Data Representativity and Granularity in Spanish Syntax

Subjecthood across Dialects and Methodologies in Spanish

Iván Ortega-Santos

First published 2024

ISBN: 978-1-032-43765-1 (hbk)
ISBN: 978-1-032-44011-8 (pbk)
ISBN: 978-1-003-36993-6 (ebk)

Chapter 2

Theoretical Syntax at the crossroads

Acceptability judgments on trial

DOI: 10.4324/9781003369936-2

The Open Access version of this chapter was funded by University of Memphis.

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

2 Theoretical Syntax at the crossroads

Acceptability judgments on trial

This chapter presents a critical review of the debate on the validity of introspective judgments gathered without an experimental data collection protocol and analyzed without statistical analysis in Generative Grammar both from the point of view of the experimental literature (Edelman and Christiansen 2003, Wasow and Arnold 2005, Ferreira 2005, Gibson and Fedorenko 2010, 2013, Sprouse et al. 2013, a.o.), Section 2.1, and from the point of view of the variationist literature (e.g., Cornips and Poletto 2005, Henry 2005, Ocampo 2010, Cornips and Gregersen 2016, a.o.), Section 2.2. This division into separate sections is made for presentational purposes as research may involve more than one subdiscipline.

2.1 Challenges to the use of acceptability judgments in Theoretical Syntax: The view from Experimental Syntax

This section is structured as follows: Subsection 2.1.1 focuses on (i) the influence of processing costs on the perception of acceptability and, therefore, the importance of controlling for this factor when developing syntactic theories based on acceptability contrasts; (ii) the relevance of experimental research to tease apart unacceptability caused by the said processing costs as opposed to violations of grammatical properties; and (iii) linking theories of the grammar and the parser (e.g., Miller and McKean's [1964] Derivational Theory of Complexity or Phillips' [1996] view that the grammar is the parser). The discussion is exemplified by Spanish examples when pertinent. Subsection 2.1.2 includes an overview of the challenges to the use of acceptability judgments for syntactic research. From the experimental perspective, this issue has been addressed recently through so-called data assessment, in other words, the study of the convergence rate between data collected through experiments and published data collected through introspective judgments without formal data collection protocols or statistical analysis (e.g., Sprouse et al. [2013] and related work). This chapter presents the results from an ongoing data assessment project for Spanish (Ortega-Santos 2020, 2021) emphasizing the following issues:

(a) The need to go beyond English in data assessment research, since convergence rates are subject to variation depending on the property under study

DOI: 10.4324/9781003369936-2

and the said properties are subject to crosslinguistic variation (see also Haider 2016, Linzen and Oseki 2018).

(b) The lack of detailed information concerning sociolinguistic factors potentially present in Theoretical Syntax data, which hides the origin of the data in a sort of black box effectively preventing a straightforward interpretation of data assessment research.

(c) The limits of acceptability judgments to gather certain kinds of data, irrespective of whether experimental data collection protocols and statistical analysis are used (e.g., for stigmatized data, information structure, heritage languages, or idiolectal differences; see Cornips and Poletto [2005], Ocampo [2010], Polinsky [2018], and Henry [2005], respectively).

(d) The virtues and limits of data-gathering marketplaces such as Amazon's Mechanical Turk for the study of Spanish syntax and beyond, where these marketplaces have been championed as part of the solution to help Theoretical Syntax adhere to higher quantitative standards (Gibson et al. 2011).

Methodological recommendations are made for 'traditional' generativists as well as editors, the gatekeepers of published research and research standards, while providing a rationale for how those recommendations help improve the resulting research (Subsection 2.1.3; see also Featherston 2007, a.o.). It is argued that even in cases where the object of study does not call for an experiment, the adoption of higher methodological standards (or, in some cases, being explicit about those standards) would help the interdisciplinary dialogue. Subsection 2.1.4 discusses the limits of acceptability judgments beyond the data assessment research, and Subsection 2.1.5 focuses on the use of marketplaces such as Amazon Mechanical Turk for data collection. In turn, Subsection 2.1.6 presents a brief overview of replication research in other fields, particularly from the domain of Cognitive Science, e.g., Psychology and the Social Sciences, as a valuable counterpoint to understand the ongoing paradigm change in Linguistics.

2.1.1 Acceptability vs. grammaticality: A first look at the relevance of experimental research to the field of Generative Syntax

Acceptability judgments, that is to say, judgments concerning the well-formedness of sentences or phrases, tend to provide the empirical basis for theoretical research, which understands syntactic knowledge as a mental grammar (competence) able to generate the sentences that are part of the language while excluding those which are not. When evaluating the pertinence of using acceptability judgments as a source of data, we are essentially asking ourselves whether this usage turned Generative Grammar into a procrustean bed in which language data do not necessarily fit perfectly but rather are forced to, e.g., through confirmation bias which allows us to focus on competence while dismissing certain kinds of data as language use. In keeping with this idea, the use of acceptability judgments as evidence has been challenged through criticisms focused on a variety of issues, ranging, for

instance, from the way the judgments traditionally tend to be collected (without an explicit experimental protocol or statistical scrutiny) to the difficulty in controlling for processing costs, which may affect the perception of acceptability. Following Ortega-Santos (2020), the next section focuses on these issues with an emphasis on Spanish. It is shown that experimental linguistics, on the one hand, has embraced acceptability while advocating for the improvement of the data collection standards (see Section 2.1.2). Furthermore, it is particularly suited to detect not only acceptability understood binarily (acceptable/unacceptable) but also gradience in the judgments or processing costs. On the other, it has gone beyond introspection (e.g., through a variety of research methods, such as eye-tracking or brain-imaging research; see Chapter 5).

2.1.1.1 Factors that affect the perception of acceptability

Early in the development of Generative Grammar, it was noted that using acceptability judgments for syntactic theorizing was a risky enterprise in the sense that determining whether the unacceptability of a sentence is determined by a violation of the grammar or by processing difficulties is not an easy task. Still, the use of acceptability judgments provided the empirical basis of Generative Grammar, allowing the field to make progress while attempting to move away from messy usage-based factors like memory limitations or performance errors. For critics, this constitutes some form of *original sin* of the field, albeit with some qualifications. For instance, it has been claimed that it was an appropriate first step, but Generative Grammar failed to fulfill its destiny to incorporate other forms of evidence and/or to avoid oversimplifying the object of study. In the words of Guy (2015: 49; see also Brandner 2012, a.o.), keeping grammar and usage separate as part of the competence/performance distinction is 'a typical step in the early stages of a scientific field: it is easier to work out generalizations and models if we can start by ignoring some of the complexity of reality.' According to this line of reasoning, the need to go beyond such an early data collection heuristic and to integrate the research with other subfields (e.g., Variationism) was not met. Moreover, it has been argued that this bias in the choice of evidence has led Generative Grammar to be ignored by other branches of linguistics or that it has misled the field with inaccurate generalizations. Research on the latter issue will be discussed in Section 2.1.2.

The importance of developing a clear understanding of the difference between the perception of acceptability and grammaticality (a theoretical construct) can be seen in well-known cases such as garden path sentences (Bever 1970), (1), center embedding (Chomsky and Miller 1963), (2), and grammatical illusions, (3) (where possible, Spanish examples are used following Ortega-Santos [2020]). In the first two cases, the structures are grammatical, and yet they are perceived as unacceptable due to processing difficulties. In the third case, an ungrammatical structure is perceived as acceptable (example translated from Herman Schultze, see Montalbetti [1984:6]):

(1) The horse raced past the barn fell.
(2) El perro que el gato que el ratón vio escuchó ladró.
the cat that the dog that the mouse saw.3sg heard.3sg barked.3sg
'The cat that the dog that the mouse saw heard barked.'

(3) Más gente ha ido a Rusia que yo he ido.
more people have.3sg gone to Russia than I have.1sg gone
'More people have been to Russia than I have.'

In the case of (1), the intended interpretation is 'The horse that was raced past the barn fell' and is grammatical. Still, the parser is biased toward an interpretation in which there are two main verbs in the sentence. Replacing 'raced' for an unambiguous past participle helps avoid the issue as in 'The horse seen at the barn fell.' So, (1) is grammatical but unacceptable, meaning it can be generated by the grammar and yet it is judged as deviant. In a similar vein, the intended interpretation of (2) is that the dog barked; the dog is individuated through a relative clause stating that the cat listened to that dog and the cat is individuated through a relative clause stating that the mouse saw that cat. The grammar can generate the structure, but it is judged as deviant. The sentence arguably improves when varying the number and semantic features of the DPs (see Bever's [1974] original observation; my data):

(4) El perro que los veterinarios que el ratón vio estirilizaron ladró.
the dog that the veterinarians that the mouse saw.3sg sterilized.3pl barked.3sg
'The dog that veterinarians that the mouse saw sterilized barked.'

Finally, no matter how acceptable (3) sounds, that is not a properly formed comparison. You can compare the number of times somebody has gone to Russia to the number of times you have gone, or you can compare the number of people who have gone to Russia to the number of people who went to Belgium, but you do not compare the number of people that have gone to Russia to you. The perception of acceptability is affected by various factors, e.g., 'the comprehensibility of a structure and the correctability of violations' (Featherston 2007), which may determine the gradience in the judgments. In a similar vein, Cowart (1997: 47) mentions that

> syntactic factors such as clausal structure, the specific devices that are used to implement various syntactic roles (e.g., choice of complementizer or relative pronoun), the complexity of a structure, the familiarity or frequency of a structure and parsability can all influence judgements.

Processing costs may add up causing a processing overload (Gibson [1998, a.o.]; see Phillips, Gaston, Huang and Muller [2021] for recent discussion). Crucially for the present purpose, determining which processing factors may affect the perception of acceptability of a structure under study is crucial for syntactic research. This is

feasible, but it entails a high degree of unification between psycholinguistic and generative research. Syntactic research, through the use of minimal pairs, aims at controlling for processing factors. However, it might not be well suited to specifically investigate this issue, as the study of locality restrictions has shown (see Chapters 3 and 5).[1] Thus, a cursory look at the origins of Generative Grammar and its standard practices already allows us to address one of the myths concerning the challenges for interdisciplinary dialogue spelled out in Chapter 1, namely, that experimental research does not bear on the issues of interest to Theoretical Syntax. As can be seen, *experimental research has been and is in fact crucial to the generative enterprise*.

While providing an exhaustive overview of the relation between the grammar and the parser is beyond the scope of this chapter, a selection of relevant works will be presented very briefly (e.g., Phillips [1996]; for early attempts to link grammatical theory and processing difficulties, see Miller and McKean's [1964] Derivational Theory of Complexity) as well as various proposals on processing costs (e.g., Gibson 1998, Lewis and Vasishth 2005). These theories are not only important as general knowledge when designing minimal pairs or experiments, but they are also relevant for syntactic theorizing as seen in the discussion of locality in Chapters 3 and 5.

The Derivational Theory of Complexity was an important milestone in the study of the relationship between Generative Grammar and the parser. It was hypothesized that transformations determined the way a sentence was processed: The higher the number of transformations needed to generate a sentence, the longer it would take speakers to parse it (Miller and McKean 1964). For instance, questions or negative statements were considered to involve additional transformations and, therefore, to be more complex to parse. Seen in this light, (5b) would be more complex than (5a), due to the addition of negation, whereas (5c) would be the most complex of the three due to the addition of negation and the transformation into a question:

(5) a. John is tall.
    b. John is not tall.
    c. Isn't John tall?

