine whether a word or multiword expression should be listed in the dictionaries, the criterion basically being to a considerable extent a frequency threshold (see also Chapter 10 in this volume). For inclusion in a lexical resource such as Saldo – given that its main purpose is to provide an analysis of as many of the words in a text as possible – three main criteria

3 "Ignoring" these aspects simply means here that they played no role in the curation and processing of the lexicons containing them for inclusion in SweFN++. Regardless of whether they are used or not, all the information items present in the original lexicons have been kept.

4 Obsolete items are not necessarily removed from the lexical databases underlying the dictionaries, but may simply be marked as not to be included in a dictionary. However, unlike the lexical resources, the lexical databases are typically not open data.

are that the item appears in text somewhere, that it is deemed not to be a typo or misspelling, and that it is not a compositional compound or particle verb.

This means that the SweFN++ conceptual model where many sources of lexical information are funneled into one kind of output, as it were, will now change into one where we still have many inputs but also generate several outputs, i.e., one for automatic analysis and annotation of text, but now also one for each distinct dictionary.

3 Conclusion: looking ahead

It follows from the above that our longer-term ambition is that the content of the lexical databases built since the 1970s in Språkdata and CLL and the computational lexical macroresource built under the SweFN++ umbrella will be unified into one lexical research infrastructure that also includes an LT-based toolset for working with the lexical databases and resources. This requires harmonization of the respective information models, into what must in effect become a union of the two (or more) models, and there are enough differences between them that this will not be a trivial task. The upside of this is that in many cases the information from the dictionary databases will in fact enrich the lexical resources with information that probably would not have been available otherwise, while the (partial) introduction of the strictly formalized machine-interpretable structuring of the lexical resources makes it possible to devise formal consistency tests of the material going into the dictionaries.

We live in exciting times. Our field of enquiry is going through something of a methodological sea change brought about by the rapid development of AI in the form of LLMs, with unclear consequences for the traditional kinds of NLP system for which SweFN++ was designed in the first place. Most of the computational lexical resources described in this volume are classically structured – Saldo (Chapter 6 in this volume), the Swedish FrameNet (Chapter 7 in this volume), and the other onomasiological resources (Chapter 8 in this volume), as well as the 19th century morphological lexicon based on Dalin's dictionary (Chapter 13 in this volume) – and typically deployed in a traditional corpus import pipeline, with no LLM involvement (see Chapter 10 in this volume).

Consequently, the next big challenge – or opportunity – faced by computational lexicography has to do with its relationship to LLMs. However, as noted in Section 1, the kind of lexical knowledge found in human-produced lexicons does not lose its value in the age of LLMs. Hence, the need for lexicographical-expertise based dictionary compilation and revision will hopefully remain, and the results will be

put to good use by the various lexical resources as well as by LLM-based NLP systems. LLMs will undoubtedly also play an important role in dictionary compilation and editing; see Chapter 15 in this volume for an example of this.

References

Allén, Sture. 1970. Åtta teser om texthantering [Eight theses about text processing]. *Dagens Nyheter* (1970-09-29).

Allén, Sture. 1973. *Förslag till inrättande av ett organ för lagring och tillhandahållande av datamaskinellt läsbara texter, benämnt* logotek [Proposal for the establishment of a facility for storage and supply of computer-readable texts, designated *logotheque*]. Gothenburg: Computational Linguistics Unit, University of Gothenburg.

Allén, Sture. [1980] 1999a. The language bank concept. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 302–310. (Originally published in Joseph Raben & Gregory Marks (eds.), *Data bases in the humanities and social sciences*, 171–176. Amsterdam: North-Holland). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allén, Sture. [1983] 1999b. En forskningsstrategi för språkvetenskaplig databehandling: Perspektivskiss [A research strategy for computational linguistics: A perspective sketch]. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 222–233. (Originally published as a Språkdata internal research report). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allén, Sture. [1984] 1999c. Skandinavisk datalingvistik [Computational linguistics in Scandinavia]. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 168–199. (Originally published in *The Nordic languages and modern linguistics 5*, 11–42. Aarhus: Nordisk Institut, Aarhus Universitet). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allén, Sture. 2002. Nordic language history and computer-aided lexical research. In Oskar Bandle, Kurt Braunmüller, Ernst Håkon Jahr, Allan Karker, Hans-Peter Naumann & Ulf Teleman (eds.), *The Nordic languages: An international handbook of the history of the North Germanic languages (Volume 1)*, 268–271. Berlin: De Gruyter.

Borin, Lars. 2021. Multiword expressions: A tough typological nut for Swedish FrameNet++. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 221–259. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Borin, Lars, Dana Dannélls & Karin Friberg Heppin. 2021. Introduction: Swedish FrameNet++. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 3–35. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Borin, Lars & Benjamin Lyngfelt. 2025. Framenets and constructiCons. In Mirjam Fried & Kiki Nikiforidou (eds.), *The Cambridge handbook of construction grammar*, 71–100. Cambridge: Cambridge University Press.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (eds.). 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Gellerstam, Martin, Yvonne Cederholm & Torgny Rasmark. 2000. The Bank of Swedish. *International Conference on Language Resources and Evaluation (LREC) 2000*. np.

Gellerstam, Martin & Christian Sjögreen. 1994. *Språkbanken: En språklig referensdatabas* [Språkbanken: A linguistic reference database]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Guo, Ruohao, Wei Xu & Alan Ritter. 2024. Meta-tuning LLMs to leverage lexical knowledge for generalizable language style understanding. *Annual Meeting of the Association for Computational Linguistics (ACL) 2024 (Volume 1: Long Papers)*. 13708–13731. DOI: 10.18653/v1/2024.acl-long.740.

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 41–106. Amsterdam: John Benjamins.

NFO 1. 1970. *Nusvensk frekvensordbok baserad på tidningstext: 1. Graford, homografkomponenter* [Frequency dictionary of present-day Swedish based on newspaper material: 1. Graphic words, homograph components]. Stockholm: Almqvist & Wiksell.

NFO 2. 1971. *Nusvensk frekvensordbok baserad på tidningstext: 2. Lemman* [Frequency dictionary of present-day Swedish based on newspaper material: 2. Lemmas]. Stockholm: Almqvist & Wiksell.

NFO 3. 1975. *Nusvensk frekvensordbok baserad på tidningstext: 3. Ordförbindelser* [Frequency dictionary of present-day Swedish based on newspaper material: 3. Collocations]. Stockholm: Almqvist & Wiksell.

NFO 4. 1980. *Nusvensk frekvensordbok baserad på tidningstext: 4. Ordled, betydelser* [Frequency dictionary of present-day Swedish based on newspaper material: 4. Morphemes, meanings]. Stockholm: Almqvist & Wiksell.

Pedersen, Bolette S., Sanni Nimb, Ida Rørmann Olsen & Sussi Olsen. 2019. Merging DanNet with Princeton WordNet. *Proceedings of the 10th Global Wordnet Conference (GWC) 2019*. 125–134.

Pedersen, Bolette S., Nathalie C. Hau Sørensen, Sussi Olsen & Sanni Nimb. 2024. Evaluering af sprogforståelsen i danske sprogmodeller: Med udgangspunkt i semantiske ordbøger [Evaluation of language understanding in Danish language models: Based on semantic dictionaries]. *Nydanske Sprogstudier* 65: 8–40.

Ralph, Bo, Jerker Järborg & Sture Allén. 1977. *Svensk ordbok och lexikalisk databas: Förstudierapport* [The dictionary *Svensk ordbok* and the lexical database: A pilot study report]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Swartz, Merryanna L. 1992. Issues for tutoring knowledge in foreign language intelligent tutoring systems. In Merryanna L. Swartz & Masoud Yazdani (eds.), *Intelligent tutoring systems for foreign language learning*, 219–233. Berlin: Springer.

Lars Borin and Markus Forsberg

6 Saldo: the hub of Språkbanken's lexical research infrastructure

Abstract: *Saldo* is the "pivot resource" of our lexical macroresource for computational text processing applications (formerly Swedish FrameNet++, now Språkbanken's Lexical Research Infrastructure). Saldo's origins are in a particular onomasiological lexicon, a form of thesaurus devised in order to test a lexicological hypothesis about the semantic structure of the vocabulary. More or less serendipitously it became the basis for a full-size Swedish computational lexical resource to be used in lexically-informed automatic text processing in Språkbanken's corpus import pipeline. In the process, much effort was spent on designing the formal structure of Saldo, in particular its data model, where a system of carefully designed persistent identifiers was devised for the various entities and relations making up the lexicon. The fruits of this work could be reaped some years later, when Saldo could be adopted as the hub (pivot) resource of Swedish FrameNet++, with very little additional work needed.

Keywords: computational linguistics, language technology, lexical resource, onomasiological lexicon, research infrastructure, semantic lexicon, thesaurus

1 Introduction

The computational lexical resource *Saldo* (Borin, Lönngren & Forsberg 2017) is one of the workhorses in the Språkbanken Text research infrastructure, a position it has held for about a decade and a half. It is both the main source of lexical information

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161). Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Markus Forsberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: markus.forsberg@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-006

in the text corpus import pipeline and the hub through which other computational lexical resources are connected.

Saldo has played a central role in Språkbanken's development since the turn of the millennium from being primarily a lexicographical R&D unit relying on quite useful but fairly elementary corpus exploration tools, to a research infrastructure based on state-of-the-art language technology for supporting research based on written language data in a variety of disciplines. During this time, the lexicographical and language technological strands of the former Department of Computational Linguistics parted ways, but have recently been reunited (see Chapter 2 in this volume).

Version 1.0 of Saldo was released in the spring of 2008, the next stable version (2.3) appeared in 2015, and the most recently (in 2025) released current version is numbered 3.3.

In this chapter, we describe the development of Saldo from its beginnings in the 1980s to the latest version, released in 2025, coinciding with the 50th anniversary of Språkbanken. Parts of this story have been told before in other publications (notably Borin 2005; Borin, Forsberg & Lönngren 2008; 2013; Borin et al. 2021), but the most recent version of Saldo is described here for the first time including an account of how it differs from the previous versions.

2 Background

The first author (LB) came to Språkbanken in late 2002, and being a computational linguist with an educational background in Slavic and Finno-Ugric linguistics rather than a lexicographer or even a scholar of Nordic linguistics, almost immediately started looking for ways of upgrading the capabilities of Språkbanken to reflect recent advances in computational linguistics, notably to provide its linguistic corpora with automatic linguistic annotations in order to enhance their usefulness to Swedish lexicographers and other students of language.

The initial aim of this work was to provide at least part-of-speech (POS) tagging and lemmatization as standard annotations in our modern text corpora. The former objective could be straightforwardly accomplished using off-the-shelf general (language-independent) statistical POS taggers. There were fairly accurate POS taggers for Swedish at the time, and Språkbanken already offered access through an online concordancing interface to the 20-MW POS-tagged Parole corpus (Gellerstam, Cederholm & Rasmark 2000), although the corpus could not be offered for download in its entirety for intellectual property rights reasons.

No lemmatizer was available at the time, and for this a lexical resource was needed, however, and one large enough to be used for processing of unrestricted Contemporary Swedish text. Gellerstam, Cederholm & Rasmark (2000) describe a morphological database – corresponding to the inflectional information in the SAOL dictionary (see Chapter 3 in this volume), and primarily built as a component in a dictionary editing aid for humans – and also present plans for using it in corpus annotation, but these plans had not yet been realized by late 2002. Ideally, the in-house lexical databases should have been adapted for this purpose, and initially LB tried to pick up this thread, which seemed like the path of least resistance at the time, and investigated the possibility of using the SAOL morphological database as a component in a computational morphological analyzer. This turned out to be not feasible, however, because of strong restrictions with regard to who would be allowed to use the database and for which purposes, something that the funder reserved the right to determine on an individual basis through a written agreement. This is understandable, since the information contained in the database was to be included in several commercial dictionaries.

This closed-source nature of the in-house lexical data was a major impediment to using it as a component in automatic corpus analysis tools. Our strong preference was for a resource that could advance Swedish computational linguistics in the most general way, by being freely shareable for both scientific and commercial uses.

3 From SAL to Saldo 1.0

LB's hunch was that an existing experimental dictionary called SAL (*Svenskt associationslexikon* 'Swedish Associative Thesaurus') with which he was familiar could perhaps be made to fit the bill with some development effort. He had worked in the same research group at Uppsala University where this dictionary had been compiled, and contacted the originator of SAL, Lennart Lönngren, who readily agreed to allow Språkbanken to use SAL as the basis for such a free resource, and made the data comprising the second edition of SAL available to Språkbanken at the end of 2003.

Thus, the story of Saldo actually begins already in 1987, when Lönngren, a Slavicist at Uppsala University, started work on "quite a new kind of dictionary" (Lönngren 1998: 467): SAL. Saldo is based on this dictionary, an experimental onomasiological lexicon created by Lönngren in the years 1987–1992.[1] It was originally published (in hardcopy only) in two editions as research reports from the Center for Computa-

1 In addition to Lennart Lönngren who initiated and coordinated the work on SAL, much of the day-to-day lexicographical work was carried out by Gunilla Fredriksson and all programming and

tional Linguistics (Lönngren 1988) and Department of Linguistics (Lönngren 1992) at Uppsala University. Because of the way they were compiled, both editions were born digital.

Word lists extracted from several small text corpora and subsequently manually curated and complemented furnished the initial set of SAL entries, and in order to capture the core vocabulary of Swedish, all the entries from the frequency dictionary by Allén (1972) (based on the Press-65 news text corpus; see Chapter 2 in this volume) were included. The first edition of SAL (Lönngren 1988) contained 16,631 entries.

Eventually the complete entry lemma list of *Svensk ordbok* (SOB 1986) was purchased from Språkdata (the former Department of Computational Linguistics at the University of Gothenburg, the host of Språkbanken; see Chapter 2 in this volume), which put SAL on a solid lexicographical footing and later made it especially suitable as lexical pivot resource of Språkbanken (see Chapters 5 and 13 in this volume). Consequently, the second edition of SAL contained all the lemmas from SOB (1986), and entries from some other sources as well, but importantly, no other information from that dictionary was included, meaning that the lemma list fell under database protection (rather than copyright), that had expired by 2003, when Lennart Lönngren gave his permission to turn SAL into a free computational lexical resource, an endeavor in which he participated actively himself.

The original SAL in its second hardcopy edition (Lönngren 1992) contained 71,750 word senses, listed in four volumes.

As in the first edition, all entries were represented by bare lemmas with numerical indices in the case of colexification, but no information (beyond the lemma itself) about their formal properties, not even part-of-speech labels.[2] Two representative original SAL entries (*alias* 'alias' and *alibi* 'alibi') are shown in (1).

(1) _&alias&namn&&annan& 'alias' : 'name' + 'other'
_&alibi&bevisa&&annanstans& 'alibi' : 'prove' + 'elsewhere'

Lönngren (1998) refers to SAL as an "associative thesaurus", but it is in fact fairly differently organized compared to a prototypical thesaurus such as Roget (1852) or Bring (1930); see Chapter 8 in this volume.

Each SAL entry is semantically characterized – "defined" in a loose sense – by one or two other SAL entries, referred to as its *primary* – *namn* 'name' and *bevisa* 'prove' in (1) and (optional) *secondary* descriptor – *annan* 'other' and *annanstans* 'elsewhere' in (1).

---

other computational work in the project was done by Ágnes Kilár, both employed at the Uppsala University's Center for Computational Linguistics at the time.

2 In fact some lemmas were intentionally intended to refer to more than one part-of-speech; see Section 4.

As the entries in (1) demonstrate, part-of-speech boundaries are freely crossed in choosing descriptors. The primary descriptor is supposed to be the closest more central "lexical-semantic neighbor" of the entry, for instance a synonym, superordinate sense, typical object (of a verb), derivational base form, and so on. The semantic organization of SAL is strictly hierarchical: following the primary descriptors, we move upwards in the hierarchy towards more central, more basic word senses.[3] In order not to be forced to find an actual "simplest word sense" applicable to the whole vocabulary of a language, Saldo introduces an artificial unique-beginner word sense, called *PRIM* (see Figure 1).

The optional secondary descriptor bears a looser relationship to the entry and is mainly intended for disambiguation among entries with the same primary descriptor.

Even though SAL was made available to Språkbanken at the end of 2003, the real work on turning it into at computational lexical resource started only towards the end of 2004.

The initial efforts focused on defining a suitable data model for the new computational lexical resource based on SAL, and also on subjecting the SAL dataset itself to various formal controls, which revealed a few instances of circularity, i.e., where following the primary descriptors from an entry would eventually lead back to the same entry rather than to PRIM. Such checks have since been become part of the editing and updating procedures of Saldo.

At the same time, we investigated various solutions for the morphology. For a number of reasons, this work did not progress in any significant way before 2007. At that point, the second author, Markus Forsberg (MF), joined Språkbanken shortly after defending his PhD thesis that included a computational morphological analysis system called *Functional Morphology* (FM; Forsberg 2007). Using the Swedish morphology that he had built as part of his dissertation work as a basis, MF was able to extend it to account for all of the inflectional patterns represented in Saldo in only a few months' time.

Thus, the work that produced Saldo from SAL consisted of, on the one hand, the smaller but still significant effort of specifying the data model for the resulting resource and implementing it as a concrete computational lexicon, and on the other hand, the much larger undertaking of designing and building a complete computational inflectional morphological description of Contemporary Swedish.

The latter was a major effort and comprised the main contribution of Språkbanken to the first version of Saldo, and one that made the lexicon usable for text

3 Relative centrality is determined in several ways in Saldo. Relative frequency, morphological complexity, semantic dependence, and stylistic value all play a role in determining this (Borin, Forsberg & Lönngren 2013; Borin et al. 2021).

```
alias..1        namn..1         annan..1        alias..ab.1
alibi..1        bevisa..1       annanstans..1   alibi..nn.1
bevisa..1       visa..1         PRIM..1         bevisa..vb.1
namn..1         PRIM..1         PRIM..1         namn..nn.1

alias..ab.1     alias   ab      ab_i_aldrig
alibi..nn.1     alibi   nn      nn_5n_saldo
bevisa..vb.1    bevisa  vb      vb_1a_laga
namn..nn.1      namn    nn      nn_6n_blad

alibi..1        word sense identifier
PRIM..1         the "unique beginner" top-level artificial word sense
alibi..nn.1     lemgram (lexeme) identifier
nn              POS identifier
nn_5n_saldo     paradigm/inflectional class identifier
```

Figure 1: The contents and structure of Saldo v. 1 (top group: word sense entries; middle group: lemgram entries; bottom group: the identifiers explained)

processing purposes. In fact, at that point, the fact that Saldo originated in a lexicographically well-motivated, large entry set – that of SOB (1986) – was of greater importance than its semantic organization, that only came into its own later, in connection with the SweFN++ project (see Chapter 8 in this volume and Dannélls, Borin & Heppin 2021).

The first version of the data model concerned primarily the semantic component of the resource – since there was not yet a morphology – i.e., the word sense identifiers, and how the two semantic relations and the artificial top-level sense should be represented. We introduced a system of persistent, human-readable – since Saldo is a computational resource curated by human experts – identifiers, that also should be valid XML names. Semantic-web knowledge representation formalisms were very much in vogue at the time, and we decided that all identifiers in Saldo should be valid XML names, since XML was the language of choice for rendering semantic-web formats such as the Resource Description Format and Web Ontology Language of the World Wide Web Foundation. Non-XML strings were allowed in actual written-language forms, of course, including lexical entry lemmas. Figure 1 summarizes the formats of the various identifiers used in Saldo.

Saldo is first and foremost an onomasiological lexicon, i.e., its lexical items are word senses. These are linked to other word senses by two hierarchical relations, and in order for the lexicon to be practically usable for computational text analysis, the word senses are also linked to form units. These are *lexemes*, represented as a lemma together with its part of speech and inflected forms. In practice, the term "lexeme" is used in more than one sense in the literature. For this reason in accounts of the structure of Saldo we have frequently used the neologism *lemgram*.

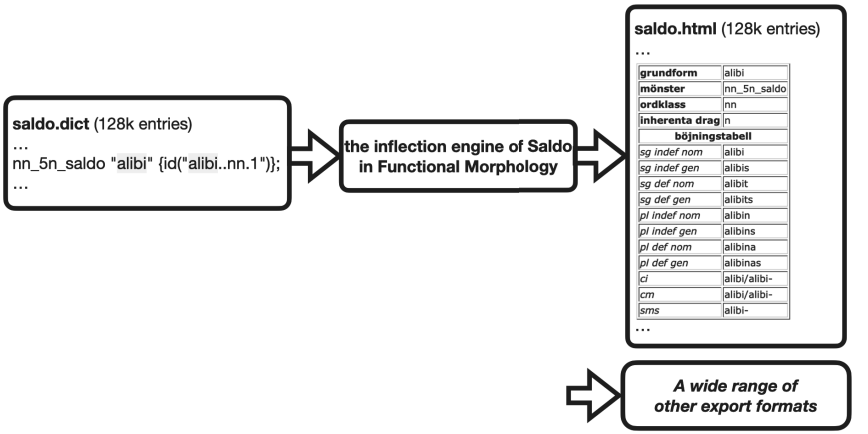


Figure 2: The first version of the Saldo morphology, implemented in FM

Functional Morphology is a computational framework for defining word-and-paradigm morphologies (Hockett 1954). The input of Functional Morphology is a dictionary with a list of lemmas (words) paired with their inflectional specifications (paradigms), and the output is a fullform lexicon, organized as a list of inflection tables for the input lemmas.

Functional Morphology uses a set of higher-level programming concepts available in the functional programming language Haskell (Marlow et al. 2010), such as algebraic data types and higher-order functions, to speed up the development of a large-scale morphology that can be exported into many different data formats. Functional Morphology also provides many other functionalities related to morphology engineering beyond just data exporting, such as data validation, to validate the formal properties of the morphology, and paradigm prediction, to find all paradigms that generate a particular set of word forms, given a lemma.

Figure 2 illustrates the general functionality of Saldo's inflection engine in Functional Morphology. On the left-hand side is the dictionary, with the example word *alibi* 'alibi' and its inflectional specification, the paradigm, is given by the identifier nn_5n_saldo. The identifier is mnemonic and specifies that the paradigm is a neuter noun of the fifth declension where *saldo* 'account balance' is an example member of that paradigm, and finally there is a unique id for the inflection table of *alibi*, alibi..nn.1. In the middle is Saldo's inflection engine, defined in Functional Morphology. And finally, on the right-hand side we have an example export format (HTML), where the inflection table for *alibi* is rendered.

With all the important components now in place, Saldo v. 1.0 was released in 2008. The sense set in the first version of Saldo was a straightforward adaptation of SAL, with 72,039 word senses expressed through 68,344 lexemes. The accompanying fullform lexicon (Språkbanken Text 2017) contained some 760,000 pairings of text words and morphosyntactic descriptions.

As a computational onomasiological lexicon, Saldo is distinctly differently conceived and organized from both the best known such lexicon, Princeton WordNet and from its predecessor at Gothenburg, the Semantic Database (see Chapter 8 in this volume).

4 From Saldo 1.0 to 2.3

The next stable version of Saldo was v. 2.0, which was released in 2010, and contained 100,051 word senses and 96,742 lexemes. Apart from many added word senses, the concept of word sense itself was modified compared to v. 1.0. Starting in version 2.0, we decided to not allow word senses to be shared across different parts of speech. Consequently, the single SAL and Saldo 1.0 entry *kollektiv* 'collective' with the primary descriptor *grupp* 'group' was split into the adjective *kollektiv*^{1} (primary: *grupp*) and the noun *kollektiv*^{2} (primary: *kollektiv*^{1}). In Saldo 2.0 the possibility was added for an entry to have more than one secondary descriptor, as in (2).

(2) hbt-person : person + bisexuell homosexuell transsexuell
'HBT person' : 'person' + 'bisexual' 'homosexual' 'transsexual'

The Saldo morphology did not undergo any major changes for the new version, except that the description of compounding forms was made more detailed, now distinguishing between initial and medial compounding forms, as in, e.g., *båtbyggare* [båt+byggare] 'boat builder' vs. *segelbåtsbyggare* [segel+båt-s+byggare] 'sail boat builder', where the compounding form of *båt* 'boat' varies according to its position in the word.

Version 2.3, that was the production version for a decade, and the version with which all of the modern corpora in Korp have been analyzed (see Chapter 10 in this volume), appeared in 2015 with 131,019 word senses and 128,036 lexemes, i.e., almost doubled in size in comparison to the original SAL dataset acquired 10 years earlier.

5 From Saldo 2.3 to 3.3

Saldo 3 represents a thorough revision of the overall organization and several linguistic-form aspects of the resource, notably the POS classification of some classes of entries and the treatment of compounding forms, while its semantic structure remains unaltered from previous versions, preserving the original intention of SAL.

As a consequence of several years of practical experience of working on Saldo as well as using it for most of the automatic word-level analysis tasks in Språkbanken's corpus import pipeline (see Chapter 10 in this volume), version 3 of Saldo has been subdivided into four largely pragmatically defined non-overlapping subsets, called pne, iie, uie, and foe, which we explain below. These form a kind of coarse hierarchy structured by two superordinate categories, FOE and UIE.[4] All entries in Saldo belong to the top-level superordinate category FOE (fixed-order expressions), reflecting the fact that in Saldo we record fixed-order linguistic expressions with distinct noncompositional senses. The FOE are then subdivided into UIE (uninterrupted expressions) and foe ([other] fixed-order expressions), viz. the multiword verbs as well as a few conjunctions and adpositions (for example particle verbs, such as *bjuda till*) 'make an effort', literally 'bid to'), that allow intercalation of other lexical units at specific positions. The UIE consist of three sets: pne (proper noun expressions, for example *Platon* 'Plato' or *Förenta staterna* 'the United States'), iie (internally inflected expressions), comprising a portion of the multiword nouns, viz. those including a modifier with agreement inflection (for example *svart hål* 'black hole', with the definite plural *svarta hålen*).[5] Finally, there is the largest class, uie ([other] uninterrupted expressions), that contains all single-word items as well as the non-interruptible multiword expressions (the "words-with-spaces" of Sag et al. 2002: 2–4), a category dominated by adverbs (for example *till exempel* 'for example').

The Saldo morphology models inflection and compounding behavior of single- and multiword lexical units, but not derivational morphology. Ultimately, this rests upon an explicit definition of what is to be counted as derivational and what as inflectional categories in Swedish, sometimes bringing Saldo into conflict with other descriptive traditions.

4 Capitals are used here for more abstract, superordinate categories, while lowercase labels denote terminal lexical sets, both set in a sans-serif typeface.

5 Strictly speaking, some multiword pronouns also belong formally here, but since this is a closed category with idiosyncratic behavior – there are even some single-word pronouns exhibiting internal inflection, such as *varannan* ∼ *vartannat* 'every second [NON-NEUTER ∼ NEUTER]' – we include them in the uie category, for largely pragmatic reasons.

On the one hand, the morphosyntactic descriptions should by and large conform to what the users of our corpus tools expect, i.e., in effect they should follow traditional lines. But on the other hand it is a fact that this varies somewhat in practice, so that there is not *one*, but a number of different descriptions of at least parts of Swedish morphology in sources that we could call traditional. A case in point: the large Swedish reference grammar commissioned by the Swedish Academy (SAG; Teleman, Hellberg & Andersson 1999) and the somewhat more recent and considerably more concise Swedish Academy grammar handbook by Hultman (SAS; 2003) differ not only in their depth of description, but also in how certain aspects of Swedish grammar are described, with SAS generally coming out as more traditional than SAG. Compared to Saldo 2, the Saldo 3 morphology has generally changed in the direction of SAG.

Although the lemgrams of Saldo 3 share their identifiers with those of Saldo 2, there are some clear differences in their behavior. The first and most pervasive difference is that compound forms are no longer included in the inflection table of a lemgram. Now only inflected forms belong there. In Saldo 3 compounding behavior is considered to belong on the word sense level: the non-final parts of a compound are Saldo word senses, not lemgrams as in Saldo 2.

The second major difference concerns verb inflection, where participles are no longer considered inflected forms of verbs. This contrasts with traditional grammatical descriptions of Swedish (e.g. Almqvist 1840; Thorell 1973) but agrees with SAG. In contrast to SAG, however, participles are not considered a separate part of speech in Saldo 3. Instead, they are classified as adjectives. Consequently, every added verb in Saldo should be accompanied by one or two participles, depending on the transitivity of the verb (as well as a verbal noun in *-(a)nde*; see below).

The third significant difference consists in proper nouns having far fewer inflection classes in Saldo 3 than in Saldo 2, since the name type information that was formerly part of the paradigm identifier is now found separately in the feature set introduced in Saldo 3.

There are also a few minor differences:

– The *infinitive marker*, that traditionally, including according to SAG, forms a POS of its own, is classified as a subjunction in Saldo.
– *Adverbs* coinciding in form with neutral adjectives in *-t* are classified as adjectives in Saldo, in accordance with their treatment in SAG. This is also applied to other cases where the neuter form is indistinguishable from an adverb, e.g., indeclinable adjectives. Notably, consistency and parsimony consequently require that indeclinable adjectives, even if etymologically derived from original adverbs, be given in Saldo only as adjectives.
– *Verbal nouns* in *-(a)nde* (as well as deadjectival nouns in *-het* and some other productive derivations) are rarely listed in dictionaries despite normally being

classified as derivations rather than inflected forms. Since Saldo should be usable for automatic analysis of arbitrary text and since it explicitly excludes derivational morphology from its morphological component, such items must be included in Saldo.

As we are all too painfully reminded of every now and then, software does not live forever, or even close to forever, at least not without major maintenance efforts. The Functional Morphology application has been maintained on a kind of basic "status quo" level since the Saldo 2.3 morphology was published in 2015, meaning basically that new Saldo entries can be added, provided that they are covered by one of the existing inflectional paradigms. The major planned change in Saldo 3 as compared to v. 2.3 was a thorough revision of the inflectional morphology, but there was no scope for doing this using Functional Morphology, for primarily two reasons: Functional Morphology is today, more than twenty years down the road, a legacy system that would require substantial technical work to modernize; and what was previously considered a strong argument for Functional Morphology, namely that implementation of the paradigms are defined with all the bells and whistles of a high-level programming language, has turned out to be the main argument against it, since the exact behavior of a paradigm is hidden in the source code. So today we want to move away from these kinds of intricate black-box paradigm implementations and towards treating paradigm specifications as lexical data, not code.

Instead, the new morphological component of Saldo 3, responsible for generating text word forms (and compounding forms) from the Saldo 3 lemgrams and their paradigm identifiers, had to be looked for elsewhere. The adopted solution became *foma* (Hulden 2009), a mature open-source program package implementing finite-state transducers (FST) for morphological description and processing (Hulden 2022). The implementation work turned out to be fairly painless, mainly thanks to the existing well-documented Functional Morphology description of Saldo 2.3, that on the whole could be fairly mechanically (even if manually) converted into the new format. Only multiword expressions have required a more substantial effort, because of more fundamental changes in the way that they are paradigmatically represented.

Using foma, an inflectional paradigm is typically implemented as a (named) LEXICON in the *lexc* formalism (Beesley & Karttunen 2003) available through foma. Common morphophonological and orthographical alternations are captured by FSTs specified using foma's regular expression formalism. Figure 3 illustrates the description format of the Saldo 3 morphology. In the figure, we see three foma rules, all dealing with orthography. The first rule states that the genitive clitic *=s* should not be written after words ending in a sibilant (<s>, <z>, <x>, <sch>, <sh>). The two following rules deal with the orthographic consequences of adding the neuter suffix *-t* to an adjective with a stem ending in certain consonant combinations. The first

```
# no genitive s after s etc. (shouldn't affect superlative)
define gens %+ s -> 0 || [s|z|x|s (c) h] _ ;

# dd -> tt in klädd, sydd
define avdd d d %+ t -> t t || _ [ .#. | %+ ] ;

# d -> t in röd, rund; nn -> nt in tunn; t -> 0 in smart
define avdt d %+ t -> t t || Vwl _ [ .#. | %+ ] ,,
d %+ t -> t || Cns _ [ .#. | %+ ] ,,
n %+ t -> t || n _ [ .#. | %+ ] ,,
t %+ t -> t || Cns _ [ .#. | %+ ] ;

LEXICON 3av_0_medelstor
% %:%:av% pos% utr:0 3avnn_case;
% %:%:av% pos% neu:+t 3avnn_case;
% %:%:av% pos% dfp:+a 3avnn_case;
% %:%:av% pos% mas:+e 3avnn_case;

LEXICON 3av_1_gul
3av_0_medelstor;
0:+ 3av1_comp;

LEXICON 3vb_2r_hyra
% %:%:vb% prs% akt:i #;
% %:%:vb% imp:i #;
% %:%:vb% inf% akt:0 #;
3vb24_npast_sfo;
% %:%:vb% prt:ide 3vb_snprs;
% %:%:vb% sup:it 3vb_snprs;
```

Figure 3: The Saldo 3 morphology: three foma FST rule definitions and three lexc minilexicons

two lexc minilexicons in Figure 3 describe the inflection of a large class of regular adjectives, and the third minilexicon captures a second-conjugation verb paradigm. Figure 4 shows the sets of word forms with grammatical descriptions generated by foma for the adjective *röd* 'red' (paradigm 3av_1_gul) and the verb *hyra* 'rent' (paradigm 3vb_2r_hyra). At the time of writing (early 2025), Saldo 3.3 contains some 148,000 word senses, expressed by about 142,500 lexemes/lemgrams.

6 Saldo: a research infrastructure component for automatic lexical analysis

The Saldo version employed in Språkbanken's analysis plattform Sparv that is used for corpus import (see Chapters 9 and 10 in this volume) is still (in early 2025) v. 2.3. While details of the word-level linguistic analysis will be different, the basic word analysis machinery of the corpus import pipeline will not need to be redesigned

```
röd..av.1      av kmp gen     rödares    hyra..vb.1     vb imp  hyr
röd..av.1      av kmp nom     rödare     hyra..vb.1     vb inf akt     hyra
röd..av.1      av pos dfp gen  rödas      hyra..vb.1     vb inf sfo     hyras
röd..av.1      av pos dfp nom  röda       hyra..vb.1     vb prs akt     hyr
röd..av.1      av pos mas gen  rödes      hyra..vb.1     vb prs sfo     hyres
röd..av.1      av pos mas nom  röde       hyra..vb.1     vb prs sfo     hyrs
röd..av.1      av pos neu gen  rötts      hyra..vb.1     vb prt akt     hyrde
röd..av.1      av pos neu nom  rött       hyra..vb.1     vb prt sfo     hyrdes
röd..av.1      av pos utr gen  röds       hyra..vb.1     vb sup akt     hyrt
röd..av.1      av pos utr nom  röd        hyra..vb.1     vb sup sfo     hyrts
röd..av.1      av spr def gen  rödastes
röd..av.1      av spr def nom  rödaste
röd..av.1      av spr idf gen  rödasts
röd..av.1      av spr idf nom  rödast
```

Figure 4: Sets of word forms with morphosyntactic descriptions generated by foma for the Saldo lemgrams röd..av.1 3av_1_gul 'red' (adjective) and hyra..vb.1 3vb_2r_hyra 'rent' (verb)

with the switch to Saldo 3.3. Since the identifiers of Saldo are not only used as the pivot for the lexical macroresource of Språkbanken, but also in the lexical analyses of Språkbanken, most language data of Språkbanken are connected via Saldo.

The lexical analysis of Sparv draws on all levels of Saldo, including lemmatization and lemgram lookup, word sense lookup, word sense disambiguation, and compound analysis.

Further, since Saldo is the pivot of the lexical macroresource of Språkbanken, the lexical analysis can be followed by a simple lookup of information in other lexical resources, such as Swedish FrameNet (see Chapter 7 in this volume), lexical information that can also be lifted up to the text level.

7 Conclusion

Summing up and looking ahead: Saldo comprises a kind of parallel lexicographical trajectory to the large conventional dictionaries produced in our department during the same period. With the merger of the two strands of computational lexicography under the common umbrella of Språkbanken Text, there are many opportunities for fruitful synergy between the language-technological and the conventional lexicographical approaches to lexical resource building and dictionary compilation, some of which are already being explored (see Chapter 11 in this volume).

References

Allén, Sture. 1972. *Tiotusen i topp: Ordfrekvenser i tidningstext* [Top ten thousand: Word frequencies in newstext]. Stockholm: Almqvist & Wiksell.

Almqvist, Carl Jonas Love. 1840. *Svensk språklära* [Swedish grammar]. 3rd edn. Stockholm: M. Wirsells förlag.

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite state morphology*. Stanford: CSLI Publications.

Borin, Lars. 2005. Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat [The man is the grandmother of the father: The *Swedish Associative Thesaurus* reincarnated]. *LexicoNordica* 12: 39–55.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2008. The hunting of the BLARK: SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf & Beáta Megyesi (eds.), *Resourceful language technology: Festschrift in honor of Anna Sågvall Hein*, 21–32. Uppsala: Department of Linguistics & Philology, Uppsala University.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. DOI: 10.1007/s10579-013-9233-4.

Borin, Lars, Markus Forsberg, Lennart Lönngren & Niklas Zechner. 2021. Swedish FrameNet++: Lexical samsara. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 69–95. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Borin, Lars, Lennart Lönngren & Markus Forsberg. 2017. *Saldo*. [Data set]. DOI: 10.23695/s80w-2517.

Bring, Sven Casper. 1930. *Svenskt ordförråd ordnat i begreppsklasser* [Swedish vocabulary arranged in conceptual classes]. Stockholm: Hugo Gebers förlag.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (eds.). 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Forsberg, Markus. 2007. *Three tools for language processing: BNF converter, Functional Morphology, and Extract*. Göteborg University & Chalmers University of Technology. (PhD thesis).

Gellerstam, Martin, Yvonne Cederholm & Torgny Rasmark. 2000. The Bank of Swedish. *International Conference on Language Resources and Evaluation (LREC) 2000*. np.

Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10(2-3): 210–234.

Hulden, Mans. 2009. Foma: A finite-state compiler and library. *Proceedings of the Demonstrations Session at the 12th Conference of the European Chapter of the ACL (EACL) 2009*. 29–32.

Hulden, Mans. 2022. Finite-state technology. In Ruslan Mitkov (ed.), *The Oxford handbook of computational linguistics*, 230–254. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199573691.001.0001.

Hultman, Tor G. 2003. *Svenska Akademiens språklära* [The Swedish Academy grammar handbook]. Stockholm: Norstedts.

Lönngren, Lennart. 1988. *Svenskt associationslexikon (Rapport UCDL-R-88-2)* [Swedish associative thesaurus (Report UCDL-R-88-2)]. Research report. Uppsala: Uppsala University, Center for Computational Lingistics.

Lönngren, Lennart. 1992. *Svenskt associationslexikon: Del I–IV* [Swedish associative thesaurus: Volume I–IV]. Research report. Uppsala: Uppsala University, Dept. of Linguistics.

Lönngren, Lennart. 1998. A Swedish associative thesaurus. *Proceedings of the European Association for Lexicography (EURALEX) 1998: Vol. 2*. 467–474.

Marlow, Simon et al. 2010. Haskell 2010 language report. *Available online http://www. haskell. org/(May 2011)*.

Roget, Peter Mark. 1852. *Thesaurus of English words and phrases, classified and arranged so as to facilitate the expression of ideas and assist in literary composition*. London: Longman, Brown, Green, & Longmans.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *Computational linguistics and intelligent text processing: Third international conference: Cicling-2002*. 1–15.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Språkbanken Text. 2017. *Saldo's morphology*. [Data set]. DOI: 10.23695/agcm-ny22.

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik* [The Swedish Academy grammar]. Stockholm: Norstedts.

Thorell, Olof. 1973. *Svensk grammatik* [Swedish grammar]. 2nd edn. Stockholm: Esselte Studium.

Dana Dannélls, Niklas Zechner, and Shafqat Mumtaz Virk

7 Swedish FrameNet: a lexical semantic resource for Swedish

Abstract: *Swedish FrameNet* (SweFN) is a lexical semantic resource that has been developed in the *Swedish FrameNet++ project*, along the lines of the Berkeley FrameNet database. It is a result of a long lexicographic work with the aim to define semantic frames whose core elements are lexical units that evoke them. During the last decade SweFN has expanded its scope to support Swedish computational lexicography and today it is a fundamental part of Språkbanken's lexical research infrastructure. More recently, it was updated to fit into a larger global framenet initiative where framenets for other languages are connected through semantic frames. In this chapter, we present SweFN 2.0 and discuss its role in advancing the Swedish language technology infrastructure. We further present our vision for the future use of FrameNet in general and Swedish FrameNet in particular.

Keywords: domain specific framenet, FrameNet, frame semantic parsing, lexical-semantic resource, semantic role labelling

1 Introduction

A lexical semantic resource records the relationship between words and their meanings within a language system. It is a network built around word meanings, where the core elements to describe these meanings are the semantic roles, describing the

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161).

Dana Dannélls, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: dana.dannells@svenska.gu.se
Niklas Zechner, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: niklas.zechner@svenska.gu.se
Shafqat Mumtaz Virk, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: shafqat.virk@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-007

Table 1: Four distinct senses of the verb *paint* categorised according to frames in FrameNet

Frame	Core Elements	Example
Filling	AGENT, CAUSE, GOAL, THEME	They painted the wall.
Create_representation	CREATOR, REPRESENTED	She painted her brother.
Create_physical_artwork	CREATOR, REPRESENTATION	He painted a fresco.
Communicate_categorization	CATEGORY, ITEM, MEDIUM, SPEAKER	The book painted me as happy.

behaviour of the word, which is also connected to the word's syntactic manifestation. Many theories for formalising lexical semantic relations in a computational network have been proposed in the literature (Fillmore 1968; Cruse 1986; Berk 1999). In this chapter we are concerned with frame semantic theory (Fillmore 1976), the theory behind the computational lexical semantic resource, FrameNet. According to Charles J. Fillmore:

> A particularly important notion, figuring especially in recent work in linguistics, cognitive psychology, and artificial intelligence, is the notion that goes by such names as "frame," "schema," and "scenario." Briefly, the idea is that people have in memory an inventory of schemata for structuring, classifying, and interpreting experiences, and that they have various ways of accessing these schemata and various procedures for performing operations on them. Some of the schemata may be physiologically built in (such as various aspects of the body schema, the identity of the focal hues in the color spectrum, and perhaps what the gestalt psychologists call "good figures" – see Rosch), others may owe their existence to perceived constant cause-effect relationships in the world, while still others may depend for their existence on symbolization. (Fillmore 1976: 25)

To exemplify, consider the transitive verb *paint* and its four distinct senses represented according to FrameNet in Table 1. As the table shows, the verb *paint* occurs in four different frames, each representing a distinct meaning.

From a computational linguistics point of view, an important aspect of a lexical semantic resource, aside from its size, is a comprehensive representation of the inheritance of word meanings as well as their relations to syntactic structures (Jackendoff 1997). These are in particular important for linguistic research and for developing language technology (LT) tools and applications. FrameNet meets all of these requirements, which is perhaps not surprisingly the reason why it has become popular among researchers working with computational lexicography.

Table 2: `Create_representation` frame

Entity	Name/Type	Description
Frame	`Create_representation`	A CREATOR produces a physical object which is to serve as a REPRESENTATION of an actual or imagined entity or event, the REPRESENTED.
Core frame elements	CREATOR	An individual or individuals that bring the REPRESENTATION into existence.
	REPRESENTED	The entity–which may be a thing, an action or a state–that is represented by the REPRESENTATION.
Lexical units	*Verb*	carve, cast, draw, paint, photograph, sketch
Sentences	*paint.v*	[$_{Represented}$ The Gerichtsstube] was [$_{LU}$ PAINTED] [$_{Creator}$ by Kuhn] [$_{Time}$ in 1767].

2 FrameNet

FrameNet (FN) is a computational lexical semantic resource that is based on the theory of *frame semantics* (Fillmore 1976; 1982) – a theory of meaning in natural language. It is built on the idea that humans develop mental structures about real life scenarios, objects, entities, and relations in their brains. In frame semantics, a frame is the cognitive structure that organises human experiences in terms of conceptual models and describes how these models are expressed when producing and interpreting linguistic utterances.

FrameNet organises both the conceptual system and the linguistic system by virtue of encoding semantic roles and utterances in the frame. For example, consider the frame `Create_representation`, defined according to the description in Table 2. It has a definition and two key semantic roles, called Frame Elements (FE). It also contains *lexical units* (LUs)[1] that identify or, in frame semantic terminology, "evoke" the frame (Fillmore 1982). In Berkeley FrameNet, LUs are taken from and linked to word senses in Princeton WordNet (Fellbaum 1998). Lexical units are evidence-based in a sense that they have been derived from corpus data, more specifically, the British National Corpus (BNC Consortium 2007) containing sentences annotated with morphological and grammatical information. Sentences have been annotated with an additional layer of frame elements, providing a comprehensive representation and analysis of corpus data. Moreover, connections between frames are established through frame-to-frame relations, such as Inheritance, Inchoative Of and Causative Of.

1 LU is the headword in one of its senses. In Saldo it is marked by a number, see Chapter 6 in this volume.

FrameNet was created as a linguistic resource for frame semantic information of English (Ruppenhofer et al. 2016) and developed within the Berkeley FrameNet project; hence therefore is referred to as Berkeley FrameNet (BFN; Fillmore, Johnson & Petruck 2003). It started off as a relational database with a user-friendly interface providing access to its content by navigation over its hierarchical structure.

Embarking on a lexicographic project, a team of lexicographers has worked actively to find evidence from corpus examples, define semantic frames and frame elements, and determine the relation between them. This has resulted in the current version, BFN 1.7 which is available for download in XML (Extensible Markup Language) format. This version has undergone significant feature and content changes since the earlier versions 1.3 and 1.5.[2] During these years BFN has been used in various downstream natural language processing (NLP) applications such as question answering (Taniguchi, Hoshino & Kano 2019) and information extraction (Marzinotto et al. 2018). Perhaps the main application area is *semantic role labelling* (Das et al. 2014). Each application has contributed to the advancement of NLP in unique ways, and therefore increased global interest in developing framenets for other languages.

2.1 Framenets for languages other than English

Following the success of the Berkeley FrameNet, the important role it played in enabling machines to understand and process human language more accurately, and its contribution to LT applications and e-lexicography (L'Homme 2014), framenet resources for languages other than English have been developed.

Like other NLP resources, the adequacy of a lexical semantic resource for its purpose depends on its coverage, underlying structure, consistency, accuracy, accessibility for other applications, and the detailed documentation level of its content. At the same time, creating a new framenet resource means making at least three decisions. The first one concerns the approach to be taken. Two of the most prominent approaches discussed in the literature are: (1) top-down, starting from the semantic structure of the domain, or (2) bottom-up, starting from the linguistic properties of the language. The second decision is about the applied methodology for creating the resource, either manual, automatic, or semi-automatic. The third decision concerns the language resources, i.e., lexicons and corpora, from which the lexical units and example sentences will be extracted.

Framenet-based lexical resources for languages other than English have applied different strategies depending on the availability of resources for the language in

2 https://github.com/clingergab/Bert_frame_semantic_parsing/blob/master/data/fndata-1.7/docs/GeneralReleaseNotes1.7.pdf (last accessed: April 4, 2025)

question (Burchardt et al. 2009). Common to all framenets is that they all contain links to BFN. Torrent et al. (2018) initiated a collaborative annotation project aimed at comparing semantic frames and frame elements across multiple languages. They started with a small subset of frames and frame elements in the Brazilian Portuguese FN and compared them to the BFN. They found that over 80% of the frames in BFN could be harmonised across both framenets. This finding has sparked a new effort called *Global FrameNet* that aims to gather framenet resources under one umbrella.[3] One of the purposes of the Global FrameNet is to align existing framenets in several languages and, through alignment, investigate and test the cross-linguistic aspects of FrameNet and its underlying theoretical basis. Alignment is standardised across all framenets and is referenced against BFN 1.7. The framenets that have been included so far are: Brazilian Portuguese, Chinese, French, German, Hebrew, Japanese, Korean, Latvian, Spanish, and Swedish.

This growing international effort, as well as interest in electronic resources enriched with lexical and semantic information (Dalpanagioti 2024) is one of the reasons why framenet resources have been expanded to cover specific domains.

2.2 The domains of FrameNet and beyond

Although framenets and frame-annotated data have proved to be useful both to the linguistic and NLP communities, they have often been criticised for their lack of cross-linguistic applicability and limited coverage. A proposed reasonable-effort solution to the coverage issue is to develop domain-specific (sublanguage) framenets to augment and extend the general-language framenet. In the literature, we can find such initiatives where domain-specific framenets are being developed, e.g.: (1) medical terminology (Borin, Toporowska Gronostaj & Kokkinakis 2007); (2) *Kicktionary*,[4] a soccer language framenet; and (3) the *Copa 2014* project, covering the domains of soccer, tourism and the World Cup in Brazilian Portuguese, English and Spanish (Torrent et al. 2014). In addition, the Global FrameNet initiative aims to cover new domains, such as biology and finance, as can be seen in Figure 1.[5]

The naming of frames and their relational structures has also evolved over the years. This development can be compared to the creation of a framenet for the linguistic domain. Over centuries, linguists have established a rich set of domain-specific terms and concepts (e.g. *inflection*, *agreement*, *affixation*, etc.) through the study, investigation, and documentation of various linguistic characteristics across

3 https://www.globalframenet.org/ (last accessed: April 4, 2025)

4 http://www.kicktionary.de/ (last accessed: April 4, 2025)

5 Figure taken from https://webtool.frame.net.br/grapher/domain (last accessed: April 4, 2025)

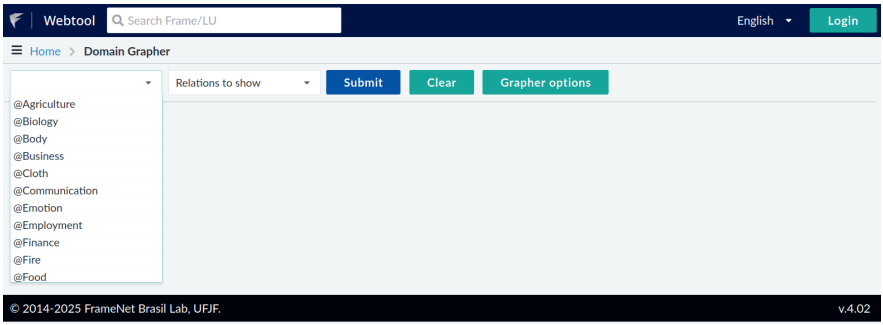


Figure 1: Domains in Global FrameNet

different languages, encompassing phonological, morphological, syntactic, and semantic levels. For various computational purposes, attempts have been made to create inventories of such terms and keep a record of them, e.g.: (1) the *GOLD*[6] ontology of linguistic terms; (2) the *SIL glossary of linguistic terms*;[7] (3) the *CLARIN concept registry*;[8] and (4) *OLiA* (Chiarcos, Nordhoff & Hellman 2012).

A minority of the terms in these collections are used only in linguistics (e.g., the noun *tense*), and in many cases, non-linguistic usages are either rare (e.g., *affixation*) or specific to some other domain(s) (e.g., *morphology*). Others are polysemous, having both domain-specific and general-language senses. For example, in their usage in linguistics the verb *agree* and the noun *agreement* refer to a particular linguistic (morphosyntactic) phenomenon, viz. where a syntactic constituent by necessity must reflect some grammatical feature(s) of another constituent in the same phrase or clause, as when adjectival modifiers agree in gender, number and case with their head noun. This is different from the general-language meaning of these words, implying that their existing FN description (if available) cannot be expected to cover their usage in linguistics. For a more comprehensive coverage, we need to define new frames, identify their LUs and FEs, and find examples that could be added to the general framenet. This was one of the major objectives of the work reported by Malm et al. (2018), where Linguistic FrameNet (LingFN) was introduced. The other objective was to investigate the relational aspects of the resulting linguistic frames. In the GOLD ontology, attempts were made to divide and organise the linguistic concepts into various groups. This organisation is not without problems. GOLD seems to lack a theoretical foundation, and the validity of the organization remains untested. Also

6 http://linguistics-ontology.org/ (last accessed: April 4, 2025)
7 http://glossary.sil.org (last accessed: April 4, 2025)
8 https://www.clarin.eu/ccr (last accessed: April 4, 2025)

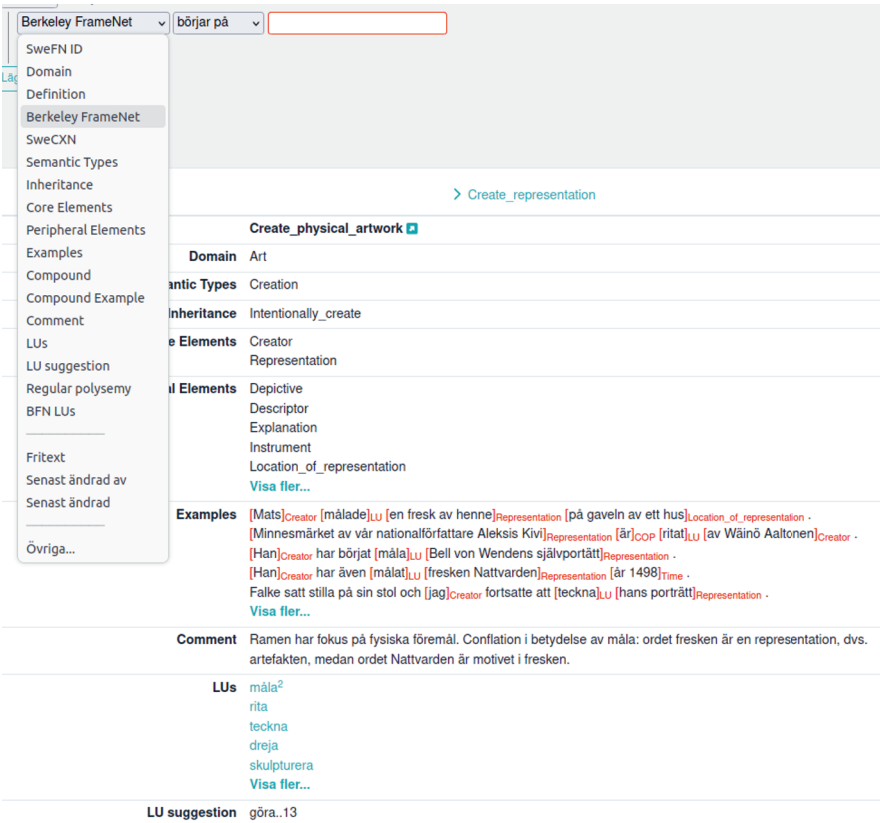


Figure 2: Screenshot of the frame `Create_physical_artwork` in the Karp editor

there is only one type of default relation – IS-A – between the terms/concepts. LingFN extended the relational structure of linguistic frames by exploring new relation types between the linguistic terms/concepts and building a network (viz. LingFN) of them.

3 Swedish FrameNet

The Swedish FrameNet (SweFN) was developed at Språkbanken Text in accordance with the principles of frame semantics (Fillmore 1982), and more precisely, following the conceptual backbone of the English FrameNet project (Baker, Fillmore & Lowe 1998). When the SweFN project started in 2009 (Borin et al. 2010), first by applying

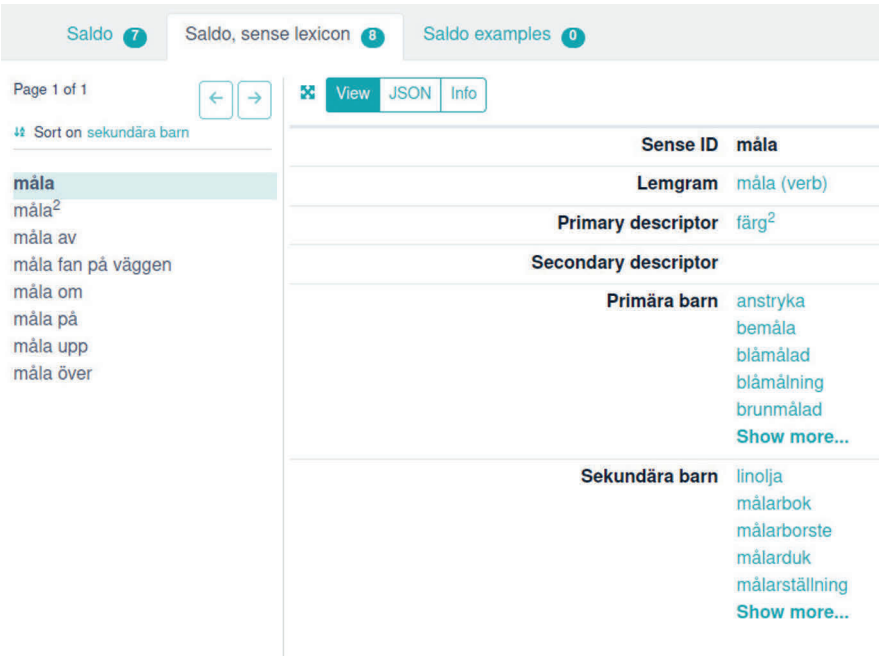


Figure 3: Screenshot of the verb *måla* in Saldo viewed in the Karp editor

manual approaches and then accelerating development with automated methods, it combined both bottom-up and top-down methods (Dannélls, Borin & Heppin 2021).

