# Open Linguistics: A platform for sharing informal judgment data and analysis

Masoud Jasbi, Sebastian Schuster, Philip Weiss

February 2019

**Abstract**

Judgments of acceptability or grammaticality provided by trained linguists constitute a major and invaluable source of data for linguistic theories. However, the informal methods used for collecting and reporting such judgments have received considerable criticism (Wasow & Arnold 2005; Ferreira, 2005; Gibson & Fedorenko, 2010; 2013). A recent study suggests that informal judgment may be particularly problematic for languages other than English (Linzen & Oseki 2018). Here we briefly discuss the issues with informal judgments and then introduce a novel solution: an online platform that enables linguists to record examples and informal judgments online, share them with other linguists, search for examples, and cite them in papers. We discuss the main features of the platform, and how it addresses the current issues with informal judgment data.

## 1 Introduction

Modern linguistics has mainly relied on informal methods of data collection for hypothesis testing and theory building. For example, in theoretical syntax and semantics, it is standard practice to construct a "minimal pair" of sentences and judge their acceptability. The sentences of the pair must be identical in every way but one linguistic factor that is manipulated between the pair. If one member of the pair is judged as acceptable while the other is unacceptable, the difference in acceptability is causally attributed to the linguistic manipulation. Consider example (1). Suppose we want to test the hypothesis that Farsi (Indo-European, Iranian) has subject-verb agreement. We construct a sentence with agreement on the verb (1a) and an identical one without agreement (1b). Then we ask a native speaker to judge whether these sentences are "acceptable". The asterisk represents the fact that (1b) was judged as unacceptable and (1a) as acceptable. Since (1a) and (1b) only differ in agreement marking, we can attribute their difference in acceptability to their difference in agreement. Therefore, (1) provides evidence for the hypothesis that Farsi manifests subject-verb agreement. Similar to (1) which was judged by the first author of this chapter, acceptability of an example is often judged by the linguist/author of the example themselves.

(1) a. man keik xord-am
1.SG cake ate-1.SG
"I ate cake."

b. * man keik xord
1.SG cake ate
"I ate cake."

Informal acceptability judgments have various advantages over more formal methods of data collection. First, they are extremely quick: native-speaker linguists can construct an example and test their own intuition immediately. Second, they are extremely cheap: no need to pay participants or acquire expensive equipment. Third, they are easy to collect: linguists can provide judgment data themselves or collect them from friends and colleagues. Fourth. they do not require elaborate procedures: often no IRB protocols are required, no need for item selection, randomization, blocking, counterbalancing, etc. Fifth, they allow for simple and quick data analysis: no need for complex data analysis techniques or statistical methods. Furthermore, they have some advantages over naturalistic data from corpora: they allow for testing constructions that rarely occur in spontaneous language use. They also provide negative evidence, i.e., evidence that an utterance is not part of the target language. Corpus data can only provide positive evidence. Finally, spontaneous speech involves production errors and slips of the tongue that do not constitute acceptable utterances of the target language. Acceptability judgments allow linguists to detect such production errors. These properties have made informal acceptability judgments extremely useful and a major force behind fast theoretical development in the past (see Schütze (1996) for the history and role of informal acceptability judgments in linguistics).

However, these blessings may come with a curse: informal acceptability judgments are sometimes not as reliable as the ones collected using more formal methods (Wasow & Arnold 2005; Ferreira 2005; Gibson & Fedorenko 2013). It is easy to see why. Intuitions of one linguist may not be shared by other linguists or non-linguists. Judgments in example (1) may be true of the first author of this chapter but not of other Farsi speakers. It may also be true of the particular lexical items used such as the verb "eat" but not of other lexical items in Farsi. Consequently, hypotheses and theories that are based on a few linguists' intuitions or a few items may not generalize to other speakers or items. Informal judgments are also limited in scope. The binary acceptable vs. unacceptable distinction does not allow generalizations based on subtle and gradable effects. Furthermore, the whole process of data collection and reporting is subject to the biases of the theoretical linguist. For example, there is no systematic way of reporting all tested examples and papers almost always include a selected sample of what the linguist actually considered for hypothesizing and hypothesis testing. Such a practice makes selective and confirmation-biased reporting of data possible and perhaps likely.