Unfortunately, the overall predictions of the theory were not fulfilled, and this result was interpreted as supporting the competence vs. performance dichotomy (see Towsend and Bever [2001] for discussion; for later attempts to provide experimental evidence for the psychological reality of generative constructs, see Bever [2014] for syntactic traces and Pablos et al. [2018] for c-command, a.o.). This dichotomy effectively separated Theoretical Syntax from Psycholinguistics, though other attempts to link the grammar and the parser and/or predict the timing of the processing of linguistic structures have routinely emerged. Particularly influential has been Phillips' (1996) view that the grammar is the parser. This author assumes left-to-right derivations based on syntactic, phonological, and parsing facts. In turn, Gibson (1998) and Lewis and Vasishth (2005), among others, developed theories of the processing costs involved in sentence processing, specifically those associated with maintaining a category in memory and integrating it into existing structure. The said processing costs may reach a certain threshold, causing a

processing overload (see center embedding and garden path sentences), which may affect the perception of acceptability. Thus, a mapping between processing factors and acceptability is provided (see Chapters 3 and 5 for a fuller discussion of this issue with an emphasis on theories of locality, e.g., so-called syntactic islands). As a result of this complexity found in the relationship between acceptability and grammaticality, among other factors, the reliance on acceptability judgments as evidence to develop grammatical theories has been called into question.

The next section reviews the ongoing debate on the reliability of syntax data in the generative literature with an emphasis on Spanish, building on work by Ortega-Santos (2020, 2021).

2.1.2 Research on the reliability of non-quantitative nonexperimental data collection methods in syntax: A look at the debate on data assessment with an emphasis on Spanish

Lately, the debate on acceptability judgments has focused on the pertinence of adopting data collection standards pervasive in closely related fields like Psycholinguistics or Second Language Acquisition. In particular, two data-gathering methods are commonly used in syntax: (i) non-quantitative nonexperimental methods and (ii) experimental methods, with an explicit data collection protocol and statistical analysis, as used in other Cognitive Science domains. While the traditional approach in (i) is found throughout, the second ("Experimental Syntax") has become widespread only recently. As part of this shift, the use of informally-gathered acceptability judgments as a source of evidence for theoretical research has been questioned (Edelman and Christiansen 2003, Wasow and Arnold 2005, Ferreira 2005, Gibson and Fedorenko 2010, 2013). For instance, Wasow and Arnold (2005: 1484) claim that Theoretical Syntax 'journals are full of papers containing highly questionable data, as readers can verify simply by perusing the examples in nearly any syntax article about a familiar language,' because of the 'overreliance' on acceptability judgments and the way they are collected (e.g., potentially with few participants, no data collection protocols, no statistical scrutiny and, possibly, cognitive bias).[2] In turn, Gibson et al. (2013: 10) argue that judgments by professional syntacticians found in the literature are not 'data.' Specifically, they view 'expert linguist judgments as essentially *expert predictions*,' useful to design formal experiments but not data. According to these researchers, this 'lack of validity of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to incorrect generalizations and unsound theorizing' (Gibson and Fedorenko 2010: 233). Unfortunately, existing attempts to provide evidence for these methodological issues tend to have what could be described as a cherry-picking tendency, in that they focus on highly specific case studies – maybe rightly so, in order to be able to provide an in-depth understanding of a specific phenomenon. Be that as it may, to address these concerns, large-scale comparisons of published data collected through non-quantitative nonexperimental methods (lacking formal data collection protocols or statistical analysis), on the one hand, and data collected experimentally and analyzed statistically, were first carried out

a decade ago. This line of research is commonly referred to as data assessment. For instance, Sprouse et al. (2013) found a 93% convergence rate between traditional judgments and data gathered experimentally for a random sample of a ten-year period of research on English published in *Linguistic Inquiry* (Likert scale results using a one-tailed null hypothesis test; see their work for other experimental designs and statistical tests). In turn, Sprouse and Almeida (2012) found a 98% convergence rate for an English syntax textbook using a different experimental design (magnitude estimation and yes–no judgment tasks; see Chen et al. [2020] for a similar piece focusing on Chinese).[3]

Questions arise concerning the representativeness of the published syntax data for languages other than English. Why? First, typological variation might affect the results. This is the case because replication rates vary according to the phenomena under scrutiny – note that we refer to replication of the results, not of the methods and procedures used when collecting and analyzing the data. Second, the number of syntacticians working on languages other than English is smaller (per language). Therefore, there could be comparatively more noise in the data and the corresponding theoretical developments (Haider 2016). For instance, Linzen and Oseki (2018: 18) argue that the quality of the peer-review process can be affected and that this

> issue is more acute in articles published in journals that are not language-specific; the editors of those journals might not be able to find reviewers who are simultaneously native speakers of the language and experts on the theoretical topic of the article.

The need to go beyond English in data assessment research is illustrated with preliminary results of data assessment in Spanish. Ortega-Santos (2020) focused on a random sample of syntax articles published in *Probus* (2006–2017) and high-impact articles/book chapters (100+ quotes in Google Scholar). The data were tested in Venezuelan Spanish, and the following convergence rates were found, respectively: 82.3% and 79% statistically significant results in the right direction (meaning that acceptability contrasts found in the publications were also found in the corresponding experiment). In principle, this convergence rate is on the low side, at least when compared to English. This being said, on top of the factors already mentioned (e.g., the low(er) number of Spanish-speaking syntacticians or the fact that *Probus* is not a language-specific journal, but rather it focuses on Romance languages), various aspects may influence the results for Spanish: Prescriptive grammar tends to be more present in the educational system of the Spanish-speaking world when compared to the US (see also the prestige of the Real Academia de la Lengua, RAE, an institution unattested in the English-speaking world), and this might have affected the results. Furthermore, Latin-American Spanish is different from American English in that it evolved in a multinational context, with intensive language contact with indigenous languages (depending on the country, area, etc.). Thus, microvariation is, all else being equal, expected to be found in Latin-American Spanish to a greater degree than in English. Moreover, a significant part of the data tested was labeled as 'Spanish' – without any further

information, particularly for the high-impact data set – or 'Latin-American Spanish', as opposed to say, Venezuelan Spanish. Thus, the current results comparing data reliability in English vs. Spanish, while relevant as a contribution to ongoing debates in the field, also highlight the limitations inherent in data assessment research, e.g., lack of details when identifying dialects in the original publications and the potential role of microvariation. In other words, the data assessment enterprise, while valuable, presents certain inherent limitations: For data assessment to be meaningful, explicit information regarding the origin of the published judgments needs to be available, e.g., beyond labeling the Spanish as Latin-American or even Argentinian or Iberian Spanish, one would like to know the region speakers come from, age, gender, bilingual skills, and even linguistic training. In other words, one would like to know all the information that Sociolinguistics has revealed to be important in the study of variation. Syntax papers, however, rarely provide such information. Instead, references to the origin of the judgments are included in the acknowledgments, typically a list of individuals who provided judgments. No details are available concerning which judgments they provided, the exact number of speakers who were tested for each sentence and their background (the dialect they may speak, educational level, training in linguistics, etc.). In the absence of such detailed information, we cannot be sure what exactly we are assessing: Are we targeting the same population as the published papers? Just to illustrate the issue: Ortega-Santos (2020) found that when judgments are subject to variation, syntacticians may discuss the issue by referring to the grammar of 'some speakers,' a fairly vague notion. The following examples quoted in Ortega-Santos (2020) help illustrate this point (my emphasis). Note that similar instances can be found in my own work, and there is no implication that the quoted works are in some sense faulty – these are, in fact, works of highly respected researchers, and the publications were quoted 100+ times. In other words, this is not about the practices of a specific researcher, but rather the practices in the field:

- However, both searches in corpora and native speaker's intuitions show that there is nothing unusual in this construction in Spanish. *Some speakers* also allow… (Fábregas 2007: 169)
- As pointed out by an anonymous reviewer, the following example appears possible for *some speakers* of Spanish. (Ordóñez y Treviño 1999: 43, n. 4)
- llustrates that, without Asp_{SE}, an *in*-adverbial or a *for*-adverbial is grammatical (although the for-adverbial is marked for *some speakers*). (MacDonald 2016: 75)

In the absence of information on microvariation, the lack of convergence in the data gathered through both methodologies might be just due to linguistic variation. Thus, Ortega-Santos (2020, 2021) suggests replacing the concept of data assessment with the notion of data representativeness.[4] Data that do not converge across methodologies might not be representative of the dialects/sociolects we targeted (see also Grieve 2021). That being said, the current goal goes beyond providing a critical understanding of the said literature. Thus, in the next section, a call to

adopt higher methodological standards or else being explicit about the standards followed is issued, including but not limited to the need to provide details concerning the origin of the data.

It should also be noted that we are comparing the results of two different data collection methods; that is to say, we are looking at the replication of the results but not the data collection method or the analysis (quantitative vs. non-quantitative). Thus, a replication result in the 79–82.3% range is perhaps not unexpected. As has been noted in the literature, replication rates vary across disciplines (e.g., physics vs. economy) and so do the goals of replication research (e.g., certain aspects may differ between the original experiments and the replication research or else researchers may aim at matching the same exact conditions, etc.; see Fanelli [2018] and Fidler et al. [2018] for perspective). For one thing data assessment experiments have explicit instructions, a training period for the participants, are either paper-based or computer-based and may use a Likert scale (among other alternatives). We don't know whether those details match the way the published data were collected. This being said, those differences are motivated by the interest to control for artifacts (e.g., conflicts of interest, satiation effects) and to determine the generalizability of the data (see Fiddler and Wilcox [2008] for a summary of the replication literature across fields).[5]

Additionally, the tendency to categorize results in terms of statistical/non-statistical significance – as if 'they were categorically distinct' – is also controversial, and a call has been issued 'for a stop to the use of *P* values in the conventional, dichotomous way – to decide whether a result refutes or supports a scientific hypothesis' (Amrhein et al. 2019). Moreover, it should be emphasized that non-convergent results in Ortega-Santos' (2020) study are overwhelmingly non-statistically significant; there were only two false positives – statistically significant in the unpredicted direction – across both experiments, which is 2.564% of the total sample of 78 data points taken into account. Furthermore, according to Fanelli (2018: 2630), the 'science in crisis narrative' does not promote better science, but rather an 'inspiring' narrative of transformation and empowerment would. In fact, it is not uncommon to hear complaints by syntacticians about the aggressiveness of the criticisms against their field. Clearly, inspiring syntacticians to improve data collections standards would be beneficial (e.g., see Featherston's [2007] reference to the stick and, most importantly, the carrot when discussing data collection standards in Generative Grammar).

So, there is a low replication rate, and we have found a culprit: potential sociolinguistic variation. At this point, it is worth stopping and reflecting. One recurrent topic in the literature aiming to improve data collection standards in Generative Grammar is that discrepancies concerning data are dismissed as dialectal variation. For instance, in the words of Labov (1970), 'when challenges to data arise on the floor of a linguistic meeting, the author usually defends himself by stating that there are many "dialects" and that the systematic argument he was presenting held good for his own "dialect".' In a similar vein, Bisang (2011) claims that Chomsky's ideal speaker-listener and homogenous community implies a level of abstraction

that prevents researchers from falsifying any claims or reproducing (replicating) research results. In Bisang's (2011: 242–243) words,

> the ideal speaker-listener simply does not act in a concrete context or speech situation with individuals with their specific mental states and their specific social positions. From such a perspective, Chomsky's ideal speaker-listener is context-free and thus beyond reproducibility as far as social factors and factors of form-function mapping are concerned.[6]

While Bisang's remarks are relevant, the current research is narrower in nature in that it addresses the concerns with data collection methods. For current purposes, Bisang's perspective underscores precisely the point made in this section: Without taking into account sociolinguistic factors, the results of data assessment research are hard to evaluate. This section is not meant as an excuse (see Labov's point), but rather as a call of attention for the need to incorporate enough detail in the description of the data. It is hoped that this would lead to a higher degree of convergence among the results of the various data collection methods, though this is ultimately an empirical question.

2.1.3 Why adopt formal data collection methods?

A brief discussion of the rationale behind experimental designs is included in this section. Researchers are encouraged to consider those relevant factors when deciding whether an experiment is necessary.