As with any other lexicographic project, whether for compiling a printed or an electronic lexicon, a variety of language resources were consulted. The primary resource for creating SweFN is the large modern computational lexicon Saldo (Borin, Forsberg & Lönngren 2013), see Chapter 6 in this volume. Saldo has played an important role in the creation of SweFN. In Friberg Heppin & Dannélls (2015), the authors explain how Saldo's sense inventory guided the lumping and splitting of semantic frames. The implementation was carried out in Karp, which not only provides access to other lexical resources but also is the SweFN editing interface through which it is possible to navigate among other Swedish lexical resources. Figure 2 shows the content and database fields that are available for a specific frame and the categories, in bold, that can be searched for in Karp.[9] See Chapter 11 in this volume.

9 https://spraakbanken.gu.se/karp/?mode=swefn&lexicon=swefn&show=swefn (last accessed: April 4, 2025)

Table 3: Three distinct senses of the verb *måla* 'paint' categorised according to frames in SweFN

Frame	Core Elements	Saldo sense
Filling	AGENT, CAUSE, GOAL, THEME	måla..1
Create_physical_artwork	CREATOR, REPRESENTATION	måla..2
Create_representation	CREATOR, REPRESENTED	måla..3 (Suggested)

As might be expected, language-specific lexicons do not always cover all the senses found in other lexicons. In the case of Saldo, because the level of granularity of Saldo's sense inventory is coarser compared to Princeton WordNet, not all the senses of a particular LU were aligned with the English one. Figure 3 shows that the verb *måla* 'paint' has only two senses listed in Saldo: måla..1 and måla..2.

Consequently, the verb *måla* only appears in two frames, compared to four in Table 1. The third sense, måla..3, see Table 3, has been added to the database field "LU suggestion" of the semantic frame `Create_representation`. Whereas the fourth sense listed in Table 3 is conveyed by another Swedish verb, viz. *utmåla* 'paint (as)', 'portray (as)', 'depict'.

3.1 SweFN in a solid infrastructure soil

SweFN has been a fundamental part of the Swedish language technology infrastructure since the first day of its development. In the early stages of its development, SweFN was integrated in Karp and was linked to all other lexical resources through Saldo, as mentioned in the previous section.

SweFN is also one of the resources for performing automatic lexical analysis across all corpus resources. It is, by default, included in the *Sparv* annotation pipeline (Borin et al. 2016). This means that all the tokens processed with Sparv and found in SweFN are enriched with information about the frame they belong to. Since Sparv underlies all the corpus resources in Korp – Språkbanken's Text corpus infrastructure (Chapter 10 in this volume) – all the corpus data is enriched with information about semantic frames. See also Chapter 10 in this volume.

Consider the results for the word *demokrati* 'democracy' in Korp, shown in Figure 4. On the right-hand side, under Text Attributes, we find the frames that have been recognised in the entire document: `Confronting_problem`, `Reforming_a_system`, and `Alliance`. The target frame for the highlighted word is `Leadership`, shown under Word Attributes. It is further possible to click on the frame and access the frame in Karp directly, see Figure 5, from where the user can further navigate to Saldo and the other lexical resources.

Figure 4: Korp view showing search results for the word *demokrati* 'democracy'

Lexical semantic enrichment of words in text enables automatic indexing, helping researchers in examining interdisciplinary research questions. In this regard, SweFN is an invaluable resource for researchers, particularly when the analysis results are available on a user-friendly platform, such as Strix, see Chapter 15 in this volume. To exemplify, consider two search results in Strix in Figure 6. The figures show the search results from one of the documents in the corpus *Swedish party programs and election manifestos* (Språkbanken Text 2024b) for the frames `Building` and `Medical_professionals`. For each frame, the number of occurrences in the document is counted. All the words that evoke the respective frame are highlighted in yellow.

3.2 From SweFN 1.0 to 2.0

The first version of the Swedish FrameNet was based on Berkeley FrameNet (BFN) 1.5. When the first version was created, some frames were added compared to the English version – both entirely new frames (such as `Countries` and `Furniture`) and more specific subframes added to existing ones. One example of the latter is that several frames were given positive and negative versions. One of those is the frame `People_by_morality`, which was given the two subframes `People_by_morality_positive` and `People_by_morality_negative` (Dannélls et

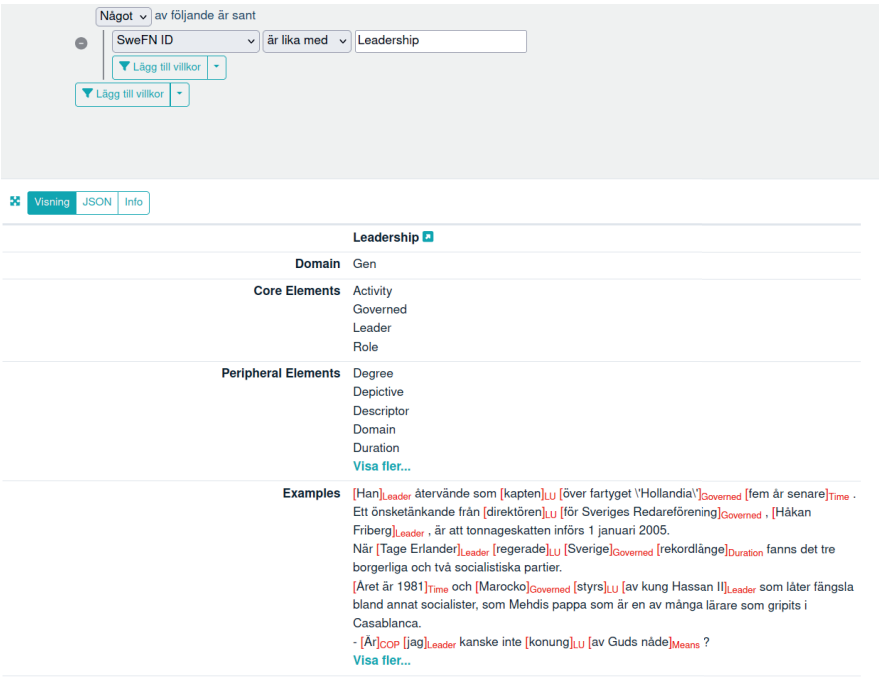


Figure 5: Karp view showing results for the `Leadership` frame

al. 2021). In other cases, frames or elements thereof were renamed, either to better fit the frame's description in Swedish context or because the new names were seen as more fitting. An example is the common frame element name EXPLANATION, which was considered vague and was replaced in the first version of SweFN by terms such as CAUSE or REASON. Dannélls et al. (2021) provide a detailed description of how Swedish FrameNet was constructed, what it contains, and the basic assumptions behind the semantic annotations of its content. Here, we focus primarily on the updated version, SweFN 2.0, which has not been described before.

SweFN 2.0 has been updated to align with BFN 1.7. Our approach to extend SweFN 1.0 combined both automatic and manual efforts. First, we automatically extracted a list that concatenates all frames from SweFN 1.0, BFN 1.5 and BFN 1.7, then identified mismatches between them. Next, we manually reviewed the list, noting how to correct mismatches by adding, replacing, or adjusting frame element names. Finally, we developed a program to automatically generate SweFN 2.0 based on these manual corrections. The second version of the Swedish FrameNet, SweFN 2.0 (Språkbanken Text 2024a), was created with a stronger emphasis on maintaining compatibility with the English FrameNet. All frames present in BFN were included,

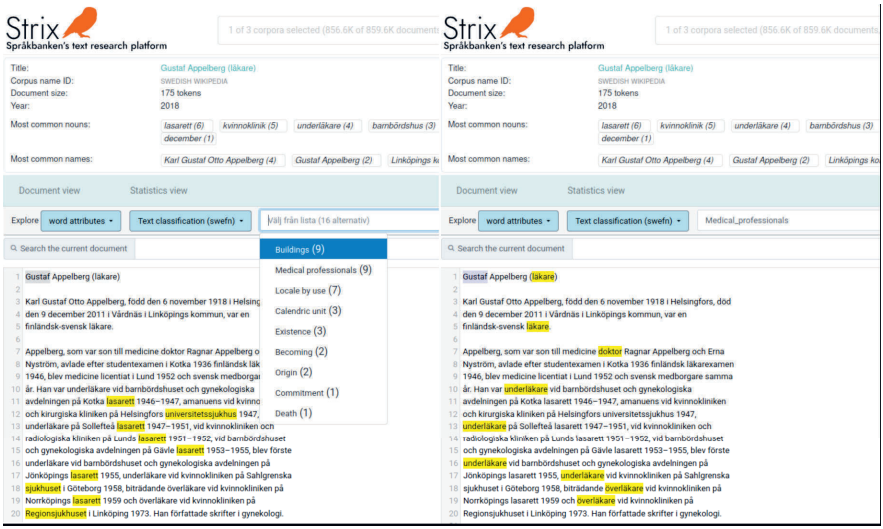


Figure 6: Screenshot of SweFN in Strix, showing search results for the frames `Building` (left), and `Medical_professionals` (right)

along with all frame elements from BFN. However, many additional frame elements were retained, and several frame element names that had been altered in earlier versions were realigned with BFN 1.7.

One of the key challenges in creating a new version of the Swedish FrameNet was balancing compatibility with the English FrameNet and other international counterparts. On one hand, the resource should be compatible with international resources, such as parallel corpora or translation methods; on the other hand, it should be the best possible resource for Swedish without compromising too much. Therefore, a one-to-one translation of frames and frame elements may not be optimal for the Swedish language. In short, the conclusion was to make the Swedish FrameNet essentially an expansion of the English Framenet, with additions allowed but no subtractions. This approach ensures that any analysis based on the English FrameNet can be translated into the Swedish FrameNet, even if the reverse is not true.

In total, the number of frames in BFN has increased from 1,020 in version 1.5 to 1,222 in version 1.7. In SweFN 2.0, the number of frames reached 1,329. The total number of lexical units and example sentences remained unchanged at 39,212 and 9,018 respectively, as shown in Table 4. In terms of lexical unit coverage, SweFN 2.0 arguably remains the largest framenet in the world.

Table 4: SweFN 2.0 in numbers

	Current status	Changes since previous version
Number of frames	1,329	134 added, 7 renamed
Number of elements	12,423	266 added, 392 renamed
Number of lexical units	39,212	—
Number of example sentences	9,018	—

3.3 SweFN for semantic role labelling

Semantic role labelling (SRL) is the task of automatically identifying semantic and morpho-syntactic information of unannotated sentences. It is an area where FN annotations have been successfully applied to solving core NLP tasks requiring natural language understanding, such as information extraction and question answering (Das et al. 2014). However, training SRL models requires large amounts of annotated semantic data. Until recently, this was challenging for languages with relatively small annotated semantic datasets, such as those with fewer than 100,000 sentences (Johansson 2021). This situation has changed thanks to the significant advancements in deep neural network and transformer technologies, allowing the development of an end-to-end Swedish SRL model based on only 9,018 annotated sentences (Dannélls, Johansson & Yang Buhr 2024).

By allowing the development of an end-to-end Swedish SRL model, we aim to bridge the current gap, as Swedish SRL models have not yet been utilised for downstream NLP applications. In addition, SRL models hold significant potential for addressing future challenges, such as semantic change detection, which we elaborate on in the next section.

4 Future application areas

As already mentioned, BFN has been exploited in numerous practical NLP downstream applications that require knowledge not only of linguistic form but also of word meanings and their relationships in specific contexts. In this section, we showcase FrameNet, particularly SweFN, and its potential to improve semantic change detection.

Languages naturally evolve over time, influenced by cultural, societal, and technological changes. One significant aspect of this evolution is semantic change, which refers to the gradual transformation in the meanings of words, phrases, and

expressions over time. For example, the word *girl* historically referred to a young person of any gender but later narrowed to mean specifically a young female.

Understanding semantic change is crucial for two primary reasons: (1) it underpins foundational studies in the social sciences and humanities, especially those exploring the interplay between language and culture; and (2) it contributes to the development of advanced language models that account for the temporal and evolutionary dynamics of natural languages.

Existing approaches to semantic change detection predominantly rely on statistical methods, which, while enabling large-scale analysis, are inherently unsupervised and lack precision. These methods struggle to capture complex, context-dependent phenomena, are difficult to adapt to domain-specific applications, and typically focus on single words without extending to broader conceptual structures. Crucially, they fail to identify the type of change (e.g., narrowing, broadening, metaphorical shift) or the reasons behind the change, limiting their usefulness for qualitative studies or hypothesis-driven research.

To overcome these challenges, we can leverage frame semantics and FrameNet. As mentioned previously, frame semantics organises meaning around conceptual structures called frames, which represent prototypical situations involving participants, props, and actions. FrameNet offers a repository of frames and their associated lexical units, along with frame elements that describe the roles played by entities in a given context. This structured representation is particularly well-suited for analysing the evolution of meaning over time.

Framenet-based semantic change detection involves parsing historical and contemporary corpora using an SRL to extract the realisation of specific frames. Suppose we have a frame `Market`, to refer to a concrete, physical location where buyers and sellers exchanged goods. Over time, this frame is expected to expand to include abstract and metaphorical meanings, such as in phrases like *love market* or *the market reacted to the news.* Such a frame does not currently exist in the English, Swedish, or Global FrameNet (although one could envision the frame in one of the Global FrameNet's domains as discussed in Section 2.2). Given two frames representing the physical and concrete senses of `Market`, with an SRL, we could:

1. Extract occurrences of the `Market` frame from diachronic corpora.
2. Identify and analyse the frame elements (e.g., Buyer, Seller, Goods, Location, System) associated with each instance of the frame.
3. Compare the distribution and contextual usage of these frame elements across different time periods.

For example, in earlier texts, frame elements like Buyer, Seller, and Location might dominate, reflecting the physical sense of `Market`. In more recent texts, frame elements like Opportunity or System might emerge, highlighting the abstract and

metaphorical extensions of the frame. This shift in frame elements provides direct evidence of semantic change.

Frame-semantic analysis not only can detect semantic change but also help identify its type and potential causes. For the `Market` frame, we can observe a process of *broadening*, where the meaning expands from a concrete, physical space to encompass abstract systems or concepts. By aligning these changes with historical and cultural shifts, such as the rise of digital economies and globalisation, we can infer causal relationships between societal developments and linguistic evolution.

The structured methodology provided by FrameNet allows for more precise and interpretable analyses of semantic change. Beyond detecting and explaining changes, this approach could support diverse research questions, such as:

- What are the patterns and processes underlying different types of semantic change (e.g., metaphorical shifts, narrowing, broadening)?
- How do cultural and societal changes drive linguistic evolution, and vice versa?
- Can semantic change be predicted by cultural or technological developments?

Moreover, understanding semantic change has practical implications for mitigating societal harm. By identifying and addressing linguistic shifts that perpetuate biases or negative societal impacts, we can design proactive interventions, particularly in the era of large generative language models (LLMs) like ChatGPT.

5 Summary

Swedish FrameNet (SweFN) is a valuable lexical-semantic resource developed as part of the Swedish lexical infrastructure at Språkbanken Text. The first version, created based on the principles of Berkeley FrameNet (BFN), was introduced over a decade ago. Since then, BFN has undergone several updates, leading to inconsistencies between SweFN and the latest version of BFN. As part of the Global FrameNet initiative, it became essential to harmonise the semantic content of SweFN with BFN 1.7, ensuring compatibility with existing multilingual framenet resources.

In this chapter, we present SweFN 2.0 – a freely available resource, which is also accessible via Karp, Språkbanken Text's data editing platform. We demonstrate how it is integrated into Språkbanken Text's infrastructure, remaining an accessible resource for researchers and supporting progress in both current and future application areas. We also discuss how SweFN, and particularly FrameNet-based semantic models, can advance Swedish language technology, with a specific focus on the area of semantic change.

References

Baker, Collin F., Charles J. Fillmore & John B. Lowe. 1998. The Berkeley FrameNet project. *International Conference on Computational Linguistics (COLING) 1998*. 86–90.

Berk, Lynn M. 1999. *English syntax: From word to discourse*. New York: Oxford University Press.

BNC Consortium. 2007. *The British National Corpus, XML Edition*. [Online resource] http://www.natcorp.ox.ac.uk/XMLedition/. Accessed on 2025-03-24.

Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj & Dimitrios Kokkinakis. 2010. The past meets the present in Swedish FrameNet++. *Proceedings of the European Association for Lexicography (EURALEX) 2010*. 269–281.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. *Proceedings of SLTC 2016*. 17–18.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. DOI: 10.1007/s10579-013-9233-4.

Borin, Lars, Maria Toporowska Gronostaj & Dimitrios Kokkinakis. 2007. Medical frames as target and tool. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA) workshop FRAME 2007: Building Frame Semantics resources for Scandinavian and Baltic languages*. 11–18.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó & Manfred Pinkal. 2009. FrameNet for the semantic analysis of German: Annotation, representation and automation. In Hans C. Boas (ed.), *Multilingual FrameNets in computational lexicography: Methods and applications*, 209–244. Berlin: De Gruyter Mouton.

Chiarcos, Christian, Sebastian Nordhoff & Sebastian Hellman (eds.). 2012. *Linked data in linguistics: Representing and connecting language data and language metadata*. Berlin: Springer.

Cruse, David Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Dalpanagioti, Thomai. 2024. Integrating frame semantic resources in EFL instruction with a focus on deliberate metaphor. In Annette Klosa-Kückelhaus & Martina Nied Curcio (eds.), *New challenges in a multilingual, digital and global world*, 271–298. Berlin: De Gruyter. DOI: doi:10.1515/9783111373294-012.

Dannélls, Dana, Lars Borin, Markus Forsberg, Karin Friberg Heppin & Maria Toporowska Gronostaj. 2021. Swedish FrameNet. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 37–65. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (eds.). 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Dannélls, Dana, Richard Johansson & Lucy Yang Buhr. 2024. Transformer-based Swedish semantic role labeling through transfer learning. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 16762–16769.

Das, Dipanjan, Desai Chen, André F. T. Martins, Nathan Schneider & Noah A. Smith. 2014. Frame semantic parsing. *Computational Linguistics* 40(1): 9–56.

Fellbaum, Christiane. 1998. Introduction. In Christiane Fellbaum (ed.), *WordNet: An electronic lexical database*, 1–19. Cambridge: MIT Press.

Fillmore, Charles J. 1968. The case for case. In Emmon Bach & Robert Harms (eds.), *Universals in linguistic theory*, 1–88. New York: Holt, Rinehart, & Winston.

Fillmore, Charles J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280(1): 20–32. DOI: https://doi.org/10.1111/j.1749-6632.1976.tb25467.x.

Fillmore, Charles J. 1982. Frame semantics. In Linguistic Society of Korea (ed.), *Linguistics in the morning calm*, 111–137. Seoul: Hanshin Publishing Co.

Fillmore, Charles J., Christopher R. Johnson & Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3): 235–250.

Friberg Heppin, Karin & Dana Dannélls. 2015. Polysemy and questions of lumping or splitting in the construction of Swedish FrameNet. *Proceedings of the Workshop on Semantic resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at the Nordic Conference of Computational Linguistics (NODALIDA) 2015*. 12–20.

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge: MIT Press.

Johansson, Richard. 2021. Semantic role labeling. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 263–280. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.04ade.

L'Homme, Marie-Claude. 2014. Why lexical semantics is important for e-lexicography and why it is equally important to hide its formal representations from users of dictionaries. *International Journal of Lexicography* 27(4): 360–377. DOI: 10.1093/ijl/ecu019.

Malm, Per, Shafqat Virk, Lars Borin & Anju Saxena. 2018. LingFN: Towards a framenet for the linguistics domain. *Proceedings of the International FrameNet Workshop at LREC 2018: Multilingual Framenets and Constructicons*. 37–43.

Marzinotto, Gabriel, Jeremy Auguste, Frederic Bechet, Geraldine Damnati & Alexis Nasr. 2018. Semantic frame parsing for information extraction : The CALOR corpus. *International Conference on Language Resources and Evaluation (LREC) 2018*. 986–993.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, Collin F. Baker & Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*. Research report. Berkeley: ICSI.

Språkbanken Text. 2024a. *Svenskt frasnät 2.0 (SweFN 2.0)* [swedish framenet 2.0 (swefn 2.0)]. [Data set]. DOI: https://doi.org/10.23695/f1sz-0572.

Språkbanken Text. 2024b. *Swedish party programs and election manifestos*. [Data set]. DOI: 10.23695/nc55-gd27.

Taniguchi, Ryosuke, Reina Hoshino & Yoshinobu Kano. 2019. Legal question answering system using FrameNet. In Kazuhiro Kojima, Maki Sakamoto, Koji Mineshima & Ken Satoh (eds.), *New frontiers in artificial intelligence*, 193–206. Cham: Springer.

Torrent, Tiago Timponi, Michael Ellsworth, Collin F. Baker & Ely Edison da Silva Matos. 2018. The Multilingual FrameNet shared annotation task: A preliminary report. *Proceedings of the International FrameNet Workshop 2018: Multilingual Framenets and Constructicons*. 62–68.

Torrent, Tiago Timponi, Maria Margarida M. Salomão, Fernanda C. A. Campos, Regina M. M. Braga, Ely E. S. Matos, Maucha A. Gamonal, Julia A. Gonçalves, Bruno C. P. Souza, Daniela S. Gomes & Simone R. Peron. 2014. Copa 2014 FrameNet Brasil: A frame-based trilingual electronic dictionary for the football world cup. *International Conference on Computational Linguistics (COLING) 2014: System Demonstrations*. 10–14.

Lars Borin

8 Semantic (onomasiological) lexical resources

Abstract: By *semantic* (or *onomasiological*) *lexical resources* we understand computational lexicons (designed to be used in automatic language processing and linguistic analysis applications) where the basic lexical items designate word senses or general concepts rather than lexemes, and further where the word senses are explicitly interrelated by specific lexical-semantic relations, such as *hyponymy* or *synonymy*. The most well-known such resource is the Princeton WordNet. Språkbanken's Lexical Research Infrastructure includes several onomasiological lexical resources, interestingly different in their organization, in particular with regard to the lexical-semantic relations posited among word senses. They include the pivot lexicon Saldo, the wordnet-like resource Swesaurus, the Swedish Roget-style thesaurus Bring, and some other resources. In this chapter we describe these onomasiological lexical resources, their similarities and differences and their (potential) uses in computational language processing.

Keywords: conceptual lexicon, emotion lexicon, framenet, lexical semantics, onomasiology, semantic lexicon, sentiment lexicon, thesaurus, wordnet

1 Introduction: onomasiological and semasiological lexical resources

In lexicographical parlance, an *onomasiological* lexicon is one organized by content – word senses – rather than by form. In fact, the onomasiological dictionary organization goes back further in history than its opposite, the *semasiological* one (Civil 1990; Hüllen 1999: 28–39). Nevertheless, the opposite lexical organization – by form

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161). Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded to me in 2024 for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-008

(alphabetically or the equivalent with non-alphabetic writing systems) rather than by meaning – is the more common and familiar one nowadays.

The onomasiological lexical resources described in this chapter have different characteristics and different provenances (see the individual sections for details), but they have the following in common. Their inclusion in Språkbanken's Lexical Research Infrastructure (SBLRI) is always through Saldo, an onomasiological lexical resource described below and in Chapter 6 in this volume, by linkage of their entries to Saldo's persistent identifiers (word senses or lexemes).[1] These then automatically provide a connection to the whole lexical macroresource that is at the heart of the SBLRI; notably, word senses are associated with the corresponding lexemes and their full inflectional paradigms, a prerequisite for automatic linguistic analysis of texts.

Some of the resources presented in this chapter have been described in other publications (see references in the corresponding sections), but never in relation to each other in the way that this is done here. The Semantic Database (SDB) has been most fully described in internal research reports written in Swedish and mentioned in passing in a few international conference papers; hence, this interesting resource receives its first more detailed treatment in English for an international audience in this chapter. The newest version of SenSaldo (v. 0.2) as well as the SB-RID are presented for the first time here (see Section 4.4).

2 The long onomasiological tradition of Språkbanken

Språkbanken has a long and somewhat disconnected history of compiling onomasiological lexicons. One – early – branch grew out of the ambitious lexical database project initiated in the 1970s as part of the effort resulting in SOB (1986), the so-called *Gothenburg Lexical Database* (GLDB; Ralph 1977; Ralph, Järborg & Allén 1977). The aim of GLDB was to systematize Swedish lexical knowledge in a database where the information found in ordinary printed dictionaries would be explicitly formally structured – as opposed to the implicit structure often ambiguously conveyed by the typography of printed dictionaries. The ambitious aim was to include the whole

1 The exceptions are the older Språkdata lexical databases that were compiled pre-Saldo, viz. the semantic database (SDB) and SIMPLE (see Section 3), and also the Swedish WordNet. These are still in the pipeline for inclusion in our lexical infrastructure, pending the resolution of some remaining intellectual property rights (IPR) issues, but are described here for completeness' sake, as important elements of the history of computational lexicography at the University of Gothenburg.

Contemporary Swedish vocabulary,[2] as evidenced in corpus data, while any actual dictionary would then constitute a principled selection from the GLDB, compiled using the rich lexical information and metadata included in the database. Note that this means that entries would not be deleted from the GLDB, only additionally marked as obsolete, old-fashioned, or the like (for instance, by using time-series corpus data. So, as an added value, the GLDB would also provide a cumulative, fine-grained record of lexical change in Swedish over the time period covered by the underlying corpora (thereby complementing the more coarse-grained approach represented by the SAOLhist initiative, whose "vocabulary snapshots" are typically years or even decades apart; see Chapter 13 in this volume). Relevant in our context, it should mean that the GLDB would be neutral with regard to the semasiological–onomasiological distinction, that would only then be different views on the same underlying set of data.

The GLDB was never realized quite as envisioned. Instead, what ensued was a series of dictionary-specific databases – one for SOB (1986), one for SAOL 11 (1986), one for NEO (1995), and so on – whose content constituted a very small superset of the entries of the particular dictionary and whose information models were generally not completely mutually compatible.

3 The beginning: SDB, SIMPLE, and SENSEVAL

In a series of projects focusing on formal semantic description of lexical items (Järborg 1996), a version of GLDB with enhanced and corrected semantic descriptions was compiled. This *Semantic Database* (SDB) was built from a subset of the entries of the GLDB, in parallel with another database version that was used to compile the (paper) dictionary *Nationalencyklopedins ordbok* (NEO 1995).

The SDB was then further developed in the framework of a project titled *Lexical Sense and Sense in Context* (1998–2003; also referred to as *SemTag*; Järborg, Kokkinakis & Toporowska Gronostaj 2002; Järborg 2003).

The purpose of SDB was not primarily computational in the sense of enabling automatic processing of Swedish text. Rather, its compilation was conceived as an exercise in formal lexical semantics, where the purpose of adopting a formal database format for the data was to promote consistency in the description.

In parallel with the work on SDB, a subset of the colexified lexemes in the one million word Stockholm–Umeå Corpus (SUC; Ejerhed et al. 1992) were annotated

2 According to the conventional historical periodization of Swedish, Contemporary Swedish extends from 1906 until the present day.

Table 1: Semantic features in the Swedish SIMPLE lexicon entry *avskeda* 'lay off' (see Lenci et al. 2000 for an explanation of the SIMPLE feature set)

Semantic feature	Value
Template type:	Cause_constitutive_change
Domain:	general (default value)
Semantic class:	Change
Predicates type:	lexical
Multilingual:	No
Argument_list:	Arg0 Arg1
Arguments selectional restrictions:	Human
Arguments obligatory status:	Check
Correspondence between syntactic and semantic arguments:	isomorphic bivalent

for the SDB word sense of each instance, a total of some 137,000 tokens. Again, the purpose of this exercise was not primarily to serve natural language processing (NLP) aims, but to test the efficacy of formal semantic description in practice, as it were. Järborg (1999) reports that the agreement among the (several) annotators involved in the project on the whole was satisfactorily high (although no interannotator agreement was calculated).

However, it is probably true to say that the work on SDB and corpus annotation in combination with the general lexicography-oriented corpus linguistic activities in which this work was embedded paved the way for the involvement in some relevant NLP research and development (R&D) activities.

The EC-funded PAROLE project (1996–1997) resulted in a large corpus, annotated for part of speech (POS) and morphosyntactic description (MSD), corpus and a computational lexical resource with morphological and syntactic information, the PAROLE lexicon (Språkbanken Text 2024a) with some 30,000 entries. This lexicon was specifically intended as an NLP resource.

The PAROLE lexicon is not a semantic lexical resource and hence not in scope here, but – relevant in the context of the present chapter – this computational lexical strand of research was continued in a follow-up EU project, SIMPLE, where parallel semantic lexical resources were compiled for the 12 project languages on the basis of the PAROLE lexicons (Lenci et al. 2000). The Swedish SIMPLE lexicon (Språkbanken Text 2024c) provides detailed semantic information for about 9,000 entries, encoded using a project-specific SGML format. For example, the semantic information about the verb *avskeda* 'lay off' contains the semantic features shown in Table 1. There are also versions of both the PAROLE and SIMPLE lexicons where the entries have been linked to Saldo word senses, referred to as *PAROLE+* (Språkbanken Text 2024b) and *SIMPLE+* (Språkbanken Text 2024d), although without disambiguation, so that

for instance the two SIMPLE (and PAROLE) entries for the adjective *varm* 'warm' are each linked to two Saldo word senses: varm..1 (about a physical quality or bodily sensation) and varm..2 (about a personality or behavior).

In the SIMPLE project, experiments were conducted on automating the acquisition of the requisite semantic knowledge for entries considered for inclusion into the lexicon using NLP tools and extensive text corpora (Kokkinakis, Toporowska Gronostaj & Warmenius 2000; 2001a).

The work on onomasiological lexical resources in the SemTag and SIMPLE projects without doubt was an important contributing factor when the Språkdata research group undertook to arrange and also participate in a Swedish exercise within the SENSEVAL-2 shared task on word sense disambiguation (Kokkinakis, Toporowska Gronostaj & Warmenius 2001b; Kokkinakis, Järborg & Cederholm 2001).

The SemTag project funding ended in 2003 and at the same time the lexical and computational linguistic branches of the erstwhile Språkdata went their separate ways (see Chapter 2 in this volume). In that connection the IPR of both the SDB and SIMPLE turned out to be overly restrictive for the envisioned uses to which an onomasiological computational lexical resource would be put in Språkbanken (see Chapter 6 in this volume). For this reason both resources got sidetracked in the further development of onomasiological lexical resources at the University of Gothenburg.

The IPR situation has changed considerably for the better in the more than 20 years that have passed since then, and both SDB and SIMPLE contain valuable lexical-semantic information that is not available through other lexical resources in Språkbanken Text, information that was manually added and vetted by expert lexicographers. Consequently, we see as a longer-term goal to include both SDB and SIMPLE in Språkbanken's Lexical Research Infrastructure, even if this is not completely straightforward, since these resources are not documented to the desired extent in all particulars, especially regarding their formal structuring: the lexical-semantic conceptual framework is much better documented than the lower-level data model implementing this framework as a concrete database.

4 Semantic resources in SweFN++ and beyond

The *Swedish FrameNet++* (SweFN++) was the name of a focused R&D activity pursued at Språkbanken for upwards of a decade, with the general aim of integrating and upgrading existing lexical resources, as well as extending them with some new resources, primarily a Swedish framenet (see Chapter 7 in this volume), but also some of the resources described in this chapter. The SweFN++ initiative is described

```
alias..1        namn..1         annan..1        alias..ab.1
alibi..1        bevisa..1       annanstans..1   alibi..nn.1
bevisa..1       visa..1         PRIM..1         bevisa..vb.1
namn..1         PRIM..1         PRIM..1         namn..nn.1

alias..ab.1     alias   ab      ab_i_aldrig
alibi..nn.1     alibi   nn      nn_5n_saldo
bevisa..vb.1    bevisa  vb      vb_1a_laga
namn..nn.1      namn    nn      nn_6n_blad

alibi..1        word sense identifier
PRIM..1         the "unique beginner" top-level artificial word sense
alibi..nn.1     lemgram (lexeme) identifier
nn              POS identifier
nn_5n_saldo     paradigm/inflectional class identifier
```

Figure 1: The various persistent identifiers used in Saldo, here in the Saldo 1.0 format (top group: word sense entries; middle group: lemgram entries; bottom group: the identifiers explained)

in more detail in Chapter 5 in this volume. A book-length treatment that goes into considerable detail regarding all aspects of the project and its various stages is also available (Dannélls, Borin & Heppin 2021).

4.1 Saldo and Swedish FrameNet

Saldo and the *Swedish FrameNet* (SweFN) were the core resources of the Swedish FrameNet++ lexical macroresource, now referred to as Språkbanken's Lexical Research Infrastructure. For this reason, they are accorded chapters of their own. See Chapters 6 (Saldo) and 7 (Swedish FrameNet) in this volume where these resources are described in detail.

Both Saldo and SweFN are onomasiological lexical resources, and consequently form the core of the lexical macroresource also in that regard. Saldo was designed from the beginning to provide the "glue" holding all the onomasiological lexical resources together. The work on SweFN – the completely new lexical resource to be compiled in the original Swedish FrameNet++ project (2011–2014) – was the first "acid test" of Saldo's information model as suitable to serve as the pivot of SweFN++. As mentioned in Chapter 6 in this volume, Saldo is structured through a set of persistent identifiers (PIDs), intended to be both manipulated by machines and handled by humans. Importantly, SweFN was built largely manually, by assigning Saldo word senses to existing – or sometimes newly created or modified – Berkeley FrameNet frames. This work was facilitated by not having to convert between machine-generated formal numerical identifiers and human-readable word sense labels, which used to be the normal working mode for example in the case of the lexical databases developed

for the dictionary projects being pursued separately in the Center for Lexicology and Lexicography in the same department. The Saldo model has turned out to be a very sensible design decision. Non-human-readable PIDs – such as numerical codes, handle identifiers, or DOIs – are of course formally equivalent. However, with well-designed mnemonic identifiers for key entities, all manual and computer-assisted lexicographical work becomes much more efficient. Figure 1 (from Chapter 6 in this volume and repeated here for convenience) summarizes the formats of the various identifiers used in Saldo.

Through its information model, Saldo defines the "lexical grid" by which all our other lexical resources must orient themselves. It provides persistent identifiers for Swedish word senses and lexemes, to which the corresponding entities in other resources are linked. Focusing our attention on the word sense identifiers, in the simplest case this is identity, i.e., a word sense in a lexical resource is identified as a Saldo sense, which requires identity of form as well, i.e., that we are dealing with (variants of) the same lexeme. If the referred concepts are identical but expressed by different lexemes, we have different word senses in a (full) synonymy relation. In other cases, there is a set of lexical-semantic relations available for describing how the item in question should be linked to Saldo. Typically this will be as a broader concept (a hyperonym or superordinate sense), a narrower concept (a hyponym or subordinate sense), or a near-synonym. More refined relationships, such as cohyponymy, can also be expressed (see Section 4.2).

4.2 Swesaurus: towards a Swedish wordnet

One of the most popular and most used onomasiological resources in language technology is the English-language Princeton WordNet (PWN; Fellbaum 1998) and corresponding resources in other languages built according to the same lexical-semantic model as PWN.[3] While Saldo, the pivot resource of the Swedish FrameNet++, is also an onomasiological lexical resource reminiscent of a wordnet, its structure is significantly different from that of a wordnet (Borin & Forsberg 2009; Borin, Forsberg & Lönngren 2013; see also Chapter 6 in this volume), and adding a Swedish wordnet to our onomasiological resources was seen as an important goal when the plans for the SweFN++ project were drawn up, and work on an open Swedish wordnet was initiated as part of the project activities.

In the project, in a sense our focus was on sustainability, finding a way of recycling as much as possible relevant lexical-semantic information already present in

3 Global WordNet Association (2025) provides a partial list of wordnets in many languages.

Table 2: Lexical-semantic relations used in Swesaurus and their logical properties (used for inferring missing links among lexical items)

Relation	Operator	Logical properties
synonymy degree	60–00	symmetric, transitive(?)
full synonymy	ss	(= 00) symmetric, transitive
near synonymy	cc	(= 90) symmetric, transitive(?)
antonymy	aa	symmetric
related sense	rr	symmetric, transitive(?)
hyponymy/subordinate sense	iu	transitive, inverse of hyperonymy
hyperonymy/superordinate sense	ui	transitive, inverse of hyponymy
cohyponymy	iuui	symmetric, transitive
partonymy	pt	transitive(?), inverse of holonymy
holonymy	tp	transitive(?), inverse of partonymy

existing available resources, so that the considerable lexicographical effort spent on compiling this information would not have been spent in vain. Thus it was natural to investigate if the existing (partial) Swedish WordNet (Viberg et al. 2002) could be included in SweFN++. However, the responses to our inquiries made it clear that that resource would not be made openly available in a way that would make it useful in a language technological setting.[4]

Recycling existing lexical-semantic information included both defining a mapping among the different ways that the same relation could be expressed in different resources, and deciding on a standardized set of relations and terminology for expressing them for the resulting wordnet-like resource. For completely independent reasons, we need to cater to the fact that different lexical resources that should be interlinked recognize partly different sets of word senses for the same lexeme. This is perhaps most obvious when we include historical lexical resources in the macroresource (see Chapter 13 in this volume), but is also noticeable in contemporaneous resources. Table 2 lists the semantic relations currently in use in SweFN++ (a subset of those used in PWN).

The wordnet-like resource constructed in the framework of SweFN++ is called *Swesaurus* (Borin & Forsberg 2010; 2011; 2014; Språkbanken Text 2017). In its current form, Swesaurus is both less and more than a wordnet. It is less in the quite trivial sense that it is still under construction, but also more significantly because of a fundamental theoretical difference regarding the definition of *synonym(y)*. In the lexical-semantic model adopted in the SweFN++ work, the basic atomic lexical-semantic entity of all our lexical resources is the word sense, basically a Saussurean

4 This position has changed in the meantime; see further below.

linguistic sign with both form and content. The content could be referred to as a concept, if you like, but in that case it is important to stress that concepts are extralinguistic entities according to this view, referred to by the content side of a word sense. In the (fairly rare) case when two or more linguistic forms – lexemes – identify the same concept, we have synonymy, but still two word senses. Not only synonymy, but all lexical-semantic relations hold among word senses on this view, which is the one chosen for SweFN++ including Swesaurus. Contrary to this, synonymy receives special treatment in the lexical-semantic model underlying WordNet. The fundamental units constituting WordNet are called *synsets*, defined as "sets of synonyms that serve as identifying definitions of lexicalized concepts" (Miller et al. 1990: 240). This means that synonymy – defined in a way that deviates from how this term is conventionally understood in lexicography – holds a special place in WordNet, different from and more basic than other classical lexical-semantic relations, where the latter with a few exceptions are understood in WordNet to hold among synsets, not word senses. Additionally, at present only a small subset of the entries and potential entries in Swesaurus are linked to PWN synsets.[5]

At the same time Swesaurus is also more than a wordnet in at least the following respects:

- it includes word senses from all parts of speech, not only open/lexical items as PWN-style wordnets;
- it provides a traditional, flexible notion of synonymy, with the possibility of indicating the degree of synonymy between word senses.

The information about lexical-semantic relations is extracted or inferred from various resources forming part of Språkbanken's Lexical Research Infrastructure, expressed in several different formats. Because of the logical relations obtaining among some of the lexical-semantic relations, it is often possible to infer ("fill in") information that is not explicitly stated in a resource. The present version of Swesaurus records lexical-semantic relations among slightly over 15,000 word senses, compiled by reusing information about lexical-semantic relations in a number of freely available lexical resources for Swedish. It lists about 57,300 (word-sense) *relational triples*,[6] that constitute the basic information units in Swesaurus, and whose three components are: (1) a source word sense; (2) a lexical-semantic relation (see Table 2); and (3) a target word sense. In addition, each triple has provenance information,

5 According to Global WordNet Association (2025) "the wordnet design" must include "links to WordNet (Princeton or others that are linked to PWN) [and] WN structure (minimally: synset, hyponymy)".

6 Each triple cooccurs with its mirror version in the Swesaurus dataset, i.e., one with the inverse relation where the source and target word sense are reversed.

i.e., from which resource it originates and whether it is primary or derived. All relations except related-sense are generally taken to hold only within a part of speech, i.e., source and target word senses must belong to the same part of speech. The provenance of the triples in the current version of Swesaurus is as follows:

- Synlex (Kann & Rosell 2006) – a crowdsourced list of Swedish synonym and near-synonym pairs with degree of synonymy (25,528 explicit and 25,246 derived relations);
- Wiktionary – a web-based project for collaboratively creating a free lexicon (7,714 explicit relations);
- the Swedish version of the Princeton Core WordNet (4,676 explicit relations).

The work on Swesaurus has long been on hold awaiting the release of Saldo 3. Now that this release has happened, wordnet integration is a top priority. Concretely, all of the following datasets will be made part of the next version of Swesaurus:

- all of the present version of Swesaurus;
- all of the Swedish IDS/LWT list (see Section 4.5);
- all of the Swedish WordNet (Viberg et al. 2002), which has now explicitly been released as an open resource;
- as much of Saldo 3 as is feasible;
- all the lexical-semantic relation information available in the SDB.

The highest yield is expected from Saldo and SDB, easily around an additional 110,000 triples. The Swedish WordNet and the Swedish IDS/LWT list will contribute additional PWN synset links, but perhaps not so many additional relational triples.

4.3 Bring's thesaurus resurrected

As mentioned in Section 4.2, the English Princeton WordNet and its clones in other languages are the best-known and most frequently used onomasiological resources in the field of language technology. However, in most other contexts the most well-known lexical-semantic resource for English is without doubt Roget's *Thesaurus* (Roget 1852; Hüllen 2004), which appeared in its first edition in 1852 and has since been published in numerous editions all over the English-speaking world.

Roget's thesaurus, although sometimes referred to as a "synonym dictionary" (e.g., by Hüllen 2004), is organized quite differently from a wordnet, and the *synonymy* referred to of a very different kind from that used to define the synsets of a wordnet. Consequently, a lexical resource based on Roget arguably offers a valuable complement to WordNet, one that has been used in language technology both to address other kinds of lexical-semantic analysis tasks than a wordnet and also has

been shown to be more effective for some of the tasks where wordnets are normally used, e.g., *lexical cohesion*, *synonym identification*, *pseudo-word-sense disambiguation*, and *analogy problems* (Morris & Hirst 1991; Jobbins & Evett 1995; Jarmasz & Szpakowicz 2004; Kennedy & Szpakowicz 2008; 2014).

A Swedish adaptation of Roget's thesaurus was published in 1930, Sven Casper Bring's *Svenskt ordförråd ordnat i begreppsklasser* 'Swedish vocabulary arranged in conceptual classes' (Bring 1930). The digitized content of Bring's thesaurus forms the basis for two lexical resources made available by Språkbanken Text:

1. *Bring* (v. 1.0), providing the full contents of the original 1930 book version (148,846 entries) in a formally structured format but without any linkages to other resources;
2. *Blingbring* (v. 0.3), a version of Bring that has been curated to remove obsolete items. This version contains 126,911 entries (∼85% of the original), each linked to all possible corresponding Saldo sense identifiers, i.e., without disambiguation.

The initial linking to Saldo senses in Blingbring did not involve a disambiguation step since the Zipfian distribution of word senses over lexemes means that most entries have only one corresponding Saldo sense when assigning these on the basis of matching lemma–POS combinations from the two resources. Blingbring includes slightly over 21,000 entries with more than one Saldo sense (∼17%), or about 4,800 ambiguous word sense assignments (out of about 43,000 unique lemma-POS combinations: ∼11%).

The formal structure of Bring is a more shallow version of that in the original Roget. At the highest level, there are 1,015 numbered *conceptual classes*. Each class comes with a label indicating a broad semantic characterization of the lexical items listed in the class. Following Roget, most classes come in pairs of opposite semantic fields, e.g., classes #17 *likhet* 'similarity' and #18 *olikhet* 'dissimilarity'. These semantic fields are often quite abstract, and hence a particular Saldo word sense may be included in more than one Bring class. For example, the Saldo entry *nirvana*[1] 'nirvana' is listed in class #981 *himmel* 'heaven', but also in #2 *intighet* 'inexistence' and #360 *död* 'death'.

Again following Roget, classes are further subdivided into parts of speech, with one division each for nouns, verbs and a third category containing words of other parts of speech, mainly adjectives and adverbs, but also idioms and some function items. The lowest-level unit above the individual lemmas – or *entries* – in Bring is the *group* (marked by a final semicolon in the printed version). The entries in a group are often further arranged so that the distance between their positions in the text corresponds coarsely to the semantic distance between their senses. Consequently synonym clusters can be discerned within groups, although these clusters are not formally indicated in any way.

Table 3: The Blingbring lexical resource, with Bring entry identifiers and Roget class identifiers. The mapping to Saldo has not yet been disambiguated

Bring	Roget	Entry	Saldo ID(s)
b0980/n/03/01	r8/5/2/2/980/Demon	spöke	spöke..1
b0980/n/03/02	r8/5/2/2/980/Demon	spökelse	spökelse..1
b0980/n/03/03	r8/5/2/2/980/Demon	spökeri	spökeri..1
b0980/n/03/04	r8/5/2/2/980/Demon	spökdjur	spökdjur..1
b0980/n/03/05	r8/5/2/2/980/Demon	spökhistoria	spökhistoria..1
b0980/n/03/07	r8/5/2/2/980/Demon	spökskepp	spökskepp..1
b0980/n/03/09	r8/5/2/2/980/Demon	spöktimme	spöktimme..1
b0980/n/03/10	r8/5/2/2/980/Demon	vålnad	vålnad..1
b0980/n/03/11	r8/5/2/2/980/Demon	skugga	skugga..1
b0980/n/03/12	r8/5/2/2/980/Demon	hamn	hamn..2
b0980/n/03/13	r8/5/2/2/980/Demon	maner	maner..1
b0980/n/03/15	r8/5/2/2/980/Demon	gengångare	gengångare..1
b0980/n/03/17	r8/5/2/2/980/Demon	andeuppenbarelse	andeuppenbarelse..1
b0980/n/03/18	r8/5/2/2/980/Demon	andeskådare	andeskådare..1
b0980/n/03/21	r8/5/2/2/980/Demon	vision	vision..1
b0980/n/03/22	r8/5/2/2/980/Demon	visionär	visionär..2
b0980/n/03/23	r8/5/2/2/980/Demon	syn	syn..1:syn..2:syn..3:syn..4
b0980/n/03/24	r8/5/2/2/980/Demon	varsel	varsel..1:varsel..2
b0980/n/03/25	r8/5/2/2/980/Demon	drömbild	drömbild..1
b0980/n/03/26	r8/5/2/2/980/Demon	hallucination	hallucination..1
b0980/v/01/01	r8/5/2/2/980/Demon	trolla	trolla..1
b0980/v/01/03	r8/5/2/2/980/Demon	gastkrama	gastkrama..1
b0980/v/01/04	r8/5/2/2/980/Demon	spöka	spöka..1
b0980/v/01/05	r8/5/2/2/980/Demon	varsla	varsla..1:varsla..2
b0980/v/01/06	r8/5/2/2/980/Demon	se i syne	se_i_syne..1
b0980/v/01/07	r8/5/2/2/980/Demon	hallucinera	hallucinera..1
b0980/a/01/01	r8/5/2/2/980/Demon	trolsk	trolsk..1
b0980/a/01/02	r8/5/2/2/980/Demon	infernalisk	infernalisk..1
b0980/a/01/03	r8/5/2/2/980/Demon	demonisk	demonisk..1
b0980/a/01/05	r8/5/2/2/980/Demon	besatt	besatt..1

Table 3 shows the structure of the Blingbring resource. Since the same word sense can appear in more than one Bring class, each position in the resource has its own identifier. Thus, `b0980/n/03/01` refers to Bring class 980 (with the label *Oknytt* 'malevolent supernatural being(s)'), the first word in the third group of nouns (`n/03/01`) in that class. There is also a pointer to the corresponding class in Roget (1852), in this case also numbered 980, *Demon*. The last two columns in the excerpt in Table 3 list a Bring lemma and the corresponding Saldo sense identifier(s), respectively.

Bring was published in 1930, and its most recent entries were first attested in print around 1920, although the influx of new words starts to peter out already around

1914 (Lange 2007: 11), i.e., its vocabulary is over a century old. Even if vocabulary changes faster over time than most other aspects of a language, there is still a good deal of continuity even over a century; see Chapter 13 in this volume for an empirical indication of this. As seen above, the obsolete items constitute only about 15% of the original Bring, corresponding to about 23% of all lemma–POS combinations (since a comparatively larger share of the obsolete items occur in only one class). This means that it makes sense to use Bring as the point of departure for compiling a modern Swedish thesaurus resource. In order to accomplish this, we need to: (1) disambiguate the ambiguous linkages; and (2) develop good methods for adding modern vocabulary to Bring from Saldo or some other SweFN++ component resource, placing each item in its most appropriate Bring class(es).

For both objectives we have pursued automatic, language-technology based methods, reported on in earlier publications (Borin, Nieto Piña & Johansson 2015; Zechner & Borin 2020; Borin et al. 2021). Regarding the first objective, this is a closed task, since the number of ambiguous items is fixed, but the effort spent on automating this task will hopefully also take us some ways toward fulfilling the second objective. The second objective is more difficult. Rather than just a small number of options, we now need to distinguish between a very large number of target classes, and an even larger number of groups within the classes. This is also an open-ended objective, in that we would ideally like any new sense added to SweFN++ to also be assigned its proper class(es) in Blingbring.

In an initial set of experiments applying machine learning approaches, both a corpus-based and a lexicon-based classifier were applied to the first objective, the disambiguation problem, reaching accuracies of 69% and 78%, respectively, measured on an evaluation dataset consisting of 1,308 manually disambiguated gold-standard entries (Borin, Nieto Piña & Johansson 2015). However, simply choosing the first listed sense in Saldo results in 63% accuracy, making the corpus-based method barely viable. Following up on the more promising lexicon-based approach, that utilized only one of several possible aspects of the lexical structure of Saldo, we have conducted a more detailed investigation of if and how more of Saldo's structure could be used for this purpose. The hypothesis was that the macrostructure of Saldo will correspond to that of Bring at some level. This hypothesis was tested in a series of experiments reported by Zechner & Borin (2020), where it was concluded that some aspects of the topology of the Saldo lexical-semantic network gave a quite high disambiguation accuracy (80% on average).

The second objective is more difficult, being both open-ended and underdetermined, in the sense that a particular Saldo sense could in principle belong in more than one Bring class. Even when artificially restricted to predict only one class, the accuracy of the lexicon-based approaches tried so far (Zechner & Borin 2020; Borin et al. 2021) is not sufficiently high for automatic linking to be attempted, but possibly

good enough to work as pre-filtering mechanism for manual Bring class assignment of Saldo word senses.

Like the wordnet compilation described in Section 4.2, the work on modernizing Blingbring has been on hold for different reasons, but mostly awaiting the completion of Saldo 3.

4.4 SenSaldo and SB-RID

A particular breed of popular computational onomasiological lexical resources is represented by sentiment and emotion lexicons. With an increased volume of online shopping for goods and services, the ability to mount a rapid and adequate response to customer opinion – especially negative opinion – has become a matter of priority to goods and service providers. It has also become societally important to have the ability to monitor social media platforms for, e.g., smear campaigns and aggressive fake news. In a research context, social and political scientists are interested in having reliable computational tools at their disposal for finding and characterizing instances of evaluative language in public discourse. See, for example, the overview by Pang & Lee (2008).

A component language technology aiding these goals is called *sentiment analysis*, i.e., automatically analyzing texts with regard to their expressed sentiment (also called *tenor*, *polarity*, or *valence*). This can be done in various ways, where many of the methods rely on sentiment lexicons, lists of words or lemmas annotated for (degree of) positive, neutral, or negative polarity (see, e.g., Devitt & Ahmad 2013).[7]

The theoretical and methodological issues that arise in connection with sentiment analysis of texts lie partly at the intersection of linguistic pragmatics and lexical semantics. Depending on your view of the scope of these linguistic subdisciplines, you may end up with very different thoughts about the prior polarity aspects of lexical sentiment information. Thus we find many different proposals in the literature for how sentiment polarity should be represented (if at all) in the lexicon (whether for use by humans or by machines), to which kinds of lexical entities (lemmas, lexemes or word senses) it should be ascribed, and how contextual information is to come into play when inferring the sentiment of a text passage from its constituent parts.

The methodological position taken here is that (prior) sentiment polarity forms part of a word's sense, and that a word sense only has one prior polarity. Connotations are also considered to form part of the word sense (as opposed to, e.g., the practice

7 See, e.g., Benamara, Taboada & Mathieu (2017); van der Veen & Bleich (2025) for comparisons of the relative pros and cons of lexicon-based methods for sentiment analysis, machine-learning methods, and methods using large language models.

in PWN). Hence information about both sentiment polarity and connotation belongs in the lexicon with the same right as a word-sense definition or information about lexical-semantic relations. From this follows that if a word appears in text with two different sentiment values, it must either represent two senses of this lexeme or, alternatively, reflect a (pragmatic) contextual effect, such as irony, for instance.

Språkbanken's Lexical Research infrastructure (SBLRI) offers access to two lexical resources for Swedish that contain information about sentiment and emotion, SenSaldo and SB-RID.

4.4.1 SenSaldo

The SBLRI includes a sentiment lexicon containing polarity-marked Saldo word senses: *SenSaldo* (Språkbanken Text 2019). The compilation of the first version of SenSaldo (v. 0.1) in 2018 was organized as a four-stage process (Rouces, Borin, et al. 2018; Rouces, Tahmasebi, et al. 2018a,b).

First, an initial sampling from Saldo (v. 2.3; Borin, Lönngren & Forsberg 2017) was done according to the estimated frequency distribution of Saldo word senses in part of the Gigaword corpus (Rødven-Eide, Tahmasebi & Borin 2016; Rødven-Eide 2016), a one-billion-word mixed-genre corpus of written Swedish, complete with linguistic annotations for lemma, POS, MSD, dependency syntax, and Saldo word sense. We used only the part of the corpus covering the period from 1990 onwards (approximately 940 million words), in order to avoid sampling dated words. The Saldo lexicon was designed to cover the vocabulary of late Contemporary Swedish – the language of the period after about 1950 – and this frequency-based sampling was done in order to collect a vocabulary representative of modern written language. A total of 1,998 word senses were sampled in this way, restricting the sampling criteria so that only single-word open-class items – adjectives, interjections, nouns, and verbs – with a lemma two letters or longer were sampled.

Second, the sampled items were coarsely annotated for sentiment polarity – negative, neutral, or positive – by three annotators, all annotating all items independently of each other and all having varying levels of linguistic training. In preparation for this step 200 additional word senses were first sampled and annotated in a joint exercise for training purposes and for harmonizing the annotation criteria. True to the methodological principle, stated above, that polarity and connotation are seen as inherent features of lexical items, this sentiment annotation exercise considered isolated word senses without (textual) context, but with their lexical-semantic descriptors from Saldo (see Section 4.1 and Chapter 6 in this volume).