The problems with informal methods of collecting acceptability judgments were understood by most theoretical linguists since the early years of generative linguistics (Schütze 1996). For example, Chomsky (1962; cited in Schütze 1996) said: "I dislike reliance on intuition as much as anyone. [...] We should substitute rigorous criteria just as soon as possible, instead of clinging to intuition". However, the argument for using informal judgments was that in clear cases, the use of formal methods would be a waste of time and resources. Therefore, formal methods should be saved for cases where informal methods fail to be conclusive. For example, Chomsky (1969, p.81) said: "I have no doubt that it would be possible to devise operational and experimental procedures that could replace the reliance on introspection with little loss, but it seems to me that in the present state of the field, this would simply be a waste of time and energy." In short, linguistic investigation and theoretical development would be hampered significantly if all judgments of linguistic examples were "subject to statistically rigorous experiments on naive subjects" (Culicover & Jackendoff 2010).

Based on the arguments summarized above, the current best practice in theoretical syntax and semantics seems to be the following: use informal methods for clear cases to develop syntactic and semantic theories; where informal methods prove not helpful or prove inconclusive, use formal (experimental) methods to adjudicate competing hypotheses. However, this approach faces two problems. First, how do we know whether a case constitutes "a clear case" or not? How do we know whether informal methods are inconclusive in that case or not? In other words, if formal methods are reserved for when informal methods fail, how can we know when informal methods have failed? So far, the detection of "clear" and "unclear" cases have also been done informally. In cases where informal judgments of a phenomenon were disputed and agreement seemed hard to achieve, linguists started to use more formal methods (see Philips 2009 for example cases). However, there is no systematic way of deciding between clear and unclear, or conclusive and inconclusive cases. There is also no systematic way of measuring the level of agreement on informal judgment data without employing more formal methods.

Second, there is a relatively large methodological gap between informal and formal methods of collecting acceptability judgments. Informal methods allow for piecemeal exploration: constructing examples and testing them one example at a time. Formal methods require a carefully selected sample of examples, controls, and fillers. Informal methods allow linguists to collect data using simple tools (often pen and paper). Formal methods rely on more advanced data collection tools such as survey software and crowdsourcing platforms such as Amazon Mechanical Turk. Informal methods allow data collection from the linguists themselves or from friends and acquaintances. Formal methods require systematic recruiting of naive participants. Informal methods often do not require an IRB approval while formal methods require it in most cases. Finally, informal methods do not involve sophisticated data analysis while formal methods require knowledge of statistical methods. Bridging this gap requires considerable investment of time and resources, which in turn makes successful detection of cases in need of formal treatment even more important.

Due to this methodological gap, many theoretical linguists find themselves at a crossroads. They have to either stick to the informal methods and risk the possibility that their findings are not generalizable; or invest considerable time and money in pursuing more formal methods, but risk the possibility that the findings do not add much to their theories above and beyond what could have been found informally. This is not ideal and raises the question of whether there is a middle ground. Is it possible to increase accuracy and reliability of informal methods, without giving up the benefits of ease, speed, and low cost?

We argue that this is possible. Recent developments in sharing and accessing scientific data on online platforms allow for a more systematic collection and monitoring of informal judgments, without adding any burden to the current workload of a theoretical linguist. In fact, we argue that it is possible to make informal data collection easier, more transparent, and more accessible, while adding to its reliability. The online platform we present in this chapter aims to do exactly that. It is inspired by similar online platforms such as the Open Science Framework (osf.io) in experimental fields. In section 2, we briefly survey the literature on acceptability judgments and list some of the main issues brought up with their use in theoretical linguistics. In section 3, we introduce our online platform that allows linguists to register their projects, and enter their examples and acceptability judgments. In section 4 we provide a brief example workflow for a project on Farsi differential object marking. In section 5, we explain how the features of this platform help address the issues raised in section 2.

## 2 Issues with Informal Methods

The use of informal acceptability judgments as evidence for linguistic hypotheses has received considerable criticism (see Scholz, Pelletier, & Pullum, 2016 for a recent discussion). The gist of the criticism is that informal collection and reporting of judgments is not rigorous enough to form a reliable foundation for linguistic theory. The criticism often involves comparison of informal judgment collection with more formal methods in fields like experimental psychology. Schütze and Sprouse (2014) discussed five ways informal methods differ from formal ones. First, informal judgments typically involve few speakers. Second, informal judgments often include the linguist/theorist themselves as a (or even "the") subject. Third, the response options are limited (typically acceptable and unacceptable) with not much room for gradient judgments. Fourth, relatively few tokens of the structures of interest are tested, and finally the data analysis used is quite simple and sometimes unsystematic. Each of these differences have been subject of criticism in the past. Schütze (1996) discusses some of the criticism and debates in the 20th century. Here, we focus on the more recent developments of the debate.