Following quantitative standards and formal data collection protocols is a sign of the maturity of any science. Still, for researchers planning to follow the traditional data collection method in the interest of time, it is recommended that they at least systematically disclose the origin of the judgments: how many speakers were polled, from which dialect(s), what training in linguistics their participants had, and their age and educational background. This alone would go a long way to help readers evaluate whether a conflict of interest might have been present. If the judgments are not from the researcher and the participants do not have anything at stake in the results, it is unlikely that a conflict of interest might be present. In other words, reaching a compromise between the requirements of a formal experiment and the more traditional approach would be in everybody's best interest. Inspiring editors to request such a level of transparency would be crucial (see Fanelli [2018] on the importance of 'inspiring' researchers, editors, etc.). Editors (and reviewers), as the gatekeepers of published work, care about transparency and, thus, may request not only the said information, but also the systematic inclusion of human subject approval details for experimental work in the publications. Likewise, requesting that the experimental results be fully posted in online databases (e.g., raw scores as opposed to only making *z*-scores or aggregate analyses available in the publication), and encouraging the publication of the full stimuli set is also highly recommended.

Beyond the fact that formal experiments allow us to control for potential conflict of interests, other interesting features include the use of multiple lexicalizations of the stimuli to avoid that, for instance, a pragmatically infelicitous choice in the data may affect the results. A researcher may, in fact, control for word frequency, sentence length, order of presentation of the sentences, and other factors. The inclusion of fillers among the target structures is relevant to avoid the possibility that the participants may sense what the researcher is aiming at (see the Clever Hans Effects, Wasow and Arnold [2005: 1483], or Gibson and Fedorenko [2010: 233]) or plainly to avoid satiation effects potentially present when the author uses their own judgments or when the participants are overexposed to a particular structure (see Snyder 2000).

To sum up the discussion so far, reaching a compromise between a nonexperimental data collection and an experimental data collection is possible. That said, I would like to outline very briefly why experiments are valuable from a methodological point of view (see also Section 2.1.1.1, and Chapters 3 and 5 for arguments concerning the kind of questions that experimental work allows us to ask). In particular, formal experimental designs allow for various strategies to control for data quality, e.g., training the participants to become familiar with acceptability judgments, so-called gold standard controls or honey pots (stimuli whose rating is known to the researcher and, thus, might be used to identify and exclude sloppy participants), or checking for the time participants took to complete the experiment, as particularly fast participants might not pay enough attention to the stimuli (see Ortega-Santos [2019] and references therein).

In turn, the use of statistics allows us to generalize across speakers and help make sure that an effect found when pooling few speakers is not found by chance, e.g., due to error variance, that is to say, 'a random scatter of individual observations around a more or less stable mean' (Cowart 1997: 31). But just how much statistics is needed? Can't we just calculate average response scores? Not quite. For instance, participants may naturally show biases in the way they use acceptability scales (so-called Likert scales); e.g., some participants may tend to use only the extremes of the scale, while others may make more fine-grained distinctions. Z-scores allow us to control for this issue (see Cowart 1997: 130–131). Furthermore, say we test a sentence with two participants. They judge the sentence using a 7-point scale, and both rate the sentence with a 3.5. The mean rating is 3.5. Let's consider an alternative world where those two speakers rate the sentence with a 1 and a 7, respectively. The mean is 3.5, just as in the first scenario. However, those two results are completely different. Statistics will let us see through these differences.[7] The downside of the use of statistics can be seen in the exclusion of idiolectal variation from the analysis. Indeed, according to Henry (2005), idiolectal variation is a prominent feature of natural language and averaging sentence ratings across speakers would prevent us from achieving descriptive adequacy.[8] Still, researchers concerned about this issue will not only analyze the aggregate data of the participants but also the data of each participant separately.

The inclusion of a large number of participants in the research project, irrespective of whether experimental or sociolinguistic methodology is used, has one

further advantage: It would make researchers seek the approval of human subject research review boards and be explicit about this approval in the corresponding publications. Syntacticians who work with a relatively small number of participants or who show a tendency to not be explicit about their data collection methodology will naturally be less likely to seek this approval. While seeking the approval of the human subject review board might be time-consuming, particularly the first time, once the author has developed a consent form, instructions for the participants, etc., it becomes significantly easier.[9]

Last but not least, the large number of participants and the need to compensate them and/or travel to gather the data allow researchers to justify requests for funding in a way that 'armchair linguistics' projects may not, depending on the specifics of the funding program.

2.1.4 Challenges in the use of acceptability judgments in experimental settings and beyond

Still another issue relevant to the discussion of data quality in Generative Grammar has to do with the very nature of acceptability judgment tasks: It has been argued that their exam-like flavor may make them less appropriate for the analysis of stigmatized linguistic features, as speakers may consciously or unconsciously give lower ratings to the less prestigious forms, irrespective of the exact instructions associated with the task (e.g., Labov 1970 or Adli 2015, a.o.). In my own research, for instance, when researching the syntax of comparative constructions in Spanish, I heard a speaker of Chilean Spanish use naturalistic examples of the nonstandard variant in (6)b alongside the standard variant, (6)a. When attempting to establish the source of the Case in the nonstandard variant, the relevant speaker – a family member – denied using that structure, a result found repeatedly in the sociolinguistic literature for other linguistic features (Labov [1970]; see Gutiérrez-Rexach and Sessarego [2014] for gender agreement in Afro-Andean Spanish). Thus, a formal experiment was deemed necessary (see Ortega-Santos 2013), in particular, an indirect acceptability judgment task: Speakers were asked to judge not whether they would use a certain variant, but rather whether the relevant variants could be used by somebody in their social network, as a way of decreasing the pressure of normative language (see Labov 1972, Barbiers and Cornips 2000).

(6) a. Pedro es más inteligente que yo. *Standard Spanish*
Pedro is more intelligent than I.
' Pedro is more intelligent than me.'

b. Pedro es más inteligente que mí. *Nonstandard Chilean Spanish*
Pedro is more intelligent than me.

Additionally, the literature has reflected the fact that there are also limits to the usage of formal data collection protocols to gather acceptability judgments (e.g., for the so-called yes bias in acceptability judgment tasks in heritage speakers, see Polinsky [2018]). Moreover, as discussed by Leivada et al. (2019), there

are linguistic varieties for which the inclusion of a high number of participants is plainly not possible (e.g., endangered languages; see also Wasow and Arnold [2005: 1484]). Likewise, relevant factors for experimental work, such as controlling for word frequency when designing the stimuli might not be possible either, due to lack of information on those frequencies. This observation led Leivada et al. (2019) to argue that adhering to data collection standards is highly desirable, but it is not possible in certain cases. Furthermore, as noted by these researchers, the high degree of variation in small, young, or nonstandard varieties means statistical analysis might not be possible. Needless to say, this will vary from case to case; sometimes it is statistical analysis that helps us reach novel observations (Cognola et al. 2019), particularly when a non-trivial number of sociolinguistic variables are incorporated in the design (Sheehan et al. 2019), precisely to see through this variation. Still, Leivada et al.'s (2019) note of caution regarding the usefulness of statistical analysis for certain varieties is well taken.

2.1.5 Data collection marketplaces as a tool to increase quantitative standards: Advantages and limitations

The availability of convenient data collection services has been welcomed as a way of enhancing the quantitative standards in Generative Grammar (Gibson et al. 2011, Gibson and Fedorenko 2013). Still, the usefulness of these platforms varies from language to language. While Amazon Mechanical Turk has been praised as a fast and reliable service to gather English data (see Gibson and Fedorenko [2013]; see Sprouse [2011] for a study of the reliability of the data gathered through this service), Ortega-Santos' (2019) evaluation of the suitability of this marketplace for research on Spanish revealed that it is helpful primarily to study US Spanish (e.g., heritage Spanish or language contact, whether it is in contact with English or among varieties of Spanish) and Venezuelan Spanish. Recruiting enough speakers from other countries can prove challenging.[10] Moreover, within the field of linguistics, little attention has been paid to ethical concerns inherent to the use of marketplaces such as Amazon Mechanical Turk, as workers do not receive, for instance, any benefits (Fort et al. 2011). Clearly, it needs to be part of the discussion, particularly as more ethical alternatives are viable depending on the variety under study, e.g., embedding experiments within course requirements (as part of the student's introduction to research methods, etc.), offering extra credit to students for participating in experiments or the gamification of data collection tools. That being said, it is not my purpose to recommend or reject any data collection tool. For one thing, Amazon Mechanical Turk provides fast access to a pool of participants that is more diverse than higher education students recruited at the university setting (Berinsky et al. 2012: 352), though Spanish-speaking Turkers in general tend to be young, highly educated, and urban (see Ortega-Santos 2019). Still another marketplace worth considering is Prolific (https://www.prolific.co/), which is specifically focused on behavioral research and includes a minimum reward an hour (as opposed to Amazon Mechanical Turk). In fact, Peer et al. (2017) researched alternatives to Amazon Mechanical Turk and found that workers or participants in Prolific were

more honest than in the former marketplace when determining whether they were entitled to receiving a bonus. Data quality is similar in Amazon Mechanical Turk and Prolific, though the response time was slower for the latter service. Participants were more naïve in the latter case, too. Moreover, the participants in Prolific were more diverse in their geographical origin (though at the time Peer et al. [2017] were writing, only 4% of the users were Latino/Hispanic). Whether the demographics of these platforms have been changed by the pandemic is an open question.

2.1.6 Replication in closely related fields

An understanding of data assessment research in Linguistics would be slightly incomplete without an overview of related research in other fields. As noted by Fidler and Wilcox (2018), concerns about data quality across fields have led to a growing literature on the topic, known as meta-science or meta-research. Similarly, this worry has been fueling an Open Science movement that champions transparency (as well as the availability of research outcomes to society in general and open access publishing). Needless to say, the debate takes a slightly different form in each field. Experimental research, for instance, has suffered from so-called publication bias which disfavors the publication of replication studies while favoring the publication of statistically significant results as opposed to null results. The Open Science movement has focused on avoiding these and other closely related issues by advocating the pre-registration of experiments prior to carrying them out, publishing the complete data sets and stimuli as opposed to just the results of the statistical analysis and some illustrative examples, etc. (Fidler and Wilcox 2018; countries and institutions may have their own repositories). The data assessment research in Linguistics is part of this concern about data quality. In particular, while data assessment research in Linguistics has emphasized the need to adopt scientific standards from the domain of Cognitive Science, e.g., Psychology, and the Social Sciences, these fields are not immune to concerns about data quality, in spite of having adopted appropriate standards (in fact, not even the so-called hard sciences, e.g., biomedical research, are immune to these concerns, see Fidler and Wilcox [2018] for discussion). Specifically, the reproducibility of psychological research has been shown to be a meager 36% of significant results for a body of data of 100 experiments (Open Science Collaboration 2015). In turn, the Social Sciences Replication Project attempted to replicate 21 experiments published in *Nature* and *Science*, and found a 62% replication rate (Camerer et al. 2018).

While this state of affairs should not serve as an excuse, it highlights the fact that experimental standards, by themselves, are not enough. To be clear, researchers advocating the use of quantitative/experimental data collection methods do not claim that quantitative standards are the solution to everything; rather, care needs to be exerted. It is worth noting that an effort as thorough as the Open Science Collaboration involving a non-trivial number of languages is still lacking in the field of linguistics. I suggest implementing a version of this initiative for syntax research: Let's build a network able to test each data point out of 100 papers including a variety of languages and focusing on the most important generalizations/

theoretical developments in the field. Such an effort would help address a particularly important issue: Even if one assumes that informally-gathered data might show a certain degree of unreliability, one would want to avoid throwing the baby out with the bath water. In other words, one would want to know which generalizations do hold. This is particularly important as the potential unreliability of the theoretical literature may be one of the factors hindering the unification of the language sciences and the status of Generative Grammar among the language sciences (Marantz 2005, Gibson and Fedorenko 2013, a.o.). While the option of creating an online database with data to crowdsource judgments has been put forward by Linzen and Oseki (2018), the current proposal – which includes a more controlled environment – could have a more straightforward impact. In fact, as noted by Ortega-Santos et al. (2019), data assessment is inherently present in L2 syntax research due to the methodological need to include L1 control groups (though see Rothman et al. [2022] for a critique of the use of monolingual control groups for research on bilingualism). Thus, a non-trivial body of data assessment results is already available.