Third, all items that had received a non-neutral label by at least two out of the three annotators – a total of 278 items – were used to construct 572 4-tuples

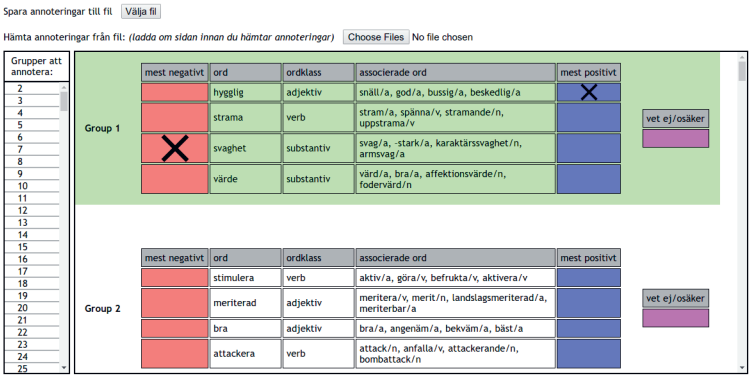


Figure 2: Screenshot of the Best-Worst Scaling annotation interface. Labels (left to right): 'most negative', 'word', 'part of speech', 'associated words', 'most positive', 'don't know/uncertain'

to be used in an annotation setup known as *Best-Worst Scaling* (BWS; Kiritchenko & Mohammad 2016). With BWS, annotators are presented with tuples (usually 4-tuples) of items to annotate, and they select the highest and lowest according to the score at hand (in this case, the items with the most positive and the most negative polarity). Provided that certain statistical properties hold of the composition of the tuples, then the number of times an element is chosen as most positive minus the number of times it is chosen as most negative can be used as a sentiment score. Four individuals were employed to make the BWS annotations using a web interface especially designed for this annotation task (see Figure 2). Since the word senses in Saldo do not come with definitions, and since a lemma may correspond to more than one word sense, the annotation interface provided a set of associated lemmas for each shown item, chosen among its close semantic neighbors as determined by the network of lexical-semantic relations recorded in Saldo.

Fourth, these items were used as a gold standard for evaluating three different methods for assigning polarity labels to Saldo word senses. The best-performing method turned out to be one that combined corpus-derived word2vec embeddings with semantic network information from Saldo (Johansson & Nieto Piña 2015). Subsequently, all positive and negative items and the top 2,500 neutral items were manually checked and corrected, resulting in version 0.1 of SenSaldo, containing a total of 7,618 Saldo word senses marked for polarity on a continuous scale $[-1, +1]$.

The current version of SenSaldo (v. 0.2) is the result of extension by a semi-automatic procedure of the previous version. Specifically, SenSaldo v. 0.2 constitutes the "transitive closure" of v. 0.1, as it were. This means that for any text word form derivable from any word sense in this version, all its possible word sense assignments according to Saldo 2.3 are also present in the dataset, together with their polarity

```
övertalningsförmåga..1  1        överutnyttja..1         -1
övertramp..1            0        överutnyttjande..1      -1
övertramp..2           -1        övervakningssamhälle..1 -1
övertrassering..1      -1        övervåld..1             -1
övertro..1             -1        överväga..1              0
överträda..1           -1        överväga..2              0
överträdande..1        -1        övervägande..1           0
övertydlig..1          -1        övervägande..2           0
övertydlighet..1       -1        övervägande..3           0
övertyga..1             0        överväldiga..1           1
övertygad..1            1        överväldigande..1        0
övertygande..1          1        överväldigande..2        0
övertygande..2          0        överväldigande..3        1
```

Figure 3: The final 26 items in the SenSaldo base lexicon (sentiment polarity: –1/0/1)

values (negative, neutral, or positive). SenSaldo v. 0.2 was released in 2019 and contains 12,287 word senses.

Like Saldo itself and all other derivatives of Saldo, SenSaldo comes in two versions. The primary, or base, version is the list of Saldo word senses, where each entry is accompanied by sentiment polarity information: -1, 0, or 1, for negative, neutral, or positive sentiment respectively (Figure 3). The other format is a more practical fullform list generated from the word senses via their associated Saldo lexemes. Each fullform entry is linked to its word senses and their sentiment polarity information, which in this case may be ambiguous, since the same lexeme – or sometimes only the same text form – can correspond to more than one word sense. For instance, the verb lexeme *diska* has one neutral sense, 'do the dishes', and one negative sense, 'disqualify (in a competition)', which means that each of the forms in this lexeme's paradigm as well as in that of its derived verbal noun in *-ande* comes with two possible sentiment polarities (Figure 4). The fullform version of SenSaldo 0.2 comprises almost 85,000 items.

With the recent release of Saldo 3, the plan is now to make a similar extension of SenSaldo in order to sync it to the current stable Saldo version (3.3).

4.4.2 SB-RID

The other computational onomasiological lexical resource available through Språkbanken falling under the rubric of sentiment and emotional lexicons is SB-RID, the Språkbanken version of the Swedish *Regressive Imagery Dictionary* (RID). The RID comes out of a long tradition of compiling specialized dictionaries/word lists for specific kinds of psychological and psycholinguistic studies. Such lists are often referred to as "norms" by researchers in these fields. It was originally compiled by Martindale

```
diskades        vb      diska..1:0|diska..2:-1
diskade         vb      diska..1:0|diska..2:-1
diskads         vb      diska..1:0|diska..2:-1
diskad          vb      diska..1:0|diska..2:-1
diskandena      nn      diskande..1:0|diskande..2:-1
diskandenas     nn      diskande..1:0|diskande..2:-1
diskande        nn      diskande..1:0|diskande..2:-1
diskanden       nn      diskande..1:0|diskande..2:-1
diskandens      nn      diskande..1:0|diskande..2:-1
diskandes       nn      diskande..1:0|diskande..2:-1
diskandes       vb      diska..1:0|diska..2:-1
diskandet       nn      diskande..1:0|diskande..2:-1
diskandets      nn      diskande..1:0|diskande..2:-1
diskande        vb      diska..1:0|diska..2:-1
```

Figure 4: Ambiguous word forms in the SenSaldo fullform lexicon

(1975), and intended to capture psychological differences between authors of texts – in a wide sense; the texts may be transcribed dialogues – as revealed in their choices from among a specific set of content words in composing their text or their utterances.

The basic distinction made in the RID is between vocabularies associated with an assumption made in (Freudian) psychology about the existence of a fundamental dichotomy in human thought between *primary process* and *secondary process* thought (Svensson, Archer & Norlander 2006; Wilson 2011). The RID vocabulary is further subclassified into 43 more fine-grained categories. In addition to the 30 primary and six secondary process categories, there are also seven emotion categories. See Table 4.

The RID has been translated into a number of languages, and the point of departure for the SB-RID is a previous Swedish translation (Svensson, Archer & Norlander 2006). Like translations of the RID into other languages, the previous Swedish version was a fullform list, manually expanded from translated lemmas, although not all possible forms had actually been supplied. Thus, genitive forms of nouns and adjectives and s-forms of verbs are systematically missing in the previous Swedish translation. Also, the translators had followed the original RID principle of leaving ambiguous words out of the dictionary altogether, although not completely successfully: the ambiguous wordform *bär* is left out as the indefinite singular and plural of the noun *bär* 'berry' (in the RID category `primary/drive/orality`) but is left in the dictionary as the present tense and imperative of the verb *bära* 'carry' (in the RID category `secondary/instrumental_behavior`). This is probably simply a slip on the part of the translators, but it illustrates the importance of framing lexical generalizations on the right linguistic level. Further, if "a word is used in so many ways that it is best left out of the [RID] dictionary" (Martindale 1975: 115), we are probably talking

Table 4: The SB-RID category tags (legend: number of Saldo word senses, main category/subcategories)

# items	category	# items	category
117	emotion/affection	80	primary/regressive_cognition/brink-passage
553	emotion/aggression	149	primary/regressive_cognition/concreteness
108	emotion/anxiety	173	primary/regressive_cognition/consciousness_alteration
97	emotion/expressive_behavior	45	primary/regressive_cognition/narcissism
140	emotion/glory	42	primary/regressive_cognition/timelessness
124	emotion/positive_affect	78	primary/regressive_cognition/unknown
142	emotion/sadness	60	primary/sensation/cold
278	primary/abstract_thought	68	primary/sensation/general_sensation
52	primary/defensive_symbolization/chaos	59	primary/sensation/hard
71	primary/defensive_symbolization/diffusion	46	primary/sensation/odor
158	primary/defensive_symbolization/passivity	41	primary/sensation/soft
162	primary/defensive_symbolization/random_movement	179	primary/sensation/sound
109	primary/defensive_symbolization/voyage	50	primary/sensation/taste
190	primary/drive/anality	104	primary/sensation/touch
351	primary/drive/orality	233	primary/sensation/vision
231	primary/drive/sex	355	secondary/instrumental_behavior
110	primary/icarian_imagery/ascend	81	secondary/moral_imperative
35	primary/icarian_imagery/depth	91	secondary/order
71	primary/icarian_imagery/descent	288	secondary/restraint
109	primary/icarian_imagery/fire	459	secondary/social_behavior
73	primary/icarian_imagery/height	78	secondary/temporal_references
144	primary/icarian_imagery/water		

about an item with relatively high text frequency, which then consequently could contribute significantly to the analysis, in particular if we consider the "one sense per discourse" principle proposed to hold for colexified lexemes (Gale, Church & Yarowsky 1992).

In the SB-RID, the Swedish vocabulary items listed under each RID category are Saldo 3 word senses, that by definition have only one meaning each, which is considered to include their connotations, as mentioned above. There will consequently be no ambiguity at this linguistic level. The problem of colexified lexemes and ambiguous text word forms is deferred to the applications using the SB-RID for actual text analysis. If so desired, a preprocessing step can be applied to filter out all ambiguous items from the fullform version of the SB-RID before deploying it for the analysis of corpus texts, thereby simulating the original mode of RID application. Alternatively, a more linguistically oriented approach could be to include word sense disambiguation as a step in the text processing used for the analysis. This is a standard processing module in Språkbanken Text's analysis platform Sparv (see Chapter 10 in this volume).

SB-RID is a Saldo (v. 3.3) word sense lexicon augmented with POS labels. While the point of departure for SB-RID was the previous Swedish RID translation, it has been corrected and substantially revised and extended with synonyms.[8] "Pseudolexemes" in the previous Swedish translation such as *vara-oense* 'be-of-different-opinion' *göra-omtyckt* 'make-popular' have not been included, since neither the copula *vara* 'be' nor causative *göra* 'do; make' are recognized as support verbs in Saldo (see further Chapter 6 in this volume). In both cases the complement adjective or verb is included in SB-RID instead, in the intended sense(s) in cases of colexification.

SB-RID provides RID classifications for about 6,000 Saldo 3 word senses. A fullform version is under preparation, which should comprise upwards of 40,000 text word forms. The fullform version will also provide information about all the possible Saldo senses expressed by each form, which will make it possible to check and compensate for ambiguities in word classification.

4.5 The IDS/LWT core vocabularies

Another onomasiological lexical resource with numerous applications is a large so-called *core vocabulary* (Borin 2012) intended for comparative linguistic research. This is the IDS/LWT series of dictionaries. The *Intercontinental Dictionary Series* (IDS)

8 The synonyms have been added manually and unsystematically. We foresee that a more systematic new revision of SB-RID will be made after the completion of the first major version of the Swedish wordnet Swesaurus (see Section 4.2).

was originally conceived in the 1970s by the American linguist Mary Ritchie Key, who in 1984 set in motion an international collaboration where an adapted version of the concept list of Buck (1949) is used for the purposes of recording core vocabulary items of languages all over the world.[9] This list contains 1,310 entries distributed over 22 thematic chapters. The entries are identified using English words as labels, but they are intended to represent concepts reflecting roughly the following categories: (1) universal concepts ('speak', 'head', 'mother', etc.); (2) environmental phenomena ('river', 'frog', 'tree', etc.); and (3) cultural concepts ('beer', 'loom', 'tailor', etc.). The concepts are identified using English words together with a (disambiguating) POS label, that of course will not necessarily be the appropriate label for the best translation in some other language. In their *Loanword Typology* (LWT) project Haspelmath & Tadmor (2009) later extended the list with some modern concepts and added a chapter containing mainly functional items. The resulting IDS/LWT list contains 1,460 concepts.[10]

The Swedish IDS/LWT list iwas compiled in conjunction with of a set of such lists prepared for a number of South Asian languages in an extensive collaborative international research effort aiming to elucidate genealogical and areal connections among languages spoken in the western Himalayas (Borin, Comrie & Saxena 2013; Saxena 2022; Saxena, Sagar & Devi 2022).

Unlike other language versions of IDS/LWT, where the lexical items are represented by lemmas, the Swedish version translates the source concepts into Saldo word sense identifiers, thus integrating this valuable core vocabulary fully into our lexical infrastructure and thereby making it into a full-fledged computational lexical resource for automatic linguistic analysis. In this connection we have also made a complete mapping to Princeton WordNet 3.0 synsets of all the IDS/LWT entries; see Table 5.

5 Looking ahead: lumpers, splitters, and LLMs

Lately, large language models (LLMs) have brought about a sea change in NLP. For many kinds of linguistic analysis tasks they trounce more traditional approaches thoroughly, seemingly – in a so-called "zero-shot" mode – without being provided any explicit linguistic knowledge at all. Very relevant in the context of this chapter,

9 At present, the official IDS website offers access to 324 IDS lists; see https://ids.clld.org/ (last accessed: April 4, 2025).

10 The official World Loanword Database website offers access to 41 LWT lists; see https://wold.clld.org/ (last accessed: April 4, 2025).

Table 5: The IDS/LWT core vocabulary mapping to Saldo sense identifiers and to Princeton WordNet 3.0 synsets ("PWN SS"), the latter via the lexical-semantic relations ("R") synonymy ("ss") and hyperonymy ("ui": the IDS/LWT concept is superordinate to the PWN SS)

LWT ID	Saldo ID(s)	Gloss	R	PWN SS
S22.110	religion..1	the religion	ss	religion%1:14:00::
S22.120	gud..1	the god	ss	god%1:18:01::
S22.130	helgedom..1 kyrka..1 tempel..1	the temple	ss	temple%1:06:00::
S22.1310	kyrka..1	the church	ss	church%1:06:00::
S22.1320	moské..1	the mosque	ss	mosque%1:06:00::
S22.140	altare..1	the altar	ss	altar%1:06:00::
S22.150	offer..1	the sacrifice	ss	sacrifice%1:04:00::
S22.160	dyrka..1	to worship	ss	worship%2:37:01::
S22.170	be..2	to pray	ss	pray%2:32:00::
S22.180	präst..1	the priest	ui	priest%1:18:00::
S22.180		the priest	ui	priest%1:18:01::
S22.190	helig..1	holy	ss	holy%3:00:00::
S22.220	predika..1	to preach	ss	preach%2:32:02::
S22.230	välsigna..1	to bless	ss	bless%2:32:00::
S22.240	förbanna..1	to curse	ss	curse%2:32:01::
S22.260	fasta..1	to fast	ss	fast%2:34:01::
S22.310	himmel..2	the heaven	ss	heaven%1:09:00::
S22.320	helvete..1	the hell	ss	hell%1:09:00::
S22.350	demon..1	the demon	ss	demon%1:18:00::
S22.370	avgudabild..1 beläte..1	the idol	ss	idol%1:06:00::
S22.420	trolldom..1	the magic	ss	magic%1:09:00::
S22.430	häxa..1 trollkarl..1	the sorcerer or witch	ui	sorcerer%1:18:00::
S22.430		the sorcerer or witch	ui	witch%1:18:00::
S22.440	fe..1 älva..1	the fairy or elf	ui	fairy%1:18:00::
S22.440		the fairy or elf	ui	elf%1:18:00::
S22.450	gengångare..1 spöke..1 vålnad..1	the ghost	ss	ghost%1:18:00::
S22.470	förebud..1 omen..1 järtecken..1	the omen	ss	omen%1:11:00::
S22.5000	omskärelse..1	the circumcision	ss	circumcision%1:04:00::
S22.5100	initiationsrit..1			

one of the first linguistic analysis tasks at which LLMs outshone other approaches concerns determining meaning relations among words. It was demonstrated for some of the earliest LLM approaches that their underlying context-vector based word representations – or *word embeddings* – contained the analogues of lexical-semantic features, demonstrable through simple algebraic manipulations of the vectors. Thus, subtracting the vector representation for *man* from that of *king* and then adding the vector representation of *woman* to the result yields the vector representation for *queen* (e.g., Mikolov, Yih & Zweig 2013).

It is not hard to imagine how, not too far down the road, a refined version of this feature could replace manually compiled onomasiological lexical resources completely. However, as Pedersen et al. (2024) have noted, there are still significant differences between the lexical-semantic knowledge inductively obtainable from the kinds of input data available to LLMs and that encoded in dictionaries and onomasiological lexical resources by highly trained lexicographers.

In the meantime, onomasiological resource creators as well as lexicographers aiming to compile traditional dictionaries should aim to get as much as possible out of these models. See Chapter 15 in this volume for an example of how LLM text representations of entries in the definition dictionary *SO* (described in Chapter 4 in this volume) can be used to improve cross-references in the dictionary.

A particularly promising and intriguing line of investigation arises from the inductive character of LLMs: that they carve nature – or language – by its joints, as it were. Thus, they may provide a new perspective on the question of a general lumping vs. splitting strategy in lexical-semantic description by allowing us to explore in a systematic fashion which (if any) model parameters correlate with this distinction.

References

Benamara, Farah, Maite Taboada & Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics* 43(1): 201–264. DOI: 10.1162/COLI_a_00278.

Borin, Lars. 2012. Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Diana Santos, Krister Lindén & Wanjiku Ng'ang'a (eds.), *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, 53–65. Berlin: Springer.

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The Intercontinental Dictionary Series: A rich and principled database for language comparison. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 285–302. Berlin: De Gruyter Mouton.

Borin, Lars & Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA) 2009 workshop WordNets and other Lexical Semantic Resources — between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. 7–12.

Borin, Lars & Markus Forsberg. 2010. *Beyond the synset: Swesaurus – a fuzzy Swedish wordnet*. Presentation at the workshop *Re-thinking synonymy: Semantic sameness and similarity in languages and their description*. Helsinki.

Borin, Lars & Markus Forsberg. 2011. Swesaurus: Ett svenskt ordnät med fria tyglar [Swesaurus: A Swedish wordnet with free reins]. *LexicoNordica* 18: 17–39.

Borin, Lars & Markus Forsberg. 2014. Swesaurus; *or*, The Frankenstein Approach to Wordnet Construction. *Proceedings of the Global Wordnet Conference (GWC) 2014*. 215–223.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. DOI: 10.1007/s10579-013-9233-4.

Borin, Lars, Markus Forsberg, Lennart Lönngren & Niklas Zechner. 2021. Swedish FrameNet++: Lexical samsara. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 69–95. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Borin, Lars, Lennart Lönngren & Markus Forsberg. 2017. *Saldo*. [Data set]. DOI: 10.23695/s80w-2517.

Borin, Lars, Luis Nieto Piña & Richard Johansson. 2015. Here be dragons? The perils and promises of inter-resource lexical-semantic mapping. *Proceedings of the Workshop on Semantic resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at the Nordic Conference of Computational Linguistics (NODALIDA) 2015*. 1–11.

Bring, Sven Casper. 1930. *Svenskt ordförråd ordnat i begreppsklasser* [Swedish vocabulary arranged in conceptual classes]. Stockholm: Hugo Gebers förlag.

Buck, Carl Darling. 1949. *A dictionary of selected synonyms in the principal Indo-European languages*. Chicago: University of Chicago Press.

Civil, Miguel. 1990. Sumerian and Akkadian lexicography. In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand & Ladislav Zgusta (eds.), *Dictionaries: An international encyclopedia of lexicography. Second volume*, 1682–1686. Berlin: De Gruyter.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (eds.). 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Devitt, Ann & Khursid Ahmad. 2013. Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources and Evaluation* 47(2): 475–511. DOI: 10.1007/s10579-013-9223-6.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt & Magnus Åström. 1992. *The linguistic annotation system of the Stockholm–Umeå corpus project: Description and guidelines*. Research report. Umeå: Department of Linguistics, Umeå University.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Gale, William A., Kenneth W. Church & David Yarowsky. 1992. One sense per discourse. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. 233–237.

Global WordNet Association. 2025. *Wordnets in the world*. [Online resource] https://globalwordnet.org/resources/wordnets-in-the-world. Accessed on 2025-01-13.

Haspelmath, Martin & Uri Tadmor. 2009. The Loanword Typology project and the World Loanword Database. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 1–34. Berlin: De Gruyter Mouton.

Hüllen, Werner. 1999. *English dictionaries 800–1700: The topical tradition*. Oxford: Oxford University Press.

Hüllen, Werner. 2004. *A history of Roget's Thesaurus: Origins, development, and design*. Oxford: Oxford University Press.

Järborg, Jerker. 1996. *Formaliserad lexikologi: Rapport från ett långtidsprojekt (Preliminär version)* [Formalized lexicology: Report from a long-term project (Preliminary version)]. (Research Reports from the Department of Swedish No. GU-ISS-96-3) Gothenburg: Department of Swedish, University of Gothenburg.

Järborg, Jerker. 1999. *Lexikon i konfrontation* [Lexicons in confrontation]. (Research Reports from the Department of Swedish No. GU-ISS-99-6) Gothenburg: Department of Swedish, University of Gothenburg.

Järborg, Jerker. 2003. *Formaliserade semantiska samband mellan enheter i GLDB* [Formalized semantic relations among items in GLDB]. (Research Reports from the Department of Swedish No. GU-ISS-03-1) Gothenburg: Department of Swedish, University of Gothenburg.

Järborg, Jerker, Dimitrios Kokkinakis & Maria Toporowska Gronostaj. 2002. Lexical and textual resources for sense recognition and description. *International Conference on Language Resources and Evaluation (LREC) 2002*. 1492–1497.

Jarmasz, Mario & Stan Szpakowicz. 2004. *Roget's Thesaurus* and semantic similarity. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova & Ruslan Mitkov (eds.), *Recent advances in natural language processing III: Selected papers from RANLP 2003*, 111–120. Amsterdam: John Benjamins.

Jobbins, Amanda C. & Lindsay J. Evett. 1995. Automatic identification of cohesion in texts: Exploiting the lexical organization of Roget's Thesaurus. *Proceedings of Rocling VIII*. 111–125.

Johansson, Richard & Luis Nieto Piña. 2015. Embedding a semantic network in a word space. *Proceedings of NAACL-HLT 2015*. 1428–1433.

Kann, Viggo & Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 105–110.

Kennedy, Alistair & Stan Szpakowicz. 2008. Evaluating *Roget's* thesauri. *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*. 416–424.

Kennedy, Alistair & Stan Szpakowicz. 2014. Evaluation of automatic updates of *Roget's Thesaurus*. *Journal of Language Modelling* 2(2): 1–49.

Kiritchenko, Svetlana & Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. *Proceedings of NAACL 2016*. 811–817.

Kokkinakis, Dimitrios, Jerker Järborg & Yvonne Cederholm. 2001. SENSEVAL-2: The Swedish framework. *Proceedings of SENSEVAL-2*. 45–48.

Kokkinakis, Dimitrios, Maria Toporowska Gronostaj & Karin Warmenius. 2000. Annotating, disambiguating & automatically extending the coverage of the Swedish SIMPLE lexicon. *International Conference on Language Resources and Evaluation (LREC) 2000*.

Kokkinakis, Dimitrios, Maria Toporowska Gronostaj & Karin Warmenius. 2001a. Corpus-based extension of semantic lexicons in large scale. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. np.

Kokkinakis, Dimitrios, Maria Toporowska Gronostaj & Karin Warmenius. 2001b. Swedish SENSEVAL: A developer's perspective. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. np.

Lange, Sven. 2007. *Thesaurus Lex: Ett hyperlexikon med rötter hos Locke, Roget och Bring* [Thesaurus Lex: A hyperlexicon with roots in Locke, Roget, and Bring]. http://www.thesauruslex.se/artiklar/Roget.pdf.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas & Antonio Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography* 13: 249–263. DOI: 10.1093/ijl/13.4.249.

Martindale, Colin. 1975. *Romantic progression: The psychology of literary history*. Washington: Hemisphere.

Mikolov, Tomáš, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies (NAACL/HLT) 2013*. 746–751.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4): 235–245.

Morris, Jane & Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17(1): 21–48.

NEO. 1995. *Nationalencyklopedins ordbok* [The Dictionary of the National Encyclopedia]. Höganäs: Bra böcker.

Pang, Bo & Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.

Pedersen, Bolette S., Nathalie C. Hau Sørensen, Sussi Olsen & Sanni Nimb. 2024. Evaluering af sprogforståelsen i danske sprogmodeller: Med udgangspunkt i semantiske ordbøger [Evaluation of language understanding in Danish language models: Based on semantic dictionaries]. *Nydanske Sprogstudier* 65: 8–40.

Ralph, Bo. 1977. Projektet lexikalisk databas [The Lexical Database project]. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 79–81.

Ralph, Bo, Jerker Järborg & Sture Allén. 1977. *Svensk ordbok och lexikalisk databas: Förstudierapport* [The dictionary *Svensk ordbok* and the lexical database: A pilot study report]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Rødven-Eide, Stian. 2016. *The Swedish Culturomics Gigaword Corpus*. [Data set]. DOI: 10.23695/3wmv-1z09.

Rødven-Eide, Stian, Nina Tahmasebi & Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A one billion word Swedish reference dataset for NLP. *From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts*. 8–12.

Roget, Peter Mark. 1852. *Thesaurus of English words and phrases, classified and arranged so as to facilitate the expression of ideas and assist in literary composition*. London: Longman, Brown, Green, & Longmans.

Rouces, Jacobo, Lars Borin, Nina Tahmasebi & Stian Rødven Eide. 2018. Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN)*. 219–227.

Rouces, Jacobo, Nina Tahmasebi, Lars Borin & Stian Rødven Eide. 2018a. Generating a gold standard for a Swedish sentiment lexicon. *International Conference on Language Resources and Evaluation (LREC) 2018*. 2689–2694.

Rouces, Jacobo, Nina Tahmasebi, Lars Borin & Stian Rødven Eide. 2018b. SenSALDO: Creating a sentiment lexicon for Swedish. *International Conference on Language Resources and Evaluation (LREC) 2018*. 4192–4198.

SAOL 11. 1986. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 11th edn. Stockholm: Norstedts.

Saxena, Anju. 2022. *The linguistic landscape of the Indian Himalayas: Languages in Kinnaur*. Leiden: Brill.

Saxena, Anju, Padam Sagar & Suari Devi. 2022. Kanashi basic vocabulary. In Anju Saxena & Lars Borin (eds.), *Synchronic and diachronic aspects of Kanashi*, 257–315. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110703245-009.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Språkbanken Text. 2017. *Swesaurus*. [Data set]. DOI: 10.23695/w5ww-x964.

Språkbanken Text. 2019. *SenSaldo*. [Data set]. DOI: 10.23695/vdax-hp87.

Språkbanken Text. 2024a. *PAROLE lexicon*. [Data set]. DOI: 10.23695/CACY-SB38.

Språkbanken Text. 2024b. *PAROLE+*. [Data set]. DOI: 10.23695/Q1FY-TG26.

Språkbanken Text. 2024c. *SIMPLE lexicon*. [Data set]. DOI: 10.23695/8RX4-D548.

Språkbanken Text. 2024d. *SIMPLE+*. [Data set]. DOI: 10.23695/56MQ-SH64.

Svensson, Nina, Trevor Archer & Torsten Norlander. 2006. A Swedish version of the Regressive Imagery Dictionary: Effects of alcohol and emotional enhancement on primary–secondary process relations. *Creativity Research Journal* 18(4): 459–470. DOI: 10.1207/s15326934crj1804_5.

van der Veen, A. Maurits & Erik Bleich. 2025. The advantages of lexicon-based sentiment analysis in an age of machine learning. *PLOS ONE* 20(1): 1–19. DOI: 10.1371/journal.pone.0313092.

Viberg, Åke, Kerstin Lindmark, Ann Lindvall & Ingmarie Mellenius. 2002. The Swedish WordNet project. *Proceedings of the European Association for Lexicography (EURALEX) 2002*. 407–412.

Wilson, Andrew. 2011. The Regressive Imagery Dictionary: A test of its concurrent validity in English, German, Latin, and Portuguese. *Literary and Linguistic Computing* 26(1): 125–135. DOI: 10.1093/llc/fqq028.

Zechner, Niklas & Lars Borin. 2020. Towards a Swedish Roget-style thesaurus for NLP. *Proceedings of the Globalex Workshop on Linked Lexicography*. 53–60.

Part IV: **A computational infrastructure for dictionary making and lexical research**

Markus Forsberg, Dana Dannélls, Lars Borin, and Aleksandrs Berdicevskis

9 Background: Språkbanken Text

Abstract: Språkbanken Text is a division of the national digital research infrastructure Språkbanken, whose mission is to support research based on language data, in the case of Språkbanken Text specifically written language data and computational analyses for working with such data. Throughout its 50-year long history, the research and development work in Språkbanken Text has been characterized by a strong insistence on the crucial importance of lexical knowledge for developing computational text analysis tools. Hence, computational lexical resources play a central role in the Språkbanken Text research infrastructure.

Keywords: corpora, language models, language resources, language technology, lexical resources, research infrastructure

1 The 50-year long history of Språkbanken Text

A proper account of the history of Språkbanken Text must start with the groundbreaking work of Sture Allén (1928–2022), a pioneer in introducing corpus linguistics in Sweden for Swedish. His 1965 PhD thesis appeared in two parts, one where he described the computer-supported method that he had used (basically an automatic

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*. Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Markus Forsberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: markus.forsberg@svenska.gu.se
Dana Dannélls, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: dana.dannells@svenska.gu.se
Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Aleksandrs Berdicevskis, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: aleksandrs.berdicevskis@gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-009

concordancer) – after first learning to program it himself in machine code – in order to investigate a text corpus of 17th-century letters (Allén 1965a), and the other a scientific edition of these letters (Allén 1965b).

After defending his thesis, Allén initiated a project aiming to prepare the way for corpus-based lexicography for Swedish. The most immediate result of this project was the one-million word *Press-65* corpus of Swedish news text, which provided the raw stuff for a series of Swedish dictionaries. See Chapter 2 of this volume.

As professor and scientific leader of the Computational Linguistics Unit established in 1972 at the University of Gothenburg, Allén took the initiative for an undergraduate program in computational linguistics, which started at the University of Gothenburg in 1984. However, his own main focus remained on the development of corpora and corpus tools in support of Swedish lexicography, and he initiated a systematic effort to build a computational research infrastructure which could further this aim.

Such an infrastructure was envisioned and eloquently argued for in an op-ed piece written by Allén for the Swedish daily *Dagens Nyheter* in September 1970 (Allén 1970). In 1973, the Computational Linguistics Unit submitted a formal proposal to the Ministry of Education, requesting earmarked funding for what was to become Språkbanken (Allén 1973). Two years later, this research infrastructure became a reality, when the *Logotheque* (as it was called initially) was established with national funding in 1975.

The focus of Språkbanken shifted noticeably around the turn of the millennium, when for various reasons the lexicographical and the language technology activities parted ways organizationally, the former being pursued in the *Center for Lexicology and Lexicography* established in 2003, while Språkbanken widened its language technological sphere way beyond lexicographical considerations (see Chapter 2 of this volume).

Since then, Språkbanken Text has grown into a nationally and internationally recognized research and development (R&D) unit for Swedish language technology and language resources. It coordinated the Swedish activities in the European CLARIN ERIC research infrastructure in the years 2014–2024, and is the coordinating node of the national research infrastructure Språkbanken, making up one of its four nationally distributed divisions, the other three being:

- *Språkbanken Tal*, the speech technology division at the Royal Institute of Technology (KTH) in Stockholm;
- *Språkbanken Sam*, the cultural heritage and language policy division at the Institute of Language and Folklore with branches in Uppsala, Stockholm, and Gothenburg; and
- *Språkbanken CLARIN*, the division coordinating the Swedish CLARIN activities, at Uppsala University.

Figure 1: Språkbanken: a self-renewing research infrastructure (image in public domain; source https://commons.wikimedia.org/w/index.php?curid=2856329 [last accessed: April 4, 2025])

As a research infrastructure, Språkbanken is quite unique in the sense that many of the research results coming out of the research it supports will to a considerable extent contribute to the further development of the infrastructure itself. Språkbanken supports research in language technology (text, speech, and sign) with an infrastructure which is itself built on language technology (text, speech, and sign), much like the mythological Ouroboros snake of antiquity (Figure 1).

2 Språkbanken Text's research infrastructure of today

If we were to summarize the ideology of Språkbanken Text of today with one word, it would be *openness*: our data sets and our software are always released as soon as possible and with as open a license as possible. Very little is kept internal and

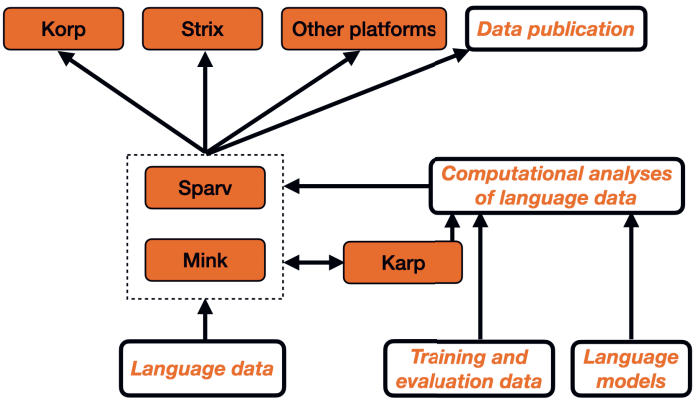


Figure 2: A bird's-eye view of the research infrastructure of Språkbanken Text

only when there is a good reason for doing so, such as privacy or personal integrity concerns. We are consequently strong supporters of the *open science* movement, and we are always trying to push our openness boundaries even further.

If we look into the machinery of the research infrastructure of Språkbanken Text, which has been designed with the openness principle in mind, there are two metaphorical beating hearts: the *language data* and the *computational analyses*, which reciprocally feed into each other. The computational analyses are commonly based on the language data that, in turn, are commonly enriched by the computational analyses. Again, much like the mythological Ouroboros snake in Figure 1.

Our data and analyses are, when possible, made accessible through one or more of our five language research platforms, as illustrated in Figure 2: *Korp* and *Strix* for our text collections; *Sparv* for our computational analyses; *Karp* for our lexical resources and other structured data sets; and finally *Mink*, which is our gateway into the other platforms. Moreover, they are also published via our data pages,[1] and analyses pages [2].

In the following sections (2.1–2.2) we will go more into details about the different components of the research infrastructure of Språkbanken Text, as outlined in Figure 2.

1 https://spraakbanken.gu.se/en/resources (last accessed: April 4, 2025)
2 https://spraakbanken.gu.se/en/analyses (last accessed: April 4, 2025)

2.1 Language data

Anything that is *digital* and *written* is a candidate for being integrated into our research infrastructure. The language data of today consists of large text collections, so called corpora, described in Section 2.1.1, structured data, including computational lexical resources, described in Section 2.1.2, training and evaluation data, described in Section 2.1.3, and language models, described in Section 2.1.4.

Some of our data sets are multimodal as well, but one of the modalities is always written language, such as transcriptions of video or audio, in order for it to qualify for inclusion into Språkbanken Text. We will focus on the written part here.

2.1.1 Corpora – large text collections

Corpora are large, formally organized collections of written or spoken texts, either modern or historical, in one or more languages. They are usually gathered and designed for a particular purpose, and can therefore take many forms. Some corpora are designed to be representative for the particular investigations they were created for, balanced with respect to text type, genre, style, and various other linguistic features (e.g., vocabulary range or level), and enriched with relevant metadata encoded according to standard schemes (Wynne 2005), while others are designed to be of more general use.

Corpora are valuable language research data for studying, analyzing, and understanding languages and have been an integral part of Språkbanken since its conception, as described in Section 1. They gained popularity in the 1990s when computational linguistics began to gain momentum, in particularly manually annotated corpora, which can also be used to train computational models. One of the first heavily used large-scale manually annotated corpora for Swedish was created in 1997, and is known as the *Stockholm-Umeå Corpus* (SUC; Gustafson-Capková & Hartmann 2006). It was created by Gunnel Källgren at Stockholm University och Eva Ejerhed at Umeå university, and consists of about one million words of written Swedish compiled from newspapers, fiction, and academic writing that have been manually annotated with lemmas, parts of speech, and morphosyntactic information.

An increasing number of Swedish corpora have been created since then with different sizes, genres, and content types. Thanks to collaborative efforts throughout the years and generous funding from Swedish research funding agencies, over 1,000 corpora, amounting to a total of more than 30 billion words, are freely available for download, in the form copyright legislation permits, via Språkbanken Text's data

pages[3] and further accessible through, for example, Korp, Språkbanken's word research platform, briefly described in Section 2.2.1 and more elaborately in Chapter 10 of this volume.

2.1.2 Computational lexical resources and other structured data

Computational lexical resources, or simply lexicons or thesauruses, are structured data (i.e., key-and-value data) that contain information about words and concepts and have been created for computational purposes. Their development requires experts with formally oriented mindsets that are familiar with the conceptual apparatus and formalism used in the target computational lexical resource. The downside of their production is the labor-intensive process, which tends to produce "human-type" errors, in part because the experts are not necessarily familiar with the technical details, and in part also because the technical editorial environment may not be mature enough for supporting the required constructions. However, the process can be accelerated by applying automatic methods, such as data validation including consistency checks, once the formal structure has been defined.

A rich variety of lexical resources, developed by lexicographers, grammarians, and other linguists over the years, are available for Swedish. Many of these resources can be downloaded through Språkbanken Text. See also Chapter 5 of this volume. Further, they are also accessible through Karp, Språkbanken's data editing platform, which also allows for advanced exploration of the lexical resources. See Section 2.2.3 for a brief account of Karp and Chapter 11 of this volume for a more elaborate description.

Experience has shown that computational power and advanced technologies allow these databases to become very large. This is particularly advantageous for complex lexicons with elaborated hierarchical structures and extended linguistic knowledge (Atkins & Rundell 2008). An example of such a lexicon is Saldo (Borin, Forsberg & Lönngren 2013), the largest modern computational Swedish lexicon with morphological and semantic information (see Chapter 6 of this volume), which plays a pivotal role in Språkbanken Text's infrastructure. For every entry, Saldo provides formal and semantic information with unique identifiers. Most resources developed within Språkbanken Text are linked to Saldo via these identifiers, and thereby form a large lexical network with Saldo as the pivot. And since the words in our Swedish corpora are enriched, through computational lexical analyses, with these identifiers as well, they also constitute a part of the same network of a substantial size, covering

3 https://spraakbanken.gu.se/en/resources (last accessed: April 4, 2025)

all Saldo entries in our corpora plus all compounds which can be analyzed as made up entirely of Saldo entries. In a sense this is a step toward the realization of a "word bank" like the one mentioned by Gellerstam & Sjögreen (1994: 5, 13) (see also Chapter 5 of this volume).

2.1.3 Training and evaluation data

The label *training and evaluation data* is reserved for the highest-quality resources, often called *reference data*, or *gold-standard data*, or simply *gold*. Our choice of label highlights two important usages of gold data in language technology: training computational analyzers and evaluating their performance. If the performance is judged acceptable, the analyzers can then be used to automatically annotate language data on a larger scale automatically.

Training and evaluation data are thus resources that at least in some respect are near-perfect. Typically they contain some kind of annotation, either fully manual or manually checked. What exactly has been annotated varies. *Dalin: Then Swänska Argus 1732-1734* (Språkbanken Text 2020), for instance, contains manual transcriptions of an 18th century Swedish newspaper, which can be used for training and testing OCR tools. *ABSAbank-Imm* (Berdicevskis et al. 2023) contains judgments about which sentiment a given fragment of text expresses towards immigration in Sweden, and can be used for sentiment analysis.

Some resources have been constructed in such a way that *ipso facto* makes them gold data without additional annotation. Such is, for instance, *SweDN* (Monsen & Jönsson 2023), where articles from the *Dagens Nyheter* newspaper are matched with their preambles, and a preamble is considered a gold-standard summary of the article. The dataset can thus be used for training and testing text summarization tools.

Perhaps the most prominent training and evaluation resources are SUC (described in Section 2.1.1) and *Talbanken* (Nilsson, Hall & Nivre 2006). Some of the computational analyses available in our analysis platform Sparv (see Section 2.2.4) are trained on these two corpora (both when it comes to parts of speech and morphosyntactic descriptors, Talbanken when it comes to dependency trees). Some other corpora, for instance, *Eukalyptus* (Adesam, Bouma & Johansson 2015), also enrich words with Saldo senses, which made it possible to train and test word sense disambiguation tools.

That said, even the very best resources are seldom entirely error-free. Moreover, in some cases the ground truth is extremely difficult to establish (consider, for instance, deciding what sentiment is expressed in a particular piece of text),

and resources can only capture an approximation to it. Improving and properly documenting our gold data is an important part of our work.

2.1.4 Language models

To help researchers save time and computational resources, and for reproducibility's sake we share pretrained models. The word *model* in this section is used loosely, basically denoting any resource that must be created automatically using computational means.

One important type of resources is represented by the models that are used in Sparv (see Section 2.2.4) to enrich texts, such as, for instance, part-of-speech tagging and syntactic parsing models for Stanza (Qi et al. 2020). We also share other models that we trained and tested, but currently do not use. For instance, we provide pretrained word embeddings for various corpora and various historical periods of the Swedish language.

2.2 Language research platforms

If the language data and computational analyses are the hearts of the infrastructure, then the language research platforms are its arms and legs. We will in this section briefly describe the five main platforms developed and maintained by Språkbanken Text: Korp, Strix, Karp, Sparv, and Mink.

2.2.1 Korp – Språkbanken's word research platform

The central and most used infrastructure component offered by Språkbanken Text is undoubtedly *Korp*, our word research platform. Korp could be characterized as a concordancer with bells and whistles. It has been under constant development by a dedicated team of research engineers and experts since its introduction in 2011 as a replacement of a plethora of corpus search and browsing systems available prior to this time through Språkbanken Text as results of a number of corpus linguistic projects, all with different capabilities and access to different corpora.

Korp offers several kinds of search, browsing, and statistics access to many billions of words of text, primarily Swedish – about 30 billion words of contemporary and historical Swedish – but also a number of other languages, for instance, Faroese, Somali, Xhosa, and Siberian German, to mention a few.

Most of the corpora, beyond being richly annotated with metadata, are automatically annotated with lexical and syntactic information at corpus import time using Sparv and other resources, and the various Korp interfaces provide full access to all these annotations. Korp is described in more detail in Chapter 10 of this volume.

2.2.2 Strix – Språkbanken's text research platform

Strix can be considered a sister platform to Korp, in that both platforms allow users to search the same kinds of language data, namely corpora, but with the main difference that the search hits in Strix are documents, rather than (sequences of) words, as in Korp. Strix resembles any search engine that you can find online: You input a search query and get a list of documents as the result of the query. But there are some crucial differences compared to search engines that can be found on the internet, which makes it more suitable for research:

- Strix offers complete data transparency. As a user you know exactly what data you search in. The data are often well-documented and, if copyright permits, downloadable from our data pages.[4]
- All corpora in Strix have been automatically annotated using our computational linguistic analyses available via Sparv, all pre-existing metadata annotations in the corpora have been maintained, and both kinds of annotations are available for search and compilation of the search results in relevant ways.
- Strix has a rich set of functionalities, geared towards research, not normally found in general-purpose search engines, such as visualization, including annotation highlighting.

See Chapter 15 of this volume for more information about Strix.

2.2.3 Karp – Språkbanken's data editing platform

Karp is Språkbanken's data editing platform for browsing and editing lexical resources and other structured data (see Chapter 11 of this volume). It provides an advanced and easy-to-use editing environment for nearly 30 Swedish computational lexical resources, along with links to other external lexical resources, through which users can perform searches within individual lexical resources. Karp provides a practical interface adapted to various annotation schemas and formal structures, of-

4 https://spraakbanken.gu.se/en/resources/corpus (last accessed: April 4, 2025)

fering multi-level browsing for categories such as lemmas, part-of-speech categories, paradigms, conservation status, and more.

One notable feature of Karp is its integration with Språkbanken's word research platform, Korp, which provides direct access to corpus examples linked to lexical entries.

2.2.4 Sparv – Språkbanken's analysis platform

Sparv is Språkbanken's analysis platform. In essence, it is a framework for adding new computational language technology analyzers to the research infrastructure of Språkbanken, so their output can become available via our platforms or via our downloadable resources. It is tightly connected to Mink (Section 2.2.5). In fact, Mink can be considered the web interface of Sparv, even though the current functionality of Mink is focused on allowing users to import their own language data into our research infrastructure.

The Sparv platform works mostly behind the scenes, receiving written text with metadata as input and providing the same text as output, but with several kinds of rich linguistic annotation added and the input metadata preserved. It plays a crucial role for the functioning of Korp (Section 2.2.1) and Strix (Section 2.2.2). Everything that gets into Korp and Strix is first processed by Sparv, either via Mink or by the staff of Språkbanken Text.

Some of the computational analyses in Sparv that are standardly applied to all Swedish corpora are, *inter alia*, sentence segmentation, lexical tokenization, part-of-speech tagging, dependency analysis, lemmatization, word sense disambigation, named entity recognition, geographical coordinates of place names, and more. Språkbanken Text's policy is to enrich corpora with as much metadata as possible, and Sparv was designed with this purpose in mind.

The software of Sparv is freely downloadable and can be installed as a standalone tool. It is also highly modular. The latest versions of Sparv have been specifically designed to make is easy to add new analyses, hence turning it into a platform that allows for new analyses by anyone to be made available through Språkbanken Text. For instance: at the moment Sparv offers limited support today for other languages than Swedish, which we hope will change now that Sparv has been turned into a general platform, rather than a piece of software that is mainly used internally within Språkbanken Text.

2.2.5 Mink – Språkbanken's data platform

As mentioned in Section 2.2.4, Mink, Språkbanken's data platform, provides a web interface for Sparv that allows any user to upload their own language data, analyze it using Sparv, and import it into Korp and Strix, accessible as a private data set protected by login. In the future, this functionality will be available for Karp as well.

From an organizational point of view, big data have long been considered a particular challenge for us, but for small data this point tends not to be raised. However, experience has taught us that having to deal with a large number of small datasets displaying the variety in content and processing requirements so characteristic of a research context, may be equally challenging. Consequently, when dealing with small data, our staff often become a bottleneck for the research on small datasets. For the individual researcher a few hundred of documents are already a large enough data set for it to be difficult to handle, but at the same time, too small to be considered a general priority for Språkbanken. Enter Mink, where the individual researchers can by themselves analyze and integrate their own language data into the research infrastructure of Språkbanken.

3 Lexicographic research at Språkbanken Text

As noted above in Section 1, starting around the turn of the millennium the lexicographical activities aiming at compilation of traditional dictionaries for humans were pursued in a separate lexicographical R&D unit established at that time.

Lexicography did not disappear from the agenda of Språkbanken Text, however, but became focused entirely on enabling computational lexical analyses of corpora representing all historical stages of Swedish. This work is described in more detail in several other chapters (see Chapters 5, 6, 7, and 8 in this volume).

It almost goes without saying that this work has nevertheless to a large extent been informed by the meticulous lexicographic research conducted in the Center for Lexicology and Lexicography, and drawing on the detailed dictionaries compiled there. In order to be of use to lexically oriented linguistic research on Swedish, Språkbanken Text's automatic corpus annotations naturally must be able to offer traditional analyses to the extent possible, meaning that the link to traditional lexicography has remained strong.

Since 2021, the lexicographical and lexicological R&D activities formerly pursued in the Center for Lexicology and Lexicography are again integrated in Språkbanken Text, forming a large and important part of it (see Chapter 2 of this volume), and a central concern in the medium term is to ensure that the needs of human and

computational lexicography can be accommodated with minimal duplication of effort.

4 The coming 50 years of Språkbanken Text

What lies ahead of an organization such as Språkbanken Text? Well, as Niels Bohr put it, prediction is very difficult, especially if it's about the future. But we will try anyway.

Språkbanken Text will continue to be a research infrastructure supporting research on language data using language technology, including language-based AI. We will continue to safeguard our identity as linguistically oriented language technology researchers, including computational lexicographers. We judge this to be more important than ever in these AI times, where machine learning becomes the main focus, and the object itself, language, often becomes understudied and reduced into strings of characters.

We have only started to see the added benefits of the integration of the lexicographical and lexicological R&D activities into Språkbanken, and we are convinced that we have just scratched the surface in that respect.

With a more global outlook, the research infrastructure landscape of today depends on scarce economic resources, and will most probably continue to do so, not at all on a par with the financial resources spent on today's AI models in the business world. At the same time, many research infrastructures face similar problems for which they end up implementing similar solutions, which is a healthy thing from an innovation standpoint – the best solutions grow out of different organizations trying to address the same challenges – but on the other hand, given the economic conditions, it does not really make sense. So we foresee a future where the digital research infrastructures become both more and more integrated and increasingly specialized, and that our role as computationally oriented language experts will be even more pronounced.

5 Conclusion

Språkbanken Text is celebrating its 50th anniversary this year, 2025, which makes it a good time to pause, take stock, and consider what may lie ahead in the coming 50 years. We have tried to give a brief account of the history of Språkbanken Text, where we are now, and what we foresee in the future.

Abraham Lincoln said that the best way to predict the future is to create it, and Språkbanken Text will continue to do so, together with old and new allies. The coming 50 years are looking bright for Språkbanken Text!

References

Adesam, Yvonne, Gerlof Bouma & Richard Johansson. 2015. Defining the Eukalyptus forest: The Koala treebank of Swedish. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 1–9.

Allén, Sture. 1965a. *Grafematisk analys som grundval för textedering med särskild hänsyn till Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655* [Graphemic analysis as a basis for text editing with special reference to Johan Ekeblad's letters to his brother Claes Ekeblad 1639–1655]. Gothenburg: Department of Scandinavian Languages, University of Gothenburg.

Allén, Sture. 1965b. *Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655: Utgivna med inledning, kommentar och register* [Johan Ekeblad's letters to his brother Claes Ekeblad 1639–1655: Edited with an introduction, commentary, and index.] Gothenburg: Department of Scandinavian Languages, University of Gothenburg.

Allén, Sture. 1970. Åtta teser om texthantering [Eight theses about text processing]. *Dagens Nyheter* (1970-09-29).

Allén, Sture. 1973. *Förslag till inrättande av ett organ för lagring och tillhandahållande av datamaskinellt läsbara texter, benämnt* logotek [Proposal for the establishment of a facility for storage and supply of computer-readable texts, designated *logotheque*]. Gothenburg: Computational Linguistics Unit, University of Gothenburg.

Atkins, B. T. Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Berdicevskis, Aleksandrs, Lars Borin, Jacobo Rouces & Nina Tahmasebi. 2023. *Swedish ABSAbank-Imm 1.1*. [Data set]. DOI: 10.23695/a8nh-1j87.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4): 1191–1211. DOI: 10.1007/s10579-013-9233-4.

Gellerstam, Martin & Christian Sjögreen. 1994. *Språkbanken: En språklig referensdatabas* [Språkbanken: A linguistic reference database]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Gustafson-Capková, Sofia & Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm: Department of Linguistics, Stockholm University.

Monsen, Julius & Arne Jönsson. 2023. *SweDN 1.0*. [Data set]. DOI: 10.23695/36v9-9017.

Nilsson, Jens, Johan Hall & Joakim Nivre. 2006. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. *Treebanking for Discourse and Speech: Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA) 2005 Special Session on Treebanks for Spoken Language and Discourse*. 119–132.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *Annual Meeting of the Association for Computational Linguistics (ACL) 2020: System Demonstrations*. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.

Språkbanken Text. 2020. *Dalin: Then Swänska Argus 1732–1734*. [Data set]. DOI: 10.23695/9z65-nv18.

Wynne, Martin. 2005. *Developing linguistic corpora: A guide to good practice*. Oxford: AHDS.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Louise Holmer, and Arild Matsson

10 Korp: Språkbanken's word research platform

Abstract: Corpus data have been foundational to the Swedish lexicographical research conducted at the University of Gothenburg since the beginnings of this research in the 1960s. Consequently, an infrastructure component for work with corpus data plays a central role also in our present-day work on both lexical resources for computational applications and dictionaries for human consumption. This component is *Korp*, Språkbanken's word research platform, whose first version was launched in 2011, and which now (2025) is at v. 9. Korp presents the user with a versatile and multifunctional web interface providing online access to upwards of 30 billion words of modern and historical corpora, primarily Swedish, but also many other languages, with many different kinds of linguistic annotations. Korp is one component in a modular e-research infrastructure, that also includes separate components for corpus import and annotation, for text-based research, and for working with formally structured linguistic data such as dictionaries.

Keywords: corpus linguistics, corpus infrastructure, Swedish lexicography, vocabulary research

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*. Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Markus Forsberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: markus.forsberg@svenska.gu.se
Martin Hammarstedt, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: martin.hammarstedt@svenska.gu.se
Louise Holmer, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: louise.holmer@svenska.gu.se
Arild Matsson, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: arild.matsson@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-010

1 Introduction: before Korp

The primary motivation for initiating the corpus compilation work that provided the foundation for Språkbanken was the desire to base Swedish lexicography on contemporary carefully selected empirical reference corpus data, but also to support other kinds of linguistic research (see Chapter 2 in this volume).

Before the advent of the internet, Språkbanken's corpus texts and concordances were prepared and distributed to researchers on microfiche.

Språkbanken later became an early adopter of web-based corpus exploration. Starting in the early 1990s, online access to a concordancer – through Telnet requiring login – became possible. Later in the same decade, this concordancing system became freely accessible on the world wide web (Gellerstam & Sjögreen 1994). This system, *Konk*, shown in Figure 1, served as Språkbanken's main web-based word exploration tool for about two decades, offering access to some 160 million words of corpus texts at the end of that period, primarily news text, but also modern fiction and various documents from public offices. There were also some smaller collections of older texts available, in Old Swedish (1225–1526) and Early Modern Swedish (1526–1732).

For various reasons the same period also saw the development in Språkbanken of a number of similar tools in different research projects. Towards the end of the first decade of the new millennium, users could access about 10 different web-based corpus browsers/concordancers through Språkbanken's web pages. These offered slightly different user interfaces and functionalities, and were implemented using different technologies and programming languages. Typically each tool would serve only a particular corpus or small set of corpora, with little or no overlap between tools. Figures 2 and 3 show some of these interfaces used with different corpora. The ORDAT corpus (Språkbanken Text 2017) shown in Figure 2 contains a historical news text material from the 1930s. This corpus was compiled in a project titled *Det svenska ORD-förrådets utveckling från Artonhundra till Tjugohundra* 'The Evolution of the Swedish Vocabulary from Eighteen-hundred to Two Thousand' (Malmgren 2000). In Figure 3 (top) we see a search in the Stockholm-Umeå Corpus (SUC; Gustafson-Capková & Hartmann 2006; Språkbanken Text 2024c) in the so-called *Konkplus* interface,[1] which also offered access to the Swedish Parole corpus (Språkbanken Text 2024a). The lower screenshot in Figure 3 shows a corpus search interface developed in the project *IT-based Collaborative Learning in Grammar* (ITG; Borin & Saxena 2004), with a search in a Swedish second-language learner corpus compiled in the project *Andraspråkets strukturutveckling* 'Second-language Structural Development' (ASU; Hammarberg 2010; Språkbanken Text 2022). These corpora were all different with regard to their

1 Despite its name, Konkplus was built using a different set of technologies from the Konk system.

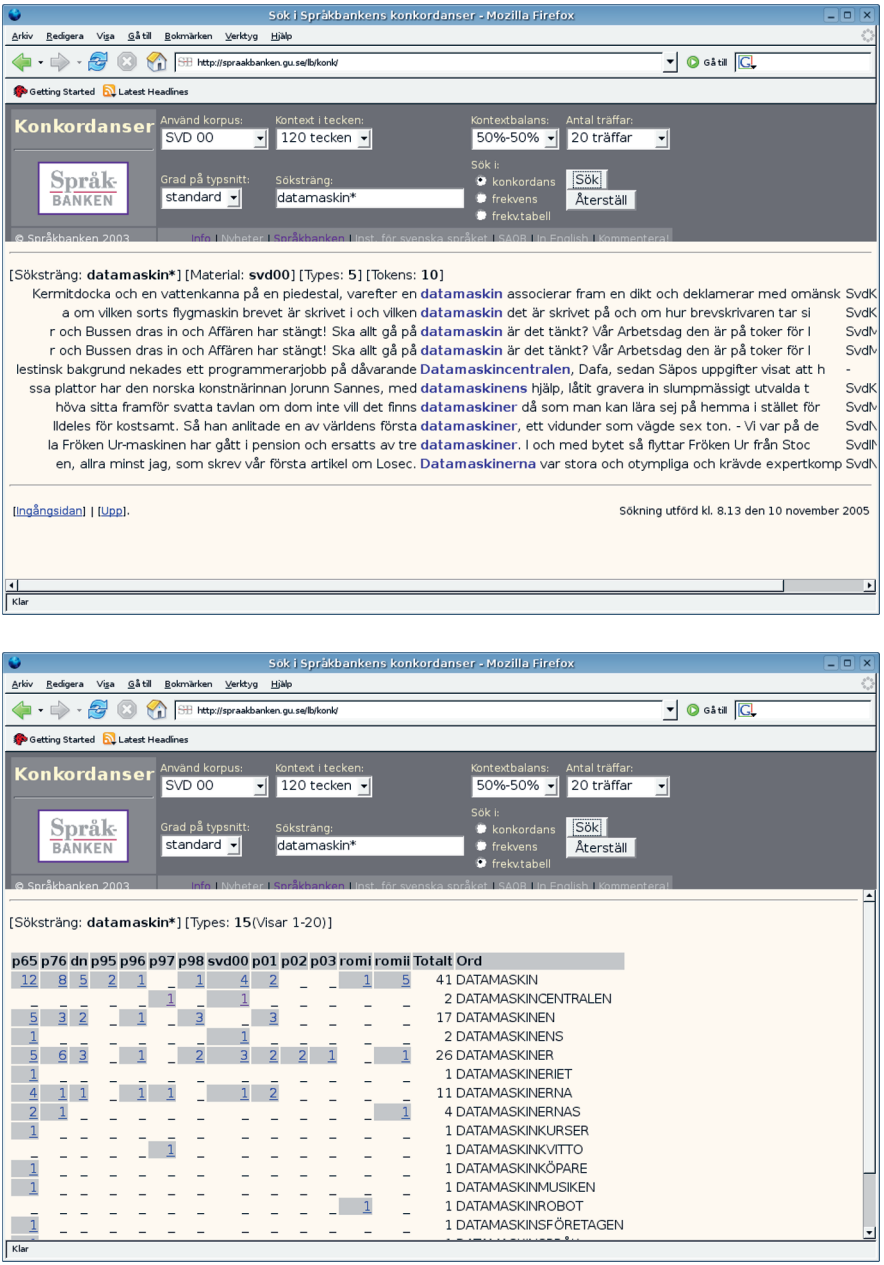


Figure 1: Konk: KWIC mode (top) and word search mode (bottom), showing a search for text words beginning in *datamaskin* 'computer'

format – a smaller problem – and their annotations – a considerably more significant headache. SUC and Parole had compatible annotations for morphosyntactic descriptions, while ASU had a completely different set of morphosyntactic annotations, and most of the corpora in Språkbanken were not linguistically annotated at all. Unlike Parole and ASU, the SUC corpus was also (manually) lemmatized. The various user interfaces were also interestingly different both in their general visual appearance and in the functionalities offered, as can be seen to some extent in Figures 2 and 3.