Wasow & Arnold (2005) argued that linguists have ignored the standards of data collection and analysis common in many natural and social sciences. They argued that judgments often vary between language users and it is common to find examples with "divided judgments": some find it completely acceptable and some completely unacceptable. They also argued that some judgments are "marginal": language users are unsure about them. Such marginal cases are susceptible to contextual factors or the linguist's cognitive biases. As a case study, Wasow & Arnold (2005) considered the roles of constituent length and syntactic complexity in explaining the position of verb particles. They argued that the results of their experiments do not support the conclusions drawn using informal judgments in theoretical papers.

Ferreira (2005) argued that due to over-reliance on informal judgments, "generative theories appear to rest on a weak empirical foundation." She explained that informal judgments do not systematically check for item-effects: idiosyncratic properties due to unique lexical content. In theoretical papers, occasionally a few items are tested but there is rarely a systematic analysis on the range of relevant factors that might be at play and must be held constant "to make sure there isn't some correlated property that is responsible for the contrast in judgments." According to Ferreira (2005), the most problematic is that "the subject who provides the data is not a naive informant, but is in fact the theorist himself or herself." Therefore, it is possible for linguists to (unintentionally) bias their own judgments. She explained that sometimes theorists check their judgments with colleagues or acquaintances (called "Hey Sally" method), but there is no agreed upon procedure for collecting and reporting such judgments, or reconciling contradictory judgments. She also used an example to show that in such cases researchers may bias the informants: "I myself have been in the situation of providing a judgment to a linguistics colleague, only to be looked at with an expression of incredulity and to be asked, 'Really? Are you sure you have the right reading?'. One occasionally even feels badgered into acceding that the data are in fact as the theorist wants them to be."

Gibson & Fedorenko (2010) discussed three types of bias that can affect informal data collection. First, confirmation bias on the part of the linguist: When asking for judgments, the linguist may have a favored hypothesis and actively seek data that support it, or disregard data that support the alternative hypothesis. Second, confirmation bias on the part of the linguist informant: For example, when a linguist asks a colleague to provide judgments, their colleague's awareness of the hypothesis being tested can affect their judgments. Third, observer-expectancy (Clever Hans) effects: Informants may want to please the linguist and rely on the linguist's subtle positive or negative cues for their judgments.

In a related paper, Gibson & Fedorenko (2013) expanded their arguments and explained that "the results obtained using this method are not necessarily generalizable because of (a) the small number of experimental participants (typically one), (b) the small number of experimental stimuli (typically one); (c) cognitive biases on the part of the researcher and participants; and (d) the effect of the preceding context (e.g. other constructions the researcher may have been recently considering)." Gibson, Piantadosi, Fedorenko (2013) added that current practices in collecting informal judgments do not allow for aggregate measures of acceptability that reveal acceptability effect sizes. To avoid these issues, they argued that every acceptability judgment must be validated in a formal experiment. However, Culicover & Jackendoff (2010) countered that this solution would unduly slow down theoretical progress and waste resources. An argument made by Chomsky (1969) as well.

Philips (2009) argued that in order for informal (intuitive) judgments to have truly harmed linguistics, one or more of the following must hold: (i) Intuitive judgments have led to generalizations that are widely accepted yet bogus. (ii) Misleading judgments form the basis of important theoretical claims or debates. (iii) Carefully controlled judgment studies would solve these problems. He argued that none of these seem to hold. Empirical claims undergo extensive vetting before they become "widely accepted generalizations". In other words, informal peer-reviews by colleagues, audience members at conferences, and reviewers of papers weed out problematic judgments and commonly result in reliable and generalizable data. Philips (2009) argued that instead he finds another issue in theoretical linguistics: the diminishing awareness of what empirical body motivates what hypotheses or theoretical choices. He also conceded that theoretical linguistics can benefit from graded acceptability judgments. Furthermore, he mentioned that as the empirical base of linguistics expanded across domains and languages, it has become increasingly hard to provide theories that cover this massive base.