2.1.7 Interim summary

An overview of the use of acceptability judgments in Generative Grammar has been provided, focusing on (a) the early discussions on the relationship between the perception of acceptability and processing costs, (b) the concerns about the potential unreliability of informally-gathered acceptability judgments, (c) the virtues and possible limitations of experimentally-gathered data, (d) the need for syntactic research to be explicit about data-gathering protocols and the origin of the published data, and (e) the use of marketplaces for linguistic research. Next, the dialogue between Generative Grammar and Variationism/Corpus Studies is considered with an emphasis on its relevance for the debate on the pertinence of using acceptability judgments as a source of data.

2.2 Acceptability judgments vs. data-intensive research using naturalistic data: Variationism and Corpus Studies

The widespread use of acceptability judgments as evidence in generative syntax has been challenged from a variationist perspective as follows: Variationism has emphasized the pervasive existence of inter- and intraspeaker variation in language and the somewhat slow speed of the competence-oriented syntactic theory when catching up with this observation (e.g., Cornips and Gregersen [2016] or Guy [2015, a.o.]). For instance, Guy (2015) notes the following: given that the standards that Chomsky outlined for the syntax field include the notion of 'descriptive adequacy', variation, which is a key property of language, should not be exempt from such standard and cannot be dismissed as performance or usage. In fact, the existence of variation – often contrasted with Chomsky's (1965) homogeneous community idealization and ideal speaker-listener – is revealed not only by Sociolinguistics/Dialectology, but also by the field of acquisition, where children allegedly entertain

different hypotheses or grammars (e.g., see Roeper [1999] and Eide and Åfarli [2010] for explicit discussion of the commonalities between the study of dialectal variation and acquisition and their relevance for the study of syntax).

That being said, the relationship between Generative Grammar and Variationism and/or Corpus Linguistics is made complex by the fact that the data used in each discipline are of a different nature.[11] Attempting to bridge the gap between both disciplines entails answering the question of how exactly usage preferences and/or frequencies fit in generative syntax, e.g., the relationship between acceptability and frequency and the hypothetical existence of optionality in the grammar (see Adli et al. 2015 for an overview). Accordingly, the following issues are discussed in light of the relationship between Variationism, Corpus Linguistics, and Theoretical Syntax in this section:

(a) The relevance of data-intensive research for theory-driven approaches such as Generative Grammar (e.g., Mendívil-Giró 2019).
(b) The dissociation between acceptability and presence/absence in corpora as well as the cases in which naturalistic data may inform generative research (e.g., Newmeyer [2013] or Adli [2011, a.o.]).
(c) The incompleteness of the data present in corpora or even the internet, again revealing that technology is helping our field attain higher levels of empiricism, though only gradually, as there are still certain limits in the availability of resources for linguistic research.
(d) The different granularity of the data, that is to say, level of detail, in variationist vs. generative research (e.g., frequency, preferences, and/or probabilities of use vs. acceptability contrasts).
(e) The methodological advantages that each field can contribute to the other.

An overview of the approaches aiming to address the relationship between these fields is provided ranging from the view that grammar and usage need to be kept separate (e.g., Kroch [1989], Adger and Smith [2005], Embick [2008], or Newmeyer [2015]; see the classic competence vs. performance distinction) to proposals that integrate frequencies of usage into the grammar (e.g., see Sankoff and Labov [1979] for early proposals or, more recently, Henry [2005]). Further topics include a brief discussion of corpora for research in Spanish syntax as well as the challenges of using the worldwide web and social media to gather data.

Section 2.2.1 discusses to what degree data intensity is relevant to Generative Grammar, Section 2.2.2 focuses on the relationship between acceptability and naturalistic data, Section 2.2.3 fleshes out the advantages and complexities inherent in using social media and the web to gather data, Section 2.2.4 provides a brief overview of Spanish corpora for syntactic research, Section 2.2.5 addresses the issue of why a fluid relation between Generative Syntax and Variationism benefits both fields, Section 2.2.6 discusses approaches aiming to bridge the gap between Sociolinguistics and Generative Grammar, and Section 2.2.7 discusses how a theoretically-informed Sociosyntax should look like.

2.2.1 Data intensity and its relevance to Generative Grammar

Variationism, when compared to traditional Generative Grammar, is an example of data-intensive research. So is Corpus Linguistics, which benefits from the increasing availability of corpora, the use of the internet or specific social media sites such as Twitter (recently rebranded as *X*) for research – in short, whatever is sometimes referred to as big data (irrespective of whether this label is actually helpful). These fields, therefore, have the potential to help Generative Grammar increase its data intensity as part of the crosspollination effort. That being said, as noted by Mendívil-Giró (2019), the potential for data intensity to have an effect on linguistic theorizing depends on whether the theory in question is functionalist or formalist in nature and the corresponding scientific method, namely, inductive or deductive. Data intensity is more crucial for the development of the former kind of theory, as it relies primarily on data to arrive at generalizations. Generative Grammar, in contrast, is an example of a formalist theory, meaning its research enterprise is theory-driven; it may derive generalizations from the study of a single language, which then can be tested crosslinguistically, as necessary for the theory to be falsifiable. So, Mendívil-Giró concludes that data intensity is relevant to generative research, but not as important as it is to functionalist research. This being said, while Mendívil-Giró (2019) has a point, so does Guy (2015), when reminding the field that descriptive adequacy has always been a goal in generative research and that inter- and intraspeaker variations should not be neglected. The study of syntactic microvariation, in fact, aims at addressing this issue. The challenge ultimately boils down to how that relationship between Sociolinguistics and Corpus Linguistics, on the one hand, and Generative Grammar, on the other, can be mutually beneficial, given the data granularity and analytical tools available to each field. This is discussed next.

2.2.2 Acceptability and naturalistic data: Presence/absence in corpora vs. acceptability

Generative Grammar has relied on acceptability to inform theoretical developments by considering that acceptable structures are part of the mental grammar of the speaker. That is to say, they are grammatical, and unacceptable structures are not, that is to say, they are ungrammatical (barring exceptional circumstances discussed in Section 2.1.1). The use of naturalistic data opens the door to the study of the grammar while avoiding the metacognitive process involved in acceptability judgments. Crucially for present purposes, the absence of a structure in a corpus or in sociolinguistic interviews cannot be equated with unacceptability. In particular, sentences deemed unacceptable might be attested, whereas unattested sentences can be grammatical (e.g., Conrad 2010). The latter case might be the result from a bias in the corpus or questionnaire, etc., say, when only a very limited variety of registers are included.[12] In this regard, it is important to emphasize that naturalistic data gathered through interviews or included in corpora include positive evidence, but not negative evidence. The same goes for the data found in traditional dialectal work. Thus, a theoretical syntactician may benefit from naturalistic data, when the hypothesized contexts where a syntactic feature occurs, as

revealed by acceptability judgments, and the contexts attested in the corpus are in complementary distribution, scenario (a) in Figure 2.1, be it complete or partial complementary distribution. In a similar vein, when the syntactic contexts where the syntactic structure is hypothesized to occur are a subset of the contexts where it is attested in the corpus, scenario (b) in Figure 2.1, corpora will be crucial as well. In contrast, when the hypothesized contexts where a linguistic feature occurs are a superset of the actually attested in the corpus, scenario (c) in Figure 2.1, corpus data may not be rich enough to the point that they may mislead a researcher, if he or she only relies on the corpus (see Adli's [2011: 398] latent constructions, that is to say, 'a form that is available grammar-wise but not used').

Multiple wh-movement constitutes a case of a structure traditionally considered unacceptable in Spanish both in theoretical and descriptive work (e.g., for the latter, see RAE [2009: 3173]) and yet attested in online sources (Bazaco [2014], his data; for discussion of multiple wh-movement in other languages, see Bošković [1997, a.o.]):

(7) a. ¿Quién con quién cruzo?
who with who breed.1sg
'Who with who do I breed?'

b. ¿Quién con quién juega?
who with who plays
'Who with who plays?'

This would at least potentially correspond to the scenarios in (a) or (b), Figure 2.1, depending on our point of view. These structures were thought to be unacceptable, but they are not (scenario (a)). Alternatively, we can look at this issue as follows: wh-movement is attested in more contexts than previously thought, namely, in multiple questions, more than one wh-element may move (scenario (b)).

With regard to the scenario in (c), Figure 2.1, the very limited presence or even absence of nonstandard comparative constructions, (6)b, in corpora is a case in point. A search in the *Corpus de Referencia del Español Actual* (CREA;

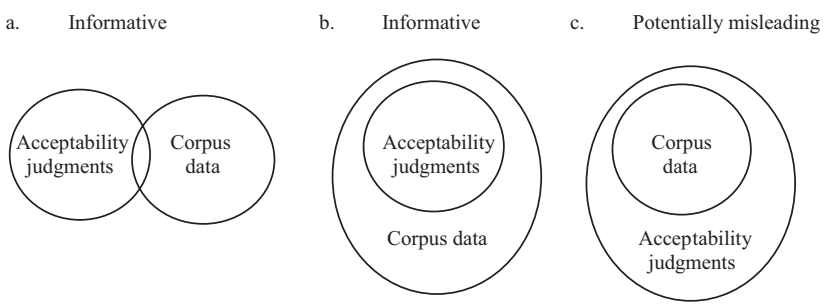

Figure 2.1 Relevance of data extracted from corpora depending on their relation to the data gathered through acceptability judgments.

Reference Corpus of Contemporary Spanish) and the *Corpus Diacrónico del Español* (CORDE; Diachronic Corpus of the Spanish Language) returned only one case in the first corpus, and revealed its productive use in only the *Biblia de Ferrara* (1553) out of the whole CORDE (Ortega-Santos 2013). Any investigation based solely on these corpora would have been somewhat misleading, at least with regard to the actual geographical distribution of this feature. This being said, as noted by Cornips and Gegersen (2016; see also Newmeyer 2013), the size of the current corpora lessens the seriousness of this issue thanks to technological advances. Nonetheless, Theoretical Syntax research is based on highly specific syntactic contexts – think of the need for minimal pairs manipulating highly specific syntactic features to build an argument while controlling for interfering factors (Newmeyer 2013, a.o.). The issue, then, is whether naturalistic data, irrespective of the size of corpora, etc., can actually meet those needs. The use of statistical generalizations may help lessen this problem but only to a certain degree.[13] That said, the importance of data triangulation cannot be emphasized enough (see Chapters 3 and 4). For instance, a corpus-assisted analysis, that is to say, an analysis using data from corpora together with other kinds of evidence (Baker 2010: 08), might be particularly pertinent for syntactic analysis.

For the sake of the presentation, the discussion so far has been put in terms of the presence/absence of a feature in a corpus. Still, corpus research can inform syntactic analysis and/or falsify syntactic analyses by studying the correlation among linguistic features (e.g., to test parametric accounts that link various properties together or to make predictions regarding the impossibility of two syntactic properties or more to co-occur). In other words, the questions we can ask are more sophisticated than what Figure 2.1 suggests, provided that those can be answered by relying on positive evidence. Variationism, in any case, goes beyond the study of rates or frequencies of use and the presence/absence of features, to analyze the relative role of conditioning factors in the occurrence of a variable. This can provide valuable insights to understand syntactic properties (see, for instance, the discussion on null subjects in English and Spanish by Torres Cacoullos and Travis [2019] in Chapter 4).