Yet, Konk remained the corpus workhorse of Språkbanken. Konk's underlying search engine (written in-house entirely in C) was heavily optimized to allow quick access to large volumes of unannotated text, and would have required substantial rewriting in order to deal with linguistically annotated – e.g., part-of-speech (POS) tagged, lemmatized, and parsed – corpora, that became increasingly available during this period. However, Språkdata in any case focused almost exclusively on the lexicographical end of things, and inclusion of language technology tools in Språkbanken was not a priority in practice.

Also typically, these corpus interfaces had been developed in order to address specific research questions in research projects with fixed-term, normally external, funding, for example, the ORDAT and ITG projects mentioned above. When the funding period expired, there was no procedure in place to ensure maintenance and upgrade of corpora or tools. Nevertheless, these *were* in many cases made available through Språkbanken, as mentioned above, but without any explicit commitment or funding on Språkbanken's part that would ensure their continued availability. At the same time this availability naturally created an expectation among the users of the corpora and tools that these would stay available indefinitely.

2 Korp

In sum, as the first decade of this century was drawing to a close, Språkbanken had some ten different corpus search interfaces to three different storage and search solutions providing different modes of access to corpora in a handful of different, not mutually compatible, format–annotation combinations. If we add to this that several different programming languages were involved, it became increasingly evident that this situation was becoming untenable. After some false starts, the Korp project was launched in late 2009, taking advantage of earmarked funding for language technology research awarded by the University of Gothenburg, in the form of a new cross-faculty research unit, the *Center for Language Technology*. With Korp we wanted to keep the good features of existing corpus tools – both the ones we had in-house and others described in the literature (e.g. Kilgarriff et al. 2004; Davies 2005;

Figure 2: The original ORDAT interface (top; search string *svenska* 'Swedish') and a variant used for the Faroese newspaper corpus (bottom; search string *kúgv* 'cow')

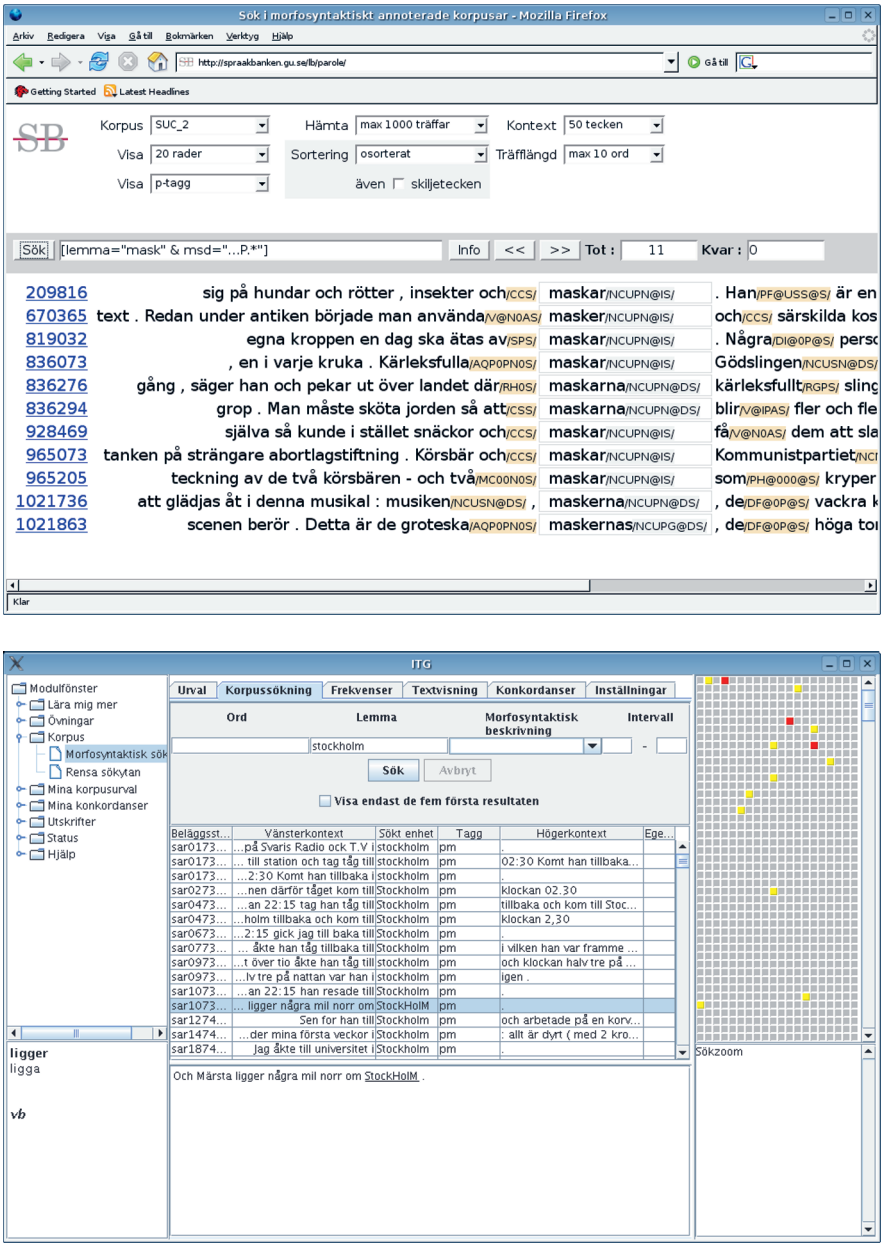


Figure 3: Interfaces for annotated corpora: the SUC corpus in the Parole interface (top; search for the lemma *mask* 'mask; worm' in the plural) and the ASU second-language corpus in the ITG interface (bottom; search for the lemma *Stockholm*)

Hoffmann & Evert 2006; Nygaard et al. 2008; Bick 2009) – and eliminate some of their more obvious limitations.

As already mentioned, we had tried out a fair number of other solutions. From our point of view, this turned out to be something of a blessing: there is no teacher like experience. We had been able to gain empirical experience of different storage and search solutions and their pros and cons. For instance, we could note that pure XML-based solutions (used among others in the ITG system shown in the lower screenshot in Figure 3) scaled poorly when the amount of text data grew beyond a few million words.

With Korp, we aimed to fulfill at least the following criteria:

- to have *one* word research platform for all corpora offered by Språkbanken; that consequently
- should accommodate several languages as well as all present and future annotations in the various corpora, with automatic annotations being added at corpus import time by a flexible ingestion pipeline;
- that completely new data annotations should become readily available throughout Korp, with little to no extra work;
- as far as possible to decouple the underlying corpus storage and search machinery (the *backend*) from the user interface(s) (the *frontend(s)*), by routing all access to the backend through a well-defined server-side application programming interface (API);
- that new functionalites in Korp are primarily added via the data first, the backend second, and the frontend (reluctantly) last, to simplify maintainability; and
- to explicitly maintain the strong connection between NLP and lexicography characteristic of Språkbanken.

Yet another requirement connected to Korp is that Språkbanken's downloadable corpora should be as close to exact copies as possible to the corpora in Korp (and other research platforms at Språkbanken). This does not only mean that the text data are kept in sync, but that the manual and automatic annotations of the corpus data are also kept in sync. For example, if you can search for named entities in Korp in a particular corpus, then you know that the downloadable version of that corpus is enriched with the same named entity annotations. All in the spirit of open science, where Språkbanken aims to shorten the distance between research data and research publications, via research platforms such as Korp.

The actual development of Korp started in 2010. As backend, the *IMS Open Corpus Workbench* was chosen, as being a mature open-source corpus storage and query engine under active development (Evert & Hardie 2011). Its Corpus Query Processor (CQP) query language could then be directly used as the Korp API. The frontend was built in-house using standard state-of-the-art web technologies (Borin,

Forsberg & Roxendal 2012). After almost two years of intensive development, the first official release of Korp was made in October 2011 at what was to become the first of Språkbanken's annual *autumn workshops*, initially held locally in Gothenburg, but since 2018 a national event.

At the time of writing in early 2025, Korp is at version 9. Even though it is not actively "marketed" in any particular way, Korp has been adopted by several centers as their main corpus exploration platform. Since Korp has been free software from its conception,[2] it can be freely used by other organizations (or individuals). The fact that is has garnered a small community of external deployers may also be due to the fact that it is maintained and continuously developed by a stable group of developers at Språkbanken Text, who respond promptly to questions and suggestions from Korp users, deployers, and contributors. Korp has been deployed at least in the following centers:

- Kielipankki in Finland;[3]
- Giellatekno, the research group for Saami language technology at UiT The Arctic University of Norway;[4]
- the Árni Magnússon Institute for Icelandic Studies, University of Iceland;[5]
- the Center of Estonian Language Resources in Estonia;[6] and
- the Department of Nordic Languages and Linguistics, University of Copenhagen, Denmark.[7]

In some cases, these sites have taken advantage of Korp's open source license and either contributed to the development of Korp (Finland) or set up their own customized Korp development repository (Denmark).

At the present time, Korp is the oldest member of an ecosystem of research infrastructure platforms built and maintained by Språkbanken Text, all supporting research based on language data in complementary ways; see Chapters 9, 11, and 15 in this volume for more details on the other platforms.

2 The source code is available under the MIT license on GitHub; see https://spraakbanken.gu.se/en/tools/korp/distribution-and-development (last accessed: April 4, 2025).
3 https://www.kielipankki.fi/korp/ (last accessed: April 4, 2025)
4 https://gtweb.uit.no/korp/ (last accessed: April 4, 2025)
5 https://malheildir.arnastofnun.is/ (last accessed: April 4, 2025)
6 https://korp.keeleressursid.ee/#lang=en (last accessed: April 4, 2025)
7 https://alf.hum.ku.dk/korp/ (last accessed: April 4, 2025)

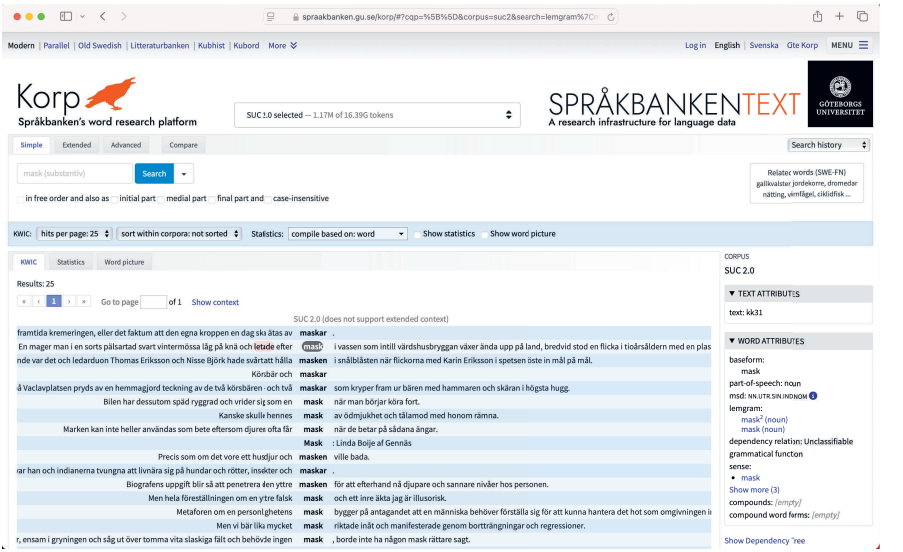


Figure 4: The Korp KWIC view showing a search for the noun *mask* 'worm' in the simple query mode

3 A well-tempered concordancer: durable design for word research

The basic functionality of a concordancer is to be able to search in corpus data and have the search hits returned in a form called *keyword-in-context* (KWIC), that is, where the search hits are centered on a single line and surrounded by their left and right context. A search hit with its context is typically a sentence in Korp. See Figure 4.

In the graphical user interface of Korp, the frontend, three query modes are available: *simple*, *extended*, and *advanced* mode (Borin, Forsberg & Roxendal 2012).

The simple mode allows searches for a word/lemma in selected corpora. It is possible to perform a search using the lexeme[8] as query expression, for searching both for its inflected forms and optionally also for compounds containing the lexeme as *initial*, *medial* or *final* part. With the lexeme query, all the inflected forms of the target word are listed in the search result with the KWIC view as default. As a service to the user, some example searches are provided via hyperlinks on the start page.

8 For historical reasons, the term *lemgram* is used as a synonym of lexeme in the context of Korp.

The extended mode is basically a graphical mirroring of conjunctive normal form (propositional logic) for full logical expressiveness. This mode allows more refined searches but still does not require knowledge of regular expressions. It is possible to combine numerous variables in a query, like words, word attributes, and text attributes, in a specific order within the sentence. For example, a query in selected corpora (press, magazines), and "POS is *adjective* + noun ends in *-ande* + sentiment is *positive*" results in NPs like *fler överlevande* 'more survivors', *stort firande* 'big celebration', *psykiskt välbefinnande* 'mental wellbeing' and *otillbörligt gynnande* 'unproper favoring', among others.[9]

The advanced mode allows custom CQP queries, i.e., it gives full access to the search capabilities of the backend, but with the cost that the user needs to know a formal and complex query language. To simplify matters somewhat, it is also possible to perform a simple and extended search and see its equivalent CQP expression that can be copied-and-pasted and modified by the user.

3.1 The data modes of Korp

The data modes of Korp subdivide the corpus collections of Korp into subcollections, e.g., a data mode with modern Swedish and another one with historical Swedish. The reason behind this division is typically that subcollections are too different to be searched together, such as corpora of different languages or with incompatible basic automatic annotations (such as part-of-speech tagging), but it could also be thematic divisions, such as a collection of all corpora of legal or religious texts in Språkbanken in one place, to ensure easy access. A corpus can appear in many (thematic) modes, but crucially, in the backend it only occurs once. The data modes are general functions, accessible on the top-level (entry) page of Korp.

3.2 The word picture of Korp

The word picture is one of few functionalities in Korp that makes strong assumptions about the underlying data, namely that the data has been analysed with a Swedish dependency parser using the Mamba-Dep format (Nilsson, Hall & Nivre 2006; Nivre et al. 2007).

9 Adding the suffix *-ande* to verbs can result in both adjectives and nouns, and without a longer context, it is not possible to know the actual POS in the case of some of the *-ande*-words. See Holmer (2022) for an elaboration on the topic.

Figure 5: A word picture of *korp* 'raven'

A dependency parse is a syntactic tree over a sentence where all words are dependents of a headword via a typed relation, except the main verb, which has the root relation to a dummy head. For example, in the sentence *The raven flies*, *the* is the dependent of *raven* through a determiner relation, *raven* is the dependent of *flies* through a subject relation, and *flies* as the main verb is the root dependent of a dummy word (see Figure 6).

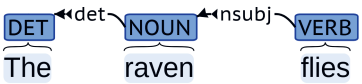


Figure 6: A dependency tree for the sentence *The raven flies*

A word picture in a particular corpus selection in Korp is created by searching for a lexeme or a word form in the simple mode. The lexeme must have been assigned one of the following parts of speech to appear as a word picture: noun, verb, adjective, or adverb.

A word picture gives a curated syntactic dependency summary of a word as it occurs in a corpus selection. In Figure 5 we have the word picture for *korp* 'raven'. From this overview we can see that typical modifiers for *korp* are *svart* 'black', *tam* 'tame', and forms of *treögd* 'three-eyed'. Typical verbs for *korp* as a subject are *flyga* 'fly', *kraxa* 'croak', and *hacka* 'pick', and typical verbs when *korp* is a object are *spela* 'play' and *se* 'see', where the somewhat surprising *spela* 'play' occurs because *korp* also has another meaning, namely that it is a colloquially shortened form of

korporationsidrott 'company sport' and typically refers to an amateur football club or league.

All words in the word pictures are hyperlinked to the KWIC view of their occurrences, which makes it possible to inspect the syntactic parses that the word pictures rely upon. And as the observant reader has already noted, the words are not ordered by their absolute frequencies (given next to the words) but with an association measure. In the literature, many different association measures have been suggested; the exhaustive survey by Pecina (2010) describes more than 80 such measures. For Korp's word picture we settled for the *lexicographer's mutual information* (LMI) measure as default, because it is a measure that captures typicality. Pure MI favors rare occurrences before the typical, which is offset by multiplying MI with the frequency to factor in what is typical as well. LMI was first used in the commercial corpus tool Sketch Engine; see the brief introduction by Kilgarriff et al. (2004), which inspired us to adopt it for Korp.

Even though it is possible to compute the word picture on the fly, it would take an unreasonable amount of time for the larger corpora to get a result. Instead, the word picture data is precomputed for every corpus. This comes with the drawback that it is not possible to create a word picture for a subcorpus, defined by, for example, a time period, without splitting the corpus data accordingly.

Since a word picture gives a bird's eye view of the syntactic behavior of a word and provides clues to its semantics, it is a popular tool for lexicographers, as illustrated in Section 4.2.

3.3 Counting words: statistics in Korp

Every data category of a search result is countable in the statistics view of Korp: text word types, word attributes, structural attributes, and text attributes. The categories can be counted individually or in combination. As an example of this, in Figure 7 we have compiled the search results of *korp* 'raven' on word forms and word senses. There are two word senses in the compilation, where the sense without an index refers to the bird and the other one (written with and without capitalization), the sense with a superscript figure "2", refers to a company sport association.

The statistics information can be viewed as tabular information in a spreadsheet-like view, as illustrated in Figure 7, but also as a *trend diagram* and as a *map*.

As an illustrative example of the trend diagram functionality we have plotted the word *korp* 'raven' in Figure 8. This functionality requires that the corpora in question have time-stamped information, which most of the corpora at Språkbanken have nowadays.

Number of rows: 23

word	sense	Total	SVT news 2004	SVT news 2005	SVT news 2006	SVT news 2007
Σ	Σ	0.9 (217)	0.0 (0)	0.3 (1)	0.0 (0)	1.3 (7)
korpar	korp	0.2 (50)	0.0 (0)	0.0 (0)	0.0 (0)	0.5 (3)
korp	korp	0.1 (28)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
Korparna	korp[2]	0.1 (24)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
Korpen	korp[2]	0.1 (22)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
korparna	korp	0.1 (21)	0.0 (0)	0.0 (0)	0.0 (0)	0.2 (1)
korpen	korp	0.1 (16)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
Korpar	korp	0.0 (12)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
Korparna	korp	0.0 (9)	0.0 (0)	0.0 (0)	0.0 (0)	0.2 (1)
Korpen	korp	0.0 (9)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
korpen	korp[2]	0.0 (7)	0.0 (0)	0.3 (1)	0.0 (0)	0.0 (0)

Figure 7: A compilation of the word-form and word-sense statistics of *korp* 'raven'

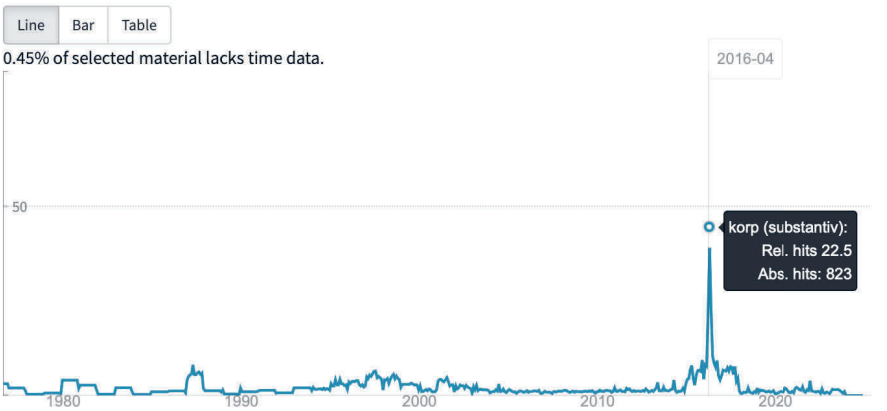


Figure 8: A trend diagram of *korp* 'raven'

Why do we see a conspicuous *korp* peak in 2016? The trend diagram is linked to the underlying data, so with a click we find the answer to this question. It was neither that the popularity of the Korp platform boomed nor that the nation was attacked by ravens. The 2016 peak reflects a trending discussion in Sweden at that time about which movies should be considered the best, where the Icelandic movie *Korpen flyger* 'The flight of the raven' (*Hrafninn flýgur*, Hrafn Gunnlaugsson 1984) came out among the top choices for many that engaged in the discussion.

Finally, we have the map functionality, which requires that the corpora contain geographic coordinates. These coordinates could either already be in the data, as in the Twitter corpus, or be the result of a computational analysis that enriches the

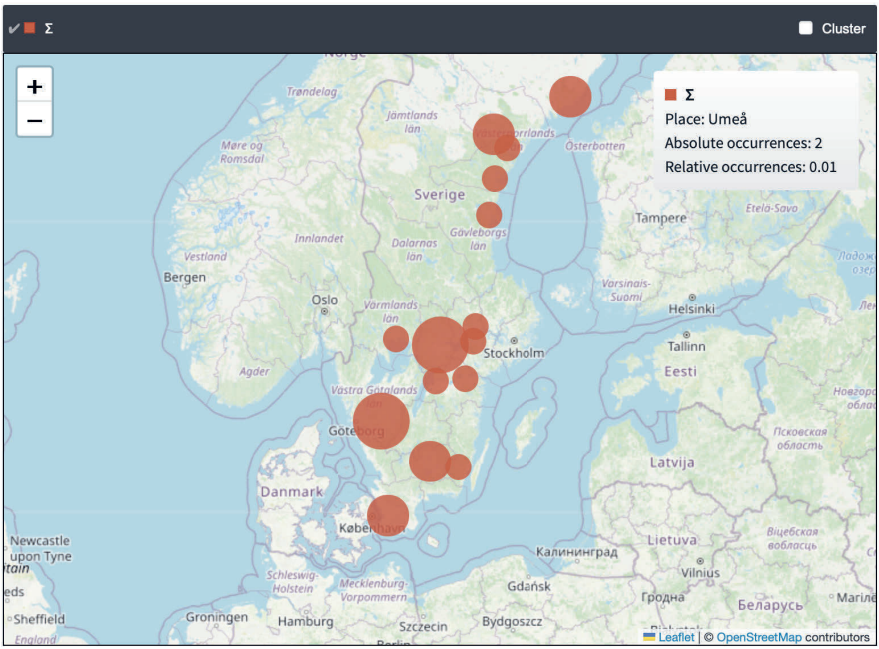


Figure 9: A map of *korp* 'raven'

data with geographic coordinates. Since most of the corpora lack pre-existing geo coordinates, it is typically the latter.

The geo coordinate analysis that is mainly used in Korp today is one that first identifies all place names through named-entity analysis and automatically provides a geo coordinate for them. Next, all words are given geo coordinates from the place names that they co-occur with in a sentence.

In Figure 9 the sentence co-occurrences of the word *korp* 'raven' with the place names in a corpus of public-service television news text are plotted on a map. Here we have clicked on the circle at the top and received the information that *korp* co-occurred with the city name *Umeå* twice. If we click on the information, we get a KWIC view of the actual occurrences.

As a final note, it is also possible to export the Korp statistics to a spreadsheet file in CSV format to enable further analyses outside Korp.

Table 1: Compounds ending in *-olja* in the subset Kubord 2 in Korp

compound with *olja*	frequency	in SAOL 14 (Y/N)
sesamolja	622	Y
motorolja	271	Y
kokosolja	248	Y
solrosolja	227	Y
tryffelolja	188	Y
frityrolja	111	N
träolja	87	N
chiliolja	70	N
jordnötsolja	67	Y
majsolja	59	Y

4 Korp as a tool for lexical research and lexicography

Korp is indispensable as a computational lexicographic tool. With Korp, the editors of SAOL (SAOL 14 2015) and SO (SO 2021) (see Chapters 3 and 4 in this volume) have the means at their disposal to conduct numerous kinds of lexical studies, such as frequency studies for single words, as well as grammatical and phraseological investigations for words in context. In the following sections, we have selected two ways of illustrating text-based research that form the basis for the scientifically grounded revisional work with SAOL and SO.

4.1 SAOL: Korp as a tool for inclusion and exclusion of headwords

For the development of SAOL, one of the main lexicographical tasks is to identify candidates for inclusion as headwords. The majority of headwords in SAOL are nouns, and the majority of these are compounds. To illustrate one of the applications where Korp is put into practice, we have chosen to investigate the word *olja* 'oil' and some of its compounds in the corpus called *Kubord 2* (Språkbanken Text 2025). Kubord 2 holds newspaper texts from the year 2010 to 2021, made available by a cooperation between the University of Gothenburg and the National Library of Sweden (see Chapter 4 in this volume).

In Table 1, a sample of the compounds ending in *-olja* are shown in descending absolute frequency order. As shown in Table 1, there are 10 compounds ending in

-olja, and seven of them are already headwords in SAOL 14 (marked "Y" in the table). Three of them, *frityrolja* 'deep frying oil', *träolja* 'wood oil', and *chiliolja* 'chili oil', are so far *not* headwords in SAOL (marked "N" in the table), but will be added in the 15th edition, mainly due to their high frequency, and also because they represent three slightly different senses of *oil* (oil for cooking, oil for furniture care, oil for direct human consumption).

The example featuring *olja* shows two things: how Korp might be used in the active search for lemma lacunae in SAOL, and to what extent existing compounds in SAOL reflect the use of *-olja* compounds in text. Pleasingly enough, the words that display the highest frequency, are already included in SAOL.

These results can also be used in reverse, as a method of identifying SAOL headwords with (too) low frequency[10] in the corpora, making them potential candidates for exclusion from the list of headwords in SAOL (Berg, Holmer & Sköldberg 2010; Diamond 2015; Holmer et al. 2024; see also Chapter 3 in this volume). Not all dictionaries exclude obsolete headwords, but the contemporary dictionaries SAOL and SO do. Excluded headwords can still be accessed through the historical project SAOLhist (see Chapter 13 in this volume), and older words are often registered in the historical dictionary SAOB (1898–2023).

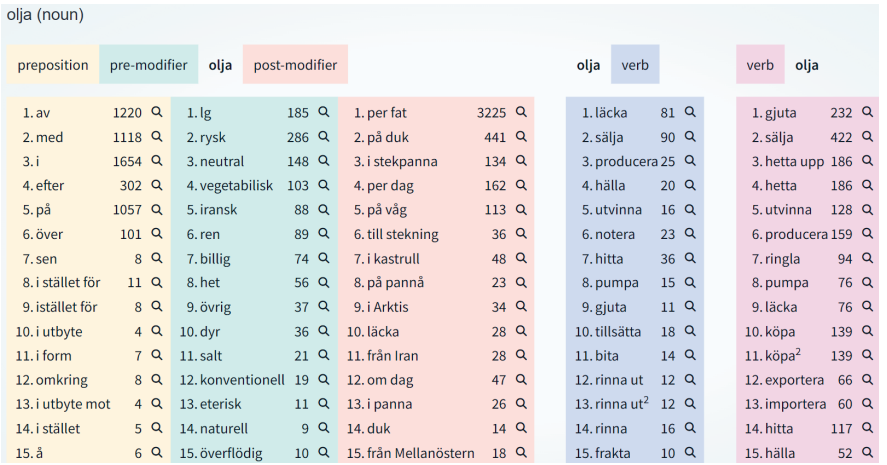
4.2 SO: using Korp to go beyond the word

In the contemporary dictionary SO 2021, there are about half as many headwords as in SAOL, approx. 65,000. In return, a typical entry in SO might display a headword's common compounds and derivatives in its entry text, along with the headword in constructions and further contexts (see Chapter 4 in this volume). The various ways of undertaking lexical studies for lexicographical purposes briefly described above can of course be used in the revision of SO as well as with SAOL. In this section, focus is on how the *word picture* in Korp (as illustrated in Section 3.2) can be used for this purpose.

As can be noted in Figure 10, the noun *olja* 'oil' forms the headword in NPs like *vegetabilisk olja* 'vegetable oil', *rysk olja* 'Russian oil', *neutral olja* 'neutral oil' and *eterisk olja* 'essential oil'.

A comparison between these search results and the current noun entry *olja* in SO, shows that the entry text in SO can be improved in several respects. Apart from most of the NP's mentioned above, phrases like *olja på duk* 'oil on canvas' and *olja på fat* 'oil per barrel' could very well strenghten the description in SO. Recurrent

10 Korp also provides relative frequency, occurrence per million words, but for convenience, only absolute frequency is listed in Table 1


olja (noun)

preposition		pre-modifier		olja		post-modifier		olja	verb		verb	olja	
1. av	1220	1. lg	185	1. per fat	3225			1. läcka	81		1. gjuta	232	
2. med	1118	2. rysk	286	2. på duk	441			2. sälja	90		2. sälja	422	
3. i	1654	3. neutral	148	3. i stekpanna	134			3. producera	25		3. hetta upp	186	
4. efter	302	4. vegetabilisk	103	4. per dag	162			4. hälla	20		4. hetta	186	
5. på	1057	5. iransk	88	5. på våg	113			5. utvinna	16		5. utvinna	128	
6. över	101	6. ren	89	6. till stekning	36			6. notera	23		6. producera	159	
7. sen	8	7. billig	74	7. i kastrull	48			7. hitta	36		7. ringla	94	
8. i stället för	11	8. het	56	8. på pannå	23			8. pumpa	15		8. pumpa	76	
9. istället för	8	9. övrig	37	9. i Arktis	34			9. gjuta	11		9. läcka	76	
10. i utbyte	4	10. dyr	36	10. läcka	28			10. tillsätta	18		10. köpa	139	
11. i form	7	11. salt	21	11. från Iran	28			11. bita	14		11. köpa[2]	139	
12. omkring	8	12. konventionell	19	12. om dag	47			12. rinna ut	12		12. exportera	66	
13. i utbyte mot	4	13. eterisk	11	13. i panna	26			13. rinna ut[2]	12		13. importera	60	
14. i stället	5	14. naturell	9	14. duk	14			14. rinna	16		14. hitta	117	
15. å	6	15. överflödig	10	15. från Mellanöstern	18			15. frakta	10		15. hälla	52	

Figure 10: Korp: Basic search with the word picture mode, with a search for *olja* 'oil'

verbs in connection with *olja* are, according to this word picture, *läcka* 'leak', *sälja* 'sell', *producera* 'produce' and *hitta* 'find', the latter rather referring to discovering an oil field (more than just finding any oil).

4.3 Korp as a general tool in lexicography

Apart from the in-house use of Korp in relation to the contemporary dictionaries SAOL and SO, the work on the historical dictionary SAOB has also been able to benefit from the platform. The SAOB editors have digitized novels in Swedish published from 1950 and onwards, in order to create a corpus with special relevance to a more historically oriented dictionary with strict requirements regarding reliable source information, etc. This particular corpus covers a time period and a text material that are otherwise not as common in Korp as for example modern newspaper text. The corpus holds material from novels from 1950 to 2007, partly manually selected to represent many different genres, and partly randomly selected (Språkbanken Text 2024b; see also Forsberg & Holmer 2024).

Korp is also a natural platform for various and varied lexical studies. Students at our department, as well as scholars associated with the University of Gothenburg, frequently use Korp for a multitude of reasons in lexical investigations.

5 The future of Korp

The main goal of the further development of Korp is to ensure that it remains a reliable and responsive research platform that can be maintained and made available in coming decades. To make this possible we need to resist the temptation to add more and more functions to Korp every year, to avoid turning Korp into a technical Frankenstein's monster that will be impossible to maintain in the long run. Instead, new functionalities should primarily be added through the data, which requires that the existing functionalities are general enough to enable such a goal.

To make this goal more concrete: say that we want to enable something new, namely the tracking of political party stances over time. Since we are making no particular assumptions about the data in the trend diagram functionality, the only thing we need to do to enable this, is to add stance information to the data, either automatically or manually.

Språkbanken is celebrating its 50th anniversary this year (2025). The (KWIC) concordance format has been around much longer than Språkbanken or even digital corpora. The first Bible concordance was worked out already in the 13th century, reputedly by mustering the efforts of 500 monks (Miller 1947: 63). Many concordances have been published in the intervening 800 years, for a large number of works in many languages. It does not defy reason, then, that a descendant of Korp may still be in use at Språkbanken's centenary celebration, fifty years from now.

References

Berg, Sture, Louise Holmer & Emma Sköldberg. 2010. Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2010*. 567–576.

Bick, Eckhard. 2009. DeepDict: A graphical corpus-based dictionary of word relations. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 268–271.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. *International Conference on Language Resources and Evaluation (LREC) 2012*. 474–478.

Borin, Lars & Anju Saxena. 2004. Grammar, incorporated. In Peter Juel Henrichsen (ed.), *CALL for the Nordic languages*, 125–145. Copenhagen: Samfundslitteratur.

Davies, Mark. 2005. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics* 10(3): 307–334.

Diamond, Graeme. 2015. Making decisions about inclusion and exclusion. In Philip Durkin (ed.), *The Oxford handbook of lexicography*, 532–545. Oxford: Oxford University Press.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*.

Forsberg, Markus & Louise Holmer. 2024. Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text [Data access, methodological development and lexicographical work at Språkbanken Text]. *LexicoNordica* 31: 61–79.

Gellerstam, Martin & Christian Sjögreen. 1994. *Språkbanken: En språklig referensdatabas* [Språkbanken: A linguistic reference database]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Gustafson-Capková, Sofia & Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm: Department of Linguistics, Stockholm University.

Hammarberg, Björn. 2010. *Introduction to the ASU Corpus: A longitudinal oral and written text corpus of adult learner Swedish*. Stockholm: Department of Linguistics, Stockholm University. https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-112130.

Hoffmann, Sebastian & Stefan Evert. 2006. BNCWeb (CQP-edition): The marriage of two corpus tools. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods*, 225–240. Frankfurt am Main: Peter Lang.

Holmer, Louise. 2022. *Neutrala substantiv på* -ande *i text och ordbok* [Deverbal neutral nouns ending in *-ande* in text and dictionary]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Holmer, Louise, Ann Lillieström, Emma Sköldberg & Jonatan Uppström. 2024. Time to say goodbye revisited: On the exclusion of headwords from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 443–452.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. *Proceedings of the European Association for Lexicography (EURALEX) 2004*. 105–115.

Malmgren, Sven-Göran. 2000. *Projektet Det svenska ordförrådets utveckling 1800–2000: Utgångspunkter* [The project The Evolution of the Swedish Vocabulary 1800–2000: Points of departure]. (ORDAT No. 1) Gothenburg: Department of Swedish, University of Gothenburg.

Miller, Donald G. 1947. Implements of interpretation: I. Concordances. *Interpretation* 1(1): 52–62.

Nilsson, Jens, Johan Hall & Joakim Nivre. 2006. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. *Treebanking for Discourse and Speech: Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA) 2005 Special Session on Treebanks for Spoken Language and Discourse*. 119–132.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2): 95–135.

Nygaard, Lars, Joel Priestley, Anders Nøklestad & Janne Bondi Johannessen. 2008. Glossa: A multilingual, multimodal, configurable user interface. *International Conference on Language Resources and Evaluation (LREC) 2008*. 617–622.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1–2): 137–158.

SAOB. 1898–2023. *Svenska Akademiens ordbok* [The Swedish Academy Dictionary]. Lund: Gleerups.

SAOL 14. 2015. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 14th edn. Stockholm: Norstedts.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

Språkbanken Text. 2017. *ORDAT*. [Data set]. DOI: 10.23695/yd1h-ag22.

Språkbanken Text. 2022. *ASU*. [Data set]. DOI: 10.23695/m34m-v368.

Språkbanken Text. 2024a. *PAROLE lexicon*. [Data set]. DOI: 10.23695/CACY-SB38.

Språkbanken Text. 2024b. *SAOB1950*. [Data set]. DOI: 10.23695/ZPH6-EN76.

Språkbanken Text. 2024c. *SUC 3.0*. [Data set]. DOI: 10.23695/wy84-ar30.

Språkbanken Text. 2025. *Kubord 2*. [Data set]. DOI: 10.23695/CKMP-PV98.

Lars Borin, Emma Sköldberg, Ann Lilliestrom, Nick Smallbone, Maria Öhrman, Jonatan Uppström, and Louise Holmer

11 Karp: Språkbanken's data editing platform

Abstract: Karp is Språkbanken's platform for the development and editing of lexical data, or more generally, formally structured data. It provides functionality for data exploration, such as support for searching, browsing, and compiling, together with support for versioned, collaborative editing. It also provides methodological support, such as data prediction, data decision support, and data validation. In this chapter we focus on how the work on the two dictionaries SO and SAOL at the University of Gothenburg is carried out in Karp, and describe the methodological support that is currently under development for SO and SAOL, such as support for formal verification, neologism detection, data-driven lexical entry generation, and empirical anchoring of lexical information.

Keywords: dictionary writing system, lexical database, research infrastructure, Swedish lexicography

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*. Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for preparing this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Emma Sköldberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: emma.skoldberg@svenska.gu.se
Ann Lilliestrom, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: ann.lilliestrom@svenska.gu.se
Nick Smallbone, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: nicsma@chalmers.se
Maria Öhrman, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: maria.ohrman@svenska.gu.se
Jonatan Uppström, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: jonatan.uppstrom@svenska.gu.se
Louise Holmer, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: louise.holmer@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-011

1 Background: from lexical resource management system to data editing platform

Karp is one of Språkbanken's research platforms, an ecosystem of research infrastructure components developed and maintained by the Språkbanken Text division at the University of Gothenburg. The platforms are designed to support scholars conducting research based on language data. Other research platforms described in this volume are the word research platform Korp and the text research platform Strix (see Chapters 10 and 15 in this volume)

structured language data, above all lexical data. The architecture of Karp mirrors that of its sister research platform Korp (see Chapter 10 in this volume). Just like Korp, Karp has a modular architecture consisting of a *backend* – a server-side component which stores, modifies, and serves lexical and other formally structured language data – and a *frontend*, a web interface by which users can search, browse, and edit such data (see Figure 1). The communication between the frontend and the backend is handled by a well-defined application programming interface (API).[1]

Karp furnishes an excellent example of the advantages of this software design principle. The first version of Karp relied on eXist, a native XML database management system, for access to lexical data expressed using the ISO standard format *Lexical Markup Framework* (LMF; Francopoulo 2013). Computer software is a moving target however. It turned out that the performance of the adopted database manager did not scale up well to deal with increasing amounts of data. We also found that our numerous onomasiological (semantic) lexical resources (including the pivot lexicon *Saldo* and the Swedish FrameNet; see Chapters 6, 7, and 8 in this volume) did not fit well into the strongly semasiological information model of LMF. Finally, over a number of years XML has gradually been abandoned as the preferred data representation formalism in language technology. For these reasons, the present version of the Karp backend uses JSON(L) as the data representation language and the database management system Elasticsearch for data retrieval and manipulation. This transition, while not completely without hurdles, could be made in a fairly painless way while keeping the appearance and functionality of the Karp frontend.

Karp is Språkbanken's data editing platform; its application area is formally The original motivation behind the development of Karp was multifaceted. In part there was certainly an aspiration to emphasize the dual nature of Språkbanken's data holdings, i.e., the originally envisioned separate and complementary "text bank" and

1 https://spraakbanken4.it.gu.se/karp/v7/ (last accessed: April 4, 2025)

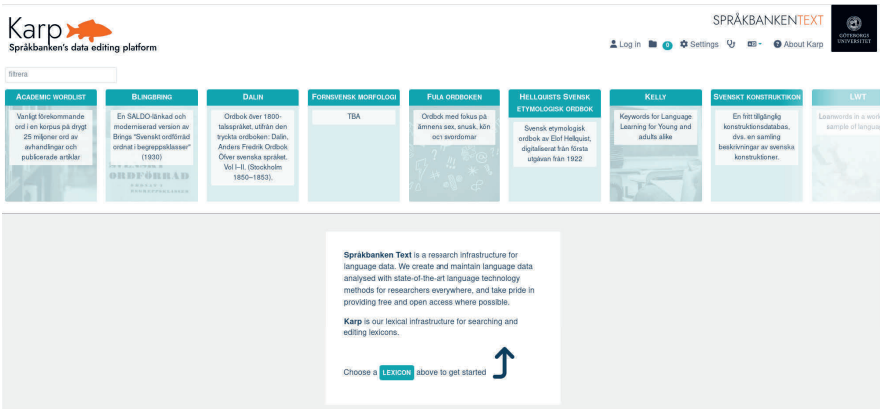


Figure 1: Karp's user interface

"word bank" aspects of Språkbanken (Allén 1970; 1973; [1980] 1999a; [1983] 1999b; [1984] 1999c; Gellerstam & Sjögreen 1994; see also Chapter 5 in this volume).

The work on Karp started in 2011, as part of the *Swedish FrameNet++* (SweFN++) project described in Chapter 5 in this volume. This coincided with the launch of Korp (see Chapter 10 in this volume), which stood as testimony to the great utility of having a unified platform for research based on text corpus data. To restore the balance between the two faces of Språkbanken, a similar initiative was needed also for Språkbanken's lexical resources. Karp became the most important means of accomplishing the integration of the many heterogeneous lexicons to be included in the lexical macroresource developed in the SweFN++ project (Borin et al. 2012; Ahlberg et al. 2013; Borin et al. 2013). The first version of Karp was officially released a year later, in late 2012.

The Center for Lexicology and Lexicography (CLL), as well as Språkdata before it (see Chapter 2 in this volume), had been developing lexical databases for many years. However, similarly to the pre-Korp concordancers built by the original Språkbanken (see Chapter 10 in this volume), the lexical databases designed by Språkdata and later CLL were specialized, tailored to support a particular task at a particular time. They did not have a common information model beyond the use of a relational database management system for data storage and access. Thus, there were (slightly) different databases for the various dictionaries. Distinct from this, Karp was designed with an explicit *general* lexical information model at its core. In essence, this model was that of Saldo, with unique mnemonic identifiers for all relevant entities, such as word senses, lexemes, inflectional paradigms, etc.; see Chapter 6 in this volume. Importantly, this model is independent of storage format, as indicated above. The locked-in nature of the lexical data in the existing lexical databases also presented a

disincentive to adapt one of them as the basis for Karp, which, just like the lexical macroresource being built in the SweFN++ project, was intended to be integrative, multi-purpose, and open-access.

Karp thus started out as a tool providing support for navigating, integrating, and curating the heterogeneous component lexicons of the lexical macroresource developed in the SweFN++ project, and in particular for building and editing the Swedish FrameNet being developed in that project (Dannélls et al. 2021).

Empirically, it turns out that there is an asymmetry between the text corpus data and the lexical data offered by Språkbanken: Korp is heavily used for all kinds of investigations of the behavior of words in text at all levels, from academic researchers to members of the public interested in language matters. Experience shows that the user base of Karp as a browsing and discovery tool is much narrower. Instead, Karp has found a niche as (lexical) data editing tool in the first hand, and also to some extent as a repository of formally structured but non-lexical linguistic data, for instance, inflectional paradigms for a large number of languages extracted from a digitized version of the *Linguistic survey of India* (Grierson 1903–1927; see Borin, Virk & Saxena 2018). In fact, much of the development of Karp has been driven by the realization that editing formally structured language data, particularly lexical data, constitutes its main use case: in addition to the Swedish FrameNet (Dannélls et al. 2021; see also Chapter 7 in this volume), Karp is used for building and editing the Swedish ConstructiCon (SweCcn; Lyngfelt et al. 2018; see also Chapter 14 in this volume), as well as for editing the two dictionaries SO and SAOL (see Section 2 below).

2 Salex: the Swedish Academy's lexical database in Karp

The most recent addition to Karp's datasets, as well as the one currently receiving most attention with regard to active development of Karp, is *Salex*, the Swedish Academy's lexical database (see Figure 2). We now turn to describing Salex and Karp primarily from the point of view of a lexicographer compiling the Swedish Academy's contemporary dictionaries, i.e., the more descriptive definition dictionary *Svensk ordbok utgiven av Svenska Akademien* 'The Contemporary Dictionary of the Swedish Academy' (SO 2021) and the more normative dictionary *Svenska Akademiens ordlista* 'The Swedish Academy Glossary' (SAOL 14 2015). See Chapters 3 and 4 in this volume, where the dictionaries are presented in detail.

Access to a *dictionary writing system* (DWS) is crucial in the creation, maintenance, and development of a lexical database from which dictionaries can be generated. Such a system simplifies and automates much of the work that was previ-

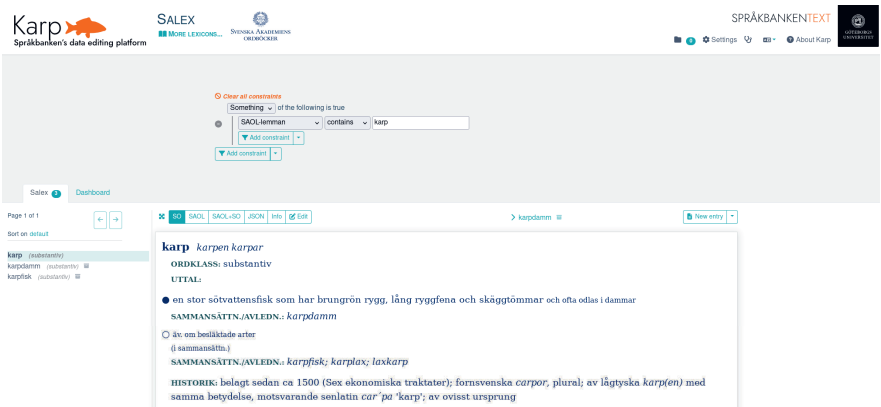


Figure 2: The Swedish Academy's lexical database Salex seen in the Karp user interface

ously done manually, which affects the pace and efficiency of the editorial work. It allows the lexicographers to focus on the lexicographic tasks, rather than worrying about the technology. It also leads to higher quality dictionaries thanks to greater internal consistency (see, e.g., Atkins & Rundell 2008; Svensén 2009 on dictionary editing systems).

An adequate and well-functioning DWS has to be intuitive and easy to navigate, making it easy for lexicographers to learn and use. Ideally, it is also sustainable, since dictionary projects are often long-term by nature. Several previous systems have not technologically withstood the test of time, causing problems for the lexicographic projects (see Sjögreen & Sköldberg 2010). In other words, the system must be flexible enough to meet both current and potential future needs.

Given the multifarious demands placed on such a system, it is easy to understand that an efficient DWS is expensive to both create or acquire and also to maintain and refine. This may explain why there are so few freely available systems on the market. This fact makes Karp particularly interesting and relevant in lexicographical contexts.

2.1 From Språkdata and CLL to Språkbanken Text

The work on the contemporary dictionaries SO and SAOL was previously carried out at the (then) department of Språkdata and within the Center for Lexicology and Lexicography (CLL) at the University of Gothenburg (see Chapter 2 in this volume). As mentioned above, much of the lexicographic work conducted within these units was based on relational databases (Sjögreen & Sköldberg 2010).

Since 2021, the work on the contemporary dictionaries of the Swedish Academy is being carried out within Språkbanken Text. In part, the project activities since then have aimed at incorporating the contents of the previous databases into Karp, which offers a hierarchical document-centric database model that is more flexible in its structure.

A key goal since the relocation of the lexicographic project into Språkbanken Text concerns methodology. One focus area has been the development of new methods for different types of lexical studies that can strengthen the dictionary material in several ways, taking advantage of the lexical data in Karp as well as the language technology expertise at Språkbanken and the existing infrastructure at the department. See Section 2.8 below for some examples, and also Chapter 15 in this volume.

The improved methods aim to support and streamline lexicographic tasks, ultimately improving the quality of the dictionaries even further. Hopefully, Språkbanken will also be able to benefit from this merger, strengthening its lexicographical competence, and getting easier access to the renowned dictionaries SO and SAOL, developed over many years.

As indicated, and unlike many other lexicographic environments in the Nordic countries, there is a tradition at the University of Gothenburg to use tailor-made, in-house systems instead of commercial ones (Sjögreen & Sköldberg 2010). One advantage of using a system like Karp is that it can be precisely adapted to the specific characteristics of the current dictionaries as well as to the needs of the lexicographers that make themselves felt in the course of the editorial work.

The further development of Salex in Karp takes place within the framework of a very close collaboration between language technologists and lexicographers. The platform's features and the interface are continuously developed, and the current version of Karp offers only a portion of all the features and functionalities that will be available for the lexicographers in the future.

As a comparison, the current work on the Swedish Academy's historical dictionary, *Svenska Akademiens ordbok* 'The Swedish Academy Dictionary' (SAOB 1898–2023), can be mentioned. The editorial team, based in Lund, has recently acquired a commercial editing software from the French company IDM for the updates of the dictionary. The software has been adapted to meet the complex structural requirements of SAOB (see Nilsson 2024).[2]

The members of the contemporary dictionary project are now (March, 2025) working on updating the content of SO and SAOL. The lexicographic work involves finding and selecting new headwords, updating and further developing existing entries, and excluding headwords that are no longer used in present-day Swedish

2 The first volume of SAOB was published in 1898, and its final, 39th, volume appeared only in 2023.

texts (Holmer et al. 2024). The work is highly data-driven and the lexicographers' access to modern corpora, such as those offered by Språkbanken, is crucial.

The next edition of SO will only be published electronically (as apps and on the web), while SAOL will be published in print as well as electronically (see Chapters 3 and 4 in this volume). The different publication formats place differing demands, not only on lexicographers and computational linguists, but also on the data in Karp, which on the one hand will be exported to typesetters, and, on the other, to external developers of apps and web interfaces.

A significant strength of Salex in Karp is that the two dictionaries are becoming more tightly integrated with each other. By having the data interconnected and displayed in the same interface, it becomes easier for lexicographers to see similarities and differences between the sets of headwords, how the same headwords are described in the two dictionaries, etc. This is advantageous, especially in the continued work of refining these two different types of dictionaries, their perspectives, and the functions they are intended to fulfill among the users. Unjustified differences between the two dictionaries become more visible, and these differences can be addressed more easily, as the lexicographers are editing both dictionaries simultaneously rather than one at a time, as was done before (see, e.g., Blensenius 2023; Sköldberg 2023).

2.2 Salex in Karp

To illustrate the size and complexity of Salex, here are some details about its content: As of March 2025, Salex includes approx. 214,000 so called *super entries* (on super entries, see below). Around 140,000 of them include SO entries and SAOL entries that are planned to be published in the next editions of the dictionaries.[3] The remaining approx. 74,000 super entries are "hidden" and will not be published at present. Among these are entries that have been removed from previous editions of the dictionaries.

Furthermore, the approximately 65,000 entries included in SO contain around 70,000 main senses, 27,000 subsenses, approx. 130,000 examples (morphological and syntactic), and 4,600 idioms.

As for SAOL, the number of entries amount to 129,000. As SAOL is a more trimmed-down dictionary, its entries do not always contain a meaning description, and therefore are not always sense separated. SAOL provides, above all, orthographical and inflectional information and less fine-grained sense distinctions than SO. About

3 There will be a minor update of SO and a new (15th) edition of SAOL. See Chapters 3 and 4 in this volume.

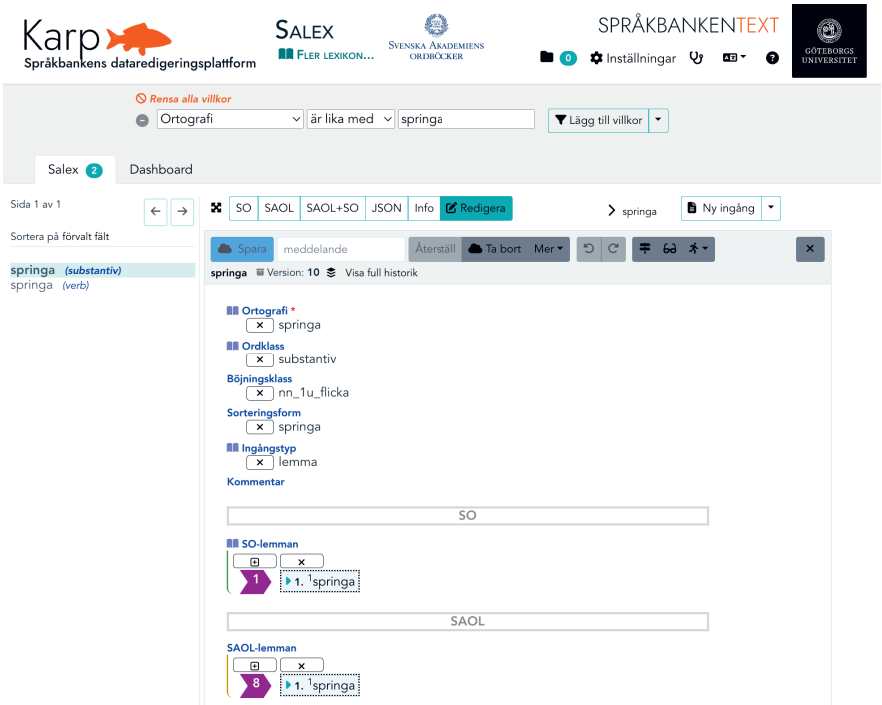


Figure 3: The super entry *springa* (noun) in Salex in Karp (March 2025). Labels in bright blue (top to bottom): 'orthography', 'word class', 'inflection class', 'sorting form', 'entry type', 'comment'. Labels in green (top to bottom): 'SO entry', 'SAOL entry'

73,000 entries do however have information on their main sense. Headwords, senses, etc., in both dictionaries are provided with unique IDs, a prerequisite for the cross-references found in each dictionary.

Headwords in SO and SAOL that share the same spelling, word class and inflectional properties, are combined into super entries, e.g. *vän* 'friend', *unik* 'unique' and *skratta* 'laugh'. As a comparison, entries that currently do not constitute super entries are the noun *bikini* 'bikini' in SO and in SAOL as they are provided with different inflectional forms. In SO, the dictionary user finds one plural form: *bikinis*. The same plural form is listed in SAOL, but another plural form, *bikinier*, more in accordance with traditional recommendations about how to write Swedish, is recommended (see Blensenius 2023; Holmer & Blensenius 2023).

2.3 Karp's search interface: the noun *springa* 'opening, slot'

In the following, some examples from the lexicographic work in relation to the platform are provided. Figure 3 shows Karp's user interface when searching for the string "springa", which corresponds to both a noun ('opening, slot') and a verb ('run'). On the left side of the figure, the two super entries *springa* ("substantiv" 'noun') and *springa* (verb) are listed. As indicated by the light blue color in the list, the super entry *springa* (noun) has been chosen.