Sprouse & Almeida (2012) and Sprouse, Schütze, & Almeida (2013) successfully replicated informal English judgments in a linguistic textbook (Adger 2003) and in the journal *Linguistic Inquiry* using formal/experimental methods. These results provided some support for Philips's (2009) argument that informal methods have not truly (or seriously) harmed linguistics. However, Linzen & Oseki (2018) showed that replicating the original reported effects may still be a serious issue in languages other than English. They collected experimental acceptability judgments for published Hebrew and Japanese examples which they deemed questionable. They reported that half of the acceptability contrasts did not replicate. They suggested that this issue can be addressed using "a simple open review system, and that formal experiments are only necessary in controversial cases." As we will see later, Open Linguistics provides such an open review system that allows separating uncontroversial and controversial cases.

Judgments may also differ systematically based on whether they are provided by linguists or non-linguists. For example, Dabrowska (2010) tested linguists' and non-linguists' judgments of questions with long distance dependencies. She found that linguists' judgments diverged from those of non-linguists: while non-linguists showed graded judgments, linguists found grammatical sentences more acceptable than non-linguists. However, linguists and non-linguists judged ungrammatical sentences as unacceptable to the same degree. More generally, Scholz, Pelletier & Pullum (2016) argued that judgments may differ based on whether they are provided by: “(i) linguists with a stake in what the evidence shows; (ii) linguists with experience in syntactic theory but no stake in the issue at hand; (iii) non-linguist native speakers who have been tutored in how to provide the kinds of judgments the linguist is interested in; and (iv) linguistically naive native speakers.” Current methods of collecting informal judgments do not distinguish among these types of judgments and weigh them equally in building linguistic theories.

We would like to also discuss some issues that we have not seen discussed in the literature. First, it is common for theoretical linguists, especially native speakers of the target language, to test many sentences informally, yet report only a subset that they deem interesting or noteworthy. This practice creates the potential for “selective reporting” and amplifies cognitive biases of the researcher. What is deemed interesting or noteworthy might be subject to confirmation bias. Indeed selective reporting is highly problematic in experimental domains and if we consider informal acceptability judgments as small and informal experiments, we should strive to report all judgments relevant to our theorizing, as much as possible.

Second, current methods in collecting informal acceptability judgments do not allow for systematically tracking reliability of judgments for the same speaker. A common informally reported phenomenon by some linguists is that they “lose sensitivity” to the (un)acceptability of particular constructions over time as they study them more and more (e.g., the constructions become more acceptable). Finally, sometimes linguists rely on examples or judgments in a language using personal communication with a linguist (abbreviated as p.c.). While such data helps linguists in forming and testing hypotheses, they are not recorded and often lack the details needed for further assessment and validation.

As the discussion in this section shows, there are many issues with informal acceptability judgments, some more important than others, and some raised repeatedly by many authors in the past decades. In this chapter, we have grouped the issues that are related to each other under the same heading and included them in Table 1. The table also presents the solutions our platform provides. In the next section we introduce the main features of our platform.

## 3 Open Linguistics: An online platform for sharing linguistic data and analyses

We introduce Open Linguistics (openlinguistics.org)[1], a free and publicly available online platform for linguists to record their informally collected judgments. Open Linguistics is inspired by successful platforms in experimental domains such as the Open Science Framework (osf.io) which facilitate open sharing of data and analysis with other experimental researchers. Open Linguistics strives to:

- Connect linguists around the world and facilitate access to research on similar topics and languages.
- Provide a free and publicly available platform for linguists to record their examples and provide acceptability judgments.
- Facilitate systematic sharing and reviewing of data in linguistics.
- Facilitate detection of examples that informal methods cannot properly adjudicate and require formal experiments.
- Separate data and theory by making it possible to record data separately in an open-access platform and reference it.
- Encourage native-speaker linguists in understudied languages to contribute more examples and judgments.

[1] Note that the platform itself is currently still under development and does not yet provide all the features that we describe here. We expect that all features will be implemented by May 2019.

Similar to other platforms for sharing data, our website provides users with a personal account. In what follows, we describe the main features of an account and more generally the website. We start with introducing the profile and the notebook tab.