That being said, frequency of use is relevant in that it is one of the factors that affects the perception of acceptability (Featherston 2007). For instance, Sampson (2007: 16) argues that 'speakers cannot be expected to make reliable judgments about the status of the unusual word sequences.' In fact, Sampson, as well as Conrad (2010), emphasizes that teasing apart extremely rare constructions from performance errors is a challenge as both are highly infrequent.[14]

For the sake of completeness, it is worth noting that the use of acceptability judgments has also been challenged in the following terms which further emphasize the importance of naturalistic data as well as rare or infrequent constructions: Sampson (2007: 14) challenges the use of negative evidence by arguing that 'it is startling to find 20th- and 21st-century scientist maintaining that theories in any branch of science ought explicitly to be based on what people subjectively "know" or "intuit" to be the case, rather than on objective, interpersonally-observable data.' As part of the evidence he cites, he includes a case of a heavily discussed data

point, which was inconsistent with naturalistic data. Still, an effort as thorough as Sprouse and colleagues' in the data assessment literature would have been beneficial for the field. Furthermore, he rejects idiosyncratic variation (cf. Henry 2005) considering that each language gives us a space of syntactic possibilities and that some might be reinforced by the educational system or those around us; certain options become a 'habit', but other options are still available unbeknown to us. Thus, some patterns might be highly frequent, while others might be highly infrequent and yet they might become widespread in the future. Ungrammaticality would not be the right notion to capture this variation, nor would idiolects. In a similar vein, Barbiers (2009: 1622) claims that

> when an informant says that a certain variant does not occur in his dialect, this can either mean that the variant is excluded by his grammatical system or that it is grammatical but happens to be absent in his dialect. (…) If the distinction between ungrammatical and unrealized structures is real, then many ungrammaticality claims in the literature may have to be reconsidered, and this may have important consequences for syntactic theory.

As noted by this researcher, the tension between absence and ungrammaticality of a feature is present in other subfields (where the concept of 'grammaticality' is adapted to each subfield), e.g., phonotaxis or morphology. While the concern about the relationship between acceptability and frequency or the absence of a feature might still be an issue, the dichotomy between acceptable and unacceptable or anomalous sentences in some respect seems to be real. For instance, brain-imaging techniques, eye-tracking, etc. can also index acceptability without an introspective judgment (see Chapter 5 for discussion). Moreover, not all infrequent structures are unacceptable (see Adli's [2011] latent constructions), so it is unclear to what extent approaches based on naturalistic/observable data may do the work of acceptability-based proposals; rather, crosspollination and data triangulation seem more promising.

At this point, it is worth addressing an implicit assumption about the relation between Generative Grammar and Sociolinguistics/Corpus Linguistics, namely, that the use of statistics in the latter fields allows researchers to arrive at more objective or solid generalizations. It is not my purpose to challenge this view. Data quality is crucial in research, as discussed throughout this book. Still, beyond the issues noted so far, it is important to bear in mind that the aggregated data subject to statistical scrutiny also include an interpretation component. Whether linguistic structure X constitutes a counterexample to generalization Y and should be coded accordingly in the analysis, is something that has to be determined – interpreted. The resulting analysis will only be as solid as the interpretation. For instance, dialectometric analyses of phonetic properties in dialectal atlases may have to consider whether all the interviewers followed the same protocols and standards to a tee and whether potential individual variation when implementing the protocols, transcribing, etc. may have had an effect on the data. Again, it is not my intention to question the virtues of data-intensive approaches. Quite the opposite. Everything

else being equal, statistical analysis will allow us to produce reliable results. Still, the results will only be as good as the interpretation inherent in the coding of the data to be analyzed statistically.

2.2.3 Using social media and the web to gather data: Challenges and limitations

Social media and the web provide easy and fast access to a truly diverse and massive amount of linguistic data. Still, their relevance is determined by the object of study. For instance, if a researcher is focusing on a specific dialect or ethnolect, the relevance will be determined by the ease of access to the web and social media that this specific population has. This ease of access may vary according to socioeconomic status, age, or, plainly put, political realities (e.g., internet access and access to social media in general are not as widespread in Cuba or Equatorial Guinea as in other Spanish-speaking countries). That being said, the use of social media and the web, in general, constitutes a unique opportunity for linguistic research, and yet it also includes some challenges or highly specific properties that set it apart from the use of corpora. Below, I summarize various works on the pros and cons of using the web as a corpus, while illustrating the discussion with examples from the syntax of Spanish.

With regard to the challenges, the language used in social media has certain features that should be taken into account: A case in point is the use of nonstandard spelling (abbreviations, capitalization, inclusion of emojis, etc.), as noted by Hundt et al. (2007), a.o. For instance, in Chapter 3, the crossdialectal use of wh-inversion will be studied through searches for the sequence *qué + tú* as in *Qué tú haces?* 'What are you doing?' *Qué* might be written as *ke* in Spanish weblish – which is why the searches in that chapter were restricted to corpora where such weblish would not be present. Moreover, the relative shortness of social media posts may cause frequency counts of grammatical features to be unreliable; everything else being equal, the chances that a specific linguistic feature is used are proportional to the length of the text (see Berber Sardinha [2021] on so-called frequency normalization to control for this issue). In a similar vein, while we may study written language in social media, the distinction between oral and written language is blurred (Hundt et al. 2007), a factor that might be an issue for certain projects, at least when attempting to determine the register of a specific linguistic feature. Last but not least, computer-mediated communication may well show some biases – the author of this book tends to use the mobile speech to text software when posting in social media. It happens to be the case that the software refuses to accept my so-called *leísmo* and automatically corrects it. It even imposed Argentinian verbal endings for a while, till the algorithm started accepting my non-Argentinian forms. Thus, the researcher may run the risk of interpreting biases in the speech to text software, the spell-checker, etc. as representative of a specific variety or register.

Further challenges come not from the linguistic features of weblish per se, but from the way we can access the data and the limited information associated with it in certain cases, e.g., whether the geographical or dialectal origin of the data may

be determined. This might be (mostly) doable in the case of social media, but it might be harder for other online environments. For instance, in Chapter 3, when working with the Corpus del Español NOW, which is a corpus consisting of online news outlets, repetitions were found in the search results across media outlets and even countries. It is worth noting that the challenge of using the web to extract data is not specific to syntactic research. Computational linguists face those same challenges when extracting and processing the data automatically (cf. text mining; see Kern et al. [2016] for discussion on the latter point as well as ethical considerations in the use of social media language). Furthermore, determining sociolinguistic variables might be a challenge, just as in (part of) the regular corpora.[15] Moreover, the results might be influenced by biases in the search algorithm (e.g., previous search history of the user, location of the user, etc.; see Hundt et al. 2007).

Other than that, when using the web as a corpus, the lack of annotations limits the nature of the searches one can run (see Leech [2007] for discussion; see also Gilquin's [2002] classic work). Let's see how. Standard corpora not only determine which kind of texts might be included (which allows for a more controlled environment, but might be a virtue or a limitation depending on the research project). Most notably for present purposes, corpora may include annotations. Thanks to those annotations, the access to syntax data might not be necessarily restricted to searching for surface strings such as the *qué* + *tú* case. In contrast, when a researcher may only search for surface strings due to the lack of annotation, this kind of search will be informative, generally speaking, as long as it involves close-class words. Why? A comparable study of the *qué* + verb sequence will be more time-consuming without annotations; the researcher may have to go through every single question using the question word *qué*. Depending on the size of the corpus, this might not be an easy task to accomplish. It is not my purpose to discourage anybody. Under certain circumstances, a search with open class words might be valuable. Still, let's compare this situation to an annotated corpus: The Syntactic Atlas of Spanish (ASinEs, Gallego 2019) allows users to retrieve linguistic examples by searching for specific phenomena, e.g., so-called *dequeísmo*, or searching for the linguistic features of a specific geographical area. In other words, annotations provide faster access to data, though the said access is only as good as the match between our research needs and the annotations.

This absence of annotations in the web has led Hundt et al. (2007: 4) to advocate for the use of the web to 'complement evidence' from standard corpora and to use the web for corpus building as opposed to using the web as a corpus. Moreover, as noted by Kern et al. (2016: 521), social media users are not a 'representative sample of the population' in that older people are underrepresented. Thus, digital gaps affecting older speakers may well render the use of social media less pertinent for certain research projects. Still, social media users are more diverse than university students (Kern et al. 2016: 517). Note that the choice of social network for linguistic research might be determined by the easiness of the availability of data. Twitter has been fairly popular among researchers, as it allowed for free limited searches by language and location. Unfortunately, rising costs in the use of Twitter (now X) for research purposes are an issue.

2.2.4 A brief excursus on Spanish-language corpora and their relevance for syntactic research

This section provides a non-exhaustive overview of various Spanish-language corpora briefly noting how they may inform syntactic research. The following corpora stand out as particularly useful, due to their annotations and/or emphasis on syntax:

- The *Syntactic Atlas of Spanish* (*Atlas Sintáctico del Español*) (ASinEs, Gallego 2019, http://asines.org/) includes data from a wide variety of sources (ranging from descriptive grammars to traditional corpora) and allows researcher to access the data by syntactic phenomena, geographical area, or source of the data, among other options.
- The *Base de Datos Sintácticos del Español Actual* (The Syntax Database of Contemporary Spanish, my translation, https://www.bds.usc.es/) includes annotations regarding verbal properties (subcategorization frames); see also the closely related *Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* (ADESSE; Database of Spanish Voice Alternations and Syntax-Semantics Schemes, my translation, http://adesse.uvigo.es/) for syntactic functions (subject, object, etc.), theta roles, and syntactic categories (finite/non-finite clause).
- The *Audible Corpus of Spoken Rural Spanish* (*Corpus Oral y Sonoro del Español Rural*, COSER, http://www.corpusrural.es/) consists of interviews carried out in rural areas in Spain and allows for searches, for instance, by province or gender of the informant. The mean age of the informants is 74.

There are also various corpora which may not necessarily have an emphasis on syntax or else the kind of linguistic annotations available only allow researchers a less direct access to syntactic properties (e.g., researchers may have to search for surface strings as noted in Section 2.2.3). These kinds of corpora, however, are relevant when attempting to gather naturalistic data of a certain phenomenon:

- The *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA, https://preseea.linguas.net/corpus.aspx) includes interviews conducted in cities across the Hispanic world using the same methodology. Thus, it is particularly well suited for dialectal comparisons of urban speech. Furthermore, it includes information relevant to sociolinguistic research (e.g., gender, age, educational level, or employment). Not all materials are available online, but if interested in a specific location, the researcher may contact the research team responsible for that location to ask for the rest of the interviews. Audios of the interviews are available.
- The *Corpus del Español NOW* (https://www.corpusdelespanol.org/now/) includes news from Spanish-speaking online media (7.6 billion words). As noted in the previous section, researchers may want to check for duplicates, particularly across countries, if interested in the dialectal origin of the data.

- The *Corpus de Referencia del Español Actual* (CREA; Reference Corpus of Contemporary Spanish, http://corpus.rae.es/creanet.html) and the *Corpus Diacrónico del Español* (CORDE; Diachronic Corpus of the Spanish Language, https://corpus.rae.es/cordenet.html) include contemporary and diachronic materials, respectively. In both cases, you can search by country, kind of media (journal, book, magazine, oral), or topic (theater, science, geography, etc.).

Additionally, dialectal atlases are available, sometimes online for various areas (provinces, countries, etc.). Still, they tend to not emphasize syntax, as opposed to the lexicon or the pronunciation (see Brandner [2012] or Ortega-Santos [2021, a.o.]; see Chapter 4 for discussion of the need for crosspollination among Theoretical Syntax and Dialectology).