At the top of Figure 3, it is shown that the interface allows the lexicographers to search for headwords that match ('are equal to') the string entered in the search field. The users can also search for entries that begin or end with the same letter sequence (e.g., *springare* 'knight; steed', *dörrspringa* 'door slot, door gap' and *småspringa* 'run lightly'), for entries where the search string appears or is absent, as well as search with regular expressions. In addition, the lexicographers can search within the dictionary entries, for example, for entries including specific pragmatic comments like "vardagligt" 'colloquial'. However, this type of searches can currently only be performed in another system that is connected to Karp. The different search functions allow the lexicographers to see and deal with groups of headwords more consistently, e.g., headwords with similar formal characteristics or similar emotive charge.

Furthermore, the lexicographers have access to rendered views of the Salex data. These views show how the SO entry, the SAOL entry, or both together will be presented in future publications. Additionally, the lexicographers can view the entries in JSON format. Figure 3 shows the current editing view. The different display modes of the data are used in different editorial tasks and accommodate different preferences and needs among the project members.

Just below this, the lexicographers see the orthography of the super entry, i.e. *springa*. The word class is identified as a noun and the super entry belongs to the inflection class nn_1u_flicka, meaning it is inflected like the word *flicka* 'girl' (definite singular form *flickan*, plural form *flickor*). Moreover, the headwords that constitute this super entry should be sorted in the list of headwords as *springa*, and they constitute proper headwords and not, for example, references to other entries. In this section of the user interface, there is also a field where project members can leave comments about the super entry for other project members.

Below this, the user finds two headings, "SO-lemman" 'SO entries' and "SAOL-lemman" 'SAOL entries', where the content of both entries can be expanded. The superscript number *1* in ^{1}*springa* indicates that this is the first of two or more homographs in each dictionary (^{2}*springa* refers to the verb). The purple arrows with the numbers 1 and 8 indicate that there is a reference to the SO entry *springa* and eight references to the corresponding entry in SAOL. More precisely: there is a cross-reference to the SO entry from the semantically related noun *spricka* 'crack' in

SO, and from compounds ending in *-springa* in SAOL's list of headwords (for instance, *dörrspringa* 'door gap').

2.4 The verb *springa* 'run': editing view

The super entry *springa* ('run'; verb) belongs to the group of words that has the inflection class vb_4n_brinna, meaning it is inflected similarly to the strong verb *brinna* 'burn'.

The entry *[2]springa* in SO is representative of the dictionary in that it contains a variety of information categories (such as inflection, pronunciation, main senses, subsenses, morphological and syntactic examples, cross-references, idioms and historical data). The entry structure is hierarchical as two of the three main senses have subsenses.

In Figure 4, the editing view for the corresponding entry in SAOL is shown. The information provided in the SAOL entry *[2]springa* is relatively limited, due to SAOL's primary task which is to give recommendations on spelling and inflection (see Chapter 3 in this volume). SAOL also provides information on word segmentation, which also indicates hyphenation points. In cases like this, pronunciation is not specified, as the user is expected to deduce it from general pronunciation rules of Swedish.

The entry shows the spelling of this verb, clarifies that it is a simple word rather than a compound, and lists its inflectional forms (under "Böjning" 'inflection' in Figure 4). The verb is said to have two main senses: 'to run' and 'to burst forth suddenly; to break into pieces suddenly'. The latter main sense, which in turn is divided into two nuances of meaning by a semicolon, is illustrated with the syntactic language example *springa i luften* 'blow up, explode (intransitive)', literally 'burst into the air'. A comment on word formation is found at the end of the entry: it states that most compounds beginning in *spring-* are related to the first sense of the verb.

In the editing view, the lexicographers can easily modify the entry text content, as is typical in a DWS. They can add or delete information (such as a sense) from the database by clicking on the icons "+" or "x", and they can also indicate information that should not be visible in the published dictionary. In that case, the information remains stored in the database. The lexicographers can also reorder the senses, etc.

2.5 Assignment of inflection class to new super entries

Examples of new super entries in Salex are the more or less synonymous adjectives *kontantfri* 'cash-free' and *kontantlös* 'cashless'. These words, whose usage has in-

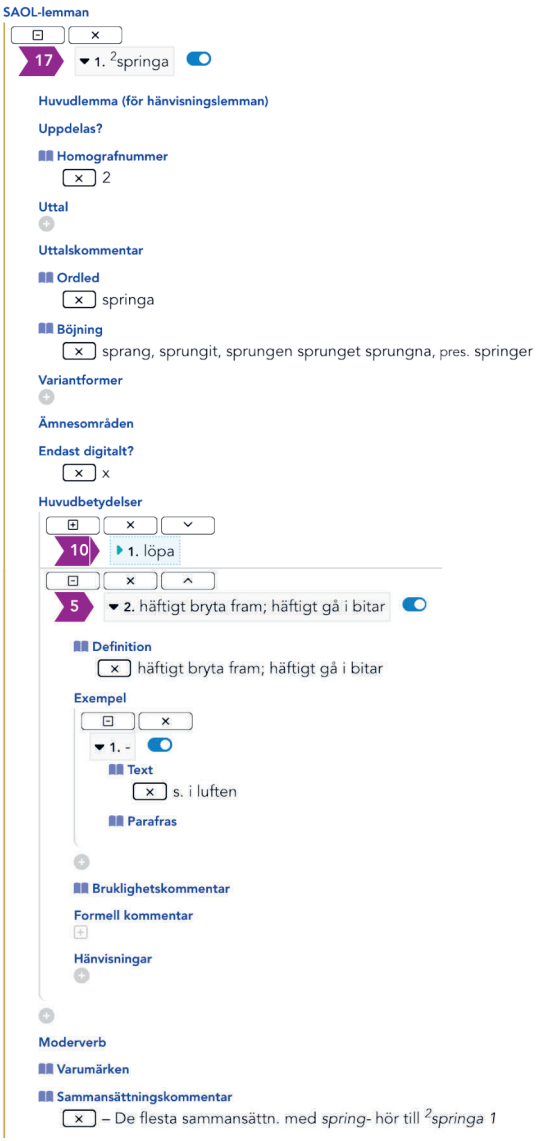


Figure 4: The SAOL verb *²springa* in the editing view in Karp (March 2025). Labels in bright blue (top to bottom): 'SAOL entry, 'main entry (for reference entries)', 'should the headword be divided?','homograph number', 'pronunciation', 'pronunciation comment', 'word parts', 'inflection', 'variant forms', 'subject field', 'digital only?', 'main senses', 'definition', 'example', 'text', 'paraphrase', 'usage label', 'formal comment', 'references', 'main verb', 'trademarks', 'compound comment'

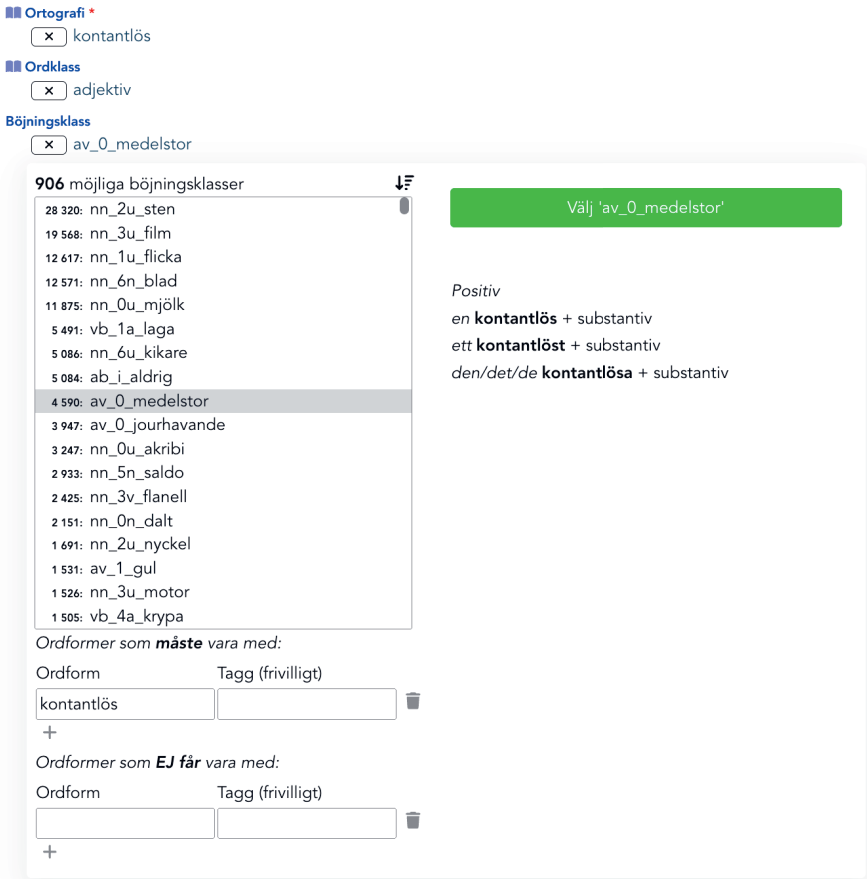


Figure 5: Selection of inflection classes in Salex (the super entry *kontantlös* 'cashless'). Labels in bright blue (top to bottom): 'orthography', 'word class', 'inflection class'

creased significantly in the last decade, have been discovered thanks to the neologism tool developed within the dictionary project (see Chapter 4 in this volume).

When incorporating these compounds, the lexicographers choose among the 906 different inflection classes listed in order of frequency (see Figure 5; on inflection classes in SAOL, see Berg 2009).

The adjective *kontantlös*, like 4,590 other super entries in the database, is assigned to the inflection class av_0_medelstor. To the right in the interface, it is shown which inflectional forms of the current adjective will be presented in a future e-publication of SAOL. The dictionary user will in the next edition of SAOL be informed

about the adjective's agreement inflection. Since this adjective does not undergo comparison, only the positive forms are provided.

2.6 Inclusion of cross-references

An example of an entry added to SO for the upcoming update is the adjective *plånboksvänlig* 'wallet-friendly'. The word was already in the database as it is included in SAOL since the edition from 2006, but the information provided about the word is very brief. In SO, the description becomes more comprehensive with corpus-based language examples, pronunciation indication, date of first appearance in Swedish texts, etc. The newly compiled SO entry contains three cross-references, including the entry *prisvärd* 'worth the price', which lexicographers were alerted to via Kubord fastText (see Chapters 3 and 4 in this volume). The tool, developed within the project, is very valuable, especially for clarifying semantic connections between different Swedish words in the dictionaries and in corpora. This tool also complements the SO findings in Språkbanken's text research platform Strix (see Chapter 15 in this volume). Thanks to Strix, the lexicographers can find semantic connections within the dictionary, such as between synonyms, antonyms, and hyponyms that are already incorporated.

2.7 Reordering of senses among polysemous headwords

As previously noted, more parallel work with both SO and SAOL in the same database can lead to better descriptions and fewer unjustified differences between the dictionaries. For instance, lexicographers have been provided with lists of entries that are polysemous, and where the listed senses in the dictionaries appear in different orders.

One example is the noun *energitjuv* 'energy thief', referring both to a device that consumes a lot of electricity, and a person perceived to drain energy from others around them. At the time of writing (early 2025), the two senses are presented in opposite order in SO and SAOL on Svenska.se (Svenska.se 2025), which may confuse dictionary users. In Karp, the lexicographer can easily change the order of the described meanings. Additionally, the lexicographer is informed (via the purple arrows) about cross-references, etc., related to these senses, which can lead to other editorial changes in both dictionaries. In upcoming editions, the two senses will appear in the same order; the sense related to devices will be listed first, as it is the oldest in Swedish.

Figure 6: A comparison between the original entry and the revised version in Salex

2.8 Continuous data validation and versioning

Salex in Karp is continuously and automatically validated, leading to quality improvements in both technical and lexicographical work. Tests have revealed typos, incomplete information, broken links, etc. and these issues, some of which were present in previous editions and others are newer, can be easily and continuously corrected by the project members.

Large-scale dictionary development sometimes involves several editors working in parallel – in some cases even making changes to the same lexical entry simultaneously. This may lead to collisions or even contradicting data in the lexicon. Karp provides several mechanisms to help overcome these problems.

When an editor starts working with an entry, the data is copied to the memory of the user's web browser. When the editor eventually hits *save*, the data is transferred back to the server (along with an optional comment that describes the edit). If the entry has been updated by another user in the meantime, the system rejects the save operation but offers to help merge the two versions into one. If it is not possible to automatically merge the two edits, the editor may choose to override the other user's edit, to discard the local version, or to do a manual merge of the two with the help of a special comparison interface. This is arguably a more practical alternative to using a locking mechanism or the like. See Figure 6.

Furthermore, Karp saves a version history for each entry, which helps in keeping track of changes and to avoid back-and-forth editing. It is also possible to list the latest edits performed by a particular user or by all editors.

3 Conclusion

From its humble beginnings as a practical project-internal tool developed in order to manage the heterogeneous lexicons of the SweFN++ macroresource, Karp has grown into one of Språkbanken's research platforms. As our general data editing platform, Karp is now successfully used to support the editorial work on two of the most renowned reference dictionaries of present-day Swedish.

References

Ahlberg, Malin, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal & Jonatan Uppström. 2013. Korp and Karp – a bestiary of language resources: The research infrastructure of Språkbanken. *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*. 429–433.

Allén, Sture. 1970. Åtta teser om texthantering [Eight theses about text processing]. *Dagens Nyheter* (1970-09-29).

Allén, Sture. 1973. *Förslag till inrättande av ett organ för lagring och tillhandahållande av datamaskinellt läsbara texter, benämnt* logotek [Proposal for the establishment of a facility for storage and supply of computer-readable texts, designated *logotheque*]. Gothenburg: Computational Linguistics Unit, University of Gothenburg.

Allén, Sture. [1980] 1999a. The language bank concept. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 302–310. (Originally published in Joseph Raben & Gregory Marks (eds.), *Data bases in the humanities and social sciences*, 171–176. Amsterdam: North-Holland). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allén, Sture. [1983] 1999b. En forskningsstrategi för språkvetenskaplig databehandling: Perspektivskiss [A research strategy for computational linguistics: A perspective sketch]. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 222–233. (Originally published as a Språkdata internal research report). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Allén, Sture. [1984] 1999c. Skandinavisk datalingvistik [Computational linguistics in Scandinavia]. In Sture Allén (ed.), *Modersmålet i fäderneslandet*, 168–199. (Originally published in *The Nordic languages and modern linguistics 5*, 11–42. Aarhus: Nordisk Institut, Aarhus Universitet). Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Atkins, B. T. Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Berg, Sture. 2009. Om ordböjning och SAOL Plus [On word inflection and SAOL Plus]. In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 139–165. Stockholm: Norstedts.

Blensenius, Kristian. 2023. Mot en harmonisk lemma-lexemmodell och ordklassuppsättning [Towards a harmonious lemma-lexeme model and word-class set]. *Nordiska studier i lexikografi 16*. 43–54.

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson & Jonatan Uppström. 2013. The lexical editing system of Karp. *Proceedings of the Electronic lexicography in the 21st century (eLex) 2013 conference*. 503–516.

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson & Jonatan Uppström. 2012. The open lexical infrastructure of Språkbanken. *International Conference on Language Resources and Evaluation (LREC) 2012*. 3598–3602.

Borin, Lars, Shafqat Virk & Anju Saxena. 2018. Many a little makes a mickle: Infrastructure component reuse for a massively multilingual linguistic study. *Selected papers from the CLARIN Annual Conference 2017*.

Dannélls, Dana, Lars Borin, Markus Forsberg, Karin Friberg Heppin & Maria Toporowska Gronostaj. 2021. Swedish FrameNet. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 37–65. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Francopoulo, Gil (ed.). 2013. *LMF: Lexical Markup Framework*. London: ISTE/Wiley.

Gellerstam, Martin & Christian Sjögreen. 1994. *Språkbanken: En språklig referensdatabas* [Språkbanken: A linguistic reference database]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Grierson, George A. 1903–1927. *A linguistic survey of India*. Vol. I–XI. Calcutta: Government of India, Central Publication Branch.

Holmer, Louise & Kristian Blensenius. 2023. Okynniga pluraler: Normering och bruk av s-plural speglat i SAOL och SO [Insolent plurals: Standardization and the use of s-plural reflected in SAOL and SO]. *Nordiska studier i lexikografi 16*. 153–164.

Holmer, Louise, Ann Lilliestrom, Emma Sköldberg & Jonatan Uppström. 2024. Time to say goodbye revisited: On the exclusion of headwords from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 443–452.

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 41–106. Amsterdam: John Benjamins.

Nilsson, Pär. 2024. Glancing back, looking forward: The Swedish Academy dictionary completed after 130 years – now time for revision. *Lexicographica* 40(1): 5–28.

SAOB. 1898–2023. *Svenska Akademiens ordbok* [The Swedish Academy Dictionary]. Lund: Gleerups.

SAOL 14. 2015. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 14th edn. Stockholm: Norstedts.

Sjögreen, Christian & Emma Sköldberg. 2010. Svenska ordboksredigeringssystem: Med fokus på Cronoma [Swedish dictionary editing systems: With a focus on Cronoma]. *LexicoNordica* (16): 197–210.

Sköldberg, Emma. 2023. ”Varför står det olika i SAOL och i SO?” Om (bearbetning av) skillnader mellan Svenska Akademiens samtidsordböcker [“Why do SAOL and SO say it differently?” On (the revision of) the differences between the contemporary dictionaries of the Swedish Academy]. *Nordiska studier i lexikografi 16*. 349–361.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

Svensén, Bo. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Svenska.se. 2025. *Svenska Akademiens ordböcker* [The Swedish Academy dictionaries]. [Online resource] https://svenska.se/. Accessed on 2025-01-09.

Part V: **Case studies**

Gerlof Bouma and Emma Sköldberg

12 Dalin revisited: a new digitization of *Ordbok öfver svenska språket*

Abstract: A.F. Dalin's Dictionary of the Swedish language (*Ordbok öfver svenska språket*, 1850–1853) is considered to be the first comprehensive monolingual definition dictionary for Swedish. The dictionary collects ca. 60,000 entries in two volumes. The work's many merits include high coverage in terms of vocabulary and semantic distinctions, as well as the treatment of different kinds of multiword units, such as phrasal verbs, collocations and idioms. This chapter presents some results of our ongoing efforts to create a new digital text version of this important piece of Swedish lexicographical history, motivated by the uneven quality of the existing digitization. We first briefly discuss the digitization itself: the digital form we currently have and what further derived versions we hope to create in the future. We then illustrate the potential of looking at the dictionary as a text, combining a qualitative, lexicological perspective with a more quantitative, corpus-linguistic approach. Along the way, we point out characteristics of Dalin's dictionary that potentially create problems when going from the linear, paper-based format to a database-style of access.

Keywords: dictionary, digitization, historical lexicography, lexical database

1 Introduction

Anders Fredrik Dalin's *Ordbok öfver svenska språket* (Dalin 1850–1853) is considered the first comprehensive monolingual Swedish definition dictionary. Several studies attest to the pioneering scope and high quality of Dalin's work. It is, for several reasons, valuable to examine older lexicographic works such as Dalin's dictionary. They

Acknowledgments: The redigitization of A.F. Dalin's Dictionary of the Swedish language was funded by *Meijerbergs institut för svensk etymologisk forskning*. The work on this chapter was carried out at Språkbanken Text and supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161).

Gerlof Bouma, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: gerlof.bouma@svenska.gu.se
Emma Sköldberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: emma.skoldberg@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-012

can serve as a source for language historical studies and studies of older dictionaries can provide deeper insights into prevailing traditions in modern lexicography. In addition, investigations of the pioneers of lexicography can inspire modern lexicographers (Hannesdóttir & Ralph 1988; Ralph 1992). Important contributions to the knowledge of older dictionaries have been made within the project *Lexikografisk tradition i Sverige* 'Lexicographical Tradition in Sweden', which started in the mid-1980s at the Department of Swedish, University of Gothenburg (see further, e.g. Ralph 1992; 2001). Examples of both doctoral dissertations and shorter papers related to this project are Malmgren (1988), Hannesdóttir (1991; 1998), Johansson (1997), and Rogström (1998). In these studies, the researchers have, among other things, examined the history of older dictionaries, their vocabulary, the structure of the dictionary entries, the authors' descriptions of meaning and whether it is possible to discern dependencies between different lexicographic works.

The previous studies of Dalin's work are based on printed versions of the dictionary or on digitized, electronic versions. Work using the printed version is, also for practical reasons, mostly of a qualitative and detailed nature. Quantitative and/or broader statements will be restricted to generalizations from sampled pages or entries. An electronic version opens venues to additional types of investigation of this material, by facilitating comprehensive searching of the dictionary and counting of items of interest in it. The difference is not that these things are possible in an electronic text – one could also search and count things in a paper dictionary – it is that they are easy and quick enough to support exploratory and iterative methods on the scale of the whole text.

Electronically, Dalin's dictionary has existed as a manually transcribed version which has been available at Språkbanken Text for two decades, and as several versions automatically transcribed using optical character recognition (OCR). Although all these versions are very valuable as they are, they unfortunately also show major flaws and transcription errors. As a result, they require careful handling and workarounds. For example, Perby (2010), using Dalin's dictionary, recommends free-text searching multiple OCR versions, to reduce the chance that transcription mistakes cause the researcher to miss relevant matches.

To address this situation, we have re-digitized *Ordbok öfver svenska språket*, resulting in an electronic version of the dictionary that is true to the original. In this chapter, we present this project. We will outline the digitization process and then go on to show some of the breadth of investigations one can conduct on the electronic version, by looking at the material with a lexicological/lexicographical, qualitative eye as well as from a quantitative, corpus-linguistic perspective. Despite their limited scope, these investigations are not just intended as a show case, but also to contribute to what we know about the dictionary.

Our re-digitization of Dalin's dictionary has thus far resulted in an electronic transcription that also contains information about the dictionary's layout in the form of XML annotations. We are now working on adding further annotation such as entry structure and information about cross-references between entries. The dictionary will also be prepared for inclusion in Språkbanken Text's lexical infrastructure Karp (see Chapter 11), to allow database-style access. In the investigations presented below, we will point out some of the lexicographic devices Dalin uses that are problematic in this type of non-linear use of the dictionary.

The rest of the chapter is structured as follows: Section 2 discusses Dalin's dictionary and its relevance in the history of Swedish lexicography. The digitization process is described in Section 3. In Section 4, we present a selection of investigations performed on the new electronic version. Section 5 concludes the chapter.

2 Dalin's dictionary in context

Anders Fredrik Dalin (1806–1873) compiled several bilingual dictionaries, especially between French and Swedish (see, e.g. Hannesdóttir 1991; 1998), but the work that is in focus here is his monolingual dictionary of Swedish published in the 1850s. As already mentioned, this dictionary is considered the first complete monolingual Swedish definition dictionary (Malmgren 1988). It marks a decisive step forward in the lexical description of the Swedish language. We refer to Hannesdóttir (1998: 462–493) for a detailed review and discussion of Dalin's lexicographic efforts.

The dictionary comprises two volumes and contains, according to an estimate by Holm & Jonsson (1990), approximately 60,000 headwords, not considering compounds. The headwords are mainly taken from general spoken and written language, but the dictionary also includes some technical terms and dialectal words. A large number of headwords are loanwords (Holm & Jonsson 1990: 1937).

In the preface to the dictionary, Dalin gives a presentation of the work and explains its purpose in detail. He points out that it is intended for the Swedish people and not for scholars. Dalin's intentions are clearly normative in such a way that he would like to "introduce consistency and order in the orthography"[1] (Dalin 1850: 8). Furthermore, he wants to bring clarity to the senses of the general words and suggest the correct way of their use, or, as Dalin writes, to "include all one needs to know to write correctly and in a cultivated manner"[2] (Dalin 1850: 8). The large

1 Sw.: "införa stadga och reda i rättskrifningen" – our translation.

2 Sw.: "upptaga allt det hufvudsakliga, som i och för ett språkriktigt och vårdadt skrifsätt kan vara af nöden att veta" – our translation.

number of synonyms listed can benefit "beginning writers or those who do feel they lack experience"[3] (Dalin 1850: 18). The dictionary is thus largely intended to support writing.

When it comes to the question of the extent to which Dalin used existing sources, Hannesdóttir (1998: 475) characterises the relationship between his monolingual dictionary and other monolingual and bilingual dictionaries as "intricate, to say the least". Malmgren (1988) shows that Dalin used the monolingual dictionary fragment by C.J.L Almqvist (1842–1844). Hannesdóttir (1998: 471–475) adds that clear traces of the dictionary fragment by E.F. Kindblad (1840) are found among the early entries in Dalin's dictionary. Both dictionary fragments are mentioned by Dalin in his preface. Furthermore, Dalin has been able to rely on his own extensive French–Swedish dictionary, published in different versions during the first half of the 1840s.

Ordbok öfver svenska språket has many merits. For example, its coverage in terms of headwords and senses is very good (Malmgren 1988: 202–203). Norén (1991: 137) states that Dalin, in terms of semantic analysis, has made a pioneering effort. Holm & Jonsson (1990: 1937) write: "Circular definitions are almost entirely absent. [Dalin] defines the words briefly but accurately, showing an astute instinct for shades of meaning in his semantic subdivisions". Malmgren (1988: 205) calls Dalin's sense division surprisingly consistent.

Dalin's treatment of different kinds of word combinations should also be mentioned. According to Malmgren (2008: 154), the treatment of collocations in the dictionary is "admirable in many ways". In addition, Sköldberg (2009) shows that Dalin's semantic analyses of idioms and proverbs are advanced: many expressions in his dictionary are illustrated by editorial language examples and the dictionary includes interesting cross-references between multiword expressions and simple words with similar meanings.

There are of course also weaknesses in Dalin's dictionary. Etymological information, to the extent that it is given, is frequently incorrect and sometimes has a folk-etymological character (Malmgren 1988: 211). Norén (1991: 137) comments upon the lack of precision in the grammatical analysis, especially regarding the classification of verbs. Chapter 13, this volume, also critically remarks upon the morphological analysis in Dalin's dictionary.

Although Dalin's monolingual dictionary overall is held in high regard now, it does not seem to have been honored with more qualified contemporary reviews (Malmgren 1988: 210). Dalin was, however, awarded a prize from the Swedish Academy for his work with the dictionary (see also Hannesdóttir 1998: 465–466).

3 Sw.: "nybegynnare i språkets skrifvande, eller dem, som deri äro mindre hemmastadde" – our translation.

3 The source material and its digitization

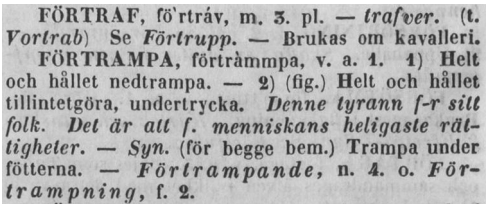
Dalin's dictionary was published in two volumes: A–K, 896 pages and L–Ö, 772 pages. In addition to the dictionary entries themselves, Volume 1 includes an introduction by the author, a list of abbreviations and a short list of errata. Volume 2 also contains ten pages of additions to the dictionary, as well as a page with errata for both volumes. At the end of Volume 2 is a list of subscribers, who purchased the dictionary at a reduced price as it was published in 34 installments between 1850 and 1855.[4]

As the primary source for our re-digitization of the dictionary, we used a copy from the Gothenburg University Library, which was available in the form of high resolution photographs. To resolve problems due to damage to the copy or artifacts of the scanning process, we consulted further copies of the book available at the Department of Swedish, Multilingualism, Language Technology.

We outsourced the conversion of photographed pages to a digital text-based format to a commercial service that uses double keying for the process. The text was extracted from the images together with information about text layout, encoded in an ad hoc XML format. The layout and typography inside the dictionary entries is relatively simple. The text is primarily set in a Didone roman type. For purposes of highlighting and indicating entry structure, the dictionary uses a combination of punctuation marks, upper case text, italic type and letter-spaced italic type. Weight is not used to highlight inside the entries, although the upper case of the used type is relatively heavy, giving the impression of bold face for upper cased words. Occasionally, comments occur inside entries, in a slightly smaller size and offset by an indented margin. The XML used to encode the text and layout is kept very light. However, it is enough to create an approximate, re-rendered digital version of the work, with the help of CSS styling. Figure 1 shows an example entry, first photographed from the original, then in XML format and finally re-rendered. And, as demonstrated in the case studies below, the XML also contains enough information to look quantitatively at a range of aspects.

An alternative route at this first stage would have been to let the double keying service encode not just the layout but also information about the structure of entries. As the process itself is carried out by people who do not understand Swedish and who are not familiar with the domain, we ruled out marking up structure on the basis of a description of the structure marking conventions in the dictionary alone as viable alternative.

4 Traditionally, the years 1850 and 1853 are given as publication dates for Dalin's dictionary, based on the publication years of the first installment of each volume. The last installment of the second volume appeared in 1855, however.



FÖRTRAF, fö`rtráv, m. 3. pl. — *trafver.* (t. *Vortrab*) Se *Förtrupp.* — Brukas om kavalleri.

FÖRTRAMPA, förtràmmpa, v. a. 1. 1) Helt och hållet nedtrampa. — 2) (fig.) Helt och hållet tillintetgöra, undertrycka. *Denne tyrann f-r sitt folk. Det är att f. menniskans heligaste rättigheter.* — *Syn.* (för begge bem.) Trampa under fötterna. — *Förtrampande,* n. 4. o. *Förtrampning,* f. 2.

```
<indent/>FÖRTRAF, fö`rtráv, m. 3. pl. – <i>trafver.</i> (t.<lb/>
<i>Vortrab</i>) Se <i>Förtrupp.</i> – Brukas om kavalleri.<lb/>
<indent/>FÖRTRAMPA, förtràmmpa, v. a. 1. 1) Helt<lb/>
och hållet nedtrampa. – 2) (fig.) Helt och hållet<lb/>
tillintetgöra, undertrycka. <i>Denne tyrann f-r sitt<lb/>
folk. Det är att f. menniskans heligaste rät-<lb/>
tigheter. – Syn.</i> (för begge bem.) Trampa under<lb/>
fötterna. – <lsi>Förtrampande,</lsi> n. 4. o. <lsi>För-<lb/>
trampning,</lsi> f. 2.<lb/>
```

Figure 1: The entries for *förtraf* 'vanguard' and *förtrampa* 'trample' as in the original, in XML, and in re-typeset form

Inspection of a random selection of pages showed that the digital text as received from the double keying service was nearly faultless, but it nevertheless contained a few issues that needed addressing before we could move forward with processing the dictionary. The copy available for text digitization was damaged in certain places and in other places the photographs showed distortion. This led to illegible passages, which were corrected with the help of further physical copies of the dictionary available to us. For some uncommon symbols (for instance runes) the appropriate Unicode character point had to be found. We also performed a systematic check of a number of error-prone cases, for instance distinguishing oe- from ae-ligatures in cursive text (*œ* vs *æ*; see Bäckerud 2014 for similar remarks). Even in the physical copy, these two ligatures can be hard to distinguish. We ended up correcting these according to the interpretation of the whole word. This is the only case where we applied "should be" corrections, in all other cases we have strived to exactly represent the content of the printed dictionary, even where this includes obvious misprints. For three symbols no appropriate Unicode character was available. These are handled by placeholders in the form of XML entities. Two of these concern rotated regular letters, which are meant to illustrate a specific shape. An XML entity is also used to handle the single instance of a figure early on in the dictionary, in the second installment, in the entry for *alfkors* 'pentagram' (lit. 'elf cross'), shown in Figure 2.

The layout-oriented XML is the source for further derived versions of the dictionary. In particular, work is ongoing on a version suitable for inclusion in Språkbanken Text's Karp infrastructure, which offers access to dictionaries as lexical databases (see Chapter 11, this volume).

ALFABETISK, allfabä'tissk, a. 2. Gående i bokstafsföljd. *A. ordning. A-t register.*

ALFKORS, ållvkårrs, n. 5. Figur, äfven kallad Druidenfot eller Trollfot, hvilken i fordna tider ansågs som ett medel emot trollskap.

AL FRESCO (italienska). *Måla a. f.*, d. v. s. med vattenfärg på våt kalk.

Figure 2: The entries for *alfabetisk* 'alphabetical', *alfkors* 'pentagram', and *al fresco*

4 Dalin's dictionary up close and in numbers

This section presents five small investigations, made on the basis of the newly digitized, layout-oriented electronic version. As our corpus, we consider all entries in the body of the dictionary, that is, we disregard the preface, section headings, the additions, the errata and the list of subscribers. The material consists of 63,487 entries,[5] for a total of 1.48 million alphanumerical tokens. The entries are divided into 29 sections, one for each of the initial letters *A*–*Z* plus *Å*, *Ä*, and *Ö*. Like the latest editions of the Swedish dictionary SAOL, starting from SAOL 13 (2006), but *unlike* the present-day Swedish orthographic tradition before that, Dalin has a separate section for *W*. Words with initial *Æ* och *É* are sorted as *AE* and *E*, respectively.

4.1 Headword references

As mentioned, there are a good 63 thousand entries in the dictionary. This number does not correspond directly to the number of headwords, however, nor to the number of full-fledged articles. The relation is muddled by the presence of homography (see Section 4.2), entries with multiple headwords (Section 4.4) and entries that just consist of references to other entries, which is the topic of this subsection.

There are 1,780 entries, that is, approximately 3%, which directly refer to another entry using the word *se* 'see', without giving word class information, as in:

(1) *TIMA, se Timme.*
'*TIMA*, see *Timme* [hour].'

We refer to Norén (1991: 136–137) for an overview of the different ways in which entries may refer to other entries.

5 This means that Holm & Jonsson's (1990) estimate of 60,000 headwords is quite accurate.

Cases like (1) typically concern minor variations in spelling, for instance *als* refers to *all**l**s* ‘at all’. But they may also concern allomorphy, for instance *förspråkare* refers to *för**e**språkare* ‘referee, advocate’; or more generally hints on where to find an article, for instance the combination of a reflexive verb with postposed particle *gadda sig tillsamman*, lit. ‘sting themselves together’, refers the reader to its prefixed particle variant *sammangadda sig* ‘conspire [against smbd], revolt’ for its definition.

References typically only give the referenced headword, although in some cases specific homographs or senses may be referenced (*oskylld* refers to the homograph *oskyld* II ‘unrelated’; *banque* refers to the sense *bank* B. 1 ‘financial bank’). References may also be bunched together into one entry (*älf, älg* refers to *elf, elg* ‘river, elk’), sometimes with the addition of *m. fl.* ‘etc.’ either in just the body or in both head and body of the entry (*exellens, exellent, exellera* ‘excellence, excellent, excell’ refers to *ex**c**ellens m. fl.*; *tjäna, tjänst, m. fl.* refers to *tjena, tjenst, m. fl.* ‘serve, service, etc.’) Although economical and hardly problematic for a reader, this use of “etcetera” poses a challenge for our further computational processing, as we do not directly know which headwords are covered by its use. If we do manage to find out this information, the use of “etcetera” furthermore interferes with a hypertext model where one would click on a referenced headword to be taken to the relevant entry. Since one “etcetera” may comprise multiple references, this single hyperlink model breaks down.

The directly referring entries are distributed fairly evenly accross the sections of the alphabetic macrostructure. Out of the 29 sections, 19 contain between 2% and 4% direct references. A notably high proportion can be found in sections *C* (212/858 ≈ 25%), *W* (6/13 ≈ 46%) and *Z* (8/56 ≈ 14%). These are “foreign” (initial) letters, indicative of loans, for which Dalin supplied preferred Swedified spellings. Headwords of referring entries beginning with *c* point at spellings with initial *k*, and to a much lesser degree with initial *s*. In many cases Dalin merely comments “see under K”. Section *C* starts with a general comment saying that “words missing under C are to be found under K or S”.[6] All of the references in Section *W* point to entries with initial *v*. The majority of references with initial *z* point to entries with initial *s*. These findings are in line with previous observations that Dalin includes many loan words but at the same time wishes to increase the consistency in Swedish orthography (see Section 2).

Only a modest number of references, about 50 or 3% of them, are problematic in that the given target is not exactly found in the list of headwords. In some cases these are simple examples of Dalin referring to a spelling variant he did not end up using for the headword (for instance, *kalfleka* refers to *kabbelleka*, but the headword

6 Sw.: “De ord, som saknas under C, återfinnas under K eller S” – our translation.

is spelled *kabbe**l**eka*; likewise *penguin* refers to *pinguin*, but given is *ping**v**in*). Other cases concern reference to an inflected form that does not match the headword form (*Necken* refers to *Näck-en* [nix-DEF], but given is uninflected *näck* [nix.INDF]) or reference to an allomorphic variant of the headword (*kungsåder* refers to *kongsåder*, whereas *kongsåd**ra*** 'main inflow stream of a lake' is given). For a person consulting the dictionary in a form that preserves its linear order, such imprecision is unproblematic, since the intended target will be quickly found in the vicinity of where the missing target would be. In a setting where the dictionary is consulted as a database, this does cause problems, however, and needs to be handled in some way, for instance by correcting hyperlinks to point at the intended rather than the specified entries. Some references are in fact harder to find. For instance, *gitarr* 'guitar' (thought to be missing by Norén 1991) does not appear in the dictionary itself but in the list of additions. The compound *storm-skratta*, lit. 'storm-laugh' V, refers to *stor-skratta* 'roar', lit. 'big-laugh', but the latter is only found inside the entry for *stor* 'big' Adj, and not as a headword of its own. Only a handful of references appear to fail completely, referring to information actually missing from the dictionary (for instance, *ärmknapp* 'cuff button', and *lånke* [a kind of plant]). This high quality of cross-referencing is impressive given the size and timespan of Dalin's enterprise.

4.2 Homography

The dictionary contains 1,222 homographic headwords, that is, headwords that occur in multiple entries. We will refer to a group of entries that share a headword as a *homography cluster*. The overwhelming majority of homography clusters (more than 1,000 cases) contain two entries. Examples of homography clusters of size two include *spricka* 'crack' (V) or 'crack' (N), and *bar* 'sandbar' (N) or 'bare' (Adj). A total of 2,668 entries are members of a homography cluster. The largest cluster contains eight entries, which is for the headword *rå*: 'raw' (Adj), 'burr' (N), 'deer' (N), 'biscuit' (N), 'stake, pole' (N), 'yard [on a sailing vessel]' (N), 'advise, control' (V), 'nymph' (N).

A mix of strategies is used to indicate homography. The explicit strategy uses roman numerals to index the different members of a cluster. This strategy is used in 947 entries, for instance in the entries for the cluster *bryna* (boldface ours):

(2) *BRYNA, v. a. 1. o. 2.* ***I.*** *(af* Brun*) 1) Göra eller förorsaka, att något blir brunt, antingen genom solens inverkan eller genom stekning, rostning.* B. i solen. B. en stek. B. smör, ost. — B. sig, *v. r. Blifva brun.*
BRYNA, v. a. 1. o. 2. ***II.*** *(af* Bryne*) Göra hvass medelst bryne.* B. en knif.
'BROWN, verb [...]. I. To make brown [...]. [...]
HONE, verb [...]. II. To sharpen with a honing steel. [...]'

The unindexed cases form the majority. They may have distinguishing traits such as word class (*bravo!/bravo* 'bravo!' Interj or 'contract killer' N) or pronunciation (*kredit* [accent on the first syllable] 'balance in favour' or [accent on the second] 'esteem'; *skyld* [long vowel] 'hidden' or [short vowel] 'related').

We have not managed to uncover a clear pattern in Dalin's use of homograph indices. It is tempting to think indices are used when there are not enough other formal distinguishing traits, but there are many exceptions to this generalization. For instance, the four entry cluster *bräcka* contains two feminine nouns: 'crack' and 'a type of hard mineral'; and two transitive verbs: 'break' and 'fry'. For the verbal entries, indices are used, but not for the nominal ones. Likewise, for *fasta* 'fast' (V), 'fasting, Lent' (N), 'legal title' (N), the two nominal entries are indexed, but the verbal one is not.[7] Furthermore, indices do not only occur on homographs of the same part of speech. For instance, the eight entries of *rå* mentioned above cover several parts of speech and are all indexed. A more straightforward example of indexed homographs with different parts of speech is found on *äga* I 'own' (V) and *äga* II 'property' (N).

Homograph indexing seems to have been an innovation for Dalin: he only starts to use it in the third installment, when the dictionary work is already underway. The first homograph indices are found in the entries for *befara*, even though there are 57 homograph clusters preceding them in the *A* and *B* sections, and one might expect around 20 of them to be indexed if we were to reason from the overall average. Starting with *befara* we have the sequence of the following four entries (abbreviated for reasons of space, our boldface):

(3) *BEFARA, v. a. 1.* ***I.*** *Motse faran, olyckan af något. […]*
BEFARA, v. a. 3. ***II.*** *[…] Fara, resa öfver ell. igenom. […]*
BEFARANDE, n. 4. […] En faras eller olyckas motseende till följe af något. […] ***Jfr*****: Befara, *I.***
BEFARANDE, n. 4. Handlingen, då man befar […]. ***Jfr*****: Befara, *II.***
'FEAR, verb […] I. Apprehend the danger, misfortune of something. […]
BE-TRAVEL, verb […] II. Travel over or through. […]
FEARING, noun […]. Apprehension of danger […]. Cf. *Befara* I.
BE-TRAVELLING, noun […]. The act of travelling over […]. Cf. *Befara* II.'

In the latter two entries, indices are used in the references to the two verbal entries to disambiguate the targets.

7 The dictionary also contains a few obvious mistakes, such as the cluster *ruter* 'diamonds [in cards]' or 'enthusiasm', in which the first entry is indexed but the second is not. It is an open question whether an electronic version should annotate the second entry as being indexed as well.

As discussed in Section 2, there is overlap with both Almqvist's and Kindblad's dictionary fragments in these early parts of Dalin's dictionary. Neither of these other two dictionary authors uses homograph indexing, and Dalin does not employ this solution in his bilingual dictionaries, either. Interestingly, looking at Almqvist's fragment, we do find very similar entries for *befara* and *befarande*, albeit with the nominalized participles collapsed into one entry (Almqvist 1842–1844: 263). As in Dalin's dictionary, Almqvist refers to the two verbal entries from the entry for the participle, but to distinguish the two, he writes 'the former *befara*' and 'the latter *befara*'. It is very likely that Dalin used Almqvist's entries for his dictionary, but we speculate that this "crossing dependency" is what prompted Dalin to introduce a more precise means of referring to homographic entries. After their introduction, the proportion of indexed homograph clusters rises somewhat towards the later parts of the dictionary.

4.3 Entries that rely on linear structure of the dictionary

Our speculation that Dalin introduced homograph indexing to avoid having to rely on the order of entries in cross-references notwithstanding, there are numerous instances where Dalin actively exploits the linear presentation of entries in the dictionary. An example of this is when one entry continues into the next, as with the definition for *krikon* 'damson'. The final colon in the first entry signifies that the following headword is to be read as part of its definition:

(4) *KRIKON, n. 5. Den lilla, mörkt rödblå, sura frukten af:*
KRIKONTRÄD, n. 5. Mindre fruktträd, liknande plommonträdet. Prunus insititia.
'DAMSON, […]. The small, dark red-blue, tart fruit of:
DAMSON TREE, […]. Small fruit tree, similar to the plum tree. […].'

Such cases are relatively rare, we counted only 47 in the dictionary. Figure 3a shows where in the dictionary these "run-in entries" can be found. They are fairly evenly distributed over the dictionary, with a small peak in the section for *C*, and a relative lack of instances after Section *R*.

A somewhat more common convention that uses the dictionary's linear presentation is found in the following example. The first entry, for *begrafningsskjorta*, lit. 'funeral shirt', is just a header. In the place of a definition is a conjunction (here marked in bold), which signals that the definition can be found in the directly following entry:

(5) *BEGRAFNINGSSKJORTA, f. 1.* ***och***
BEGRAFNINGSSKRUD, m. 2. Se Svepning.
'FUNERAL SHIRT, [...] and
FUNERAL ATTIRE, [...]. See shroud. '

There are 331 cases of this in the dictionary, distributed as shown in Figure 3b. This way of connecting adjacent entries is also found throughout the dictionary, but it is slightly more common in sections *A* and *B* and from section *N* onwards. From our current, cursory look at this distribution, we have not been able to identify possible reasons for this temporary dip in the usage of this device.

Both conventions create problems for any presentation of the entries in isolation, as for instance is common in a lexical database. The situation in which the database user is presented with an incomplete entry, especially if there is no way of quickly navigating to the next entry[8] should of course be avoided. As a workaround, in our database version of the Dalin's dictionary, we combine entries when needed, so that the reader will have everything in one place. So, for instance, the entry for *begravningsskjorta* will include the text for the entry for *begrafningsskrud*.

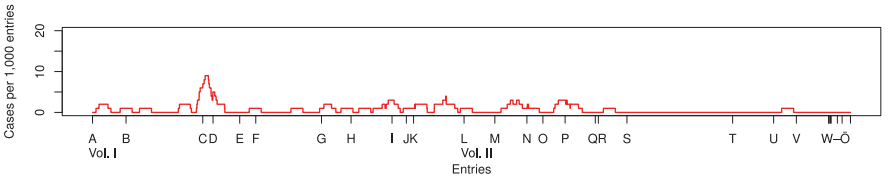
Further reliance on linear order in the dictionary is found inside definitions, by comments like *se föreg[ående]* 'see preceding' or *se följ[ande]* 'see following', as in the two examples below:

(6) *FRAMVIGT, c. 3. Öfvervägande vigt eller tyngd på framdelen af ett föremål.*
FRAMVIGTIG, a. 2. Som har framvigt; se föreg.
'FRONT WEIGHT, [...]. Overhanging weight or pressure on the front part of an object.
FRONT HEAVY, [...]. That which has front weight; see preceding.'

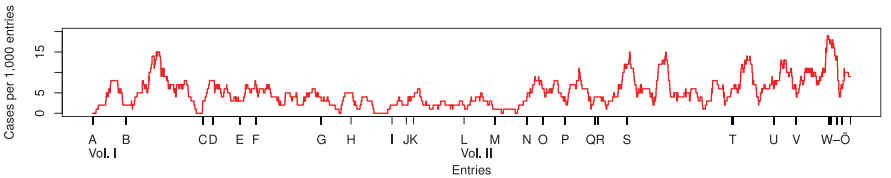
(7) *FRIHERREBREF, n. 5. Se följ.*
FRIHERREDIPLOM, - - - - plå´m, n. 3. Diplom, hvarigenom någon blifvit upphöjd till värdighet af friherre.
'BARON LETTER, [...]. See following.
BARON CHARTER, [...]. Charter by which someone was elevated to the status of baron.'

We do not yet have a clear picture of how common this is in the dictionary, since these references vary in form more than the other types of references, and are thus harder to identify on the basis of the layout-oriented digital text version we are

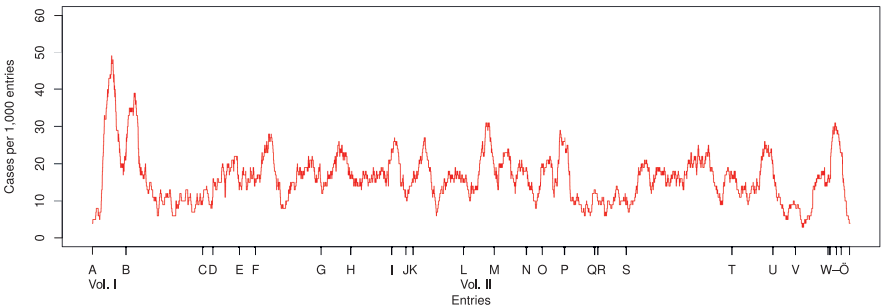
8 A consistent, single notion of a "next entry" is not always available in a lexical database setting, for instance when the database combines material from different sources. Should the next entry be the one that is alphabetically next in the database, next in whichever (ordered) view of the database the user has chosen, or next in a printed original of the source of the current entry, etc.?



Subfigure a: Entries whose definition ends with a colon and includes the next entry's headword



Subfigure b: Entries with their own header that rely on the following entry for their definition



Subfigure c: Entries that contain multiple headwords in the header

Figure 3: Distribution of different entry types over the dictionary. The x-axis lays out all the entries in the order they appear in the dictionary. The y-axis gives the number of entries of the type of interest in a window of 1,000 entries around the point at the x-axis. The x-axis is labelled with section names.

performing these investigations on. In a lexical database, these could be handled in a similar fashion as described above, by including the text of preceding/subsequent entries. Alternatively, if the database allows entries to contain hyperlinks, the text 'see following' could be made to point at the relevant entry, just as one would do with explicit references to headwords.

4.4 Entries with multiple headwords

We have thus far seen several devices that overlap in their application area: referring entries that use headwords, header-only entries, and references to following or

preceding entries. These devices can all be used to relate alternative forms to one and the same definition. As mentioned above and shown in the examples, these form alternatives can be a matter of mere spelling variants, of allomorphic variants, or even of compounds with (nearly) synonymous heads. However, Dalin's most used solution for this by far is to give multiple headwords for one entry. Some examples of this are:

(8) *ABCGOSSE, ABCPOJKE, m. 2. pl.* — gossar, — pojkar. *Gosse, som håller på att lära abc.*
'ABC-LAD, ABC-BOY, […]. A young lad learning the alphabet.'

(9) *ALLTFÖR ell. ALLT FÖR, adv. 1) För mycket, nog mycket, nog.* En a. tidig död. — *2) Ganska mycket, i hög grad.* A. söt, vacker. Ni är a. god. *[Altför.]*
'ALL-TOO or ALL TOO, […]. 1) Too, […] — 2) Very, […].'

(10) *AMORTERINGS- eller AMORTISSEMENTSKASSA, - - - - màngs - -, f. 1. Kassa, bestämd för amorteringen af vissa skulder.*
'AMORTIZATION or AMORTIZEMENT FUNDS, […]. Funds reserved for the amortization of certain debts.'

(11) *BAKSTUDS, m. 2. BAKSTUDSNING, f. 2. Studsning tillbaka i mer eller mindre horisontel riktning, t. ex. ifrån en vägg, o. s. v.*
'BACK-BOUNCE, masc. BACK-BOUNCING, fem. Bouncing back in more or less horizontal direction, for instance from a wall, etcetera.'

There are 1,319 entries with multiple headwords. As can be seen in the examples, these include cases that go beyond spelling variation or allomorphy, and may even include headwords with different grammatical properties. The distribution of entries with multiple headwords is given in Figure 3c.

From the fact that their uses overlap, one might expect to see some kind of complementarity in the distribution of head-only entries and multiple headword entries. For instance, perhaps the low incidence of the former between sections *C* and *N* aligns with a rise in the use of the latter. However, comparing Figures 3b and c, we do not find any clear signs of such complementarity. What does stand out in the distribution of entries with multiple headwords is their high frequency in Section *A* and the beginning of Section *B*. Because of the location of the peak in the beginning of the dictionary, a comparison with Almqvist's and Kindblad's dictionaries is relevant here, too. Almqvist does have entries with multiple headwords, but he uses this much more sparingly, and primarily for headwords that are very close variants of each other. For more loosely related forms, Almqvist prefers to give separate entries, of which one refers to the other. Some of these cases reappear in Dalin's dictionary as entries with multiple headwords (for instance, *aftvå*, *aftvaga* 'wash off'). Kindblad does not have any entries with multiple headwords.

4.5 Meaning structure

Until now, the focus of our investigations has been on relations between entries, that is, we have mainly focused on macrostructural characteristics of Dalin's dictionary. In our final case study, we take a closer look at the entries themselves, in particular at their meaning structure and the way Dalin indicates this structure.

Different senses of a headword can be ordered linearly – with a number of discrete senses arranged in a sequence – or hierarchically – with main senses divided into subsenses (Svensén 2009: 211). Dalin allows for hierarchical structure, here illustrated in the entry for *sur* 'sour', for which Dalin gives three main senses 'sour taste', 'bad, wet', and a third figurative sense divided into two subsenses 'difficult' and 'peevish'. The main senses are marked by arabic numbers, the subsenses by latin letters (our boldface):

(12) *SUR, a. 2.* ***1)*** *Säges om ett ämne, som i mer eller mindre mån innehåller någon syra och till följe deraf har en egendomlig stickande smak, i allmänhet ansedd som motsatsen till Söt.* S-a äplen. S-t vin. S. mjölk. S-t bröd. *(Ordspr.)* De äro s-a, sade räfven om rönnbären, *säges, då någon låtsar sig förakta det, han gerna skulle vilja äga, men ej kan få. —* ***2)*** *Betecknar åtskilliga dåliga egenskaper, t. ex.:* S. ved, *ej torr.* S. tobakspipa, *full af tobaksolja.* S-a ben, *såriga.* S-a ögon, *rinnande.* S. och våt, *genomvåt. —* ***3)*** *(fig. fam.)* **a)** *Svår, vedervärdig.* Göra en lifvet s-t. — **b)** *Tvär. ovänlig.* S. min. S-t ansigte. Göra s-a miner, *som utvisa missnöje.* Ge en s-a miner.
'SOUR, adj. […] 1) Said of acidic substance and with pungent taste […] 2) Describes several bad qualities, for instance: sour firewood, not dry […] 3) figurative, colloquial a) Difficult, disagreeable […] b) Peevish, unfriendly […].'

Although he in principle can use hierarchical ordering, Dalin often makes main senses out of figurative uses of headwords, which is a linear ordering strategy. The third sense of *sur* is an example of this. In contrast, the present-day Swedish dictionary *Svensk ordbok utgiven av Svenska Akademien* (SO 2021, henceforth SO; also see Chapter 4 of this volume) by default attaches figurative uses as subsenses to main senses. So, where Dalin has two senses for *orm* 'snake': one for the reptile and one to describe a sneaky person, SO has one main sense for the animal, and a subsense for the character description.[9] A further difference with the hierarchical strategy in SO, is that Dalin does not explicitly define senses that are divided into subsenses:

9 The structure for *orm* 'snake' in SO, consisting of one main sense with one subsense is not one that occurs in Dalin, at least not when it comes to formally marked structure: Dalin only has numbered senses when there are multiple senses, and likewise only has numbered subsenses when there are multiple subsenses for one main sense (and therefore multiple main senses, too). However, in his

Table 1: Markers of labelled divisions at different levels inside entries

Level	Examples	Number of entries
subentry	A) — B)	1
	A. — B.	8
	I. — II.	1
sense	1) — 2) — 3)	11,616
subsense	a) — b) — c)	1,268
subsubsense	α) — β) — γ)	28
subsub- / subsubsubsense	aa) — bb)	2

descriptions are only provided for the subsenses and the superordinate main sense is left implicit. A comment on style/genre, domain or constructional restrictions ('figurative', 'only in the transitive', 'ship building', etcetera) is, however, frequently attached to such main senses, like in *sour* 3 above.

We have already seen two formal means used by Dalin to mark divisions inside entries and number them: arabic numerals and latin letters, in both cases followed by a closing parenthesis and sometimes preceded by an m-dash. The full range of formal division markers is given in Table 1, together with their frequency of occurrence. The subentries labelled with roman numerals are found in the entries for *ljus* 'light', of which there are two: one adjectival and one nominal. The nominal entry is further divided into subentries 'that which makes objects visible' and 'candle', numbered with roman numerals. Another example of a subentry is found in *parti* 'party', where the different subentries correspond to different etymological origins (see Norén 1991 and Hannesdóttir 1998 for discussion). The division into (main) senses is the by far the most frequent division in the dictionary and occurs in roughly 1 out of 6 entries.

Apart from these formally explicit and numbered divisions, Dalin also regularly further divides entries on the basis of specific inflections of the headword, valency frames or constructions using the headword. Such divisions may occur at any level of the formally labelled divisions. We will not consider these any further at this point, since recognizing them reliably requires more annotation than we have available in the current electronic version.

Figure 4 plots the relation between entry length and structural complexity. The former is expressed in terms of alphanumerical tokens (that is, ignoring punctuation),

preface, Dalin writes he aims to desribe "distinct senses, numbered and ordered from the original and the directly derived to the more distant, figurative and particular" (Sw.: "särskilta bemärkelser, uppställda i en viss med siffror betecknad ordning, från den ursprungliga och de närmast derifrån afledda till de aflägsnare, figurliga och mer enskilta" – our translation). The information in the hierarchical structure of SO can thus *in principle* be recognized in the linear structure of Dalin.