2.2.5 Why Generative Syntax needs Sociolinguistics and the other way around

The complex relation between acceptability and usage does not prevent the dialogue between Generative Syntax and Sociolinguistics. As noted, Variationism has emphasized the existence of inter- and intraspeaker variation in natural language, thus contributing to other subfields of Linguistics, particularly Dialectology, but also Generative Grammar. According to Cornips and Gregersen (2016: 502), this interdisciplinary crosspollination between Generative Grammar and Variationism has led to the 'gradual empirification of armchair linguistics,' though the distinct goals of each discipline remain unmodified. In particular, the study of microvariation in syntax, that is to say, dialectal variation, is receiving increasingly more attention. For this line of study, the strategies to collect data of stigmatized varieties as developed by sociolinguistic research are particularly relevant, e.g., to overcome the Observer's Paradox by including various members of the community in a given interview and, possibly, a trained interviewer from the very community (Labov 1970, 1972, Barbiers 2009, a.o.) to avoid that interviewees may accommodate their language to the variety spoken by the interviewer. For detailed discussion of data collection methods in research on microvariation, see Brandner (2012) and references therein.

In turn, as argued by Adli (2015), variationist research would benefit from using acceptability judgments to determine 'the envelope of variation in syntax (by taking into account acceptable but scarce constructions),' given the frequency/acceptability mismatch, whereby a rare construction might be acceptable.[16] Arechabaleta Regulez and Montrul (2021; see also references therein) further enlighten the relationship between acceptability and frequency of use (production) in the following way: They focus on variation in DOM to reveal that acceptability judgments and eye-tracking, that is to say, research on sentence processing, can reveal variation and language change before production does. While Arechabaleta Regulez and Montrul used an oral narrative task and an elicited production task, resulting in a more controlled linguistic context than your average corpus, they also note the relationship between Sociolinguistics/Corpus Studies and production.

Discussions on the relationship between Sociolinguistics and syntactic research also stress the fact that the concept of variable is essential to the former field. In syntax, however, the notion of equivalence among structures is less straightforward, thus complicating the use of sociolinguistic methodology for the study of syntax (e.g., see Cornips and Gregersen [2016]; see Adger [2015] for related discussion). For this reason, Cornips and Gregersen emphasize the fact that, in order to apply such a notion to syntactic variables, sociolinguists would benefit from theoretical research.

But why is the integration of linguistic variation into Generative Grammar crucial? We have already noted that (i) the standards that Chomsky outlined for the field include the notion of descriptive adequacy, and (ii) variation, as a key property of language, should not be exempt from such a standard and cannot be dismissed as performance or usage (Guy 2015). Having a fluid relation between sociolinguistic research (including but not limited to its data collection methods) and generative research allows syntactic research to include insights into the relationship between syntax and social variables. This alone would mean that the data in generative papers would be more specific in terms of its origin, to the point that the data assessment enterprise would become more meaningful, in contrast to the black-box issue concerning the origin of the data mentioned in Section 2.1.3. By the same token, by relying on the methodological standards of variationist research (e.g., systematically being explicit about the data collection methodology, including providing information on the background of the speakers, adding naturalistic data, statistical analysis, etc., when pertinent), the unification of the language sciences and the mutual relevance across fields would be promoted.

The exact contribution of these factors for a specific investigation will vary depending on the syntactic phenomenon under investigation (e.g., the use of naturalistic data may lead to new generalizations in some cases, but not others, etc.; for instance, Kato's [2015] study of null subjects in Brazilian Portuguese revealed the role of prescriptive grammar and the educational system in the use of null subjects). Linguistic features that depend on the linguistic context (e.g., information structural properties) will naturally be more likely to benefit from the use of naturalistic data. This is precisely what led Ocampo (2010) to claim that acceptability judgments are not well suited for the study of information structure, e.g., the pre-/postverbal subject alternation (see Chapter 3, Section 3.2.2, for the case study of the VOS order). Furthermore, it is important to note that the analysis of naturalistic data in Corpus Linguistics or Sociolinguistics does not stop at a descriptive level, but rather an explanation of whichever correlations were unveiled is attempted, whether it is in theoretical or cognitive terms (see Conrad [2010] for corpus studies). For instance, following a fruitful line of research, Conrad relates the tendency for heavy phrases to appear last in the sentence to cognitive principles/properties of the parse (see Chapter 3 for discussion on its effects on the syntax of subjects). Thus, the convergence between Sociolinguistics, Corpus Linguistics and Generative Grammar is not just methodological; the respective analyses can also enter into a fruitful dialogue, even if the nature of the data is different.

2.2.6 Approaches focusing on the relationship between both fields

Approaches focused on the relationship between both fields range from the view that grammar and usage need to be kept separate (e.g., Adger and Smith 2005, Embick 2008, Nevins and Parrott 2010 or Newmeyer 2015), in keeping with the classic competence vs. performance distinction, to proposals that integrate frequencies of usage into the grammar (e.g., Labov's [1972] classic work or Guy [2015, a.o.]). In turn, the frameworks adopted in these proposals range from Minimalism (Adger 2016) to Distributed Morphology (Nevins and Parrot 2010) and Optimality Theory (Seiler 2003).

When keeping grammar and use separate, researchers do not deny the existence of complexity in the usage of language (e.g., links between rates/probabilities of usage and either linguistic or extralinguistic variables) yet put it outside syntax proper. Hence, Newmeyer (2015) argues that the grammar does not have any 'numbers' (probabilities); grammatical competence would interface with a users' manual that includes instructions on all 'external factors that affect the grammar' (Newmeyer 2015: 33). Adger and Smith (2005) and Embick (2008) fall in the same line of thought as Newmeyer: The existence of variation in the usage is not denied, but rather this variation is not the grammar per se; variation is determined by a miscellaneous set of factors, e.g., ease of lexical access, and priming, a.o.. Researchers may vary in terms of where they draw the line between those factors and properties originally thought to be part of the grammar, islandhood being a case in point. Locality restrictions were originally argued to be part of the competence of the speakers (Chomsky 1981), whereas intense research in Psycholinguistics argued at least part of these to follow from processing restrictions (see Chapters 3 and 5 for relevant discussion of wh-islands and Relativized Minimality, Rizzi 1990).

With regard to the way variation is implemented within generative theory, early approaches in terms of parallel grammars (or competing grammars) within the same language have been criticized, for instance, for increasing the number of grammars available to each speaker arbitrarily, namely, one new grammar for each property subject to variation, e.g., Seiler (2003) and Nevins and Parrott (2010). For Seiler (see also references therein), Optimality Theory provides the right framework to capture variation in one single grammar that allows for more than one output. In particular, Seiler's variable output grammar allows for overlapping constraints, consisting of both domination and, crucially, unranking; this allows the grammar to yield more than one possible output. Furthermore, as argued by Seiler, Optimality Theory can be adapted to model not only variation, but also differences in the frequency of use of the structures. In turn, Nevins and Parrott develop a variable-rules approach á la Labov framed in terms of Distributed Morphology. It assumes the existence of a single grammar where rules apply probabilistically. Usage and grammar, however, are kept distinct. In this context, Adger's (2016) Combinatorial Variability Model stands out: According to Adger, generative syntax is well suited to deal with probabilistic patterns of variation found in syntax. In particular, feature checking is argued to be responsible for variation. Differences in

uninterpretable features – irrelevant for semantic interpretation – may lead to differences in the pronunciation, thus giving rise to variation. Furthermore, in Adger's proposal, if there are three variants, the probability of occurrence of each variant is 0.333 (unless there is syncretism among any of the variants). The frequency of each variant (as opposed to the probability) would be determined by a diverse set of factors, e.g., ranging from priming effects to idiosyncratic preferences for certain words. Crucially, those factors are not part of the grammar. To sum up, while the implementation of variation in generative syntactic theory varies widely, it is fair to say that the competence/performance distinction tends to be present in various proposals. Nonetheless, the interest in modeling frequencies is also evident in the literature, thus helping meet the need for descriptive adequacy even in the case of intraspeaker variation.

2.2.7 Sociosyntax: How to navigate a research program involving both fields

A research program meant to benefit from advances in both fields would include, minimally, the following (see Cornips and Poletto [2005], Barbiers [2009], Brandner [2012], Adger [2016], a.o.; for examples related to the syntax of subjects in Spanish and beyond, see Chapters 3 and 4):[17]

- data collection techniques from both fields, e.g., both interviews and acceptability judgments with strategies to avoid the effects of normativity
- a thorough study of potential variation in the data (as determined not only by naturalistic data, but also by acceptability judgments; see Adli [2015]) and the potential contribution of sociolinguistic variables to the analysis
- tools of analysis from both fields (e.g., any statistical analysis should include theoretical notions or primitives relevant to the object of study, e.g., c-command, phases, or parameters, a.o.)

The incorporation of this perspective would go a long way to help the interdisciplinary dialogue. For discussion of specific examples of sociosyntactic research in Spanish, see Chapter 3, Section 3.3.1.

2.2.8 Interim conclusion

The relationship between Generative Grammar, on the one hand, and naturalistic data and the disciplines focusing on it (Sociolinguistics and Corpus Linguistics), on the other, has been analyzed. The emphasis has been put on how these disciplines may complement one another and on the pros and cons of available resources for research on Spanish syntax and beyond, e.g., corpora or social media. Theories aiming at bridging the gap between both fields have been presented as well. Chapters 3 and 4 consist of an investigation into these issues through specific study cases. In particular, Chapter 3 focuses on the contributions of variationist (and experimental) research to our understanding of subjecthood in Spanish. In turn,

Chapter 4 consists of a corpus study on the overt/null subject pronoun distinction in Spanish. This corpus study is designed to allow for the results to be contrasted with the generalizations present in the generative literature.

2.3 Conclusion

This chapter has discussed the ongoing dialogue between Generative Grammar, on the one hand, and experimental linguistics and Variationism, on the other. An argument has been made for the integration of the respective data collection methodologies, so as to provide a more comprehensive view of the object of study. In doing so, the ongoing discussion about the pertinence of using acceptability judgments as the main source of evidence in syntactic theory has been reviewed, e.g., the so-called data assessment literature, in which the results of experimentally-gathered syntax data are compared to published data based on acceptability judgments which were neither gathered through experiments nor analyzed statistically. The argument for data triangulation is summarized in Figure 2.2, which suggests that acceptability is influenced by a number of factors (such as processing costs and frequency of occurrence of a specific structure or feature).

Throughout this chapter, it has been argued that using various kinds of data will provide a more complete view of the object of study clarifying potential confounds. Chapters 3–5 include specific case studies that exemplify the importance of this point.

Moreover, the convergence among subfields in the data collection methods allows us to go beyond the association of Generative Grammar and experimental linguistics with both positive and negative data as opposed to the association of Sociolinguistics (and Corpus Linguistics) with only positive evidence. Clearly, acceptability judgments can be integrated in sociolinguistic research (see Chapters 3

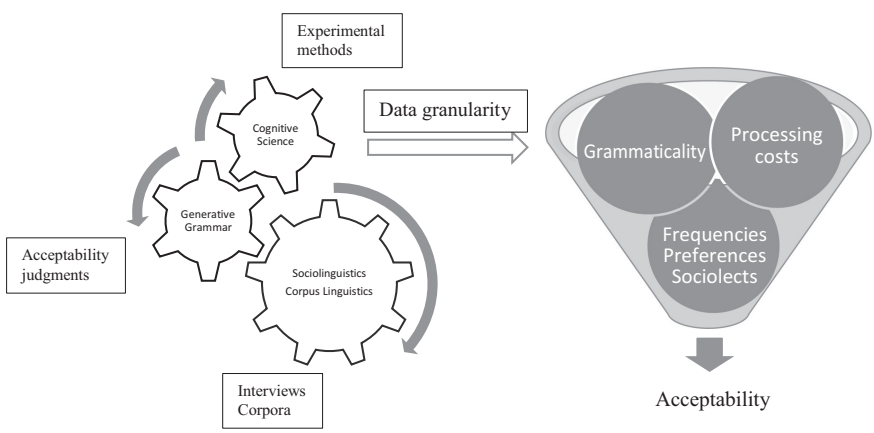

Figure 2.2 Interdisciplinary relations, data collection methodology, data granularity, and their effect on the perception of acceptability.

and 4), and, as noted in Chapter 1, the boundaries among subdisciplines are not hard boundaries as seen in Sociosyntax research programs.