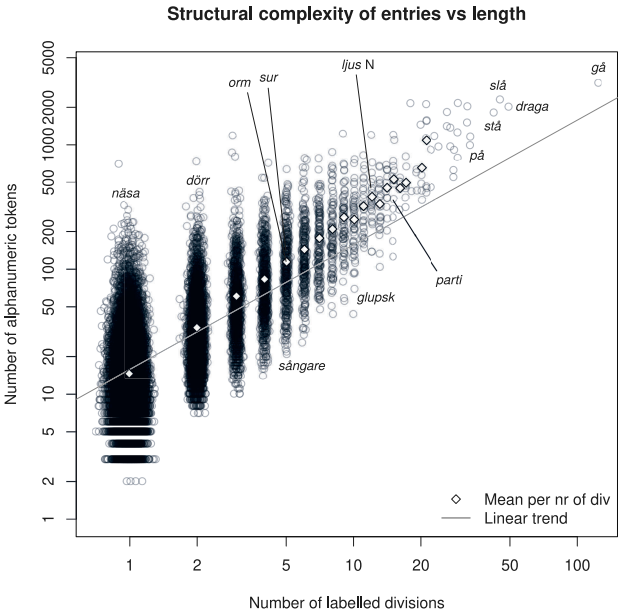


Figure 4: A subdivision in a structurally complex entry is on average *longer* than a subdivision in a simpler entry. The gray line indicates the expected entry lengths if subdivision length did not correlate with entry complexity. The diamonds show that the observed averages grow faster than this expected linear trend, at least initially. (The graph is on a log-log scale, so that differences on the lower end and differences on the higher end of the scales can be shown together. Some noise has been added to the x-positions of the dots to show the cardinality of each level of complexity. The entries discussed in Section 4.5 are labelled with their headwords.)

the latter as the number of divisions, of any level, marked by the devices listed in Table 1. Entries with no such division markers are counted as having complexity 1. We can see there is a wide range of entry lengths, ranging from just a few words (these are the referring entries discussed in Section 4.1) to several thousand tokens, that is, up to 2 full pages in the printed dictionary. We also see a wide range of complexity levels: the majority is of complexity 1 (no subdivisions), but at the extreme end we find entries with 40–50 divisions like *stå* 'stand', *slå* 'strike', and *draga* 'pull'; and even one with 123 divisions: *gå* 'go', which also is the longest entry at 4,731 tokens. The highest number of main senses is 25, for *på* 'on'. Given the markers in Table 1, the deepest possible hierarchy is of depth 5, but this is not found. Six entries are of depth 4.

There is a clear and unsurprising correlation between entry length and complexity: entries with more subdivisions are longer. But in addition, Figure 4 shows that this growth is supralinear: the subdivisions themselves are longer on average in entries with more subdivisions. We find this remarkable, and we have not been able to locate the source of this effect. The opposite – structurally more complex entries have shorter subdivisions – would be more expected, as an instance of Menzerath's law (Altmann 1980). One possible explanation is that this is a mere artifact, a result of the fact that we do not consider the unlabeled divisions in our complexity measure.

In spite of this average trend, there is still a lot of variation in entry lengths within each complexity level. On the one hand, the dictionary includes longer entries with low structural complexity, for instance *näsa* 'nose' with 325 tokens but only one sense. In this case, the length of the entry can be related to the fact that the headword, as well as many other headwords denoting body parts, is part of many idiomatic expressions, diligently listed in the dictionary as mentioned in Section 2. Another example is *dörr* 'doorway, door' with two main senses described with 419 tokens. This entry is relatively long due to the fact that both senses are illustrated with many common word combinations but also compounds with *dörr-*.

On the other hand, there are also shorter and more condensed entries with many senses. Our observations in this area support Malmgren's remark that Dalin often, perhaps too often, divides entries into many different senses (Malmgren 1988: 205). This applies, among other things, to an entry like *sångare* 'singer, cantor, bard, songbird', in which five different main senses are conveyed in just 23 tokens. The corresponding entry in SO (2021) includes three main senses. Another relatively short entry, *glupsk* 'voracious', 70 tokens, has no less than 10 subdivisions: There are three main senses, of which the first and third are divided into two subsenses. Additionally, the first subsense of the third main sense is divided into three subsubsenses. The corresponding entry in SO only has one main sense and one subsense. Using the terminology employed by Lew (2013: 287), the division into senses in Dalin's dictionary is thus more characterized by splitting than by lumping.

5 Final remarks

Dalin's *Ordbok öfver svenska språket*, the first comprehensive monolingual definition dictionary for Swedish, has many qualities from both a lexicological and lexicographical perspective. This has been made evident not in the least in previous studies of the dictionary. It continues to be a valuable object of study, and a digital version facilitates styles of investigations not accessible to earlier researchers. Unfortunately,

the digital versions that have been available until now show shortcomings, which is the motivation for the re-digitization project presented in this chapter.

We have briefly described the importance and context of Dalin's dictionary, the material itself and the digitization process. At the time writing, the project has resulted in a highly-accurate digital text version with information about the layout of the printed version. Furthermore, we have in this chapter complemented the qualitative conclusions that previous researchers have drawn about the dictionary by means of five short case studies of different information categories in it, performed using the new digitized version. We have discussed, among other things, lemma types, Dalin's treatment of homography, meaning structure, and dependencies between different entries. We hope that our chapter also demonstrates the value of combining qualitative inspection with quantitative methods in the study of historical dictionaries.

The project will continue to develop the digital version of Dalin's dictionary, for instance by adding more detailed annotation of entry structure, and by creating a version suitable for inclusion in a lexical database. In our discussion we have therefore highlighted the partially different needs that a human reader of a printed dictionary has compared to a user of a non-linear access format like a database interface to the dictionary.

Dalin's *ordbok* is extensive and many other close studies are possible, for instance into Dalin's use of various kinds of usage comments to express the style and emotive charge of headwords. We therefore have reason to make the material available in different formats to other researchers and return to it ourselves in the future.

References

Almqvist, Carl Jonas Love. 1842–1844. *Ordbok öfver Svenska Språket i dess närvarande skick* [Dictionary of the Swedish language in its present state]. Örebro: N.M. Lindhs boktryckeri.

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn (ed.), *Glottometrika 2*, 1–10. Bochum: Studienverlag Brockmeyer.

Bäckerud, Erik. 2014. Nydigitalisering av SAOB [Redigitizing the Swedish Academy dictionary]. 96–105.

Dalin, Anders Fredrik. 1850. Förord [Preface]. In *Ordbok öfver svenska språket. Vol. I*, 2–20. Stockholm: Self-published.

Dalin, Anders Fredrik. 1850–1853. *Ordbok öfver svenska språket* [Dictionary of the Swedish language]. *Vol. I–II*. Stockholm: Self-published.

Hannesdóttir, Anna Helga. 1991. Relationer mellan Dalins olika ordböcker [Relations between Dalin's dictionaries]. In Sven-Göran Malmgren & Bo Ralph (eds.), *Studier i svensk språkhistoria 2*, 78–89. Gothenburg.

Hannesdóttir, Anna Helga. 1998. *Lexikografihistorisk spegel: Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga* [History of lexicography reflected: The development of monolingual Swedish

lexicography from the bilingual]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Hannesdóttir, Anna Helga & Bo Ralph. 1988. Early dictionaries in Sweden: Traditions and influences. *Symposium on Lexicography IV. Proceedings of the Fourth International Symposium on Lexicography April 20–22, 1988 at the University of Copenhagen*. 265–279.

Holm, Lars & Hans Jonsson. 1990. Swedish lexicography. In *Ein internationales Handbuch zur Lexikographie. 2. Teilband*, 1933–1943. Berlin: De Gruyter Mouton.

Johansson, Monica. 1997. *Lexicon Lincopense: En studie i lexikografisk tradition och svenskt språk vid 1600-talets mitt* [Lexicon Lincopense: A study in lexicographic tradition and Swedish language in the middle of the 17th century]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Kindblad, Karl Eduard. 1840. *Ordbok öfver Swenska Språket*. Stockholm.

Lew, Robert. 2013. Identifying, ordering and defining senses. In Howard Jackson (ed.), *The Bloomsbury companion to lexicography*, 284–302. London: Bloomsbury Publishing.

Malmgren, Sven-Göran. 1988. Almqvist, Dalin och den svenska definitionsordbokens födelse [Almqvist, Dalin and the birth of the Swedish definition dictionary]. In Gertrud Pettersson (ed.), *Studier i svensk språkhistoria*, 195–213. Lund: Lund University Press.

Malmgren, Sven-Göran. 2008. Collocations in Swedish dictionaries and dictionary research. *Lexicographica* 24: 149–158. DOI: doi:10.1515/9783484605336.1.149.

Norén, Kerstin. 1991. Utformningen av artiklarna i Dalins Ordbok öfver svenska språket (1850–55) [The composition of entries in Dalin's dictionary (1850-55)]. *Språk och stil: Tidskrift för svensk språkforskning* 1: 109–138.

Perby, Maja-Lisa. 2010. Dalins ordbok: En spegel för gammalt medicintänk [Dalin's dictionary: A mirror of medical thinking in older times]. *Läkartidningen* 107(41): 2495–2498.

Ralph, Bo. 1992. The older dictionaries as sources for Nordic language history. *The Nordic languages and modern linguistics 7: Proceedings of the Seventh International Conference of Nordic and General Linguistics*. 493–509. DOI: 10.18602/frodskapur.vi.805.

Ralph, Bo. 2001. Orden i ordning: Den historiska framväxten av en lexikografisk tradition i Sverige [Words in order: The historical development of a lexicographic tradition in Sweden]. *Nordiska studier i lexikografi 5*. 282–321.

Rogström, Lena. 1998. *Jacob Serenius lexikografiska insats* [Jacob Serenius' influence on Swedish lexicography]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

SAOL 13. 2006. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 13th edn. Stockholm: Norstedts.

Sköldberg, Emma. 2009. Talesätt och talemaader: Om ordförbindelser hos Dalin och Molbech [Figures of speech: About collocations in the works of Dalin and Molbech]. *LexicoNordica* (16): 219–239.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

Svensén, Bo. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Lars Borin, Yvonne Adesam, and Louise Holmer

13 Investigating lexical change with diachronic lexical resources and corpora

Abstract: As part of the reintegration of the traditional and computational lexical research and development activities at the University of Gothenburg, the Swedish Academy's lexical databases are being edited for greater consistency and included in the computational lexical infrastructure of Språkbanken Text, which in turn is undergoing considerable development in order to meet the requirements of traditional lexicography. One aim of this work is to formally interlink these databases with Språkbanken's Lexical Research Infrastructure. In this chapter we focus on the opportunities for diachronic lexical research offered by the inclusion of one such dataset in this infrastructure. *SAOLhist Plus* brings together digitized versions of successive editions of the SAOL dictionary, 10 editions covering a timespan of almost a century and a half (1874–2015). Together with some other historical and modern dictionaries and combined with large corpus-extracted vocabularies from the same time interval, we can make wide-ranging and detailed investigations of lexical change in relation to the history of lexicography in Late Modern and Contemporary Swedish.

Keywords: diachronic corpora, dictionaries, historical linguistics, history of lexicography, language change, language technology, lexical change, lexicography, research infrastructure

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161). We are grateful to the Swedish Academy for granting us permission to distribute several of the SAOLhist Plus dictionaries (SAOL 1, SAOL 6–13, and SO) under a CC-BY license. Thanks also to the Royal Society of Arts and Sciences in Gothenburg for a Grez-sur-Loing residency grant awarded in 2024 to Lars Borin for working on this volume.

Lars Borin, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: lars.borin@svenska.gu.se
Yvonne Adesam, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: yvonne.adesam@gu.se
Louise Holmer, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: louise.holmer@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-013

1 Introduction

In the present chapter, we introduce a new computational lexical resource intended both to aid in the study of Swedish historical linguistics, with a focus on lexical change, and to support historiographical investigations of an important period in Swedish lexicography. Historians have long talked about the "long nineteenth century", a label that has entered public consciousness perhaps mostly due to Hobsbawm (1962). The long nineteenth century is a period bracketed by two momentous historical events, namely the French Revolution in 1789 and the outbreak of the First World War in 1914. In the same vein we could say that the history of Swedish lexicography is characterized by a *long twentieth century*, initiated by the appearance of Dalin's (1850–1853) dictionary, the first complete monolingual Swedish definition dictionary (see Chapter 12 in this volume), and concluded by the publication of the first fully corpus-based general dictionaries of Swedish, from *Svensk ordbok* ('Swedish dictionary'; SOB 1986), over *Nationalencyklopedins ordbok* ('The dictionary of the National Encyclopedia'; NEO 1995), to the latest editions of *Svenska Akademiens ordlista* ('The Swedish Academy Glossary'; SAOL 14 2015) and *Svensk ordbok utgiven av Svenska Akademien* ('The Contemporary Dictionary of the Swedish Academy'; SO 2021), that are full denizens of the digital realm.

The work described in this chapter continues SAOLhist, an initiative undertaken during the years 2009–2015 as a primarily lexicographical exercise, supporting historiographical studies of lexicography. We now complement it with rich corpus data and a computational linguistic component that enhance its usefulness considerably for the study of Swedish historical linguistics generally.

2 SAOLhist

The first edition of the Swedish dictionary *Svenska Akademiens ordlista* (SAOL)[1] was published in 1874 and this dictionary has since then appeared in a total of 14 editions, the latest in 2015. SAOL provides information about spelling and inflection of its entry words and about pragmatic aspects of word usage (style level and sometimes usage domain). The dictionary generally does not offer definitions: only about one fifth of the entries come with a brief definition or some other comment. It is the

1 The title of SAOL is conventionally but somewhat misleadingly translated into English as 'The Swedish Academy Glossary'. However, it is in fact a full-fledged (normatively oriented) dictionary of Swedish, bearing roughly the same relation to the Swedish Academy's large historical dictionary SAOB as that of the *Oxford Concise English Dictionary* to the *Oxford English Dictionary*.

largest contemporary Swedish dictionary in terms of number of lexemes (close to 127,000 in the 14th edition), mostly due to an abundance of compounds. SAOL and its history are described in detail in Chapter 3 in this volume.

The SAOLhist initiative was launched in 2009 with the aim of collecting all the SAOL editions in a database that could be used for a variety of investigations of the history of the Swedish vocabulary during the period since the appearance of the first edition of the dictionary (Holmer 2012; Holmer, Malmgren & von Martens 2016). Importantly, the most recent editions were already available in a digital format when the work on SAOLhist began. For the remainder, the digitization process comprised an OCR step using off-the-shelf commercial software followed by manual checking and correction of the OCR output. This was done with the first edition (1874), and editions 6–10. Editions 2–5 were essentially reprints of the first edition and consequently would not add any information that could motivate the considerable effort needed for their digitization.

Finally, the relevant pieces of information were extracted into an XML format, from which the information was imported into a relational database for offline processing as well as online searching and browsing.

In order to fulfill the intended comparative purpose of SAOLhist, the included dictionary data had to undergo two kinds of standardization. A new Swedish orthography was officially adopted by royal decree in 1906, i.e., between the 7th (1900) and 8th (1923) editions of SAOL (see Chapter 3 in this volume). In order to compare entries across this divide, the spelling of the entries in the older editions has been modernized to a post-1906 form.[2] This cannot be mechanically done, but requires awareness of the morphology and etymology of the words. Thus, the compound *affärd* 'departure' should be modernized to *avfärd*, whereas the monomorphemic loanword *affär* 'affair, business, shop' should not be changed: *affär* is also the modern form. Similarly, the word *golf* 'floor', of Old Swedish provenance, and the loanword *golf* 'gulf' are pronounced differently and have the corresponding modern spellings *golv* and *golf*, respectively.

The other instance of standardization concerns the nature of the lexical entries themselves, i.e., the entities that are compared across time. Starting with the 12th edition (1998), the unit making up the lexical entries is a *lexeme*, in its normal sense: a set of inflectionally related forms with a unitary meaning represented by a lemma (a conventional citation form). Before that, the lexical entries were often organized by their etymology. Thus, in the 9th edition (1950) there is one entry *ljus* 'light (a/n), bright (a), candle (n)', subdivided into a noun section and an adjective

2 Actually, to the spelling used in the 13th edition (2006) in order to also capture some other instances of diachronic spelling variation not directly connected to the spelling reform.

section. In the SAOLhist database the aim is that the lexeme uniformly should be the lexical entry.

The end result of the SAOLhist project was a relational database containing the headwords of 12 dictionaries in the standardized format described above, which however was not made available as such to researchers and the public, but could only be accessed through a web interface allowing a limited range of – essentially single-word – queries. There is no provision for downloading results for further processing offline. On the other hand, the interface provides links from query results to the corresponding scanned page images of each dictionary.

The online interface to SAOLhist was designed and built by Monica Martens (Holmer, Malmgren & von Martens 2016). Its first version became available in 2013 (SAOLhist 2013), providing access to the entries of SAOL editions 1 and 6–13. Later, Dalin's (1850–1853) dictionary (see Chapter 12 in this volume), SAOL 14 (2015), and SO (2009) were added. The online interface to SAOLhist has had its present form and content since 2015.[3]

3 Towards SAOLhist Plus

While SAOLhist provides a convenient, if limited, window on the history of Swedish lexicography over the last 170 years, it does not directly answer questions about the relationship between dictionaries and their contemporaneous language varieties. For this we also need appropriate corpus data and the means for connecting corpus word occurrences to dictionary entries, i.e., lemmatizers for the language varieties involved (historical and modern).

It also remains to investigate how it can be used for bulk (offline) processing, to try to find out about broader tendencies of lexical change. SAOLhist has not yet been used for such larger-scale studies; Holmer, Malmgren & von Martens (2016) report on two small studies using the online interface.

After SAOLhist was completed and published online, large amounts of historical Swedish texts have become available in digital form (e.g. Pettersson & Borin 2022). In addition, Språkbanken Text possesses a number of digitized older and modern Swedish dictionaries, the results of several digitization efforts undertaken since its establishment in 1975, and most recently with the aim of building a unified interlinked lexical infrastructure for language technology and other research (see Dannélls, Borin & Heppin 2021 and Chapter 5 in this volume). Some of these are clear

3 Notably, the computational infrastructure of the SAOLhist interface was adopted for a parallel Danish project, *ROhist*, initiated by the Danish Language Council (Diderichsen et al. 2015).

Table 1: The dictionaries of SAOLhist Plus and the publication year of the first edition of each dictionary, together with its size, and the number of text word types covered in the respective KB datasets for the same years (both rounded to thousands)

Dictionary	Year	Entries	KB types
Dalin	1855	63,000	1,623,000
SAOL 1	1874	34,000	3,049,000
SAOL 6	1889	42,000	5,564,000
SAOL 7	1900	72,000	7,787,000
SAOL 8	1923	79,000	2,115,000
Bring	1930	52,000	2,035,000
SAOL 9	1950	149,000	1,931,000
SAOL 10	1973	140,000	4,441,000
SAOL 11	1986	117,000	3,640,000
SAOL 12	1998	119,000	3,979,000
SAOL 13	2006	123,000	1,533,000
SO	2009	96,000	1,433,000
SAOL 14	2015	127,000	5,380,000

candidates for addition to SAOLhist, thus both extending the timespan covered and filling gaps within it.

We follow the original intent of SAOLhist and include only dictionaries describing the written standard language (roughly) contemporaneous with the time of their publication (which by no means limits our selection to SAOL editions, however). Hence, this excludes, for example, the Old Swedish[4] dictionaries compiled in the 19th and 20th centuries by Söderwall (1884–1918) and Schlyter (1887), which are also available in digitized versions through Språkbanken Text (Adesam et al. 2021), or the historical *Svenska Akademiens ordbok* ('The Swedish Academy Dictionary'; SAOB 1898–2023). Also excluded are dialect dictionaries, and various specialized dictionaries, e.g., dictionaries of slang. We further limit our selection – at least for the time being – to monolingual Swedish dictionaries, which excludes several older dictionaries such as the Swedish–Latin dictionary by Swedberg & Holm (2009). Finally, we have only been able to consider dictionaries that are easily and freely available in a well-structured digital format, or convertible into such a format with a reasonable effort.

As a result of our work, an enhanced dataset is now available, containing an additional dictionary – Bring's (1930) thesaurus (see Chapter 8 in this volume) – and corpus-based vocabulary occurrence information for each year of publication of all

4 In the commonly adopted periodization of the history of Swedish, the *Old Swedish* period extended between 1225 and 1526.

```
@source kbnews          :       58458           m       22794
@year   1855            till    49020           a       21435
@month  12              att     48786           är      20399
@feats  word            den     44488           1       18355
@size   5407081         för     41523           d       18063
.       388619          en      39363           på      17367
,       361457          ;       33366           ett     17222
och     111108          med     32979           t       17071
i       108831          de      30957           (       15581
»       98968           —       30568           sig     15076
-       70567           som     29928           o       14912
af      68696           det     23350           2       14576
```

Figure 1: The header and beginning of a KB word count file (data for December 1855)

the dictionaries. We refer to this dataset as *SAOLhist Plus*. An overview of the size of the various component resources of SAOLhist Plus can be found in Table 1.[5]

In the following sections, we describe the corpus data (Section 4) and the two morphological analyzers used for lemmatizing the corpus data, one of which was developed specifically for the present work (Section 5). Section 6 is devoted to some proof-of-concept large-scale case studies illustrating some of the possible uses of SAOLhist Plus. A summary and outlook conclude the chapter in Section 7.

4 Corpus-derived diachronic lexical data: 170 years of newspaper vocabulary

The main corpus data used for our investigations consist of text word statistics extracted from the digitized newspapers held by the National Library of Sweden (*Kungliga biblioteket* 'The Royal Library': KB; Adesam, Dannélls & Tahmasebi 2019; Dannélls, Johansson & Björk 2019). Most of the texts themselves are not available outside of KB's premises due to intellectual property rights restrictions, but various derivatives are, including the data used here, comprising monthly text word statistics for the years 1850–2021. Each file contains a header, shown in Figure 1, with information about source material (always "kbnews" in this dataset), year, month, text feature (always "word" in this dataset) and total feature count (tokens in this dataset), followed by a list of text word type counts, sorted by descending frequency.

5 The number of entries reported for a given edition of the SAOL dictionary fluctuates somewhat depending on the source of the information. The figures in Table 1 come out of the SAOLhist database and are a result of the standardization decisions described above, except for Bring (1930).

ochmed	82	ochhans	23	och2	13
ochen	47	koch	23	ochpä	12
ochi	38	ochhennes	22	störreoch	12
ochde	35	ochför	20	broch	12
tilloch	29	ochnorrige	17	ochett	12
ochandra	29	ochsom	16	bloch	12
ochden	28	swerigeoch	16	postoch	12
ochatt	27	heloch	15	toch	11
ochdet	27	ochj	14	meroch	11
ochtill	25	ochnorge	14	ochder	11
ochmindre	24	ochpå	13	barnoch	11

Figure 2: Potential unsegmented *och* instances in the KB data for December 1855

These datasets were originally produced in another project (Ingvarsson et al. 2022), and were kindly made available to us by our colleague Niklas Zechner. For our investigations, we prepared separate datasets for each of the publication years of the 13 SAOLhist Plus lexicons,[6] for a total of slightly over five and a half billion tokens distributed over slightly less than 150 million word types.

The KB newspaper collection has been digitized in a lossy automatic optical character recognition (OCR) process (Dannélls, Johansson & Björk 2019). Consequently, the word lists contain numerous non-words, especially for the older material, since the OCR has been trained on modern language. In order to work effectively with the word lists it was necessary to first filter out as much as possible of the dross generated by the OCR process, and also other uninteresting (for our purposes) text word types, such as numbers and punctuation[7] as well as any string containing digits or punctuation marks.[8]

The OCR process often runs text words together, and a separate filtering process was used to separate out initial or final *och* 'and', which – due to its characteristic spelling – is an extremely unlikely initial or final part of a well-formed Swedish word, and further is a very high-frequent word type, as Figure 1 shows, so that words containing an illicit initial or final *och* are common in the dataset (see Figure 2, show-

6 While the publication years printed in the two volumes of Dalin's dictionary are 1850 and 1853, the latter is a misprint for 1855 (Hannesdóttir 1998: 462). Although the standard bibliographical reference that we use in this chapter has 1850 and 1853 as the publication years, we actually date Dalin to 1855 and use this as the starting year for the corpus data.

7 The tokenization algorithm producing the word list items separates out initial and final punctuation from space-separated text strings and lists punctuation marks separately as tokens in their own right.

8 This does exclude some lexical items present in the SAOLhist dictionaries, such as *3G* '3G' or *95-oktanig* (adj) '95 octane-'. These are extremely marginal in the dictionaries (10 out of almost 300,000 entries).

Table 2: Coverage of the morphological analyzers on the KB datasets: tokens (legend: “a55”, etc. = ‘1855’, etc.; “n00”, etc. = ‘1900’, etc.; “t06”, etc. = ‘2006’, etc.)

dataset	tokens	removed	left	frac.	analyzed	frac.
a55-kb	66,211,566	26,608,865	39,602,701	0.598	32,417,315	0.819
a74-kb	174,332,021	71,665,934	102,666,087	0.589	85,201,797	0.830
a89-kb	414,711,425	154,224,719	260,486,706	0.628	220,812,015	0.848
n00-kb	515,024,315	193,007,902	322,016,413	0.625	268,205,893	0.833
n23-kb	160,329,409	56,593,459	103,735,950	0.647	94,550,982	0.911
n30-kb	183,183,169	67,337,090	115,846,079	0.632	105,929,431	0.914
n50-kb	165,660,158	56,890,403	108,769,755	0.657	100,493,006	0.924
n73-kb	231,399,363	95,815,454	135,583,909	0.586	119,736,142	0.883
n86-kb	269,327,213	120,489,572	148,837,641	0.553	133,451,589	0.897
n98-kb	367,587,473	187,387,937	180,199,536	0.490	157,468,103	0.874
t06-kb	225,143,701	111,778,538	113,365,163	0.504	100,695,890	0.888
t09-kb	234,460,637	123,587,451	110,873,186	0.473	98,524,569	0.889
t15-kb	2,616,969,294	1,220,632,495	1,396,336,799	0.534	1,262,635,597	0.904
sum	5,624,339,744	2,486,019,819	3,138,319,925		2,780,122,329	
avg				0.578		0.878

ing the top instances of upwards of 4,500 word types). Provided that the remainder contained at least one vowel, initial or final *och* was split off (and counted as an instance of *och* ‘and’).[9] There are many other run-on words in the lists, and it could be rewarding to try to use the fullform lexicons to split these, also in order to produce higher-quality and more useful corpora.

The two fullform lexicons together with the compounding lexicons described in Section 5 were used for generating trigram lists for 19th century (or rather: pre-1906 spelling reform) and 20th century Swedish. Any KB list item containing a trigram not present in the corresponding trigram list was filtered out.

After filtering, the remaining dataset contained upwards of three billion tokens – about 56% of the original – and 44.5 million types (reduced by some 70%; see Tables 2 and 3, columns *tokens/types*, *removed*, and *left*).

Finally, the morphological analyzers described below in Section 5 were applied to the remaining corpus word types, since the large-scale lexicon–corpus comparisons discussed in Section 6.2 can only be carried out with items that have received an analysis both in the lexicon and the corpus.

9 Note that this means that the type statistics of the derived SAOLhist Plus word lists differ from that of the original KB datasets.

Table 3: Coverage of the morphological analyzers on the KB datasets: types (legend: “a55”, etc. = ‘1855’, etc.; “n00”, etc. = ‘1900’, etc.; “t06”, etc. = ‘2006’, etc.; msds = morphosyntactic descriptions; cmp = compound analyses)

dataset	types	removed	left	frac.	msds	frac.	cmp	frac.
a55-kb	5,690,554	4,067,308	1,623,246	0.285	211,381	0.130	150,954	0.714
a74-kb	10,716,977	7,668,320	3,048,657	0.284	365,473	0.120	289,399	0.792
a89-kb	17,084,802	11,520,469	5,564,333	0.326	668,532	0.120	594,080	0.889
n00-kb	23,727,190	15,940,581	7,786,609	0.328	853,460	0.110	777,530	0.911
n23-kb	7,188,768	5,073,552	2,115,216	0.294	461,975	0.218	377,022	0.816
n30-kb	7,019,903	4,984,758	2,035,145	0.290	504,672	0.248	417,145	0.827
n50-kb	6,356,804	4,426,139	1,930,665	0.304	611,202	0.317	518,533	0.848
n73-kb	12,235,548	7,794,143	4,441,405	0.363	961,566	0.217	864,573	0.899
n86-kb	11,201,484	7,561,401	3,640,083	0.325	912,308	0.251	814,050	0.892
n98-kb	12,996,117	9,017,073	3,979,044	0.306	962,249	0.242	864,346	0.898
t06-kb	6,452,492	4,919,529	1,532,963	0.238	547,052	0.357	456,105	0.834
t09-kb	5,953,538	4,520,514	1,433,024	0.241	523,504	0.365	433,449	0.828
t15-kb	23,223,337	17,843,745	5,379,592	0.232	1,396,373	0.260	1,290,329	0.924
sum	149,847,514	105,337,532	44,509,982		8,979,747		7,847,515	
avg				0.294		0.227		0.852

Thus, using the year 1923 as an example, from Tables 2 and 3, we learn that the KB corpus dataset for this year (dataset n23-kb) contains (in rounded figures) a little over 160 million tokens (Table 2) and about 7.2 million text word types (Table 3). The initial filtering reduces these numbers to 104 million (or 65%) and 2.1 million (29%), respectively. Applying the morphological analyzers to the remaining items, almost 95 million tokens (or 91%) get an analysis, whereas only 462,000 text word types (or 22%) receive an analysis, most of which – 377,000 (or 82% of the analyzed text word types) – are analyzed as compounds. These, then, are the corpus words which will be compared with lexicon entries in Section 6.2.

5 Computational morphologies for Late Modern and Contemporary Swedish

A central component of the research infrastructure described in this chapter is the means of connecting text word forms to dictionary entries, i.e., morphological analyzers (and decompounders) for the various historical language varieties involved.

For the work described here, we have used two such analyzers and decompounders. For the period after the spelling reform (after 1906), the Saldo 3 mor-

phology has been used (referred to below as *s3m*), where the lexical basis of this modern (Contemporary Swedish) fullform lexicon is the Saldo computational lexicon, currently (in early 2025) in version 3.3 (see Chapter 6 in this volume). The *s3m* description was implemented using computational tools especially designed for building morphological analyzers and generators, the *foma* platform for defining finite-state transducers (FST) implementing morphophonological and orthographical rules, and the *lexc* formalism for specifying (inflectional and derivational) morphotactics (Beesley & Karttunen 2003; Hulden 2009; 2022).

For the Late Modern Swedish variety of the period 1855–1906, another morphological analyzer was created specifically for SAOLhist Plus based on the vocabulary and morphological – POS and inflectional paradigm – information semi-automatically extracted and extended from the recently re-digitized mid-19th century dictionary by Dalin (1850–1853; see Chapter 12 in this volume),[10] basically by modifying the already existing *s3m* description, which in fact could be recycled to a significant extent, reflecting the fact that Swedish inflectional morphology has not undergone major changes over the long twentieth century. The resulting description contains 116,517 lexemes inflected according to a total of 445 paradigms (including paradigms for 3,312 multiword expressions – which are not used in SAOLhist Plus, however – and for abbreviations). The large number of lexemes compared to the number of entries in Dalin (see Table 1) is due to the fact that participles and verbal nouns are automatically generated from the verb entries and added as separate entries in the morphology if not already present.

Often lauded by present-day lexicographers as a pioneering achievement of Swedish lexicography (Malmgren 1988), Dalin's dictionary should be a good point of departure for this endeavor. However, it quickly becomes apparent that Dalin's forte was not morphological description; the dictionary contains many inconsistencies and downright errors in inflection class assignment. Impressionistically, many of the inconsistencies are due to the three- (or four-) gender norm adopted in the dictionary clashing with the two-gender system actually in use in Late Modern Swedish (Teleman 2003: 138–144).

Also, surprisingly often a simplex noun and (some) compounds containing this noun as final member are listed with different paradigms. In Contemporary Swedish this *can* occur, but only as a decided rarity (for instance, non-neuter *blick* 'look, glance (n)' vs. neuter *ögonblick* 'moment, instant', etymologically a compound of *öga* 'eye' and the simplex *blick*). It is unlikely that things were radically different in

10 The re-digitized version is of much higher quality than our previous digital version, concerning both the text itself and the dictionary structure, which has been a great help for extracting the relevant information from the dictionary. See Chapter 12 in this volume for more detailed information.

Late Modern Swedish, and since we observe a number of other inconsistencies in Dalin's inflectional descriptions, it is highly probable that we are here dealing with mistakes. (It seems to be fairly common that a declension or conjugation number inadvertently repeats that of the preceding entry.)

Dalin (1850: 16) states that the morphological description by Enberg (1836) has been adopted in the dictionary "without exception". The contemporaneous Swedish grammar by Almqvist (1840) makes a more solid impression on the modern reader through its decidedly more analytical and descriptive slant than the Academy grammar by Enberg (1836) which comes across as quite normative. While it is clear that Almqvist fully subscribes to the normative spirit of the times – after all, his grammar is announced as a "textbook primarily intended for Youth" (Almqvist 1840: i; our translation) – he provides numerous examples of actual usage that deviates from the norm. See also Haapamäki (2002). Consequently, Almqvist (1840) has been used as the main reference in preparing the Late Modern Swedish morphology.

All in all, using Dalin (1850–1853) as the basis for the computational Late Modern Swedish morphology turned out to be a quite instructive exercise. Since the information about inflectional class is automatically extracted from the digitized dictionary text and forms the basis for selecting the corresponding lexc inflectional class, there is an unknown number of errors in the fullform lexicon. Some of these errors have been corrected when spotted, but no systematic correction has been made. Consequently, some errors undoubtedly remain. It is probably fair to say that *dm* is less reliable than *s3m*.

Figure 3 illustrates the format of the Dalin morphology. In the figure, we see two foma rules, both dealing with orthography, in the first case how the genitive clitic *=s* should be written after words ending in a sibilant (<s>, <z>, <x>, <sch>, <sh>), and the second rule deals with the orthographic consequences of adding the neuter suffix *-t* to an adjective with a stem ending in certain consonant combinations. The two lexc minilexicons in Figure 3 describe a noun and a verb paradigm, respectively. A comparison with the corresponding *s3m* description shows that the foma FST rule set is similar, but that the inflectional paradigms are more complex in the Dalin morphology, especially verbal conjugations, as can be seen in Figure 4 (repeated here from Chapter 6 in this volume for convenience).

Vocabulary-wise, Dalin's dictionary is quite complete, as we demonstrate below in Section 6.2. Hence *dm* provides fair coverage of Swedish texts published in the (mid-)19th century.

The compound analysis is done by a simplistic brute-force approach and is not exhaustive. First, only compounds with two members are recognized, although the members – being entries in Dalin or Saldo – may themselves be compounds. Second, if the computational compound lexicon allows for more than one possible segmentation of a word, the algorithm always chooses the one with the longest

```
# genitive 's after s etc. (shouldn't affect superlative)
define gens %+ s -> ' s || [s|z|x|s (c) h] _ ,,
   %+ s -> ' || [s|z|x|s (c) h] _ ;

# nn -> nt in tunn; dd > dt in stadd; t -> 0 in smärt
define avt n %+ t -> t || n _ [ .#. | %+ ] ,,
   d %+ t -> t || d _ [ .#. | %+ ] ,,
   t %+ t -> t || [Cns|a (i) |u] _ [ .#. | %+ ] ;

LEXICON dnn_35v_redskap
dnn_5n_blad;
dnn_3m_bild;

LEXICON dvb_2r_hyra
dvb_124_ndep;
% %:%:vb% inf:0 dvb_vsfo;
% %:%:vb% imp% 2sg:¹ dvb_vesfo;
% %:%:vb% prs% sg:¹ dvb_vsfo;
% %:%:vb% prt% sg:¹+de dvb_vsfo;
% %:%:vb% prt% 1pl:¹+de dvb_vsfo;
% %:%:vb% prt% 2pl:¹+den dvb_vsfo;
% %:%:vb% prt% 3pl:¹+de dvb_vsfo;
% %:%:vb% sup:¹+t dvb_vsfo;
```

Figure 3: The Dalin morphology: two foma FST rule definitions and two lexc minilexicons

```
# no genitive s after s etc. (shouldn't affect superlative)
define gens %+ s -> 0 || [s|z|x|s (c) h] _ ;

# dd -> tt in klädd, sydd
define avdd d d %+ t -> t t || _ [ .#. | %+ ] ;

# d -> t in röd, rund; nn -> nt in tunn; t -> 0 in smart
define avdt d %+ t -> t t || Vwl _ [ .#. | %+ ] ,,
d %+ t -> t || Cns _ [ .#. | %+ ] ,,
n %+ t -> t || n _ [ .#. | %+ ] ,,
t %+ t -> t || Cns _ [ .#. | %+ ] ;

LEXICON 3nn_vn_lexikon
% %:%:nn% n% sg:0 3nnn_defsg;
% %:%:nn% n% pl:²+a 3nnu_defpl;
% %:%:nn% n% pl:0 3nnn_defpl;

LEXICON 3vb_2r_hyra
% %:%:vb% prs% akt:i #;
% %:%:vb% imp:i #;
% %:%:vb% inf% akt:0 #;
3vb24_npast_sfo;
% %:%:vb% prt:ide 3vb_snprs;
% %:%:vb% sup:it 3vb_snprs;
```

Figure 4: The Saldo 3 morphology: three foma FST rule definitions and two lexc minilexicons

first element and does not backtrack to look for alternative segmentations with shorter first parts. Since our aim is simply to find compounds, this is generally not a problem, as long as the analyses in the lexicon and the corpus word lists coincide, for example:[11]

(1) a. *bild-rulle* 'picture-roll', not *bil-drulle* 'car-hog/fool/oaf'
b. *stork-hanen* 'the stork-male', not *stor-khanen* 'the great-khan'
c. *glass-skål* 'icecream-bowl', not *glass-kål* 'icecream-cabbage' or *glas-skål* 'glass-bowl'

Like all computational linguistic analyses, both the morphological analysis process and the decompounding procedure will produce errors. However, the errors are few – both processes are sufficiently accurate for our purposes – and our hypothesis is that the errors are random with regard to our research questions, i.e., that they do not favor or disfavor a particular corpus or a particular lexicon. In other words, the comparisons presented below are expected to give correct results, in the sense that their ordering is valid, although the magnitude of individual results may be lower or higher than ground truth.

The morphological analyses used in our investigations are produced using full-form lexicons and a compound splitter. The fullform lexicons are generated from lexeme lists plus inflection tables using a lexical finite-state transducer written in foma (Hulden 2009) and lexc (Beesley & Karttunen 2003).

The same approach is used to generate the compounding forms used by the compound splitter: a (different) foma transducer cascade and lexc lexicon set define possible initial and medial compounding forms for those lexical items that can form compounds.

As mentioned above, the foma/lexc morphological description of Dalin is built on and basically parallels that of Saldo 3. The main differences are that the morphology – in particular the verb inflection – is considerably richer in the older language stage described by Dalin. The compounding description is different and above all more generous in the allowed formal variation. For *dm* there is also a separate description of participles, described in both *s3m* and *dm* as separate lexical entries, in accordance with the Swedish Academy grammar (Teleman, Hellberg & Andersson 1999), but as adjectives rather than a separate POS participle. In the case of *dm* the participle (and verbal noun) generator is a kind of "metadescription" that generates entries for the main morphological component *dm*.

11 In the last case (1c) because the decompounding algorithm tries to restore a third equivalent consonant first.

As mentioned, verbs have considerably more forms in *dm* than in *s3m*. Also, *dm* (as well as its contemporaneous grammars, e.g., Enberg 1836; Almqvist 1840) assigns four gender designations to nouns. The non-neuter nouns of Contemporary Swedish are classified as masculine, feminine, or – reflecting the *de facto* already completed merger of these two in everyday speech – *realgenus*, corresponding to modern *utrum* 'non-neuter'.

There is a fair number of inflectional differences in the nominal realm as well. Thus, 19th century Swedish has many more neutral nouns forming their plural in *-er* than the modern language (where the corresponding items show zero plural). Compounding forms are often different: Dalin has a number of compounds beginning with *olj-* 'oil' (e.g., *oljpalm* 'oil palm', *oljväxt* 'oil plant', that in Contemporary Swedish can be only *oljepalm*, *oljeväxt*).

And, of course, the orthography is different because of the 1906 spelling reform (which actually played out over about half a century; see Teleman 2003: Ch. 4).

6 Proof-of-concept case studies

In the following we explore aspects of lexical change using two different approaches. First, we compare different lexica over time to find words that appear or disappear. Second, we compare lexica to diachronic corpus data.

6.1 Diachronic lexicon comparison

If we want to learn about lexical change, a reasonable assumption would be that using lexica or dictionaries from different time periods may be helpful. Unfortunately, comparing dictionaries for this purpose is not entirely straightforward. First, dictionaries generally lag behind when it comes to language change, because up until recent digital dictionaries, the process of gathering words for the dictionary was slow, and space on printed paper was limited. Most dictionaries therefore also only include the most standard patterns of language. Different dictionaries additionally have different criteria for selecting words and how to organize them, which may also change over time. However, as we will see, we can still track language change through the dictionary, learning about lexical change by comparing different lexica over time.

As we can see from Table 1, Dalin's dictionary is a bit larger than the first versions of SAOL, while something happened between SAOL 8 and 9, nearly doubling the size. The large increase in entries in the 9th edition could be explained by a general

development in society in areas like social care, sports, amusement, engineering, music, etc. In addition, the 9th edition offered a wider selection of the more colloquial vocabulary (see Gellerstam 2009).

Let us turn to comparing the entries present in the different lexica. While we can easily compare word forms, it is much more difficult to compare word meanings. To date, there is no linking between different entries, and there is no easy way of automatically determining if two entries point to the same meaning. However, we have a standardized spelling variant for all head entries, together with a standardized part of speech (POS). On a more superficial level, we can therefore compare word forms with POS over time. While we cannot compare *agn*[1] (n) 'bait' with *agn*[2] (n) 'husk' from Dalin across time, we can see that *agn* (n) is present in all the different dictionaries (and, in fact, all dictionaries have two different meanings registered).

We have extracted a list of all word forms, in their standardized spelling, together with a standardized POS label, together with the lexica they can be found in. Counting all word form–POS combinations, there are a little over 273,000 present in any of the dictionaries. Of these close to 15,000 appear in all of the dictionaries. Close to 9,000 of these are nouns (of 190,000 nouns in total for the word form–POS combinations), 3,000 are verbs (of 18,000), and 2,000 are adjectives (of 27,000), the rest being labelled as pronouns or "other".

One simple method for exploring lexical change is to list words that appear in all dictionaries before a particular breaking point, but in none after, or conversely, words that appear in all dictionaries after a certain point, but none before. This allows us to find word forms that appear or disappear in the lexica. If we arbitrarily pick 1906 as a breaking point (the year of the last Swedish spelling reform, although spelling does not affect our question), we have close to 600 word form–POS combinations that appear in all lexica from Dalin to SAOL 7, but not after, and almost 2,800 word form–POS combinations that appear in all lexica from SAOL 8 to SO, but not before. Although we cannot assume that these are new or disappearing words – they could also indicate, e.g., a change in the organization of the dictionaries – the diachronic stability indicates that they may be interesting to investigate further. It should be pointed out that since we lack links between meanings over time, we cannot easily explore semantic change in this lexical resource.

A poignant example of a word not appearing before 1906, but in all lexica published after this year, is the noun *atomkärna* 'atomic nucleus'. Since it was discovered after experiments in 1909, it is not surprising that it did not appear in the dictionaries up to 1900, although it is interesting to note that it was of sufficiently common knowledge to be included in the 1923 dictionary. Its first appearance in the KB material is in 1915, and it is mentioned less than 70 times until the end of 1922.

Another example of new inventions or scientific discoveries is the word form *bil* 'car', including numerous compounds, entering the dictionaries after 1906. While the

longer form *automobil* is present in all dictionaries from SAOL 7, compounds with *automobil* can only be found until SAOL 8 or possibly SAOL 9, with the exception of *automobilklubb* 'automobile club' and *automobilkår* 'automobile corps', that hang on for longer. While the longer form was used early on, it was not productive after the first quarter of the 20th century, when the shorter form had caught on. The alternative *lokomobil* can be found in the dictionaries from SAOL 7 to SAOL 10. Apparently, it took a while for these words to enter the dictionaries in 1900, since *lokomobil* can be found in the KB material already in the 1850s, and *automobil* appears twice in the 1870s and in the 1880s, until becoming more frequent in the 1890s.

Let us explore a word that has disappeared. The word *hjon* (approximately 'serf')[12] refers to someone who is part of a household, generally a pauper or someone dependent on others. Searching for *hjon* and all its compounds (removing *mahjong* with compounds, and the beetle *blåhjon*), we find most entries in Dalin, 28 words. In SAOL, we see a decline over time – with the exception of SAOL 7, that jumps up to 23 words – ending with 9 and 8 words respectively in SAOL 13 and SO, and only 4 in SAOL 14. We thus clearly see how the term, while still in the dictionaries, is used less and less, shown by the loss in compounds. We can assume that the main reason for the word still being present in the newest dictionaries is to point at the historical use of the word, which is what we also find, the word being marked as historical and archaic in the latest dictionaries. While this word, pointing to a division of people into categories that is not acceptable today, is still present in the most recent dictionaries, others have disappeared entirely. For example, the word *halvmänniska* 'half-human' refers to people (or probably rather human races) that are assumed to have lower intellectual capacity or morals. The word is present in Dalin, and SAOL 1 – SAOL 7, but has disappeared after 1900. Exploring similar words, we can follow when society changed, as reflected in what words were allowed in the dictionary.

6.2 Lexical saturation and lexical corpus coverage

As mentioned above, the present incarnation of SAOLhist Plus contains 13 lexicons published between 1855 and 2015, a period of 160 years (see Table 1). We have seen in Section 6.1 that the vocabulary present in the dictionaries does indeed change over time. An interesting question to pursue would then be if this change correlates with lexical change in the language, and if it does, if the former precedes the latter or vice versa. Changes in the lexicon preceding lexical change in the language would be

12 Dalin also lists the meaning 'spouse', but this sense was obsolete in practice even in the mid-19th century.

prescriptivism writ large, as it were, while the reverse situation would correspond to a case of run-of-the-mill descriptive linguistics.

Having a diachronic set of good-sized lexicons as that in SAOLhist Plus as well as a large amount of time-stamped corpus data for the same period, we can investigate various ways in which lexical and corpus data relate to each other over time, and in this way empirically and on a large scale approach one of the "questions [that] come readily to mind in this connection", namely: "To what extent is the vocabulary of a dictionary a reliable reflection of actual contemporary usage?" (Ralph 1992: 495).

One such way is the degree of *lexical saturation* with regard to the corpus data, by which we mean the fraction of lexical entries (lexemes) of a particular lexicon that can be found in a particular corpus, the "efficiency" of a dictionary, as it were.

For this we need to lemmatize the corpus data and ensure that the corresponding lexemes can be located in the dictionaries. In order to do this, we apply the two morphological analyzers (*s3m* and *dm*) on both the filtered corpus wordlists and the dictionaries (see Section 4).

A small caveat: what we can find in the lexicons and corpus wordlists using the morphological analyzers are – at the most – unique lemma–POS pairings, that we will refer to as *lemposes* here in order to avoid confusion. As mentioned in Section 6.1, one lempos can in principle correspond to more than one lexeme. In Swedish there is a fair number of lemposes – mainly nouns, but also a few verbs – that correspond to more than one lexeme, for example *mask nn* 'mask' (plural *masker*); 'worm' (plural *maskar*), or *brygga vb* 'brew' (past *bryggde*); 'bridge' (past *bryggade*). Consequently, lemma and POS are generally not sufficient information to uniquely identify a lexeme in Swedish. Incidentally, this was the observation underlying the unfortunate choice of the term (in Swedish) *lemma* for 'lexeme' of Allén's so-called lemma–lexeme model (e.g. Allén 1967; 1981), as well as the coining of the neologism *lemgram* for the form units of Saldo (see Chapter 6 in this volume).

Consequently, we analyze the corpus wordlists and the dictionary entries into lemposes. In the case of the dictionaries, the morphological analysis of the lexical entries is filtered on lemma, so that, e.g., *akter* 'stern (of a ship)' will not be analyzed as the indefinite plural form of *akt* 'act; document n.', or *agens* 'instrumental factor' not as the genitive singular definite form of *ag* 'a kind of grass'.[13]

The corpus wordlists for the years 1855, 1874, 1889, and 1900 – i.e., the period before the spelling reform – are analyzed with *dm*, and those from 1923 onwards using

13 We also do not attempt to analyze multiword expressions, that do occur in the dictionaries, e.g., phrasal verbs. Even though these are present in the two morphological analyzers, the KB news text corpus wordlists contain only single-word entries, so in this case there is no basis for comparison between dictionaries and corpora in this respect. As mentioned already, we also do not account for lexical entries containing digits.

Table 4: The coverage of the two morphological analyzers (*dm* and *s3m*) on the SAOLhist dictionaries (legend: “a55”, etc. = ‘1855’, etc.; “n00”, etc. = ‘1900’, etc.; “t06”, etc. = ‘2006’, etc.; dcmp = *dm* compounds; s3cmp = *s3m* compounds)

dict.	size	not dm	dm	cov.	not s3m	s3m	cov.	either	cov.	dcmp	fract.	s3cmp	fraction
a55-dln	62,975	2,525	60,450	0.960	14,594	48,381	0.768	60,876	0.967	501	0.008	17,697	0.366
a74-s01	34,308	3,136	31,172	0.909	4,103	30,205	0.880	32,920	0.960	1,929	0.062	4,721	0.156
a89-s06	42,367	5,397	36,970	0.873	5,196	37,171	0.877	40,619	0.959	3,370	0.091	6,462	0.174
n00-s07	72,458	17,212	55,246	0.762	9,163	63,295	0.874	67,920	0.937	16,198	0.293	18,637	0.294
n23-s08	78,524	26,420	52,104	0.664	10,015	68,509	0.872	73,311	0.934	16,795	0.322	19,761	0.288
n30-bng	51,589	15,708	35,881	0.696	6,248	45,341	0.879	48,183	0.934	6,782	0.189	5,636	0.124
n50-s09	149,133	61,410	87,723	0.588	21,349	127,784	0.857	133,442	0.895	48,862	0.557	61,047	0.478
n73-s10	139,604	59,468	80,136	0.574	19,047	120,557	0.864	125,484	0.899	42,591	0.531	49,500	0.411
n86-s11	116,949	52,676	64,273	0.550	15,108	101,841	0.871	105,305	0.900	30,709	0.478	31,614	0.310
n98-s12	119,347	54,377	64,970	0.544	13,722	105,625	0.885	108,565	0.910	32,772	0.504	34,881	0.330
t06-s13	123,031	56,238	66,793	0.543	13,020	110,011	0.894	112,589	0.915	35,091	0.525	37,139	0.338
t09-so1	96,143	43,622	52,521	0.546	7,864	88,279	0.918	88,911	0.925	24,519	0.467	16,100	0.182
t15-s14	126,900	59,870	67,030	0.528	14,944	111,956	0.882	114,165	0.900	36,040	0.538	38,634	0.345
average				0.672			0.871		0.926		0.351		0.292

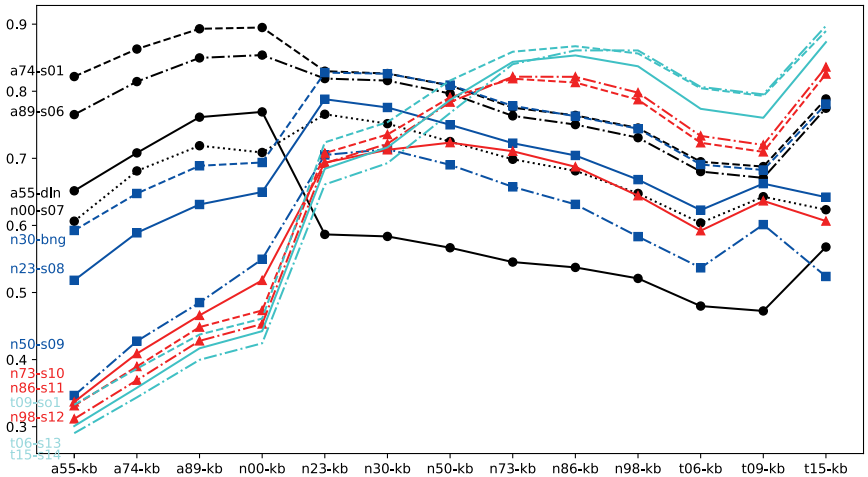


Figure 5: Lexical saturation (y axis) for the SAOLhist Plus dictionaries on KB news text data (x axis) (legend: "a55", etc. = '1855', etc.; "n00", etc. = '1900', etc.; "t06", etc. = '2006', etc.)

s3m. The dictionaries are analyzed with both morphologies, and when calculating if a corpus lempos exists in a dictionary either analysis counts as a positive result, which in some way compensates for the fact that the old and the new orthographies overlap in time (Teleman 2003: Ch. 4; see also Chapter 3 in this volume).

It stands to reason that we can only calculate lexical saturation for that share of a dictionary's entries that has received a lempos analysis. Table 4 shows how well the two morphologies handle the SAOLhist Plus dictionaries. The relevant columns in Table 4 are *either* and the immediately following *cov(erage)* column. The latter is the fraction of entries analyzed by at least one of the two morphologies, ranging between 0.967 (Dalin) and 0.895 (SAOL 9). Since *dm* was built using Dalin as its vocabulary source, this result – less than 100% coverage – shows that the vocabulary extraction algorithm used in creating *dm* is obviously different from that used to populate the Dalin component of SAOLhist. In the case of *s3m*, its vocabulary has been compiled quite independently of any of the SAOLhist dictionaries. However, due to its history (see Chapter 6 in this volume), it should be closest to SO (2009), and this is confirmed by the actual numbers in Table 4 (where the *s3m* cov[erage] of SO 2009 is 0.918, the highest in its column).