Beyond the fact that acceptability judgments might not be suited to collect every single kind of data (Section 2.1.4), it has been claimed that in the absence of detailed information concerning the dialect/sociolect under study, the results of the data assessment literature are hard to interpret, and that the label data representativeness is to be preferred to data assessment. The rationale for the adoption of formal data collection methods has been spelled out (Section 2.1.3).

In turn, the importance and limits of naturalistic data for the study of syntax have been discussed in Section 2.2 (e.g., the challenges of integrating an acceptability-based framework as Theoretical Syntax with a frequency-based framework as Sociolinguistics). The distinct contributions of each field to our understanding of syntax, as well as approaches aimed at bridging the gap between the two fields, have figured prominently in the discussion. Furthermore, the usefulness of data collection marketplaces (e.g., Amazon's Mechanical Turk, Section 2.1.5), and the pros and cons of the use of corpora (Section 2.2.2), social media, and the web to gather data (Section 2.2.3) have been presented.

Notes

1 Acceptability judgment tasks can be modified to be able to address such questions to some extent. For instance, Sorace (2011: 20), when reviewing research on bilingualism and the differences in the processing heuristics of monolingual vs. bilingual speakers, notes that 'external load […] (in the form of a concurrent tasks, for example, or time pressure, as in speeded grammaticality judgment test) can provide information about the point at which the processor breaks down under load and thus about the processing resources available to the speaker.'

2 It is interesting to note that research advocating for higher data collection standards does not necessarily adhere to those standards systematically, possibly for rhetorical reasons. For instance, Wasow and Arnold's claim that 'journals are full of papers containing highly questionable data' ideally would not be supported by 'as readers can verify simply by perusing the examples in nearly any syntax article about a familiar language,' but rather the issue should be quantified using proper data collection methods and statistical analysis.

3 There is, in fact, a separate debate on how the judgments should be collected within experimental paradigms (e.g., using a so-called Likert scale or magnitude estimation, etc.; see Sprouse 2023 for discussion).

4 Concerns for the representativeness of the data are not unique to syntax. In fact, this concern lies at the heart of the use of statistics in Sociolinguistics and Corpus Linguistics, and it plays a prominent role in the design of corpora (see Leech 2007). Thus, by adopting the label representativeness, we are underscoring a trend common to various fields or subfields, opening the door to benefiting from existing discussion in those other fields. Interestingly enough, Leech's discussion of representativeness in corpora considers it to be a scalar phenomenon. While the results of the data assessment research could also be thought of in terms of a scale, it is not immediately obvious where one should draw the line in the scale to separate appropriate from inappropriate results (e.g., the English results in Sprouse et al. [2013] vs. the Spanish results in Ortega-Santos [2020], etc.).

5 Sprouse and colleagues test the data with various judgment tasks (7-point Likert scale, Forced Choice, Magnitude Estimation), thus revealing whether the results are found

consistently across methodologies (construct validity). In turn, when studying whether the results can be generalized to different populations as Sprouse and colleagues did, too, we are focusing on the external validity (Fidler and Wilcox 2018).

6 Ironically, in Bisang's (2011: 253) view, usage-based linguistics do not fare any better in the reproducibility test: 'texts are unique in the sense that they cannot be reproduced. They have been uttered once for a specific purpose and that's it – it is extremely unlikely that they will be produced in identical form for a second time.'

7 While there is a learning curve associated to the use of statistics, there are introductions available in the market (e.g., Winter 2020), as well as open-source textbooks (e.g., Schneider and Lauber 2010). In particular, Winter (2020) provides a reader-friendly and practical introduction to statistical analysis with R – a free programming language and statistical software – in linguistics. As pointed out by Winter (2020: xiv), R is part of the Open Science movement, which positions it better in the field than pay-to-use software alternatives. Moreover, universities frequently offer statistical consulting, either free or for a reduced rate.

8 Henry (2005: 111) argues through a study of agreement in sentences with expletive *there* that the syntactician needs to work with each speaker individually to explore what factors determine the acceptability for him/her. In her view, 'this cannot be done with a predetermined list of sentences; and often speakers will volunteer information about what they can and cannot say, or what they would say instead, that enables the linguist to arrive at a more complete an understanding of their grammar than would otherwise be possible.' Henry also advocates for an understanding of the grammar in terms of rules whose application might be determined by linguistics and extralinguistic factors (e.g., register). This emphasis on idiolects contrasts with Featherston's (2007: 284) view that individual judgments are 'noisy' and that mean judgments are needed to 'remove this error variance' and to ensure that the corresponding theories show the property of generality. Cf. also Labov's complaint that potential counterexamples to syntactic generalizations may be dismissed as dialectal variation (Section 2.1.2).

9 The number of informants is determined by the hypothesized size of the effect. Smaller effects require a higher number of informants. Cowart (1997: 83–83) argues for a minimum of eight informants for the syntactic phenomena typically discussed in the 1990s. According to some researchers, e.g., Gervain (2003), there is now a tendency to focus on particularly subtle data, which would suggest that more informants are needed. Out of the two samples studied by Ortega-Santos (2022), the more recent sample, *Probus* (2006–2017), used diacritics other than * (e.g., *? / % / # / ? / **/??) in 15.909% of the data as opposed to 12.244% for the older high-impact sample (100+ quotes, 1971–2009, mean year: 1990). This suggests that indeed the data might be more subtle or that researchers are emphasizing the subtlety in the data to a greater degree.

10 The main features of Amazon Mechanical Turk are as follows (see Ortega-Santos [2019] for detailed discussion): Researchers may log into https://www.mturk.com/ using their Amazon ID and password. To design the task, researchers can use templates provided by the marketplace, though they may also use third-party software and post the task on an external site. While workers (or participants) choose freely the tasks they would like to complete, quality control is present in that the researcher decides whether to approve the work after evaluating it. If the researcher rejects the work, no payment is issued. Moreover, the researcher determines the profile of the workers to a certain degree (e.g., language, country, age, gender, or percentage of previously approved tasks), though additional fees may apply. The researcher also determines the payment, the time allotted per assignment, and the expiration date. Amazon Mechanical Turk charges a fee on the reward paid to the workers.

11 Sociolinguistic research and Corpus Studies share a number of properties such as the use of naturalistic data or their data-intensive component. Therefore, they are discussed together in this section. As noted by Baker (2010), while Sociolinguistics is a framework for the study of language, Corpus Linguistics was originally a methodology – not

unlike experimental linguistics – focused on the study of naturalistic data. Still, Corpus Linguistics has come to be regarded as a discipline.

12 For instance, when reflecting on why the competence/performance distinction was being maintained in the 1970s much to the dismay of sociolinguists, Labov (1970: 36) states: 'The data based on what speakers actually say may be adequate for the most common phonological or syntactic forms. For any deep analysis of the sound patterns of a language, it will be necessary to elicit such rare words as *adz* (the only English morpheme ending in a cluster of voiced obstruents). In the study of syntax, the inadequacy of the average corpus is even plainer. Any attempt to specify syntactic rules inevitably involves forms which one cannot expect to hear in any limited investigation.'

13 Corpus Linguistics relies on automatic processing to extract data and arrive at generalizations. Still, there are limits to this procedure (beyond the absence of negative evidence). In the words of Baker (2010: 11), 'corpus software tends to work best when counting the presence of something (such as nouns), rather than features that are manifested through absences (such as zero articles or bare infinitives). Some patterns are too complex or are based on rules which cannot be easily encoded in a search algorithm.'

14 From the perspective of language acquisition, data input and its frequency are crucial notwithstanding the classic nature vs. nurture debates. Specifically, children need to acquire core properties of the grammar and tell them apart from more peripheric properties (see Yang 2000 for discussion).

15 In fact, as noted by Bisang (2011) and Conrad (2010), following Meyer (2002), some of the variables traditionally present in sociolinguistic research (e.g., social class or age) would not be included in corpus research, generally speaking, 'due to the difficulty of compiling a large spoken corpus that is representative of these different variables.'

16 Adli's view is in stark contrast with Sampson's (2007) claim that introspective judgments are not rich enough to provide an accurate description of the grammar. In particular, Sampson (2007:10) emphasizes the existence of 'constructions which before I encounter [them] I would not have thought of as available in my language, but which after confronting a real-life example I come to see as a valid possibility which had been available to me all along.' Needless to say, both views are not incompatible. That is precisely why data triangulation is valuable.

17 Note that dialectological work is also crucial to this discussion due to the close relation between Sociolinguistics and Dialectology; see Chambers and Trudgill (1998) for discussion; for an example of this line of research, see van Craenenbroeck and van Koppen 2023 and references therein.

References

Adger, D. 2006. Combinatorial variability. *Journal of Linguistics* 42, 503–530.

Adger, D. 2016. Language variability in syntactic theory. Eguren, L., O. Fernández-Soriano, and A Mendikoetxea (eds.), *Rethinking Parameters*. New York: Oxford University Press. 49–63.

Adger, D., and J. Smith. 2005. Variation and the minimalist program. Cornips, L., and K. Corrigan (eds.), *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam: John Benjamins. 149–178.

Adli, A. 2011. On the relation between acceptability and frequency. Rinke, E., and T. Kupisch (eds.), *The Development of Grammar: Language Acquisition and Diachronic Change – Volume in Honour of Jürgen M. Meisel*. Amsterdam: John Benjamins. 383–404.

Adli, A. 2015. What you like is not what you do: Acceptability and frequency in syntactic variation. Adli, A., M. García García, and G. Kaufmann (eds.), *Variation in Language: System- and Usage-based Approaches.* Berlin: De Gruyter. 173–199.

Adli, A., M. García García, and G. Kaufmann. 2015. System and usage: (Never) mind the gap. Adli, A., M. García García, and G. Kaufmann (eds.), *Variation in Language: System- and Usage-Based Approaches*. Boston: De Gruyter. 1–28.

Amrhein, V., S. Greenland, and B. McShane. 2019. Scientists rise up against statistical significance. *Nature* 567, 305–307.

Arechabaleta Regulez, B., and S. Montrul. 2021. Psycholinguistic evidence for incipient language change in Mexican Spanish: The extension of Differential Object Marking. *Languages* 6, 131.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Barbiers, S. 2009. Locus and limits of syntactic microvariation. *Lingua* 119, 1607–1623.

Barbiers, S., and L. Cornips. 2000. Introduction to syntactic microvariation. Barbiers, S., L. Cornips, and S. van der Kleij (eds.), *Syntactic Microvariation*. Amsterdam: Meertens Institute Electronic Publications in Linguistics (MIEPiL). 1–12.

Bazaco, C. 2014. Multiple wh questions in Spanish. A new interpretation. Talk at the *Hispanic Linguistics Symposium 2014*, University of Purdue, November 13–16th.

Berber Sardinha, T. 2021. Corpus linguistics and the study of social media: A case study using multi-dimensional analysis. O'Keeffe, A., and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. New York: Routledge. 656–674.

Berinsky A. J., G. A. Huber, and G. S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 20, 351–368.

Bever, T. G. 1970. The cognitive basis for linguistic structures. Hayes, J. R. (ed.), *Cognition and the Development of Language*. New York, NY: Wiley. 279–362.

Bever, T. G. 1974. The ascent of the specious, or there's a lot we don't know about mirrors. Cohen, D. (ed.), *Explaining Linguistic Phenomena*. Washington, DC: Hemisphere. 173–200.

Bickerton, D. 1971. Inherent variability and variable rules. *Foundations of Language* 7, 457–492.

Bisang, W. 2011. Variation and reproducibility in linguistics. Siemund, P. (ed.), *Linguistic Universals and Language Variation*. Berlin: De Gruyter. 237–263.

Bošković, Ž. 1997. Superiority effects with multiple *wh*-fronting in Serbo-Croatian. *Lingua* 102, 1–20.

Bosque, I., and V. Demonte (dir). 2009. *Gramática descriptiva de la lengua española*. Buenos Aires: Espasa.