The lexical saturation figures are shown in Table 5 and illustrated graphically in Figure 5. We see there that SAOL 1 is the top-ranking dictionary for about the first half of the long twentieth century, after which SO (2009) takes over for a few years, and then SAOL 14 rises to the top. This is a sign that the compilers of the first

Table 5: Lexical saturation of the SAOLhist Plus dictionaries on KB news text data (legend: “a55”, etc. = ‘1855’, etc.; “n00”, etc. = ‘1900’, etc.; “t06”, etc. = ‘2006’, etc.; boxed value = “best” corpus wrt this dictionary; **boldfaced** value = “best” dictionary wrt this corpus)

dict.	size	a55-kb	a74-kb	a89-kb	n00-kb	n23-kb	n30-kb	n50-kb	n73-kb	n86-kb	n98-kb	t06-kb	t09-kb	t15-kb
a55-dln	60,876	0.6516	0.7081	0.7614	0.7692	0.5866	0.5836	0.5667	0.5454	0.5374	0.5210	0.4798	0.4725	0.5678
a74-s01	32,920	**0.8219**	**0.8629**	**0.8931**	**0.8950**	**0.8299**	**0.8264**	0.8085	0.7755	0.7640	0.7452	0.6949	0.6877	0.7885
a89-s06	40,619	0.7652	0.8144	0.8497	0.8539	0.8189	0.8159	0.7966	0.7633	0.7503	0.7308	0.6803	0.6708	0.7747
n00-s07	67,920	0.6063	0.6812	0.7188	0.7087	0.7657	0.7517	0.7250	0.6986	0.6814	0.6477	0.6037	0.6428	0.6234
n23-s08	72,430	0.5182	0.5890	0.6312	0.6498	0.7881	0.7758	0.7500	0.7226	0.7043	0.6683	0.6228	0.6623	0.6421
n30-bng	48,183	0.5925	0.6476	0.6889	0.6938	0.8278	0.8261	0.8092	0.7782	0.7629	0.7441	0.6908	0.6826	0.7809
n50-s09	133,442	0.3465	0.4275	0.4850	0.5496	0.7050	0.7136	0.6903	0.6575	0.6314	0.5833	0.5368	0.6012	0.5239
n73-s10	125,484	0.3369	0.4090	0.4659	0.5181	0.6935	0.7127	0.7234	0.7106	0.6874	0.6443	0.5921	0.6366	0.6065
n86-s11	105,305	0.3312	0.3897	0.4484	0.4732	0.7078	0.7355	0.7910	0.8187	0.8130	0.7874	0.7232	0.7096	0.8256
n98-s12	108,565	0.3116	0.3693	0.4279	0.4531	0.6926	0.7217	0.7838	0.8217	0.8215	0.7979	0.7334	0.7200	0.8361
t06-s13	112,589	0.3009	0.3580	0.4168	0.4424	0.6845	0.7163	0.7871	0.8438	0.8534	0.8372	0.7738	0.7604	0.8732
t09-so1	88,911	0.3328	0.3865	0.4373	0.4611	0.7238	0.7539	**0.8160**	**0.8589**	**0.8671**	0.8563	0.8045	0.7931	0.8897
t15-s14	114,165	0.2903	0.3437	0.3996	0.4243	0.6610	0.6928	0.7676	0.8403	0.8609	**0.8608**	**0.8065**	**0.7953**	**0.8975**

Table 6: SAOLhist Plus dictionary coverage of the KB news text datasets (legend: “a55”, etc. = ‘1855’, etc.; “n00”, etc. = ‘1900’, etc.; “t06”, etc. = ‘2006’, etc.; boxed value = “best” corpus wrt this dictionary; **boldfaced** value = “best” dictionary wrt this corpus)

corpus	size	a55-dln	a74-s01	a89-s06	n00-s07	n23-s08	n30-bng	n50-s09	n73-s10	n86-s11	n98-s12	t06-s13	t09-so1	t15-s14
a55-kb	211,381	.1876	.1280	.1470	.1948	.1776	.1350	**.2187**	.2000	.1650	.1600	.1603	.1400	.1568
a74-kb	365,473	.1179	.0777	.0905	.1266	.1167	.0854	**.1561**	.1404	.1123	.1097	.1103	.0940	.1074
a89-kb	668,532	.0693	.0440	.0516	.0730	.0684	.0496	**.0968**	.0875	.0706	.0695	.0702	.0582	.0682
n00-kb	853,460	.0549	.0345	.0406	.0564	.0551	.0392	**.0859**	.0762	.0584	.0576	.0584	.0480	.0568
n23-kb	461,975	.0773	.0591	.0720	.1126	.1236	.0863	**.2036**	.1884	.1613	.1628	.1668	.1393	.1634
n30-kb	504,672	.0704	.0539	.0657	.1012	.1113	.0789	**.1887**	.1772	.1535	.1553	.1598	.1328	.1567
n50-kb	611,202	.0564	.0435	.0529	.0806	.0889	.0638	**.1507**	.1485	.1363	.1392	.1450	.1187	.1434
n73-kb	961,566	.0345	.0266	.0322	.0493	.0544	.0390	.0912	.0927	.0897	.0928	.0988	.0794	**.0998**
n86-kb	912,308	.0359	.0276	.0334	.0507	.0559	.0403	.0923	.0946	.0938	.0978	.1053	.0845	**.1077**
n98-kb	962,249	.0330	.0255	.0309	.0457	.0503	.0373	.0809	.0840	.0862	.0900	.0980	.0791	**.1021**
t06-kb	547,052	.0534	.0418	.0505	.0750	.0825	.0608	.1309	.1358	.1392	.1456	.1593	.1307	**.1683**
t09-kb	523,504	.0549	.0432	.0520	.0834	.0916	.0628	.1532	.1526	.1427	.1493	.1635	.1347	**.1734**
t15-kb	1,396,373	.0248	.0186	.0225	.0303	.0333	.0269	.0501	.0545	.0623	.0650	.0704	.0567	**.0734**

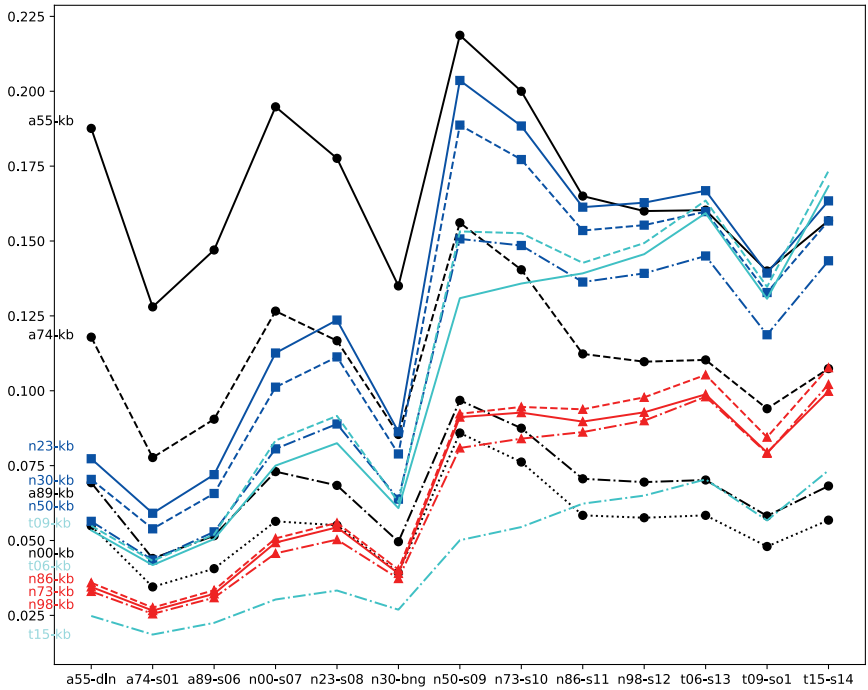


Figure 6: SAOLhist Plus dictionary (x axis) coverage (y axis) of the KB news text datasets (legend: “a55”, etc. = ‘1855’, etc.; “n00”, etc. = ‘1900’, etc.; “t06”, etc. = ‘2006’, etc.)

edition of SAOL in fact managed to capture a Swedish core vocabulary even in the absence of any kind of corpus-linguistic support, which arguably plays a crucial role in the case of SO (2009) and SAOL 14. It is noteworthy in this connection that the most voluminous dictionaries in this set, SAOL 9 and 10, rank among the lowest achievers when it comes to lexical saturation.

The inverse of lexical saturation is *corpus coverage* – how much of the vocabulary in a corpus can be found in a dictionary. Table 6 and Figure 6 show this for the KB corpora, but of course only for those text word types that have received a morphological analysis. Again taking the year 1923 (i.e., the KB subcorpus n23-kb) as example, Table 6 shows that the dictionary with the highest coverage in this case is SAOL 9 (n50-s09), at 0.2036, i.e., slightly over 20%. This is intended to indicate how well the vocabulary of the texts is reflected in the dictionaries, which is the natural question in a lexicographical context (rather than token coverage, which ought to be higher, but which we have not calculated). Note however, that this is a fraction of the *analyzed text word types* in the dataset. The latter in turn make up only about 22%

of the filtered text types,[14] which constitute about 29% of all types in the dataset.[15] The distribution of the boxed values (which corpus is best covered by a particular dictionary) in Table 6 is most likely due to the uneven corpus sizes, with the "winner" in almost all cases being the smallest – by a good margin – 1855 corpus.

With the boldfaced values (which dictionary accounts best for the vocabulary in a corpus) in Table 6, we see an interesting partial mirror image of the lexical saturation figures in Table 5, namely that the largest dictionary SAOL 9 comes out on top from 1855 up to and including its publication year 1950. For the remainder of the period (1973–2015), SAOL 14 (2015) emerges as the winner when it comes to corpus coverage.

7 Summing up and looking ahead

7.1 Concrete results

The original SAOLhist project was – like so much research – opportunistically born out of serendipity: several recent editions of SAOL were available in a useful digital format for completely independent reasons.

In a similar vein, the existence of SAOLhist and the reunification of the lexicographical and language technological strands of Språkbanken Text's activities have created a strong affordance to take this work to a new level: SAOLhist Plus.

Concretely, in SAOLhist Plus we have extended SAOLhist in two ways:

1. An additional dictionary is included among the lexical data, raising the number of lexicons to 13, adding an additional timepoint to this diachronic resource. This is Bring's thesaurus, a Swedish version of Roget's thesaurus published in 1930, that follows the structure of Roget closely. It contains 51,589 unique lemma–POS pairings;
2. Using morphological analyzers for 19th and 21st century Swedish, we add corpus lempos frequencies for the publication years of the 13 dictionaries from a large collection of digitized newspapers.

Other concrete and lasting outcomes of the work described in this chapter are *dm*, i.e., the morphological analysis and generation system for 19th century Late Modern

14 But as shown earlier, they cover 91% of the text tokens remaining after filtering.

15 Hence, this coverage figure could instead have been given as 0.0444 (of all filtered types) or 0.0130 (of all types in the dataset).

Swedish based on Dalin's dictionary, and the 19th century Swedish fullform and compounding lexicons generated using this system.

7.2 Conclusions

The work described in this chapter is more about (enabling) methodology than about presenting a piece of concrete historical linguistic research. It is very clear that the proof-of-concept investigations described in Section 6 would have been impossible to conduct without access to the full SAOLhist Plus dataset. The large-scale diachronic dictionary comparisons described in Section 6.1 could not have been made using the existing SAOLhist web interface. In the same way a prerequisite for the lexical saturation and corpus coverage calculations presented and discussed in Section 6.2 was full access to both the lexical and the corpus data included in SAOLhist Plus. Full access to the raw datasets is a general condition for being able to formulate and address new research questions by applying language technological and corpus linguistic methods to all kinds of language data. This does not mean that there is no need for web interfaces such as that available for SAOLhist or for Språkbanken Text's various research platforms described elsewhere in this volume (see Chapters 10, 11, and 15 in this volume), but such interfaces always offer a limited range of options and almost by definition are not suitable for addressing new research questions, especially not "bird's-eye-view" questions looking for broader trends in the data. For this reason the need for full data access is taken as axiomatic in language technology, not only to particular researchers, but to everybody and always, since research results should be reproducible.

7.3 For the future

In addition to a general need for a revision of the dictionary data in SAOLhist already mentioned, for instance checking part-of-speech information thoroughly, the work described in this chapter has resulted in two sets of desiderata for improvements and extensions, that are described in the following sections.

7.3.1 Morphological analysis

First, as opposed to many of the lexical resources included in Språkbanken's Lexical Research Infrastructure, the Dalin fullform lexicon is (still) what we could call a *"superlexeme" lexicon*, where lexeme distinctions are *not* made on the basis of sense

distinctions. The latter has been the strategy adopted in Saldo, where one lexeme – called *lemgram* in the Saldo context (see Chapter 6 in this volume) may exhibit a set of forms that are a proper subset of those of another lexeme, because they reflect different senses of the lemma in question. For instance, Swedish deverbal nouns in *-(a)nde* and *-(n)ing* exhibit more or less the same range of senses as the corresponding English derivations in *-ing* (Holmer 2022). Thus, the noun *målning* 'painting' conveys both an action noun and a result noun sense, and in Saldo these two senses correspond to two different lemgrams, one – the result noun – including plural forms and the other lacking them. In Dalin the more inclusive lexeme is linked to both these senses, and the information about the lack of plural forms in the action noun will consequently need to be conveyed by some other means. As already mentioned, it would generally be desirable to be able to track dictionary word senses through the long twentieth century, and not only lexemes.

Further, as mentioned in Section 5, the inflectional information in Dalin's dictionary is not entirely reliable, and consequently the *dm* morphology is in need of a thorough review, including formal evaluation of its accuracy.

Finally, even though the *dm* and *s3m* morphologies between them cover most of the lexicographical long twentieth century, there is in fact a gap in the middle, namely the period falling approximately between 1906 and 1950, when the spelling corresponded to the modern norm used in *s3m*, but number indexing in verbs (as in *dm* but not in *s3m*) remained in force in the written standard language (Teleman 2003: 144–149), i.e., there were forms like *blevo* [become.PST.PL], that are not recognized at all in *s3m*, while the corresponding *dm* form is *blefvo*, with the older spelling. We are now planning for a diachronic fullform lexicon covering the same timespan as SAOLhist Plus, tentatively named *Swelex2c*, that would combine the information present in *dm* and *s3m* with verbal paradigms for the period 1906–1950, adding the vocabulary of Bring's (1930) thesaurus, and adding period indications to paradigm slots, i.e., to which of the three periods – Dalin, Bring, or/and Saldo – a particular form belongs.

7.3.2 Corpus data

The KB news corpus data were chosen because they cover the whole timespan of the SAOLhist Plus dictionaries without gaps, but this dataset is limited to one genre. We also have used merely a very small fraction of the available data in the studies described in this chapter. Only 13 years out of a total of 172 have been included in SAOLhist Plus so far. Even so, the amount of raw data is already quite daunting: initially 150 million word types, subsequently reduced in a filtering step to 45 million. Nevertheless, one line of future work will be to add more years of KB data

to the dataset, e.g., based on hypotheses about the temporal relationships between dictionary and corpus vocabularies at different time points in history

The KB news corpus represents one genre only, and especially for the 19th century, news text has been less lexicographically important than today (cf. Teleman 2005: 1971–1972).

There are two other suitable corpora that we aim to include in a future version of SAOLhist:

SPF *Svensk prosafiktion 1800–1900* 'Swedish fiction 1800–1900' (Språkbanken Text 2024b). This corpus contains digitized versions of all works of fiction published in Swedish in the years 1800, 1820, 1840, 1860, 1880, and 1900, 295 texts in total comprising upwards of 15 million words (slightly over 16 million tokens).

SAOB1950 This is a mixed-genre corpus compiled for the work on the large historical Swedish Academy Dictionary (SAOB), with digitized books in Swedish published in the period 1950–2007, comprising about 50 million tokens in total (Språkbanken Text 2024a).

Like the KB corpus, both these corpora also contain OCR errors. Hence, corpus filtering needs improving, e.g., for separating run-together words, as mentioned above.

References

Adesam, Yvonne, Peter Andersson, Lars Borin & Gerlof Bouma. 2021. A lexical resource for computational historical linguistics. In *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, 98–121. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.04ade.

Adesam, Yvonne, Dana Dannélls & Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive KubHist. *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN)*. 9–17.

Allén, Sture. 1967. *Studier över nusvenskans vokabulärsystem* [Studies on the vocabulary system of Modern Swedish]. Research report. Gothenburg: Department of Scandinavian Languages, University of Gothenburg.

Allén, Sture. 1981. The lemma-lexeme model of the Swedish lexical data base. In Burghard B. Rieger (ed.), *Empirical semantics II*, 376–387. Bochum: Brockmeyer.

Almqvist, Carl Jonas Love. 1840. *Svensk språklära* [Swedish grammar]. 3rd edn. Stockholm: M. Wirsells förlag.

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite state morphology*. Stanford: CSLI Publications.

Bring, Sven Casper. 1930. *Svenskt ordförråd ordnat i begreppsklasser* [Swedish vocabulary arranged in conceptual classes]. Stockholm: Hugo Gebers förlag.

Dalin, Anders Fredrik. 1850. Förord [Preface]. In *Ordbok öfver svenska språket. Vol. I*, 2–20. Stockholm: Self-published.

Dalin, Anders Fredrik. 1850–1853. *Ordbok öfver svenska språket* [Dictionary of the Swedish language]. *Vol. I–II*. Stockholm: Self-published.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (eds.). 2021. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins. DOI: 10.1075/nlp.14.

Dannélls, Dana, Torsten Johansson & Lars Björk. 2019. Evaluation and refinement of an enhanced OCR process for mass digitisation. *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN)*. 112–123.

Diderichsen, Philip, Anna Sofie Hartling, Anne Kjærgaard & Anna Kristiansen. 2015. I ulige linje fra *Linje* til *linje*: Og andre nedslag i Retskrivningsordbøger gennem historien [Along different lines from *Linje* to *linje*: And other examples from orthographic dictionaries through history]. In Dorthe Duncker, Eva Skafte Jensen & Ole Ravnholt (eds.), *Rette ord: Festskrift til Sabine Kirschmeier-Andersen i anledning af 60-årsdagen*, 97–107. Bogense: Dansk Sprognævn.

Enberg, Lars Magnus. 1836. *Svensk språklära utgifven av Svenska Akademien* [Swedish grammar published by the Swedish Academy]. Stockholm: A.G. Hellsten.

Gellerstam, Martin. 2009. SAOL i många upplagor [SAOL in many editions]. In Martin Gellerstam (ed.), *SAOL och tidens flykt: Några nedslag i ordlistans historia*, 53–83. Stockholm: Norstedts.

Haapamäki, Saara. 2002. *Studier i svensk grammatikhistoria* [Studies in Swedish grammar history]. Turku: Åbo Akademi University Press.

Hannesdóttir, Anna Helga. 1998. *Lexikografihistorisk spegel: Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga* [History of lexicography reflected: The development of monolingual Swedish lexicography from the bilingual]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Hobsbawm, Eric. 1962. *The age of revolution: Europe 1789–1848*. London: W&N.

Holmer, Louise. 2012. SAOLHist: Alla upplagor av SAOL i en och samma databas [SAOLHist: All editions of SAOL in a single database]. *Nordiska studier i lexikografi* 11: 287–295.

Holmer, Louise. 2022. *Neutrala substantiv på* -ande *i text och ordbok* [Deverbal neutral nouns ending in *-ande* in text and dictionary]. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Holmer, Louise, Sven-Göran Malmgren & Monica von Martens. 2016. SAOLhist.se: För allmänt och vetenskapligt bruk [SAOLhist.se: Serving the public and research]. In *Nordiske Studier i Leksikografi 13*, 349–358.

Hulden, Mans. 2009. Foma: A finite-state compiler and library. *Proceedings of the Demonstrations Session at the 12th Conference of the European Chapter of the ACL (EACL) 2009*. 29–32.

Hulden, Mans. 2022. Finite-state technology. In Ruslan Mitkov (ed.), *The Oxford handbook of computational linguistics*, 230–254. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199573691.001.0001.

Ingvarsson, Jonas, Daniel Brodén, Lina Samuelsson, Victor Wåhlstrand Skärström & Niklas Zechner. 2022. The new order of criticism: Explorations of book reviews between the interpretative and algorithmic. *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB) 2022*. 228–234.

Malmgren, Sven-Göran. 1988. Almqvist, Dalin och den svenska definitionsordbokens födelse [Almqvist, Dalin and the birth of the Swedish definition dictionary]. In Gertrud Pettersson (ed.), *Studier i svensk språkhistoria*, 195–213. Lund: Lund University Press.

NEO. 1995. *Nationalencyklopedins ordbok* [The Dictionary of the National Encyclopedia]. Höganäs: Bra böcker.

Pettersson, Eva & Lars Borin. 2022. Swedish diachronic corpus. In Darja Fišer & Andreas Witt (eds.), *CLARIN: The infrastructure for language resources*, 561–585. Berlin: De Gruyter Mouton.

Ralph, Bo. 1992. The older dictionaries as sources for Nordic language history. *The Nordic languages and modern linguistics 7: Proceedings of the Seventh International Conference of Nordic and General Linguistics*. 493–509. DOI: 10.18602/frodskapur.vi.805.

SAOB. 1898–2023. *Svenska Akademiens ordbok* [The Swedish Academy Dictionary]. Lund: Gleerups.

SAOL 14. 2015. *Svenska Akademiens ordlista* [The Swedish Academy Glossary]. 14th edn. Stockholm: Norstedts.

SAOLhist. 2013. *SAOLhist*. [Online resource] Stockholm: Svenska Akademien and University of Gothenburg. https://spraakbanken.gu.se/saolhist/.

Schlyter, Carl J. 1887. *Ordbok till samlingen af Sweriges gamla lagar* [Dictionary accompanying the collection of Sweden's ancient laws]. Lund: Berling.

SO. 2009. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. Stockholm: Svenska Akademien.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

SOB. 1986. *Svensk ordbok* [Swedish dictionary]. Solna: Esselte studium.

Söderwall, Knut Fredrik. 1884–1918. *Ordbok öfver svenska medeltids-språket: Vol. I–III* [Dictionary of medieval Swedish: Vols. I–III]. Lund: Svenska fornskriftsällskapet.

Språkbanken Text. 2024a. *SAOB1950*. [Data set]. DOI: 10.23695/ZPH6-EN76.

Språkbanken Text. 2024b. *Svensk prosafiktion 1800–1900*. [Data set]. DOI: 10.23695/VD9R-3T41.

Swedberg, Jesper & Lars Holm. 2009. *Swensk ordabok: Utgiven efter Uppsala-handskriften, med tillägg och rättelser ur övriga handskrifter, av Lars Holm* [Swedish dictionary: Published on the basis of the Uppsala manuscript, with additions and corrections from other manuscripts, by Lars Holm]. Skara: Stifts- och landsbiblioteket i Skara.

Teleman, Ulf. 2003. *Tradis och funkis: Svensk språkvård och språkpolitik efter 1800* [Trad and mod: Swedish language cultivation and language policy after 1800]. Stockholm: Norstedts.

Teleman, Ulf. 2005. Language cultivation and language planning II: Swedish. In Oskar Bandle, Kurt Braunmüller, Ernst Håkon Jahr, Allan Karker, Hans-Peter Naumann & Ulf Teleman (eds.), *The Nordic languages: An international handbook of the history of the North Germanic languages*, 1970–1983. Berlin: De Gruyter.

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska Akademiens grammatik* [The Swedish Academy grammar]. Stockholm: Norstedts.

Kristian Blensenius and Benjamin Lyngfelt

14 Network relations in the Swedish ConstructiCon

Abstract: This chapter discusses the development of a network model for the Swedish ConstructiCon (SweCcn). SweCcn is a digital repository of Swedish construction descriptions. It is a practical implementation of construction grammar, offering over 400 entries that demonstrate a variety of constructional patterns. The constructiCon incorporates both formal and functional descriptions while also striving for simplicity for practical use in linguistics, language technology, and language education. The chapter outlines the network model being developed for SweCcn, addressing relations between constructions in several dimensions, particularly focusing on its system of association types. It also discusses the relationship between reference constructiCons and cognitive constructiCons.

Keywords: construction grammar, constructiCon, constructicography, construction networks, taxonomies, associations

Acknowledgments: The research presented in this chapter is funded by Riksbankens Jubileumsfond (grant P21-0473), and the work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161). Thank you. Parts of the content have been presented at seminars in Gothenburg, Erlangen, and Tokyo, and benefitted from helpful comments at these occasions. Thank you. The network model we present is developed in collaboration with the rest of the SweCcn team, who have also given helpful feedback on a previous version of this chapter: Maia Andréasson, Linnéa Bäckström, Steffen Höder, Peter Ljunglöf, and Jonatan Uppström. Thank you.

Kristian Blensenius, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: kristian.blensenius@gu.se
Benjamin Lyngfelt, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, e-mail: benjamin.lyngfelt@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-014

1 Introduction

ConstructiCons, specifically reference constructiCons, are digital repositories of construction descriptions.[1] They are applied instantiations of the construction grammar idea of language as a cognitive network of constructions: a mental constructiCon.[2] Thus, the term *constructiCon* is polysemous in the same way as *grammar* and *lexicon*: on the one hand, a theoretical conception of a linguistic system, on the other hand, a corresponding descriptive resource. Likewise, the enterprise of constructing reference constructiCons, *constructicography*, is to construction grammar what lexicography is to lexicology.

This parallel between constructicography and lexicography is also reflected in analytical and descriptive practice, as the emerging field of constructicography is influenced by and builds on its older and more established associate lexicography, particularly with regard to methodology and description formats. There are, however, crucial differences between the two fields, due to the nature of their respective objects of investigation. Lexicography deals with lexical units, which are described and organized in terms of head words (or *lemmas*). These head words are identified, and in some respects defined, by their written form and may therefore be organized alphabetically.

Construction entries, on the other hand, are typically multi-word units with one or more schematic elements, as illustrated by English examples in (1). Hence, there is not always a given written form to identify them by, nor any consistently applicable and transparent naming principle to adhere to. This means that an alphabetical organization is not very helpful.

(1) a. X *BY* X (examples: *day by day*; *step by step*)
b. N X:S *A* Y (examples: *two times a day*; *50 miles an hour*; *10 euros a head*)
c. *WHAT'S* X *DOING* Y? (examples: *What's that fly doing in my soup?*; *What's this scratch doing on my car?*)

Furthermore, as constructions may vary substantially in both form and function, they may be grouped in different ways and at different levels of abstraction. For example,

1 The capital C inside *constructiCon* serves to avoid confusion with the visually similar word *construction*. Another notation used for the same purpose is the hyphenated form *construct-i-con* (e.g., Goldberg 2003; Hilpert 2019).

2 The terminological distinction between *mental constructiCons* and *reference constructiCons* was coined by Herbst (2019). Mental constructiCons have also been labeled *cognitive constructiCons* (Hilpert, Lyngfelt & Torrent n.d.), and we use both terms interchangeably. For a thorough account of the constructionist notion of a mental/cognitive constructiCon, see, e.g., Diessel (2023).

should the set of rate expressions in (1b) be further divided by form ([N X:s *A* Y] vs. [N X:s *PER* Y]), by function (time proportions vs. distribution over entities), by some combination thereof (which formal variants are associated with which functions?), or simply lumped together? Such complexities, even more pertinent at higher levels of abstraction, make the ever-present issues of categorization and organization increase exponentially and go far beyond traditional lumper-splitter concerns (cf. Lyngfelt et al. 2022: 108f. on the resulting consequences for cross-linguistic comparison).

Consequently, the traditional lexicographic tools for identifying, delimiting, labeling and organizing entries are not sufficient to handle the needs of constructicography. Instead, we look to the construction grammar conception of a (mental) constructiCon as a construction network and develop corresponding tools for organizing reference constructiCons as well. In this chapter, we present the network structure being developed for the Swedish ConstructiCon at Språkbanken Text, with a particular focus on its system of association types.

The chapter is organized as follows: After briefly introducing the Swedish ConstructiCon (SweCcn) in Section 1.1, we discuss construction networks as treated in construction grammar in Section 2. The overall network model of SweCcn is presented in Section 3. In Section 4 we zoom in on the treatment of association types, and Section 5 addresses external network relations, i.e., various types of links to other resources, both within and external to Språkbanken Text. Section 6 wraps up the chapter with a concluding discussion.

1.1 The Swedish ConstructiCon

The Swedish ConstructiCon (SweCcn)[3] is an open online resource for Swedish construction descriptions, integrated into Språkbanken's open lexical infrastructure Karp (see Chapter 11 in this volume).[4] It serves as a resource for linguistics, language technology and educational applications, especially for second and foreign language pedagogy.

SweCcn comprises more than 400 entries, demonstrating a variety of constructional patterns. While the current collection tends to emphasize partially schematic constructions situated "between" grammar and lexicon, the ambition is to account for constructions across nearly the entire grammar–lexicon continuum. This resource is a practical implementation of construction grammar (CxG), integrating methodologies from lexicography. Compared to the detailed and elaborate construction analyses usually found in the CxG literature, construction entries in SweCcn are

3 https://spraakbanken.gu.se/karp/?mode=konstruktikon (last accessed: April 4, 2025)

4 For a detailed account of the relationship between SweCcn and Karp, see Lyngfelt et al. (2018).

brief, simplified approximations, in order to comply with the practical purposes of the resource. Thus, an entry in SweCcn may be seen as a hybrid of a CxG description and a dictionary entry. See Section 3.2 for an example of a SweCcn entry.

By accounting for patterns that combine grammatical and lexical properties, constructicography bridges the traditional gap between grammar and lexicon. This fusion, however, is not without challenges, as grammar and lexicography have historically evolved with differing objectives. Lexicography provides accounts of the meanings of words and to some extent multi-word expressions, along with their morphology and other features, whereas grammar typically addresses larger units and more general patterns, involving considerations of constituency, word order, etc. To be able to jointly treat both kinds of phenomena in a single resource and by the same kind of unit, *constructions*, constructicography has to incorporate these two somewhat diverging linguistic traditions. Moreover, while relationships between dictionary entries are primarily lexico-semantic, e.g., relations like synonymy and hyperonymy such as *move* vs. *walk*, relations between constructions are more complex, encompassing not only semantic but also lexical, morphological, and syntactic ones. For example, the differences between comparison constructions expressing similarity (e.g., as X as Y) and dissimilarity (X-er than Y), while fairly straight-forward, involve both lexicon, morphology and (to a lesser extent) syntax.

2 Construction networks

While the idea of language as a network of constructions has been central to construction grammar and related frameworks since the 1980s, it has evolved considerably over the decades. An early formulation is that of Fillmore (1988):

> The grammar of a language can be seen as a repertory of constructions, plus a set of principles which govern the nesting and superimposition of constructions into or upon one another (Fillmore 1988: 37).

The "repertory" that Fillmore proposed was later dubbed a constructiCon (by Jurafsky 1991). In early (Berkeley) construction grammar, its internal structure was mainly accounted for in terms of inheritance relations (see Section 2.1). Over time, the notion of a construction network became more pronounced and more elaborate, encompassing more kinds of relations (see Section 2.2).

The "nesting and superimposition of constructions", on the other hand, have not received similar attention. When addressed at all, the former has typically been treated in terms of *unification* and *instantiation* (cf. Fillmore & Kay 1995) and the latter in terms of *coercion* (e.g., Michaelis 2004). For the most part, however, the

theory development in construction grammar has been concerned with the constructiCon. The combinatory part, which may perhaps be called *constructional syntax* (Andréasson & Lyngfelt n.d.), is rarely mentioned. Hence, the overall direction of constructionist linguistics in the 21st century is better illustrated by quotes like the following:

> The totality of our knowledge of language is captured by a network of constructions: a 'construct-i-con'. (Goldberg 2003: 219)

In the gradual development of different conceptions of construction networks, one may distinguish three stages: inheritance networks, association networks, and deconstructionist networks.

2.1 Inheritance networks

Inheritance networks pertain to taxonomic relations, construed as top-down taxonomies in which more specific constructions inherit properties from more general ones. This was the main structuring principle in early construction grammar and remains a central component in present-day construction networks. In the construction grammar literature, the discussion of inheritance relations has mainly been concerned with two issues: complete vs. default inheritance and redundant vs. reductionist representation.

The original Berkeley model advocates complete inheritance: "When one construction inherits another, the first contains all the information of the second and – in the non vacuous case – more" (Kay & Fillmore 1999: 7). By contrast, inheritance by default, also called normal inheritance, means inheriting all features "that do not conflict with its own specifications" (Goldberg 1995: 70). Thus, complete inheritance only applies to (proper) subtypes, whereas default inheritance allows for inheriting constructions to have idiosyncratic properties that deviate from the (not necessarily proper) supertype.

SweCcn takes a default inheritance approach, to incorporate constructions that appear to be specific variants of more general constructions, although exhibiting some deviating idiosyncrasies. A case at hand is the family of adjective-as-nominal constructions, as exemplified by *det okända* 'the unknown' and *de äldre* 'the elderly'. These constructions are arguably a kind of noun phrases, with respect to both formal and functional properties – except for the striking feature of lacking a head noun (or pronoun). On a model of complete inheritance, they do not qualify as noun phrases. On default inheritance, they are simply a special kind of noun phrase, thus open for

adnominal modification and eligible for noun phrase slots in other constructions, etc., by inheritance.

The issue of redundant representation concerns whether shared properties are represented in all constructions they apply to or merely at the highest node where they apply. It plays out somewhat differently in linguistic theory and constructicographic practice. In the theoretical domain, it is a question of cognitive plausibility (we will return to this in Section 2.2 in relation to the so-called *fat node problem*; Hilpert 2021). In applied practice, the degree of redundancy is a matter of database design and presentation in a user interface. In SweCcn, we have opted for a fairly high degree of redundancy, in order for each construction entry to be readable and reasonably self-supporting.

The main limitation of inheritance networks is that they only account for taxonomic relations. Besides the obvious drawback that relations between constructions without a shared parent (i.e., a superordinate construction) do not fit, this may in turn incite the positing of perhaps weakly supported general constructions, in order to provide a superordinate node to inherit from (cf., e.g, Audring 2019). A lesser problem with inheritance is the top-down directionality, which is somewhat at odds with usage-based linguistics, according to which linguistic patterns are established by bottom-up generalization from individual usage events (e.g., Langacker 2000; Diessel 2019). The directionality problem, however, is easily amended by recasting unidirectional inheritance relations as bidirectional taxonomic relations – generalization upwards and instantiation downwards.

2.2 Association networks and deconstructionist networks

The natural next step from inheritance networks is to also address other kinds of relations. Over the years, various network relations have been proposed by different authors, thus gradually building towards a more elaborate, multidimensional model of construction networks: association networks. An influential figure in bringing this development together is Diessel (e.g., Diessel 2019; 2023). In Diessel (2023), the following five types of association are discussed:

- *Taxonomic relations*: associations between more general constructions and their more specific subtypes (cf. inheritance)
- *Sequential relations*: associations of linear order between units, notably between construction elements within a construction (cf. syntagmatic relations)
- *Symbolic relations*: associations between a formal pattern and a meaning or function (i.e. the constituting feature of a 'construction' in the construction grammar sense)

– *Filler-slot relations*: associations between construction elements (slots) and particular instantiations (fillers) of these elements
– *Horizontal relations*: associations between constructions "at the same level of abstraction" (Diessel 2023: 16). On a loose interpretation of "same level", this category may subsume any relation between constructions that are not hierarchically related in a taxonomy or interact in a filler-slot relation (see Section 4).

Note that this list not only includes associations *between* constructions, but also relations *within* them, such as the symbolic connection between the form and the function of a construction. The internal network structure of constructions is highlighted in Diessel (2019) and is one of the features paving the way for deconstructionist networks (see further below). For a more comprehensive overview of network relations, see Hilpert, Lyngfelt & Torrent (n.d.).

Association networks effectively subsume inheritance networks and may be considered the current standard in construction grammar. While there is extensive discussion about the nature, establishment, maintenance and change of network connections (e.g., Sommerer & Smirnova 2020; Ungerer & Hartmann 2023; Sommerer & Van de Velde 2025), there seems to be a consensus about the idea of multidimensional association networks as the overall structuring principle of (mental) constructiCons.

What has been challenged, however, is the assumption of constructions as the nodes in the network. They are allegedly too complex and too static to be nodes in a dynamic network, where individual properties change gradually and not in sync with each other. On such grounds, Schmid (2020) and others have proposed that associations are all there is and that what we perceive as holistic constructions are merely (dynamic) clusters of associations. Thus, "it's networks all the way down" (Hudson 2015: 692). This view of network structure may be called *deconstructionist*.

In a similar vein, although without abandoning constructions as analytical units, Hilpert (2021) calls attention to what he calls "the fat node problem". He observes that constructions, as holistic units combining formal and functional properties and interrelations, contain too much information to be plausible nodes in a cognitive network. This is even more the case in a model with redundant representation of shared information (cf. the discussion of inheritance in the previous section). Hilpert therefore advocates placing less information in the nodes/constructions and attributing more to the connections between them.

While they may seem incompatible, the difference between construction networks and deconstructionist networks is not so much a matter of disagreement as a matter of different foci. The deconstructionist view is more closely aligned with current knowledge about how the mind works, aiming to account for the socio-cognitive processes involved in language use and development. Constructions, on the other hand, abstract away from individual usage-events to account for conventional usage

patterns. Thus, they are to some extent idealizations, just like any notions pertaining to standard languages are.

In the terminology of van Trijp (2024), construction networks represent an *aggregate* perspective on language, whereas deconstructionist networks align with a *population* perspective. van Trijp acknowledges both perspectives as important for linguistics but suited for different purposes (see the quote below). Boas, Leino & Lyngfelt (2024) make an analogous differentiation between *system-oriented* and *usage-oriented* approaches to language (picturing the difference as a scale rather than a dichotomy).

> The aggregate perspective is well-suited for tasks for which its idealizations are useful, such as the development of reference grammars, cross-language comparisons, or language teaching. The population perspective is better suited for answering questions that involve processes such as language learning, language change, and language usage. (van Trijp 2024: 339)

In building a reference constructiCon for (standard) Swedish, we assume a system-oriented, aggregate perspective. Hence, construction networks are better suited for our purposes than the more close-up view of deconstructionist networks is. This position implies no disagreement with a deconstructionist view; on the contrary, we acknowledge the insights of Schmid, Hilpert and others, and try to take them into account to the extent possible without losing track of the overall purpose of SweCcn. Some degree of idealization may be necessary for the needs of constructicography, but it should certainly not be overdone.

3 Overall network model in SweCcn

In some respects, a reference constructiCon can be seen as an applied instantiation of the idea of a cognitive constructiCon. Analogously, a network model for a reference constructiCon may in some respects be seen as a reflection of the idea of language as a network of constructions.

However, there are also fundamental differences between building a reference database and modeling what is going on in the human mind. The construction descriptions in SweCcn are idealizations both by pertaining to standard Swedish and by being adapted to the condensed format of construction entries. The network model, in turn, is designed to be useful both as an organizing system and as a search tool, in ways that cannot be presumed to (fully) mirror the organization of a human mind or information retrieval from memory. Nevertheless, it is of course desirable to harmonize the application with its theoretical foundation to the extent practically feasible.

Table 1: Relation types and their implementation in SweCcn

Relation type	Implementation in SweCcn
Taxonomic relations	*Taxonomic relations (Inheritance)*
Horizontal relations	*Associations*
Symbolic relations	*Construction-internal information*
Sequential relations	*Construction-internal information*
Meronymic relations	*Construction-internal information*
Pragmatic relations	*Construction-internal information + Associations*
Extra-linguistic relations	*Construction-internal information + Associations*
Polysemy relations	*Construction-internal information* ***or*** *Taxonomic relations*
Filler-slot relations	*Construction-internal information + External links*
Crosslinguistic relations	*External links*

A cognitive constructiCon presumably includes all the linguistic units and patterns stored in the mind of a language user, and all the network relations between them. The ambitions of SweCcn, and hence its network model, are necessarily more modest. At the same time, it also includes some network relations which do not reflect cognitive theory but are relevant for practical purposes. We distinguish network relations in three different domains: relations *between* constructions, relations *within* constructions, and *external* relations to units outside the constructiCon.[5] An overview of relation types and their treatment in SweCcn is presented in Table 1.

In the following, we discuss relations *between* constructions in Section 3.1. In Section 3.2 we turn to relations *within* constructions and provide an overview of external links. A more detailed account of external links is presented in Section 5.

3.1 Relations between constructions

The central domain of the network concerns relations between constructions (within SweCcn). In this domain, we distinguish two major types: *taxonomic relations*, including inheritance, and *associations*. The taxonomic relations are addressed in detail in Bäckström, Höder & Lyngfelt (n.d.) and the system of associations is treated in Section 4 below. In this section we will give a more concise overview of the two and

5 Another dimension involved is constructional syntax, i.e., how constructions are combined into (sentence) constructs (cf. Fillmore 1988: 37, as quoted in the introduction to this chapter). Although not part of the construction network *per se*, it plays a crucial role for the compatibility between construction entries and the coherence of the network. The treatment of constructional syntax in SweCcn employs the syntax model presented in Andréasson & Lyngfelt (n.d.).

address their roles in the overall network model, including how they may help users to navigate the constructiCon.

Taxonomic relations hold between subtypes and supertypes and, indirectly, between subtypes to the same supertype. They apply recursively over a generality continuum, where it is often an open question which generality levels to represent as construction entries. Accounting for such relations has been on the SweCcn agenda ever since the start (Lyngfelt & Forsberg 2012). It is only recently, however, that we have gone beyond minor taxonomies of closely related constructions in a coordinated effort to build more comprehensive taxonomies, as a major part of the endeavor to turn SweCcn into a coherent construction network.

The taxonomic relations were initially implemented in terms of inheritance, since inheritance networks were the branch standard when we started. Since then, our conception and treatment of these relations have shifted somewhat: it is now less a matter of specific constructions inheriting properties from their more general counterparts and more a matter of establishing chained connections between hierarchically related constructions.

This undertaking is approached both bottom-up and top-down. The majority of entries in SweCcn are quite specific, accounting for partially schematic constructions that "bridge the gap" (cf. Janda et al. 2018) between grammars and dictionaries. Building upwards from there, we identify and characterize the more general patterns that these constructions are specific instances of. At the opposite end, we treat very general clausal, phrasal and argument structure constructions, which occupy the higher levels of the taxonomies and thus are central nodes in the overall network. A coherent overall account of these constructions is also key to the development of a syntax model (and, importantly, *vice versa*; see Section 6.2). Working from both ends, up and down the specificity continuum, the development of construction taxonomies is a dual exercise in lumping and splitting (Bäckström, Höder & Lyngfelt n.d.).

Non-taxonomic relations between constructions are accounted for in terms of **associations**.[6] These associate constructions via shared properties, connecting, for example, coordination constructions (grammar), time expressions (semantics/pragmatics), or constructions containing expletives (construction element). Each construction is associated with a number of properties and each of them thereby connects to other

6 In principle, taxonomic relations are associations as well. Hence, they are to some extent included in our system of association types (see Section 4), when relevant for practical reasons. In particular, they may be included to benefit the resource's search tool, by which construction entries may be found and selected by association (among other features).

constructions associated with the same property. Associations in this model thus correspond more or less to what Diessel (2023) calls *horizontal* relations.

In earlier versions of SweCcn, the associations were treated is terms of *construction types*; each connecting property defined a type, and constructions with a certain property belonged to that type (Lyngfelt et al. 2018). While similar in function, the types and the associations differ in perspective; the shift in terminology mirrors a shift from type classification to network connections.[7] A more detailed account of the system of associations is provided in Section 4 below.

From a user's point of view, these relations help navigate the constructiCon in several ways. Within each construction entry (in the full display mode, see Section 3.2 below), associations and taxonomic relations are presented as lists of links. Each association link will lead to a description of the property defining the association, along with a list of links to all constructions associated with that property. Additionally, taxonomic relations are shown by inheritance links to both superordinate and subordinate constructions.

The associations also function as search criteria to access all construction entries associated to a given property – or a subset thereof, by combinations with other search criteria. In addition, the user may list constructions by association type and select particular construction entries from there. Furthermore, the whole network of relations between constructions can be displayed in a graph visualization, where the user may view a relevant portion of the network by selecting a particular construction or association type and extending the connections as desired (see Section 4 below).

3.2 Relations within constructions and external relations

Relations within constructions do not necessarily have to be treated as components in a construction network; indeed, we have for many years presented this information in SweCcn without explicit reference to a network structure (e.g., Lyngfelt et al. 2018). Nonetheless, the construction descriptions have clearly benefitted from the current, more network-oriented approach. The changes made are a combination

7 It can also be noted that the type system was less systematic. It was not a part of the original design of SweCcn, where constructions were primarily organized by inheritance and grammatical category, but was introduced in response to the need for a more versatile organization. The type system grew organically, by new types being added when needed. When it was time for a systematic revision of the types, we took the opportunity to recast them as associations and, as such, an integral part of the updated network model presented in this chapter. A more comprehensive and systematic type categorization has been developed for the Russian constructiCon (Janda et al. 2020).

Figure 1: SweCcn entry (simple display mode) for the construction RECIPROK_REFL(EXIV). Screenshot from Karp (see Chapter 11 in this volume)

of reinterpreting existing information, on the one hand, and the addition of new network features, on the other. Even the reinterpretation part has made a practical difference (in addition to harmonization with contemporary constructionist theory, cf. Section 2.2) by bringing attention to the interplay between different parts and aspects of the construction descriptions.

To accommodate different user needs, the construction entries come in two display modes: simple (which is the default view) and full. Naturally, there are more network connections in the full display mode, but we will start with the simple(r) view. Besides the name and a short illustration, this view contains a prose definition, a linear structure sketch, and some annotated examples, as illustrated in Figure 1.

In Figure 1, we see the RECIPROK_REFL(EXIV) 'reciprocal_reflexive' entry, provided with the example *De förlovade sig* 'they got engaged' (lit. *they engaged themselves*) in the simple (Sw. *enkel*) view. Starting with the symbolic relation between the form and function, the morphosyntactic form is primarily shown in the structure (*struktur*) sketch (which also expresses some functional information on particular construction elements). The definition field always includes the meaning/function of the construction (the definition in Figure 1 translates to 'mutual interaction between individuals'), while formal characteristics are included to the extent they fit the prose format without making the definition overly complex. Thus, the form-function connection is on the one hand a relation between the structure sketch and the definition, and on the other hand explicitly expressed within the definition, when it benefits the description to do so.

Sequential relations are shown in the structure sketch, expressing the prototypical linear order between the construction elements. Word order variation may be presented as alternative structure sketches, if the alternative order is common enough. Alternatives that are less frequent but still common enough to be acknowledged (with respect to the limited format of a construction entry) are illustrated in the examples.[8]

Meronymic relations between the construction as a whole and its parts, or construction elements (CEs), are manifested by links between different fields of information where the CEs are represented. All example sentences are annotated for CEs, and so is the construction definition to the extent CEs are mentioned. In the examples, the construct (expression) instantiating the construction is marked by a colored background, and each construct element is delimited by brackets. Hovering over a construct element reveals a feature description of the corresponding CE. Likewise, CEs mentioned in the definition are indicated by brackets, and hovering over them reveals the same feature description of the CE.[9]

Turning to the full display mode (Figure 2), more fields of information are shown. Not all fields are specified for all constructions, and only the fields containing information are displayed.

The additional fields present the following kinds of information (for a more detailed account, see Lyngfelt et al. 2018: 82–91):

- organization: associations (Sw. *association*), (grammatical) category (*kategori*), and taxonomic relations (*undertyp till* 'subtype of', linking to superordinate constructions, and *undertyper* 'subtypes', linking to subordinate ones)
- lexical links: keywords (*nyckelord*; lexically specific CEs) and common words (*vanliga ord*; common lexical slot-fillers)
- construction elements (*konstruktionselement*): feature analyses of each CE; this is the information that is revealed by hovering over the corresponding units in the examples and definition
- external links: FrameNet and MoCCA (*Model of Comparative Concepts for Constructicon Alignment*, a model for connecting related constructions in different languages; see Section 5 below)
- additional information: status, comments, references (these fields do not directly concern the network model and will not be addressed further here).

8 For a more detailed account of how constructional variation is handled in SweCcn, see Lyngfelt et al. (2018: 62–66).

9 Notice the terminological distinction between *construct elements*, as parts of the actual expression instantiating a construction, and *construction elements*, which are corresponding parts of the conventional construction pattern instantiated by the construct.

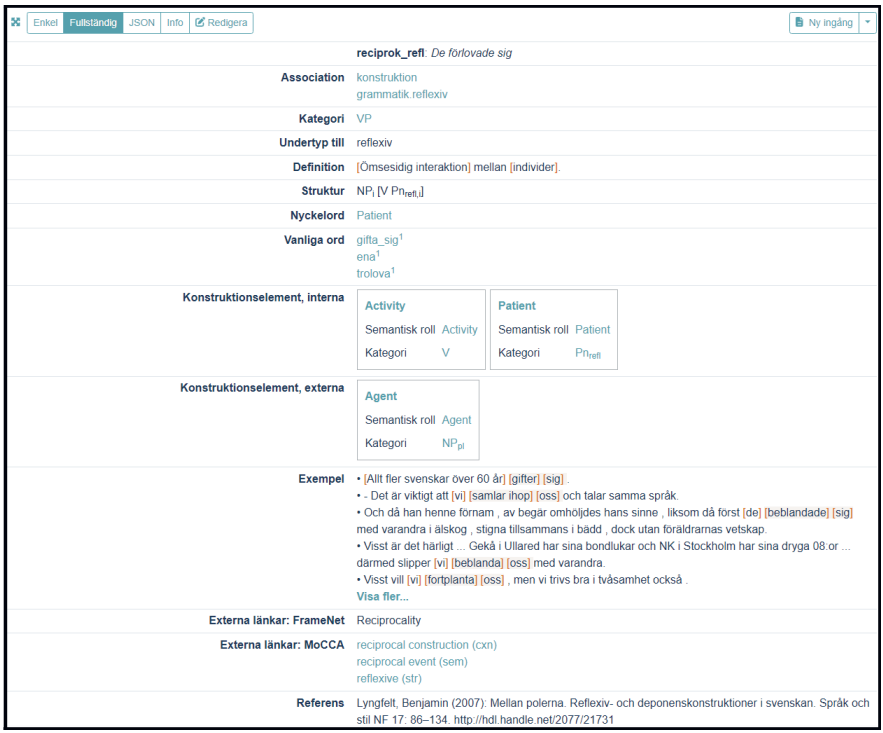


Figure 2: SweCcn entry (full display mode) for the construction RECIPROK_REFL(EXIV). Screenshot from Karp

The organization fields concern relations between constructions, as outlined in the previous section. Under *Association*, as mentioned above, we link constructions to associated properties. By these links the constructions are also connected to other constructions associated with the same property (see further Section 4). Clicking one of the links will lead to a description of the property and a list of links to all constructions associated with that property. Grammatical *category* is in principle a kind of association but is presented separately, partly for historical reasons, and is not yet linked to category descriptions. The inheritance information represents taxonomic relations, linking the construction to both superordinate and subordinate constructions.

Keywords and *Common words* represent lexical filler-slot relations by listing lexically specific CEs and common slot fillers, respectively. All lexical units listed in these fields are linked to the corresponding entries in the lexical resource Saldo and thereby to all other lexical resources in Karp, the lexical infrastructure of Språkbanken Text (see Chapter 6 in this volume). These links are external in the sense

that they connect to units outside SweCcn,[10] but internal from the viewpoint of Karp (see Section 5 below). Information about slot fillers is also provided in the annotated examples, where the first examples illustrate typical instantiations, while later examples are less typical to illustrate the variability and productivity. More systematic constraints on slot fillers are specified in the CE analyses.

To facilitate cross-linguistic comparison, the constructions are linked to *FrameNet*[11] and to *MoCCA*.[12] MoCCA is a model for aligning constructions via language-neutral comparative concepts (CCs), using the set of CCs proposed in Croft (2022). FrameNet frames relate linguistic units by their meaning in terms of frame semantics. By the frame annotation, the constructions are related both cross-linguistically (cf. Global FrameNet; Torrent et al. 2018) and to the Swedish FrameNet (see Chapter 7 in this volume). The connections to FrameNet and MoCCA are treated in Section 5.

Finally, some kinds of relations are treated differently from case to case. Polysemy relations (cf. Goldberg 1995) may be treated in the definition of a construction or by positing several subordinate constructions. Thus, the distinction between polysemy and homonymy, which is fundamental in lexicography, for example, is less so in a multigranular construction network and often boils down to a matter of level of abstraction. Pragmatic relations (cf. Schmid 2020) and extra-linguistic relations are noted in the definition where relevant. If the same relation saliently applies to several constructions, it is treated as an association type.

Some of these associations do not concern relations presumed to be relevant in a cognitive construction network but are included in SweCcn for other reasons. Such an association is 'learner focus', indicating constructions considered to be particularly relevant for second language pedagogy.

4 Association types

As stated in Section 3, constructions in SweCcn are connected via *associations*, which are defined in terms of shared properties of different kinds. The arrangement of construction entries by types of associations has been carried out for several reasons,

10 In constructionist theory, lexical constructions – words – are a subset of the set of constructions in the (cognitive) constructiCon, which in principle makes lexical filler-slot relations internal. In the case of SweCcn, however, in line with constructicographic practice in general, the lexical resources in Karp are external to the constructiCon, for practical reasons.

11 https://framenet.icsi.berkeley.edu/ (last accessed: April 4, 2025)

12 https://github.com/comparative-concepts/ (last accessed: April 4, 2025)

for example to structure the set of constructions, establish order, and identify groups of constructions. Association types also serve a practical function by making the database more accessible to users of SweCcn. For users, association types offer a way to locate constructions that carry certain properties, even when the construction names are not entirely transparent.

Association types provide a versatile grouping mechanism: associations, like their predecessor, *types*, "can be based on any salient property shared by a group of constructions" (Lyngfelt et al. 2018: 58), as long as they have some degree of establishment. *Properties* may refer to a specific constructional function (e.g., resultative) or an included constructional element (e.g., verb particle) and can be either functional (e.g., comparison) or structural (e.g., coordinated structures).[13]

While association types are not limited to a set of specific property categories, they typically reflect simplex properties. For instance, instead of assigning a complex type like "concessive subordinate clause", the constructions in question are associated with two separate properties: "concessive" and "subordinate clause". The association types are primarily intended to be language-specific, designed to primarily benefit the description of Swedish (in contrast to *comparative concepts*, discussed in Section 5).

The set of association types is expanding and undergoing revisions in the developing construction network. The association types have recently been recategorized, reflecting that the constructions are associated through certain categories. In comparison to the network model of Diessel (2019; 2023), our association types resemble what Diessel calls *horizontal relations*, which combine constructions at the same level of specificity. Diessel in turn draws inspiration from Cappelle (2006), who assumes horizontal links to define relations between constructions that *complement* each other. An example of such a link is the one between transitive verb-particle allostructions (Diessel 2023: 59), such as *turn on something* vs. *turn something on*, traditionally treated as alternations.

As mentioned, association types in SweCcn can be based on any salient property shared by a group of constructions, thus providing a highly versatile grouping mechanism. They are divided into the following categories:

– *semantik/pragmatik* 'semantics/pragmatics'
– *grammatik* 'grammar'
– *element* 'element'
– *övrigt* 'other'.

13 The groupings resulting from the association types correspond somewhat to what has been called *families*, *neighborhoods*, and *clusters* of constructions (e.g., Diessel 2023; Endresen & Janda 2020).

The *semantik/pragmatik* 'semantics/pragmatics' association type links constructions that share any salient semantic or pragmatic property. Examples include *semantik/pragmatik.kontrast*, which encompasses diverse constructions with a contrastive meaning like ADJ_MEN_DOCK 'adjective_but_still', for example *(ett) litet men dock hopp* '(a) small but still a hope', and ÖMSOM_X_ÖMSOM_Y 'alternately_X_alternately_Y' including examples like *ömsom löften ömsom hot* 'alternately promises and threats'.

The *grammatik* 'grammar' association type includes constructions that share both formal and functional properties such as *grammatik.ordbildning* 'grammar.word-formation'. This association type includes a simple compound construction, instantiated by, e.g., *husvagn* 'caravan' (lit. house-trailer), as well as a degree-adjective construction instantiated by, e.g., *jättebra* 'good, great' (lit. giant-good).

The *element* 'elements' association type (cf. *subpart links*, Hilpert 2019: 62) connects constructions that include particular construction elements such as verb particles or certain morphemes. For example, the *element.reflexiv* association connects constructions that include a reflexive element (in this case a reflexive pronoun), in principle regardless of whether the construction as a whole expresses reflexive semantics. Thus, it includes semantically less reflexive or non-reflexive constructions such as the RECIPROK_REFL(EXIV), as exemplified in (2), as well as the more conventional reflexive construction TRANSITIV_REFLEXIV 'transitive_reflexive', as exemplified in (3).[14]

(2) *De förlovade sig.*
they got.engaged REFL
'They got engaged.'

(3) *Han tvättar sig.*
he washes REFL
'He washes himself.'

The *övrigt* 'other' association type connects constructions associated for other reasons than shared linguistic properties. For example, *övrigt.inlärningsfokus* 'other.learning focus' concerns constructions considered to be of particular relevance for second language pedagogy. Such connections, which are included in the network for practical rather than linguistic reasons, will not be explored further here.

As mentioned in Section 3.1, the association network can be displayed as a visualization graph. Figure 3 shows a portion of the network in the SweCcn visualization tool[15], focusing on the association type *semantik/pragmatik.aspekt* 'semantics/prag-

14 Glosses in this chapter follow the Leipzig Glossing Rules.

15 https://spraakbanken.github.io/sweccn-graph/sweccn-graph.html (last accessed: April 4, 2025)

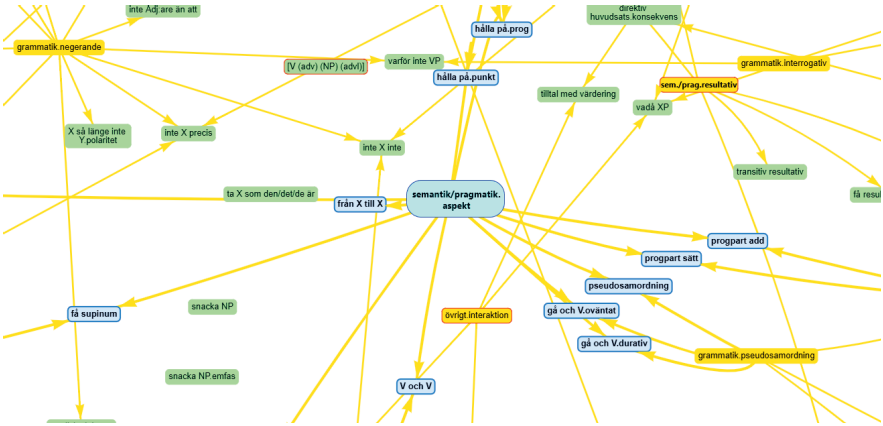


Figure 3: Visualization of a portion of the association network in SweCcn (the image is slightly modified)

matics.aspect'. This association type is located in the middle of the figure, and directly associated constructions – those that could more or less easily fit into the figure – are marked in blue. Constructions not directly associated with the *semantik/pragmatik.aspekt* association type are marked in green; other association types in the figure are highlighted in yellow.

Figure 3 illustrates that *semantik/pragmatik.aspekt* encompasses verbal aspectual constructions such as PSEUDOSAMORDNING 'pseudocoordination' (4) and HÅLLA_PÅ.PROG 'hålla_på.progressive' (5), and iteration constructions like FRÅN_X_TILL_X 'from_x_to_x' (6).

(4) *Hon sitter och skriver på en uppsats.*
she sits and writes on an essay
'She is (sitting and) writing an essay.'

(5) *Hon håller på att skriva en uppsats.*
she holds on to write.INF an essay
'She is writing an essay.'

(6) *Fågeln hoppade från gren till gren.*
bird.DEF jumped from branch to branch
'The bird jumped from branch to branch.'

These constructions are not obviously structurally related, and it is not immediately clear from their names that they can all express some kind of durativity and ongoingness.

All in all, association types capture relations between constructions, relations that are not entirely captured by other network relations in SweCcn. For example, the *element.reflexiv* association type includes constructions which are not semantically reflexive (in the sense that a participant performs an action on themselves). Also, the association type *semantics/pragmatics.aspect* connects diverse structures such as the construction TIDSAVGRÄNSNING_PERFEKTIV.PÅ 'temporal delimitation_perfective.på', which encompasses prepositional phrases functioning as adverbials indicating duration like *på fem minuter* 'in five minutes' as well as verbal reduplication in the durative V_OCH_V 'verb_and_verb' construction exemplified by *går och går* 'goes and goes'. Such associations are not captured by, for example, inheritance relations.

5 External relations

There are links from construction entries to three resources outside SweCcn: Saldo, MoCCA, and FrameNet, which will be addressed in turn in this section.

Saldo is a semantic and morphological lexicon of Swedish and the pivot of the lexical macroresource of Språkbanken (see Chapter 6 in this volume). Links to Saldo are found in the structure sketches, i.e., the linear representation of the grammatical structure of a construction. For example, the construction ADJ_SOM_NOMINAL.ANAFORISK 'adjective_as_nominal.anaphoric', e.g., *Den gula är finast* 'the yellow one is the nicest', includes the structure sketch [den^{1} AP_{def}] in which den^{1} 'the' is the Saldo entry for the definite article. The fields *keywords* (lexically specific elements) and *common words* (words commonly appearing in a certain construction) are also linked to corresponding Saldo entries (see Section 3.2).

In collaboration with constructiCon projects for other languages, for example Japanese and Brazilian Portuguese, the SweCcn project aims at developing cross-linguistic applicability for the constructiCon. The existence of constructiCons for several different languages opens possibilities for multilingual applications, including establishing relationships between similar constructions across languages. Such connections can support machine translation, second-language teaching, etc. A strategy that has previously been explored for aligning constructiCons in different languages is that of direct comparisons between individual constructions to establish equivalencies (e.g., Bäckström, Lyngfelt & Sköldberg 2014). However, apart from being overly time-consuming, such comparisons are often biased toward the source language, requiring additional analyses in the other direction, among other things (discussed in Lyngfelt et al. 2022). Instead, we have turned to a strategy of connecting constructions indirectly, via a language-neutral base of comparison. Such a base of comparison is provided by comparative concepts (CCs, Croft 2022), as used in lin-

guistic typology. By linking constructions to CCs, similarities and differences across languages can be systematically represented without implying full equivalence.

For this purpose, the alignment model MoCCA, *Model of Comparative Concepts for Constructicon Alignment* (Lorenzi et al. 2024), has been developed, based on the set of CCs in Croft (2022). Construction entries in SweCcn are linked to relevant CCs in MoCCA and thereby indirectly to other constructions linked to the same CCs. Thus, the CCs fill a similar function to that of association types in SweCcn, with the difference that our association types are defined with respect to Swedish only (see Section 4 above). The CCs in MoCCA, on the other hand, are defined to be language-neutral. They characterize cross-linguistically relevant properties organized into four categories:

– Constructions (CC-cxn): e.g., *clause*, *pronoun*
– Strategies (CC-str): e.g., *cleft strategy*, *compounding*
– Semantic content (CC-sem): e.g., *agent*, *degree*
– Information packaging (CC-inf): e.g., *contrast*, *topic*.

Constructions and Strategies are hybrid CCs, in the sense that they define combinations of form and function. Constructions define basic form-function pairings (what), while strategies specify particular ways of realizing them in different languages (how). Semantic content and Information packaging are functional CCs, defining the function side of the constructions. Semantic content is the content being expressed (what) and Information packaging concerns how the content is construed (how). Compared to the association types in SweCcn, the hybrid CCs (Constructions and Strategies) correspond to the *grammar* type, whereas the functional CCs (Semantic content and Information packaging) correspond to *semantics/pragmatics* (see Section 4).

When assigning CCs to a Swedish construction, the first step is typically to consider the functional properties of the Swedish construction, since these tend to be more cross-linguistically comparable than form. An example is the assignment of CC to the construction RECIPROK_REFL(EXIV), exemplified in (2) above. It is a construction that expresses mutual interaction between individuals, but unlike typical reflexive constructions, the reflexive element is not referential. We begin by assigning the CC-sem *reciprocal*, which connects to the CC-cxn *reciprocal construction*. Because this relation is expressed using a reflexive pronoun, the construction is also connected to the *reflexive strategy* CC.

As mentioned in Section 3.2, SweCcn also connects to FrameNet and SweFN (see Chapter 7 in this volume) through links to corresponding frames where applicable

(i.e., in the case of frame-evoking constructions; Lee-Goldman & Petruck 2018: 36; Lyngfelt et al. 2018: 68–81; Ohara 2018).[16]

6 Conclusions and outlook

In this chapter, we have presented the network model of the Swedish ConstructiCon (SweCcn), addressing relations *between* constructions, relations *within* constructions, and external relations.

Relations between constructions include taxonomic (inheritance) relations and association relations. Taxonomic relations link specific constructions to superordinate and subordinate ones, while associations connect constructions sharing common properties of different kinds. These associations are not merely of theoretical interest; they also serve a practical purpose in SweCcn. They provide a versatile grouping mechanism, allowing users to find constructions based on for example shared semantic, pragmatic, or grammatical features.

Relations within constructions include symbolic, sequential, and meronymic relations, which highlight connections between a construction's elements. External links connect the constructiCon to resources such as Saldo, MoCCA, and FrameNet. Again, the motivation and purpose are both theoretical and practical, depending on the type of relation.

In the following, we will discuss the proposed network model in relation to a broader theoretical context. In Section 6.1, we discuss if and how reference constructiCons such as SweCcn correspond to the theoretical conception of language as a cognitive constructiCon, hoping that our applied enterprise may be of some benefit to the theoretical tradition it builds on. In Section 6.2, we address how a constructiCon, as an inventory of constructions, relates to an overall construction *grammar*, as well as the mutual dependence between the construction inventory and what may be called constructional syntax (Andréasson & Lyngfelt n.d.; cf. Fillmore 1988). We also make a comment on how a system-oriented project like SweCcn might fit into, and contribute to, the tradition of usage-based construction grammar.

16 There is also an initiative to relate constructions across languages via Universal Dependencies (Weissweiler et al. 2024), as well as ongoing discussions of ways to combine MoCCA, UD and/or FrameNet.

6.1 Reference constructiCons and cognitive constructiCons

From a practical viewpoint, the network model presented in this chapter serves the purpose of organizing the constructiCon database and providing variable ways of searching for construction entries. From a theoretical viewpoint, it is an implementation of the constructionist notion of language as a cognitive network of constructions. This notion has been at the heart of construction grammar theory for decades, inspiring and guiding various kinds of research on relations between constructions. What is still lacking, however, is proof of concept of the idea that language as a whole can be fruitfully treated as a construction network. Can a network like the one proposed here provide such proof of concept?

As noted in Section 3, building a reference database is not the same thing as modeling human cognition. SweCcn is designed to be a resource for linguistics, language technology and language pedagogy. Hence, many of the design choices, regarding content as well as description format, are motivated by the purpose and practical conditions of the resource rather than by theoretical concerns. For example, it is unlikely that the linguistic knowledge of ordinary Swedish speakers includes a classification of certain constructions as particularly relevant for second language education or connections to a cross-linguistic system like MoCCA.[17]

Such practically motivated features aside, however, the network in general is devised in accordance with the constructionist literature (e.g. Diessel 2023), implementing the theory in applied descriptive practice. In that regard, it does show that a coherent global construction network is feasible. To what extent it is also cognitively plausible is up for discussion.

As argued by Schmid (2020) and others, deconstructionist networks are probably a more accurate approximation of cognitive linguistic structure (cf. Section 2). Since constructions, as representations of conventionalized usage patterns, are somewhat idealized conceptions, so are construction networks. This is nothing unique to CxG or constructicography, however. Such idealizations are standard practice in descriptive linguistics, aiming to characterize conventional language patterns in a linguistic community, rather than modeling the minds and the socio-cognitive processes of individual language users. Nonetheless, the idealizations may and should provide fairly accurate characterizations of the linguistic varieties and structures being described, grounded in solid empirical evidence. To the extent these characterizations are also compatible with current knowledge of human cognition, they may be considered cognitively plausible. Hence, although the theoretical conception of cognitive

17 Although multilingual speakers surely have knowledge of relations between structures in different languages (Höder 2018), it would hardly be of the typological scope or language-neutral setup of MoCCA.

construction networks is essentially a theory of the mind, most of the linguistic practice based on it is applied to language phenomena as perceived from the aggregate perspective of a language community.

The theoretical claims driving this research concern language as a whole, while the empirical research itself typically addresses smaller networks of closely related constructions. From the viewpoint of this linguistic practice, a network model such as the one presented here and its implementation in SweCcn does provide proof of concept, of sorts. On the one hand, it provides empirical support from an overall perspective, in contrast to the more detailed but also more narrow support provided by specific case studies. On the other hand, it substantiates the overall theory by implementing it on a general scale. Thus, the constructicographic application not only builds on constructionist theory but may also inform the theory back (Boas, Lyngfelt & Torrent 2019: 49f.).

6.2 From inventory to network to grammar – and back

As noted in Section 2, a construction grammar of a language presumably consists of an inventory of constructions, on the one hand, and "a set of principles which govern the nesting and superimposition of constructions into or upon one another", on the other (Fillmore 1988: 37). The latter set of principles may be called *constructional syntax* (Andréasson & Lyngfelt n.d.). A constructiCon, in its most basic sense, is a representation of a construction inventory. Structured as a network, it also captures relations between the constructions, but it is still all about the inventory. Therefore, it has to be complemented by a constructional syntax, which handles the interplay between constructions, e.g., how they may be combined to form utterances.

What has become increasingly obvious during the process of building a construction network for SweCcn is that the syntax model is not only a complement to the constructiCon; it has turned out to be key to achieving a coherent comprehensive network as well. To be able to connect all the various construction entries in SweCcn one has to account for the most general constructions, the top nodes of the major taxonomies and thus the supertypes of which all the other constructions are subtypes. These constructions are the central nodes of the network. They include the most schematic phrasal and clausal constructions, which constitute basic patterns for how smaller units are combined into larger configurations. Hence, one cannot account for these constructions without having a good idea of how constructions are combined.

Furthermore, for constructions to be combinable they have to be compatible, Consequently, the development of a syntax model, accounting for constructional syntax as well as basic syntactic constructions, serves as a test of the coherence of

the constructiCon. While investigating how different kinds of constructions may be combined, either by straightforward matching or by accommodation of functioning mismatches, one also notes what is required by the construction descriptions to capture relevant possibilities of combination. It is only when you work out the interplay between *different* kinds of constructions, not just comparisons between closely related ones, that you get a good picture of how the various construction descriptions fit together. Thus, the syntax model serves as a heuristic for identifying discrepancies and inconsistencies, as well as providing clues to how they may be fixed.

For instance, on this syntax model, the primary mechanism of clause formation is the combination of a clausal construction with a verbal argument structure construction (ASC; cf. Goldberg 1995; for details, see Andréasson & Lyngfelt n.d.). Thus, a sentence like *Anna likes jazz* is presumably formed by the combination of a declarative clause construction and a transitive ASC (instantiated by *likes*), plus some other constructions. The sentence *Does Anna like jazz?*, accordingly, is formed by combining the same ASC with a polar question construction.

The clausal constructions capture general morphosyntactic patterns such as constituency and word order, whereas ASCs handle lexically dependent properties such as valency. The two construction types represent different perspectives on clause structure, characterize different kinds of linguistic properties, and are rooted in different traditions. Syntactic properties are traditionally represented in terms of discrete formal categories, whereas ASCs were developed in the tradition of cognitive linguistics and functional lexical semantics. As long as the two types of constructions are treated each by themselves there is no strong incentive to harmonize their representations. When they are to be combined in a coherent model, however, they have to be accommodated. Thus, the syntax model benefits coherence and compatibility across different parts of the construction network.

Finally, let us return to the question of how SweCcn related to usage-based linguistics (UBL), since most contemporary work in CxG adopts the label *usage-based construction grammar*. According to UBL, language is a dynamic, complex-adaptive system, linguistic patterns are continuously shaped and reshaped by particular usage events, and constructional properties are partially emergent from the particular usage context (e.g., Diessel 2019; 2023; Langacker 2000). While virtually all constructionists, as far as we can tell, seem to be sympathetic to this view of language, it varies considerably how it is reflected in analytic practice. A common approach is to employ corpus linguistics, capturing patterns of usage and variation at aggregate level (van Trijp 2024; see Section 2.2 above) but abstracting away from the contextual factors and socio-cognitive processes involved in situated usage events.

SweCcn is usage-*based* in the sense that all construction descriptions build on authentic corpus data, but system-*oriented* in the sense that we aim to characterize

conventional usage patterns and relations in (standard) Swedish. There is no disagreement with UBL, only a difference in focus. We also hope and believe that these different perspectives may complement each other in the shared quest of attaining a better understanding of language from a constructionist viewpoint.

Related to the difference in focus is a difference in scope. Aside from a large literature on ASCs, following the influential work of Goldberg (1995), the main area of CxG research is concerned with fairly specific constructions, typically patterns involving both lexical and grammatical properties. On the one hand, this is the origin and the unique selling point of CxG, since such structures are difficult to account for by traditional grammar-and-dictionary model (Hilpert 2019). On the other hand, it aligns with the bottom-up approach of UBL, where concrete patterns are awarded primacy over more abstract ones and high-level generalizations are approached with caution, since the empirical support becomes less substantial the further from concrete utterances you get.

Nonetheless, UBL theory applies to all of language, not just the specific patterns. To substantiate the claim/assumption that language (as a whole) consists of an association network of constructions, one eventually has to address the higher levels of abstraction as well. And in order to do so, some degree of abstract modeling seems to be required. We therefore hope that the overall network model of SweCcn presented here, and its associated syntax model, are not seen as an alternative and incompatible approach to UBL, but rather as a complementary perspective that can make a useful contribution to usage-based construction grammar. If not, we will have to settle for the contribution of the applied resource as such.

References

Andréasson, Maia & Benjamin Lyngfelt. N.d. Clausal constructions and clause formation in Swedish: A constructionist model of clausal syntax. In prep.

Audring, Jenny. 2019. Mothers or sisters? The encoding of morphological knowledge. *Word structure* 12(3): 274–296.

Bäckström, Linnéa, Steffen Höder & Benjamin Lyngfelt. N.d. Taxonomies and argument structure in the Swedish Constructicon. In prep.

Bäckström, Linnéa, Benjamin Lyngfelt & Emma Sköldberg. 2014. Towards interlingual constructicography: On correspondence between constructicon resources for English and Swedish. *Constructions and Frames* 6(1): 9–32.

Boas, Hans C., Jaakko Leino & Benjamin Lyngfelt. 2024. Constructionist views on construction grammar. *Constructions and Frames* 16(2): 169–190.

Boas, Hans C., Benjamin Lyngfelt & Tiago Timponi Torrent. 2019. Framing constructicography. *Lexicographica* 35(1): 41–85.

Cappelle, Bert. 2006. Particle placement and the case for “allostructions”. *Constructions* 1: 1–28.

Croft, William. 2022. *Morphosyntax: Constructions of the world's languages*. Cambridge: Cambridge University Press.

Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.

Diessel, Holger. 2023. *The constructicon*. Cambridge: Cambridge University Press.

Endresen, Anna & Laura A. Janda. 2020. Taking construction grammar one step further: Families, clusters, and networks of evaluative constructions in Russian. *Frontiers in Psychology* 11.

Fillmore, Charles J. 1988. The mechanisms of "construction grammar". *Annual Meeting of the Berkeley Linguistics Society 1988*. 35–55.

Fillmore, Charles J. & Paul Kay. 1995. Construction grammar coursebook. Ms.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *TRENDS in Cognitive Sciences* 7(5): 219–224.

Herbst, Thomas. 2019. Constructicons: A new type of reference work? *Lexicographica* 35(2019): 3–14.

Hilpert, Martin. 2019. *Construction grammar and its application to English*. 2nd edn. Edinburgh: Edinburgh University Press.

Hilpert, Martin. 2021. *Ten lectures on diachronic construction grammar*. Leiden/Boston: Brill.

Hilpert, Martin, Benjamin Lyngfelt & Tiago Timponi Torrent. N.d. The constructicon: Language as a cognitive network of constructions. To appear in Elsevier Encyclopedia of Language and Linguistics, 3rd. edn.

Höder, Steffen. 2018. Grammar is community-specific: Background and basic concepts of diasystematic construction grammar. In Hans C. Boas & Steffen Höder (eds.), *Constructions in contact: Constructional perspectives on contact phenomena in Germanic languages*, 37–70. Amsterdam: Benjamins.

Hudson, Richard. 2015. Book review: The nature of rules, regularities and units in language: A network model of the language system and of language use. *Journal of Linguistics* 51(3): 692–696. DOI: 10.1017/S002222671500016X.

Janda, Laura A., Anna Endresen, Valentina Zhukova, Daria Mordashova & Ekaterina Rakhilina. 2020. How to build a constructicon in five years: The Russian example. *Belgian Journal of Linguistics* 34: 162–175.

Janda, Laura A., Olga Lyashevskaya, Tore Nesset, Ekaterina Rakhilina & Francis M. Tyers. 2018. A constructicon for Russian: Filling in the gaps. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 165–181. Amsterdam: John Benjamins.

Jurafsky, Daniel. 1991. *An on-line computational model of human sentence interpretation: A theory of the representation and use of linguistic knowledge*. University of California, Berkeley. (PhD thesis).

Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *What's X Doing Y?* construction. *Language* 75: 1–34.

Langacker, Ronald W. 2000. A dynamic usage-based model. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 1–63. Stanford: CSLI Publications.

Lee-Goldman, Russell & Miriam R. L. Petruck. 2018. The FrameNet constructicon in action. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 19–39. Amsterdam: John Benjamins.

Lorenzi, Arthur, Peter Ljunglöf, Benjamin Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig & Ely Matos. 2024. MoCCA: A model of comparative concepts for aligning constructicons. *Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable*

Semantic Annotation at the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) 2024. 93–98.

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 41–106. Amsterdam: John Benjamins.

Lyngfelt, Benjamin & Markus Forsberg. 2012. *Ett svenskt konstruktikon: Utgångspunkter och preliminära ramar* [A Swedish constructicon: Points of departure and preliminary frames]. Research report. Gothenburg. University of Gothenburg, Dept. of Swedish.

Lyngfelt, Benjamin, Tiago Timponi Torrent, Ely Edison da Silva Matos & Linnéa Bäckström. 2022. Comparative concepts as a resource for multilingual constructicography. In Kristian Blensenius (ed.), *Valency and constructions: Perspectives on combining words*, 101–129. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Michaelis, Laura A. 2004. Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive Linguistics* 15(1): 1–67.

Ohara, Kyoko. 2018. Relations between frames and constructions: A proposal from the Japanese framenet constructicon. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages*, 141–163. Amsterdam: John Benjamins.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: Usage, conventionalization and entrenchment*. Oxford: Oxford University Press.

Sommerer, Lotte & Elena Smirnova (eds.). 2020. *Nodes and networks in diachronic construction grammar*. Amsterdam: John Benjamins.

Sommerer, Lotte & Freek Van de Velde. 2025. Constructional networks. In Mirjam Fried & Kiki Nikiforidou (eds.), *The Cambridge handbook of construction grammar*. Cambridge: Cambridge University Press.

Torrent, Tiago Timponi, Michael Ellsworth, Collin F. Baker & Ely Edison da Silva Matos. 2018. The Multilingual FrameNet shared annotation task: A preliminary report. *Proceedings of the International FrameNet Workshop 2018: Multilingual Framenets and Constructicons*. 62–68.

Ungerer, Tobias & Stefan Hartmann. 2023. *Constructionist approaches: Past, present, future* (Elements in Construction Grammar). Cambridge: Cambridge University Press.

van Trijp, Remi. 2024. Nostalgia for the future of construction grammar. *Constructions and Frames* 16(2): 311–345.

Weissweiler, Leonie, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft & Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, 16919–16932. Torino, Italia: ELRA & ICCL.

Markus Forsberg, Yousuf Ali Mohammed, Emma Sköldberg,
and Maria Öhrman

15 SO in Strix: a lexicographic case study of entry vectors

Abstract: Strix is Språkbanken's platform for conducting research on text collections. It contains the subset of the data available in Korp without copyright restrictions, that hence can be provided as full texts. In this chapter we present a case study of the lexicographic benefits of integrating the contemporary dictionary SO into Strix. In particular we investigate how the document vectors of Strix can be used as a methodological support in the development of SO.

Keywords: contemporary dictionaries, definition dictionary, large language model, research infrastructure, semantic search, word vectors

1 Introduction

Svensk ordbok utgiven av Svenska Akademien (SO 2021; henceforth SO), is a definition dictionary that aims to describe present-day Swedish. The content of SO has been developed over several decades and can today be considered an intricate network of Swedish words. In this chapter, we present a case study that investigates how this lexical network can be even further developed by using computational methods available through Språkbanken's text research platform *Strix* (Strix 2024).

Acknowledgments: The work on this chapter was partly supported by two Swedish Research Council national research infrastructure grants: *Språkbanken & Swe-CLARIN* (contract no. 2017-00626) and *Språkbanken* (contract no. 2023-00161), and by a grant from the Swedish Academy to Språkbanken Text for the project *Svenska Akademiens samtidsordböcker*.

Markus Forsberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: markus.forsberg@svenska.gu.se
Yousuf Ali Mohammed, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: yousuf.ali.mohammed@svenska.gu.se
Emma Sköldberg, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: emma.skoldberg@svenska.gu.se
Maria Öhrman, University of Gothenburg, Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, e-mail: maria.ohrman@svenska.gu.se

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111577234-015

The bulk of the about 65,000 entries in SO include cross-references to other entries and these references point out synonyms, antonyms, co-hyponyms etc. to the entries' headwords. For example, the entry *nätverk* 'network' has two cross-references, one to the co-hyponym *nät* 'net' and one to the related verb *nätverka* 'to network'. But there are many other possible cross-references, such as *nätverkare* 'networker', *router* 'router', *uppkopplad* 'connected', and more, so how do we identify all possible cross-reference candidates and how do we decide on which ones should be included? These are the questions that the lexicographer who compiles or revises the *nätverk* entry will need to answer, but herein lies a methodological challenge: How can the lexicographer answer these questions without having a full overview of what may be considered all possible candidates in SO? Not even the most skilled lexicographer will be able to keep all the 65,000 SO entries in mind. To address this computationally and to provide the lexicographer with the necessary material to be able to make a qualified judgement, we have explored how dictionary entries can be turned into *document vectors* – henceforth referred to as (dictionary) *entry vectors* – using Strix, where the general idea is that if two entry vectors are close in the geometric space, they are possible candidates for being linked.

In the chapter, we conduct an in-depth study of 36 of the 65,000 entries, e.g., the adjective entry *adekvat* 'adequate', and we examine what other SO entries these 36 entries are related to according to the geometric space of the entry vectors. In short, we would like to address the following questions:

- Is it fruitful from a lexicographer's point of view to use entry vectors when deciding which cross-references the 36 selected SO entries should have?
- From the lexicographer's perspective, are the entry vectors more useful for some types of entries than others?

In the following, we briefly introduce the platform Strix and the dictionary SO (see Sections 2.1–2.2). This is followed by a data and method section (Section 3). In Section 4, we present and discuss some of the results of our case study. Finally, in Section 5, we conclude and explore possible directions for future research.

2 Background

2.1 Strix: Språkbanken's text research platform

Strix, Språkbanken's text research platform, hosts a diverse collection of modern and historical corpora. As a user, you can select one or more corpora and retrieve all documents as a list. Each entry in the list provides an overview of the document,

with an option to open and read the full document in a document view mode. The platform resembles any search engine that you can find online: You input a search query and get a list of documents from the selected corpora that satisfies the given query.

The search engine user interface is divided into two parts: *Simple Search*, where the user can search for a word or phrase and retrieve documents that contain the given query, and *Document Search*, where the input is a text that can consist of just a single word, a sentence, or a full text. This method uses vector search to return documents that are semantically similar to the input text.

Strix stands out in relation to other search engines in many ways. First of all, each document in the corpora has been computationally enriched using a wide range of language technology methods. As a user of Strix, you know exactly what documents you perform your search on. Furthermore, Strix offers functionalities not typically available in standard search engines, such as retrieving related documents that are semantically similar to a document hit, filtering, and visualization, to name a few. In this platform, you can also upload collections of texts and enrich their data with linguistic annotations at the word, sentence and text levels. Additionally, you can utilize all the tool's functionalities to analyze the data.

The main functionality of this platform, as used in this case study, is related documents. Behind this functionality are document vectors – referred to here as (dictionary) entry vectors, since the documents are SO's dictionary entries. These document vectors are generated using the KB-SBERT sentence transformer (Börjeson et al. 2023) and are used in the related document function to retrieve the documents (entries) that are close to the given entry vector.

2.2 The contemporary dictionary SO

The case study thus concerns *Svensk ordbok utgiven av Svenska Akademien* 'The Contemporary Dictionary of the Swedish Academy' (SO 2021), a comprehensive monolingual dictionary of present-day Swedish, and how the description of the SO entries can be further developed. SO constitutes a subset of the Swedish Academy's lexical database (Salex). The second edition of SO, available on the dictionary web portal Svenska.se and in the form of apps, was published in 2021 (see Sköldberg 2022 about the edition; see also Chapter 4 in this volume). However, in the study we have used a development version of the dictionary which was extracted from Språkbanken's data editing platform Karp in October 2024 (see Chapter 11 in this volume).

3 Data and methods

In this section we discuss the set of selected entries. We also give an example of how an entry is presented in the dictionary and how the content of the same entry is turned into an entry vector in Strix. Furthermore, we provide some examples of related SO entries, which are identified using the entry vectors.

3.1 The 36 SO entries

The full development version of the SO dictionary has been added as a collection of documents in Strix, where every entry is treated as a document. As previously mentioned, in the case study we have explored 36 entries in depth, e.g., *adekvat* 'adequate', *bagage* 'luggage', *baksida* 'backside', *dammig* 'dusty', *disputation* 'thesis defence', *eller* 'or', *explodera* 'explode', *fotavtryck* 'footprint', *fräsch* 'fresh', and *hagla* 'hail' (verb; 'pour down hail'). The entries describe words of different parts of speech. More precisely, the selected entries cover 12 nouns, 11 adjectives, 11 verbs, 1 interjection, and 1 conjunction. Hence, to some extent, the selection can be said to reflect the distribution of different parts of speech among the entries in SO, as nouns, adjectives, and verbs are the most common.

All selected words in this study have been examined in previous methodological studies oriented toward the semantics of SO. About half of the words have been included in case studies on the use of word vectors in lexicography (Forsberg & Sköldberg 2022; forthcoming; Bouma et al. 2024). The former studies have shown that the use of word vectors is fruitful in several ways in lexicographic contexts. Not least, the word vectors complement the information that the SO lexicographers obtain from selected corpora by using concordances and word pictures in Språkbanken's word research platform Korp (see Chapter 10 in this volume). Among the neighbors in the examined vector spaces, there are many semantically related words that only co-occur in that they have similar contexts, which is something captured neither by concordances nor word pictures. The word vectors can thus, in a relatively objective and data-driven way, clarify new connections between both existing entries in the dictionary, and between existing entries and words in the corpora that can be added as new entries, language samples for the entries, etc. The word vectors differ from the entry vectors in that they represent an open universe of words with no connection to SO or any other dictionary. The entry vectors represent a closed universe of words that are tightly connected to the current content of the SO entries. In that sense word vectors and entry vectors are complementing methodologies rather than competing ones.

The remainder of the 36 selected words appear in studies of automatically identifying lexical variation and change among Swedish contemporary words in modern corpora. The aim of these studies has been to revise the semantic descriptions of the entries in SO by using the annotation tool DURel and its underlying computational language models for semantic variation and change detection (Sköldberg et al. 2024). The experiments presented in the studies are small-scaled but, so far, the work has brought the SO lexicographers' attention to figurative uses, meaning extensions, and specializations of the headwords that should be added to the dictionary.

A review of the 36 selected entries in this case study shows that they exhibit different length and complexity. Among the shortest entries, we have *rullstolsburen* 'wheelchair-using' (80 tokens) and *vinstvarning* 'profit warning' (90 tokens). Among the longest entries, we have *röd* 'red' (908 tokens), and *hund* 'dog' (842 tokens). The length of the entries in SO varies and depends, e.g., on the number of information categories in the entry. Shorter entries include data that is mandatory for all entries in the dictionary, e.g., information about inflection, a meaning description, and a year when the word was used in written Swedish texts for the first time. Entries like *röd* and *hund* include, in addition, several subsenses, idioms, and a quotation related to the entries (see an example of an SO entry in Figure 1).

In this context, it is also worth noting that we have examined a version of SO that includes all the information available in the SO entries. This means that internal editorial information, which is not available to the dictionary users, has been incorporated. This includes the information category *Subject field*, which, as the name suggests, gives an indication of what kind of topic that the entry relates to. The entry *fasad* 'facade' has the subject field *architecture* and those like *hund* 'dog' have *zoology*. In addition, the relative order of the information categories in the entries is not exactly the same as in the public dictionary, but tests indicate that the order has no substantial consequences for the results of this case study. The order of the information categories in the entries tends to change the relative order of related entries somewhat, but that is about it.

In terms of semantic complexity, only a few of the selected entries concern words that have only one sense, e.g., the interjection *usch* 'ugh'. The rest of the entries are polysemous. Most of them have one main sense and one or more subsenses, such as the adjective *rutten* 'rotten', but there are also entries with at least two main senses, such as *organisera* 'organize' (see Ralph, Järborg & Allén 1977 and Järborg 1989 on meaning description and the division into different senses in SO).

To sum up, in this case study, we have examined a set of 36 SO entries and reviewed their related entries, as they occur as related documents in Strix. In Strix, you can currently explore up to 50 most related entries. However, to further reduce the amount of data to make it more manageable, we have restricted ourselves to the top 20.

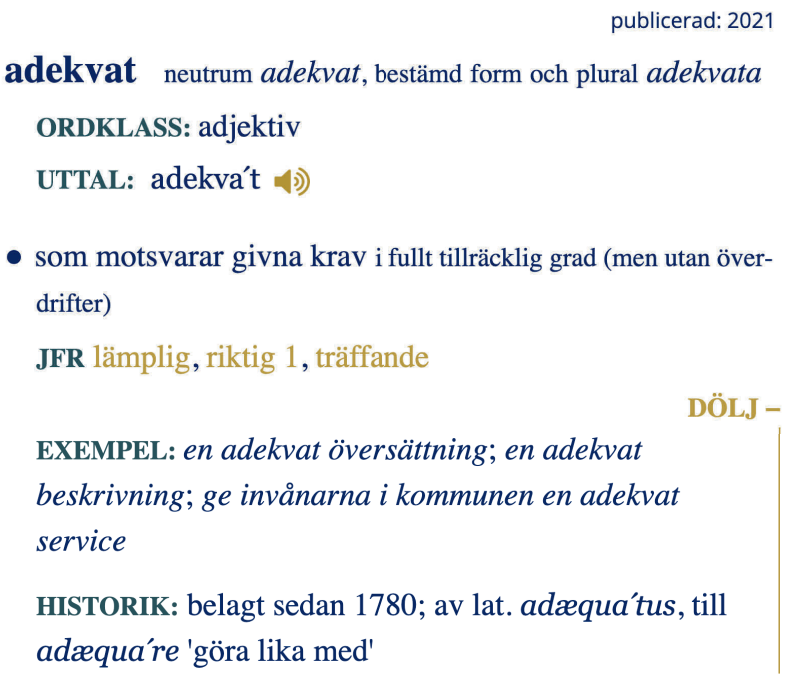


Figure 1: The entry *adekvat* 'adequate' in SO (2021) at Svenska.se

3.2 In depth example: *adekvat* 'adequate'

To clarify matters with an example, let us consider the adjective *adekvat* 'adequate'. Figure 1 shows how the dictionary entry is presented in SO (2021) at the dictionary web portal Svenska.se.

The entry *adekvat* comprises the headword together with formal information about its inflection, part of speech, and pronunciation. In addition, the entry includes semantic information in the form of a definition and definition supplement as well as three cross-references after the heading *JFR* 'cf.', that refer to the synonymous entries *lämplig* 'suitable', *riktig* 'proper', and *träffande* 'apt'. Furthermore, the entry provides phraseological information by three syntactic language samples, e.g., *en adekvat översättning* 'an adequate translation'. At the end of the entry, there is historical information: a year for when the word was first used in a written Swedish text, and the word's etymology (see more about the information categories in SO in Chapter 4 in this volume).

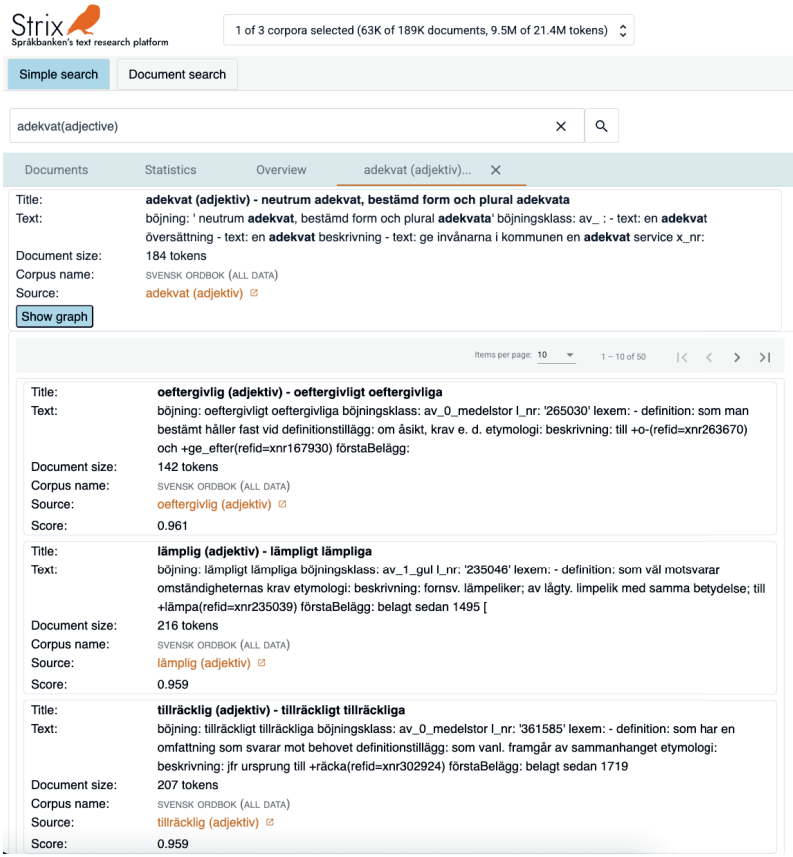


Figure 2: The related dictionary entries to *adekvat* 'adequate' in Strix

In Figure 2, we have searched for *adekvat* (adjective) in the SO collection in Strix and selected the related documents, i.e., the related SO entries.

At the top of Figure 2, to the right of the Strix logo, one of three versions of the SO collection has been selected. Below that, you see that a search has been performed and that the search query is *adekvat* (adjective), i.e., the adjective in all its inflectional forms.

Under the search field in the interface, some information about the entry *adekvat* is given. For example, the inflectional forms of the adjective included in the dictionary entry are presented as well as the length of the entry. As seen, the entry *adekvat* comprises 184 tokens, i.e., it is among the shorter ones in the survey. There is also a link to Språkbanken's data editing platform Karp, in which SO is edited (see Chapters 4 and 11 in this volume).

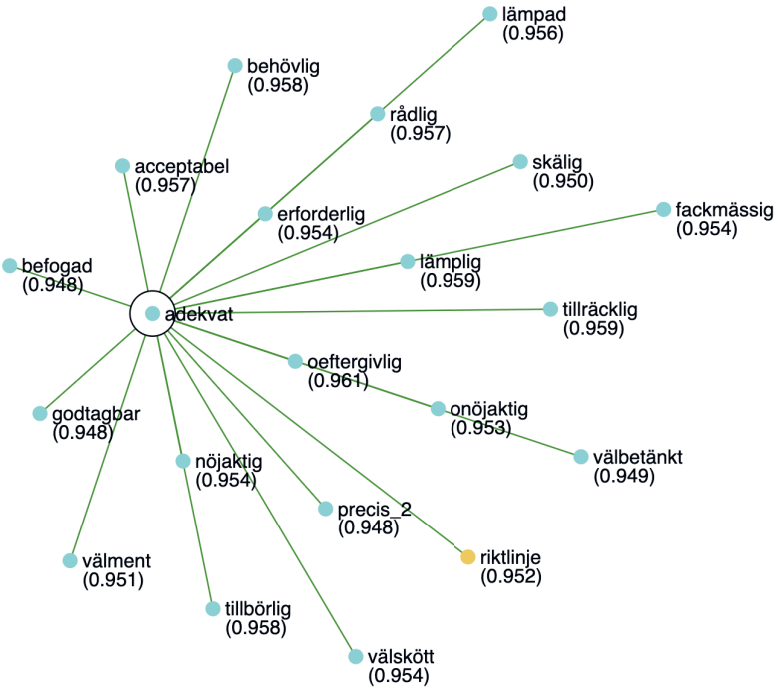


Figure 3: The relation graph of the entry *adekvat* 'adequate' and its related entries in SO (top 20)

Below this field, the related SO entries are listed. The top 3 most related entries are *oeftergivlig* 'unyielding', *lämplig* 'suitable', and *tillräcklig* 'sufficient'. As mentioned, currently the 50 most related entries are given by the tool, but in this case study, we have restricted the survey to the top 20.

In Strix, there is also a relation graph which makes it easier for the user to get an overview of the related documents, i.e., related entries in this case study. Figure 3 illustrates how this is rendered for the 20 most related SO entries of *adekvat* with an accompanying distance score.

Figure 3 shows that 19 out of 20 related entries are marked with the same color as *adekvat*, indicating that their headwords share the same part of speech. The only entry deviating is the noun *riktlinje* 'guideline'. The figure also shows the similarity, more precisely normalized cosine similarity, between *adekvat* and its related entries (see the numbers in parentheses under the words). The scale of the normalized cosine similarity measure is between 0 and 1, where 0 is the least and 1 the most

similar. In this and in other cases in the study, it can be noted that the numbers in the top 20 are often close to 1. The entry closest to *adekvat*, i.e., *oeftergivlig*, has a score of 0.961. The entry *[2]precis* 'precise', that occurs on the place 20 in the top 20, has the score 0.948. Since the numbers are close, the relative order among the top 20 is not regarded as relevant.

From a lexicographic perspective, it is noteworthy that several of the adjectives among the top 20 are synonyms to *adekvat*. However, they are not included in the current SO entry (see Figure 1). For example, cross-references to the near synonym *lämpad* 'suited' as well as to a subsense of *tillbörlig* 'proper' could be relevant. In addition, a link to an antonym, e.g., *onöjaktig* 'inaccurate', could both improve the meaning description and offer the dictionary users information on alternative ways of expression (see further Section 4). In other words, the entry *adekvat* could be improved based on the information provided by the entry vectors in Strix.

3.3 Three more examples

To give an idea of the characteristic features of the sets of related entries, we present the full top 20 for the three entries *kärlek* 'love', *skör* 'fragile', and *hagla* 'hail' (verb; 'pour down hail').

- **kärlek 'love'**: *älska* 'love' (verb), *förälska sig* 'fall in love', *förälskelse* 'infatuation', *förälskad* 'infatuated', *kär* 'in love', *svärmeri* 'infatuation', *kärlekskrank* 'lovesick', *kärleksförhållande* 'love affair', *älskad* 'beloved', *kära ner sig* 'fall in love', *käresta* 'sweetheart', *kärlekslös* 'loveless', *kärleksbetygelse* 'mark of love', *kärlekssaga* 'love story', *älskling* 'sweetheart', *svärmisk* 'romantic', *crush* 'crush', *kärleksförklaring* 'declaration of love', *kärlekshistoria* 'love story', *kärleksförbindelse* 'love affair'.
- **skör 'frail'**: *bräcklig* 'frail', *fragil* 'fragile', *spröd* 'brittle', *trasig* 'broken', *bruten* 'broken', *sliten* 'frayed', *klen* 'flimsy', *[1]knäck* 'crack' (noun), *[1]bräcka* 'crack' (verb), *[2]trasa* 'break' (verb), *brusten* 'broken', *[1]sköra* 'be torn apart', *[2]sköra* 'crack' (noun), *sönder* 'broken', *brista* 'break' (verb), *svikta* 'fail' (verb), *rubba* 'move slightly', *skranglig* 'rickety', *slitage* 'wear and tear', *slita ut* 'wear out'.
- **hagla 'hail' (verb; 'pour down hail')**: *skur* 'rain shower', *spöregna* 'rain heavily', *bedarra* 'calm down', *hällregna* 'rain heavily', *torna upp* 'pile up', *rusk* 'nasty weather', *vindkantring* 'turning of the wind', *duggregna* 'drizzle' (verb), *ösregna* 'rain buckets', *[1]dugga* 'drizzle' (verb), *torna upp sig* 'pile up', *hopa sig* 'pile up', *snöa in* 'snow in', *[2]yra* 'swirl' (verb), *mojna* 'abate', *blåsa under* 'blow down', *spöregn* 'downpour', *regna* 'rain' (verb), *slaska* 'sleet' (verb), *regna bort* 'rain away'.

In these three cases, there are slightly larger formal differences between the selected entries and their related entries, since there is a wider spread in parts of speech.

However, in all cases, there is a clear semantic connection between the examined entries and the related entries. This is especially obvious when the selected entry is a compound member in a related entry, like in the case of *kärlek* with the related entries *kärlekskrank* 'lovesick', *kärleksförhållande* 'love relationship', *kärleksbetygelse* 'love declaration', *kärleksförklaring* 'love declaration', *kärlekshistoria* 'love story', etc.

In the following, additional results of the case study are presented. The characteristics of the related entries are examined, and their potential to assist in improving SO is discussed.

4 Results of the case study and discussion

The relations between the 36 selected SO entries and their respective related entries are diverse. The reason for this may be that the selected items vary in part of speech, style, and degree of semantic content, to mention a few factors. For example, some entries concern concrete nouns while others involve abstract nouns (cf. *styrekonom* 'controller' and *vinstvarning* 'profit warning'). In addition, among the selected entries, there are function word entries like *eller* 'or'. As already stated, the 36 entries also show different length and semantic complexity in that some of the headwords, for example, are monosemous and others are polysemous. However, there are patterns that indicate the usefulness of the entry vectors in a lexicographic setting. First, the related entries often show a clear connection to one of the senses in a polysemous entry, mainly the main sense. This applies, for example, to *tvärnita*, with the main sense 'jam on the breaks'. Some of the related entries are *köra om* 'overtake', *krypköra* 'edge along', *köra över* 'run over', *tvärvända* 'turn on the spot', *samåka* 'carpool' (verb), *väja* 'give way', and *trafikera* 'traffic' (verb). The entry *tvärnita* also comprises two subsenses, one of which is figurative, but neither has a meaning description, which might be the reason why they are not represented among the related entries. Similar results are observed, for example, for *bagage* 'baggage' and *fasad* 'façade', which both have figurative subsenses, but their related entries are connected only to their main senses, i.e., 'luggage' and 'exterior of a building'.

In one case, *disputation*, one of two main senses in the entry has more impact among the related entries than the other. The entry covers the senses 'academic gathering with review and defense of doctoral thesis' as well as 'exchange of opinions'. Although the second sense is stated to be less common, it is much more present among the related entries. The reason may be that there are many near-synonymous words expressing the sense 'exchange of opinions' in the dictionary (see, e.g., *tvistighet*

'dispute', *kontrovers* 'controversy', *meningsskiljaktighet* 'difference of opinion'). This is not the case with the first sense of *disputation*.

However, sometimes a subsense in the selected entry has more impact among the related entries. This applies, for example, to *fotavtryck* 'footprint' where almost all related entries have to do with ecology and the environment.

Another noticeable pattern is that the headword of the selected entry and those of the related entries are frequently of the same part of speech. This is beneficial for the SO lexicographers when seeking new or improved cross-references, i.e., synonyms, antonyms and co-hyponyms, to include in the dictionary. The possibility of including more cross-references in the entry *adekvat* is discussed in Section 3.2. Furthermore, an entry like *vansinnig* 'insane', which currently comprises only a few links, can be provided with new cross-references to entries such as, for example, *tossig* 'silly', *rubbad* 'deranged', and *bindgalen* 'raving mad', i.e., adjectives that denote degrees of insanity. An entry like *vissen* 'wilted', with the figurative subsense '(slightly) sick', currently lacks cross-references to other entries and it can be linked to entries like *dassig*, *krasslig*, and *risig* with the same sense. The addition of more cross-references to, for example, synonymous words reinforces the semantic network between the lexical units covered by SO (cf. Blensenius, Sköldberg & Bäckerud 2021). In addition, the inclusion of more entry links can strengthen the dictionary as a writing tool. Through connections to other entries, the dictionary users are provided with alternative expressions, making it easier for them to vary their language (see further Malmgren 2009). Consequently, SO can also function somewhat like a thesaurus (see Chapter 8 in this volume).

Several of the headwords of the related entries can also serve as morphological language samples. For example, the SO entry *klimat* 'climate' could be supplemented with more morphological examples such as the compounds *fastlandsklimat* 'continental climate', *kustklimat* 'coastal climate', *klimatombyte* 'climate change', as well as derivations like *klimatologi* 'climatology'. In the same way, new compounds like the adjectives *rödbrusig* 'red nosed', *rödlätt* 'ruddy', *rödgråten* 'red with weeping', *rödsprängd* 'red bursted', and the verb *rödglödga* 'bring to red heat' could be added to the subsenses in the adjective entry *röd* 'red'. The shades of red in the compounds above are not the same as in the main sense of the word, where the color is likened to the color of flowing blood. Including these words makes the SO description of *röd* 'red' even more nuanced.

Furthermore, the entry vectors clarify the different kinds of semantic fields for an entry, which can be useful for lexicographers. This becomes clear, for example, when examining the verb entry *ventilera* 'ventilate' and some of its related entries: *ventilation* 'ventilation', *vädra* 'ventilate', *tilluft* 'supply air', *lufta* 'air' (verb), *frånluft* 'exhaust air', *friskluftsintag* 'fresh air intake', *luftväxling* 'air exchange', *luftkonditionering* 'air conditioning', *luftrenare* 'air cleaner', *fläkt* 'fan', *vädra ut* 'ventilate'.

Such words are crucial for describing the sense of *ventilera* and can also be incorporated into new syntactic language examples. From the selected entry *kriga* 'wage war' another semantic field emerges. Among the related dictionary entries, you find *utkämpa* 'wage' (verb), *strida* 'fight' (verb), *[2]här* 'army', *batalj* 'battle', *fältslag* 'battle', *krigstillstånd* 'state of war', *militarisera* 'militarize', *sammandrabba* 'clash' (verb), *krigföring* 'warfare', *stridighet* 'strife', and *drabbning* 'skirmish'. The main sense of current *kriga* in SO could advantageously be further developed with links and morphological examples based on these related entries. And, in addition to the usefulness for the SO lexicographers, the related entries of the kind we study here, may also be relevant for other dictionary resources, or in teaching with a focus on the Swedish vocabulary.

Moreover, the related entries highlight word combinations that include the selected words. Among the related entries to *bagage* 'luggage', there are the verbs like *lasta ur* 'unload', *stuva* 'stow', and *pollettera* 'have one's luggage registered'. The current SO entry *bagage* provides only two syntactic language examples, one per sense, and the inclusion of word combinations like *lasta ur sitt bagage* 'unload one's luggage' among the syntactic language samples would lead to a more detailed description of how the main sense of the noun is used.

The related entries can also give an indication of the emotional charge of a word, at least in terms of one of its senses. As an example, the main sense of *enkelspårig* 'having only one track' can be considered as neutral. However, the figurative subsense of the word, i.e., 'narrow minded, superficial', has negative connotations. Among the related entries you find: *fläng*ig 'ragged', *obändig* 'inflexible', *[2]fläng* 'crazy', *chosig* 'pretentious', *snäsig* 'snarky', *ohyvlad* 'vulgar', *osmidig* 'clumsy', *urspårad* 'derailed', *rivig* 'ragged', *oordentlig* 'disorderly', *oreflekterad* 'unreflective', *tetig* 'stilted', *okonstlad* 'unassuming', *orubblig* 'unshakable', *knipslug* 'sly', and *trulig* 'discontent', and as shown, many of them denote negative qualities. Nevertheless, it remains unclear if and how these related entries can be utilized in the lexicographic development of SO.

In addition, the entry vectors indirectly contribute information about other entries than the selected ones in this case study. For instance, *ofantlig*, which in addition to the main sense 'huge' can be used adverbially as an intensifier (e.g., *ofantligt bra* 'extremely good'), has a related adjective *enorm* 'enormous' that behaves similarly. However, the semantic description of *enorm* in SO is today relatively brief and could be elaborated in the same way as *ofantlig*.

So far, we have provided several examples where the entry vectors have been deemed beneficial for lexicographic work. However, it can be noted that results of the study are more rewarding regarding some of the entries. An examination of all the related entries reveals that the entry vectors are less useful when it comes to two of the entries: the conjunction *eller* 'or' and the interjection *usch* 'ugh'. Among

the related entries to *eller* you find *åtskillig* 'numerous', *okränkbar* 'inviolable', and *överens* 'in agreement', i.e., randomly related entries. As for *usch*, the related document functionality points out some other interjections that, at least earlier, could be used to express negative feelings (*hu* and *tvi*). Apart from them, it is again rather random, except that most of the related entries to *usch* have negative connotations. In other words, the entry vectors work better for content words than function words. At the same time, it is not completely clear what related entries would be ideal for a function word as *eller* from a lexicographic perspective.

5 Conclusions and future work

In this chapter we have presented a case study of the lexical network of *Svensk ordbok utgiven av Svenska Akademien* (SO) using computational methods available through Språkbanken's text research platform Strix (Strix 2024). We focus on 36 dictionary entries in a development version of SO and we review their top 20 related entries according to the entry vectors in Strix.

Currently, the SO lexicographer manually selects possible cross-references, etc. when compiling new or revising existing entries in the dictionary. The task is challenging, considering that the dictionary covers as many as approx. 65,000 entries. The results of the case study indicate that the entry vectors, in a more data driven and objective way, draw the lexicographer's attention to entries that are suitable as cross-references, i.e., links to synonyms, antonyms etc. Furthermore, the entry vectors clarify SO headwords that can form the basis for new morphological and syntactic language samples in existing SO entries. That is to say, the entries of SO can be expanded semantically, by using the entry vectors.

From a lexicographic point of view, the related entries often seem to have a clearer connection to the main senses than to subsenses of the selected entries. One possible reason for this is that the main sense often receives more attention in the entry. For example, the main sense is normally described in the form of a complete definition. Moreover, if the main sense is the most frequent sense as well, it will also have an influence on the entry vectors created using KB-SBERT. Subsenses, on the other hand, are not always described as clearly. Hence, the results indicate that the subsenses of the headword in SO could become more prominent in the entries, especially if they are commonly used.

Furthermore, the results indicate that the entry vectors work better for entries concerning content words than for entries concerning function words (like conjunctions and interjections).

Comparisons with previous studies, not least Bouma et al. (2024), show that the results of using entry vectors and word vectors in lexicographic work overlap to a certain extent, when it comes to data that can be included in the dictionary. However, a very important difference is that, with the entry vectors, the existing SO entries are related to other SO entries. When it comes to word vectors, there is no relation to SO, only to the corpora used to build the word vectors. In that sense, the different types of vectors provide various kinds of information and can therefore complement each other in the editorial work.

The case study presented in this chapter is qualitative in nature. In the future, we aim to evaluate the entry vectors quantitatively by using the existing co-references and comparing them with the results obtained by using the entry vectors. However, since SO cannot provide any definite ground truth, we still need to conduct a qualitive assessment in addition to the quantitative evaluation.

We also intend to investigate how entry vectors work in relation to other lexical resources in addition to and in combination with SO, in particular SAOL (see Chapter 3 in this volume). For SAOL, it may be a problem that most of the entries are much shorter than in SO, but this might be remedied by the use of the large language model KB-SBERT.

To conclude, the use of entry vectors has already proven to be a productive tool that has resulted in many new revisions of SO entries. We are currently investigating how its use can be integrated into Språkbanken's data editing platform Karp (see Chapter 11 in this volume), to improve the general lexicographic workflow while using Karp.

References

Blensenius, Kristian, Emma Sköldberg & Erik Bäckerud. 2021. Finding gaps in semantic descriptions: Visualisation of the cross-reference network in a Swedish monolingual dictionary. *Proceedings of the Electronic lexicography in the 21st century (eLex) 2021 conference*. 247–258.

Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf & Justyna Sikora. 2023. Transfiguring the library as digital research infrastructure: Making KBLab at the National Library of Sweden. *College & Research Libraries* 85(4): 564. DOI: 10.5860/crl.85.4.564.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg. 2024. Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten [The art of conducting Swedish lexical research without violating copyright]. *Proceedings of the Huminfra Conference (HiC 2024)*. 161–167.

Forsberg, Markus & Emma Sköldberg. 2022. Ordvektorer i lexikografiskt arbete [Word vectors in lexicographic work]. In Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (eds.), *Live and learn: Festschrift in honor of Lars Borin*, 37–41. Gothenburg: Department of Swedish, Multilingualism, Language Technology, University of Gothenburg.

Forsberg, Markus & Emma Sköldberg. Forthcoming. Ord med liknande kontext sökes! Om ordvektorers roll i svensk lexikografi [Words with similar context wanted! On the role of word vectors in Swedish lexicography]. *Den 17. konferansen om leksikografi i Norden*.

Järborg, Jerker. 1989. *Betydelseanalys och betydelsebeskrivning i Lexikalisk databas: Preliminär version* [Semantic analysis and semantic description in the Lexical Database]. (Research Reports from the Department of Swedish No. GU-ISS-89-01) Gothenburg: Department of Swedish, University of Gothenburg.

Malmgren, Sven-Göran. 2009. On production-oriented information in Swedish monolingual defining dictionaries. In Sandro Nielsen & Sven Tarp (eds.), *Lexicography in the 21st century: In honour of Henning Bergenholtz*, 93–102. Amsterdam: John Benjamins.

Ralph, Bo, Jerker Järborg & Sture Allén. 1977. *Svensk ordbok och lexikalisk databas: Förstudierapport* [The dictionary *Svensk ordbok* and the lexical database: A pilot study report]. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Sköldberg, Emma. 2022. Andra upplagan av Svensk ordbok: Förutsättningar och redaktionella val [Second edition of the Swedish dictionary: Conditions and editorial choices]. *LexicoNordica* 29: 139–152.

Sköldberg, Emma, Shafqat Virk, Pauline Sander, Simon Hengchen & Dominik Schlechtweg. 2024. Revealing semantic variation in Swedish using computational models of semantic proximity: Results from lexicographical experiments. *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 169–182.

SO. 2021. *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy]. 2nd edn. Stockholm: Svenska Akademien.

Strix. 2024. *Språkbanken's text research platform*. [2024-12-14].

Index

A note on terminology: in this volume, a *dictionary* and its underlying *lexical database* are intended for human consumption, while a *lexical resource* is intended for use in natural language processing systems. We further use the term *lexicon* to include all the three mentioned, i.e., dictionaries, lexical databases, and lexical resources, as well as some dual-use datasets.

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

https://doi.org/10.1515/9783111577234-016