Brandner, E. 2012. Syntactic microvariation. *Language and Linguistics Compass* 6, 113–130.

Camerer, C.F., A. Dreber, F. Holzmeister *et al.* 2018. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat Hum Behav* 2, 637–644.

Chambers, J. K., and P. Trudgill. 1998. *Dialectology* (2nd ed.). Cambridge: Cambridge University Press.

Chen, Z., Y. Xu, and Z. Xie. 2020. Assessing introspective linguistic judgments quantitatively: The case of *The Syntax of Chinese*. *J East Asian Linguist* 29, 311–336.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N., and G. A. Miller. 1963. Introduction to the formal analysis of natural languages. Luce, R. D., R. R. Bush and E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. 2. New York, NY: John Wiley. 269–321.

Cognola, F., I. Baronchelli, and E. Molinari. 2019. Inter- vs. intra-speaker variation in mixed heritage syntax: A statistical analysis. *Frontiers in Psychology* 10, 1528.

Conrad, S. 2010. What can a corpus tell us about the grammar? O'Keeffe, A., and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. New York: Routledge. 212–226.

Cornips, L., and C. Poletto. 2005. On standardising syntactic elicitation techniques. *Lingua* 115, 939–957.

Cornips, L., and F. Gregersen. 2016. The impact of Labov's contribution to general linguistic theory. *Journal of Sociolinguistics* 20 (4), 498–524.

Cowart, W. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. London: Sage Publications.

Edelman, S., and M. H. Christiansen, 2003. How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7, 60–61.

Eide, K. M., and T. A˚ farli. 2010. Dialect syntax and parallel grammars: A challenge to generative frameworks? lingBuzz/000993. [Online]. Retrieved on November 11 2022 from: http://ling.auf.net/lingBuzz/000993.

Embick, D. 2008. Variation and morphosyntactic theory: Competition fractioned. *Language and Linguistics Compass* 2, 59–78.

Fábregas, A. 2007. An exhaustive lexicalisation account of directional complements. *Tromsø Working Papers on Language and Linguistics* 34, 165–199.

Fanelli, D. 2018. Is science really facing a reproducibility crisis, and do we need it to? *PNAS* 115, 2628–2631.

Featherston, S. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33, 269–318.

Ferreira, F. 2005. Psycholinguistics, formal grammars, and cognitive science. *Linguistic Review* 22, 365–380.

Fidler, F., and J. Wilcox. 2018. Reproducibility of scientific results. Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/.

Fort, K., G. Adda, and K. B. Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Association for Computational Linguistics* 37(2), 413–420.

Gallego, Á. 2019. Atlas Sintáctico del Español (ASinEs). Retrieved September 1, 2019 from http://asines.org/.

Gervain, J. 2003. Syntactic microvariation and methodology: Problems and perspectives. *Acta Linguistica Hungarica* 50, 405–434.

Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1–76.

Gibson, E., and E. Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14, 233–234.

Gibson, E., and E. Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1–2), 88–124.

Gibson, E., S. T. Piantadosi, and E. Fedorenko. 2011. Using mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5 (8), 509–524.

Gibson, E., S. T. Piantadosi, and E. Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida. *Language and Cognitive Processes* 28, 229–240.

Gilquin, G. 2002. Automatic retrieval of syntactic structures: The quest for the Holy Grail. *International Journal of Corpus Linguistics* 7(2), 183–214.

Grieve, J. 2021. Observation, experimentation, and replication in linguistics. *Linguistics* 59(5), 1343–1356.
Gutiérrez-Rexach, J., and S. Sessarego. 2014. Morphosyntactic variation and gender agreement in three Afro-Andean dialects. *Lingua* 151, 142–161.
Guy, G. R. 2015. The grammar of use and the use of grammar. Adli, A., M. García García, and G. Kaufmann (eds.), *Variation in Language: System-and Usage-Based Approaches*. Vol. 50. Boston, MA: De Gruyter. 47–68.
Haider, H. 2016. *Incredible Syntax – Between Cognitive Science and Imposture*. Manuscript: University of Salzburg.
Henry, A. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*. Oxford: OUP.
Henry, A. 2005. Idiolectal variation and syntactic theory. Cornips, L. and K. P. Corrigan (eds.), *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam: John Benjamins. 109–122.
Hundt, M., N. Nesselhauf, and C. Biewer. 2007. Corpus linguistics and the web. Hundt, M., N. Nesselhauf, and C. Biewer (eds.), *Corpus Linguistics and the Web*. Leiden: Brill. 1–5.
Kato, M. A. 2015. Variation in syntax: Two case studies on Brazilian Portuguese. Adli, A., M. García García, and G. Kaufmann (eds.), *Variation in Language: System- and Usage-Based Approaches*. Boston, MA: De Gruyter. 91–112.
Kern, M. L., G. Park, J. C. Eichstaedt, H. A. Schwartz, M. Sap, L. K. Smith, and L. H. Ungar. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological Methods* 21(4), 507–525.
Kroch, A. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1, 199–244.
Labov, W. 1970. The study of language in its social context. *Studium Generale* 23, 30–87.
Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
Leech, G. 2007. New resources, or just better old ones? The Holy Grail of representativeness. Hundt, M., N. Nesselhauf, and C. Biewer (eds.), *Corpus Linguistics and the Web*. Leiden: Brill. 133–149.
Leivada, E., R. D'Alessandro, and K. K. Grohmann. 2019. Eliciting big data from small, young, or non-standard languages: 10 experimental challenges. *Frontiers in Psychology* 10, 313. https://doi.org/10.3389/fpsyg.2019.00313
Lewis, R. L., and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive-Science* 29, 1–45.
Linzen, T., and Y. Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa* 3(1), 100–125.
MacDonald, J. 2016. Spanish aspectual *se* as an indirect object reflexive: The import of atelicity, bare nouns, and leísta PCC repairs. *Probus* 28, 73–117.
Marantz, A. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22, 429–445.
Mendívil-Giró, J.-L. 2019. How much data does linguistic theory need? On the tolerance principle of linguistic theorizing. *Frontiers in Communication* 3, 62. https://doi.org/10.3389/fcomm.2018.00062
Meyer, C. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
Miller, G. A., and K. O. McKean. 1964. A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology* 16, 297–308.
Montalbetti, M. 1984. *On the Interpretation of Pronouns*. Ph.D. Diss., MIT, Cambridge.
Newmeyer, F. J. 2003. Grammar is grammar and usage is usage. *Language* 79, 682–707.

Newmeyer, F. J. 2004. Against a parameter-setting approach to language variation. Pica, P., J. Rooryck, and J. van Craenenbroeck (eds.), *Language Variation Yearbook*, v. 4. Amsterdam: Benjamins. 181–234.

Newmeyer, F. J. 2013. Goals and methods of generative syntax. Den Dikken, M. (ed.), *The Cambridge Handbook of Generative Syntax*. Cambridge: Cambridge University Press. 61–92.

Newmeyer, F. J. 2015. Language variation and the autonomy of grammar. Adli, A., M. García García, and G. Kaufmann (eds.), *Variation in Language: System-and Usage-Based Approaches*. Vol. 50. Boston, MA: Walter de Gruyter. 29–46.

Ocampo, F. 2010. The place of conversational data in Spanish syntax: Topic, focus and word order. *Studies in Hispanic and Lusophone Linguistics* 3, 533–543.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251.

Ordóñez, F., and E. Treviño. 1999. Left dislocated subjects and the pro-drop parameter: A case study of Spanish. *Lingua* 107, 39–68.

Ortega-Santos, I. 2013. Microvariation in Spanish comparatives. *Catalan Journal of Linguistics* 12, 175–192.

Ortega-Santos, I. 2019. Crowdsourcing for Hispanic Linguistics: Amazon's Mechanical Turk as a source of Spanish data. *Borealis: An International Journal of Hispanic Linguistics* 8, 187–215.

Ortega-Santos, I. 2020. La fiabilidad de los juicios de aceptabilidad en los estudios de sintaxis del español. *Cuadernos de la ALFAL* 12(2), 567–589.

Ortega-Santos, I. 2021. Dialect distance and data assessment in Chilean, Venezuelan and Puerto Rican Spanish. Rogers, B., and M. Figueroa Candia (eds.), *Lingüística del castellano chileno: estudios sobre variación, innovación, contacto e identidad.* Wilmington: Vernon Press. 451–477.

Ortega-Santos, I., L. Reglero, and J. Franco. 2019. Wh-Islands in L2 Spanish and L2 English: Between poverty of the stimulus and data assessment. *Fontes Lingvæ Vasconvm stvdia et docvmenta* 126, 435–471.

Pablos, L., J. Doetjes, and L. L.-S. Cheng. 2018. Backward dependencies and in-situ wh-questions as test cases on how to approach Experimental Linguistics research that pursues Theoretical Linguistics questions. *Frontiers in Psychology* 8, 2237. https://doi.org/10.3389/fpsyg.2017.02237

Peer, E., L. Brandimarte, S. Samat, and A. Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70, 153–163.

Phillips, C. 1996. *Order and Structure*. Ph.D. Diss., MIT.

Phillips, C., P. Gaston, N. Huang, and H. Muller. 2021. Theories all the way down: Remarks on "theoretical" and "experimental" linguistics. Goodall, G. (ed.), *The Cambridge Handbook of Experimental Syntax*. Cambridge: Cambridge University Press. 587–616.

Polinsky, M. 2018. *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.

Rizzi, L. 1990. *Relativized Minimality*. Cambridge, MA: MIT Press.

Roeper, T. 1999. Universal bilingualism. *Bilingualism* 2, 169–186.

Rothman, J., F. Bayram, V. DeLuca, G. Di Pisa, J. Duñabeitia, K. Gharibi, and S. Wulff. 2022. Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics,* 1–14. https://doi.org/10.1017/S0142716422000315

Sampson, G. R. 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory* 3, 1–32.

Sankoff, D., and W. Labov. 1979. On the uses of variable rules. *Language and Society* 8, 189–222.

Schneider, G., and M. Lauber. 2010. *Statistics for Linguists: A Patient, Slow-Paced Introduction to Statistics and to the Programming Language R*. Universität Zürich: Creative Commons License.

Seiler, G. 2003. On three types of dialect variation and their implications for linguistic theory. Evidence from verb clusters in Swiss German dialects. Kortmann, B. (ed.), *Dialectology Meets Typology: Dialect Grammar from a Cross-Linguistic Perspective.* Berlin, New York: De Gruyter Mouton. 367–400.

Sheehan, M., M. Schäfer, and M. C. Parafita Couto. 2019. Crowdsourcing and minority languages: The case of Galician inflected infinitives. *Frontiers in Psychology* 10, 1157. https://doi.org/10.3389/fpsyg.2019.01157

Snyder, W. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31, 575–582.

Sorace, A. 2011. Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism* 1, 1–33.

Sprouse, J. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavioral Research Methods* 43, 155–167.

Sprouse, J. 2023. Acceptability judgment methods. Sprouse, J. (ed.), *The Oxford Handbook of Experimental Syntax*. Oxford: Oxford University Press. 3–28.

Sprouse, J., and D. Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. *Journal of Linguistics* 48, 609–652.

Sprouse, J., C. T. Schütze, and D. Almeida. 2013. Assessing the reliability of journal data in syntax: *Linguistic Inquiry* 2001–2010. *Lingua* 134, 219–248.

Torres Cacoullos, R., and C. E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57: 653–692.

Towsend, D. J., and T. G. Bever. 2001. *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Van Craenenbroeck, J., and M. Van Koppen. 2023. Parameters and Language Contact: Morphosyntactic Variation in Dutch Dialects. Catalan Journal of Linguistics 22, 71–95.

Wasow, T., and J. Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496.

Winter, B. 2020. *Statistics for Linguists: An Introduction to Using R*. New York: Routledge.

Yang, C. D. 2000. Internal and external forces in language change. *Language Variation and Change* 12(3), 231–250.