## 3.1 Profile & Notebook

Each account on Open Linguistics has its own publicly accessible profile which allows linguists to record information about their affiliation, their research interests, and the languages they study. Users can also link their account to their personal web pages. User profiles are searchable using the information available in user profiles. This feature allows linguists to find colleagues that work on similar topics or languages around the world. User profiles serve the goal of connecting linguists and research on similar topics or languages.

An account also provides linguists with a "notebook." A notebook is a collection of projects that the linguist is working on. On the notebook tab, users can add projects by simply providing the name and description of the project. We encourage users to include their research question and the set of hypotheses they plan to test in their research description. Projects can be marked as public or private. Public projects are included in search results and can be accessed by other users. Projects can also be initially private and made public at a later time. The next section discusses details of a project.

## 3.2 Projects

A project has the following components: a title, a description, one or more files, a permanent URL, topic tags, language tags, and the citation info. We explained the title and description in the previous section. Users can upload files to their projects in PDF, TeX, MS Word, CSV, and similar formats. These files can contain raw linguistic data (e.g., in CSV format) or the analysis and write up of the project (e.g., as a PDF or Word document). The platform assigns each project a short and permanent URL. Therefore, Open Linguistics also acts as a pre-print repository where linguists can store their manuscripts and data before submission for peer-review and formal publication. In fact, authors can include the permanent project URL in their manuscripts to allow easy and open access to all the materials and data used in the project. Public projects are searchable using the topic tags and language tags provided by the authors. Finally, each project has citation information that includes the permanent URL to the project and allows other linguists to cite the project's manuscript and data.

## 3.3 Examples and Acceptability Judgments

Linguists can also add "examples" to their notebooks. An example consists of a transcription, an interlinear gloss, and a translation. The platform automatically checks to make sure transcription and gloss contain equal number of matching elements. Examples are displayed on the website in the same format and structure presented in theoretical linguistics papers. We also ask users to provide the language and the variety the example is from. Linguists can further optionally specify a set of topic tags for the example. For instance, an example can be tagged for definiteness, tense, aspect, differential object marking, gapping, ergative case, etc. Topic tags help increase the chance of finding examples based on the commonly known crosslinguistic phenomena or theoretical constructs. A linguist who wants to find examples of gapping constructions across languages can search for its topic tag and easily retrieve such examples.

Inside the notebook, an example can be part of one or more projects, or part of no project at all. This way, linguists can use their notebooks to record interesting linguistic examples they encounter or come up with, even before working on its theoretical implications. Examples are time-stamped and they each receive a unique identifier, which is also part of the URL to that unique example on the website. Examples can also be exported in LaTeX and text format, along with their unique identifier URLs. This allows linguists to easily copy examples from their Open Linguistics notebooks to their manuscripts without losing the connection to the original example and its associated information stored online. Most importantly, linguists can select the set of examples they would like to have included in their manuscript, and report all the examples they have actually tested using the project URL. A project page lists all the examples associated with it. This property of our platform is meant to address the issue of "selective reporting" discussed before.

Linguists can also specify the "source" of an example. They can report that the example is constructed by them (source: user), provided in another linguist's published work (source: publication), provided informally by another linguist (source: personal communication), provided by a non-linguist informant (source: informant), or encountered in a "naturally occurring" context (newspaper, blog, etc.) (source: naturally occurring). In case the source is a publication or the example is naturally occurring, users can provide citation information or URLs to where they encountered the example. Users can also add further notes on any aspect of the example or data collection procedure that they deem relevant. Finally, examples can have links to other examples that together constitute minimal pairs.

Each example has a judgment contributed by the linguist who added the example. Researchers can choose between categorical or scalar judgments. Categorical judgments are marked by one or two of the following prefixes: *, #, or ?. We adopt Scholz, Pelletier, and Pullum's (2016) guidelines on the approximate meaning of the prefixes and their combinations such as ?* or **. Likert scale judgments use numbers between 1 and 7. A judgment can also be accompanied by information on the context of of the utterance. If left blank, the judgment is considered to be context-independent. Lastly, judgments are classified into three types based on whether they are provided by: (i) a linguist familiar with the hypothesis being tested (informed linguist) (ii) a linguist unfamiliar with the hypothesis (uninformed linguist) or (iii) a non-linguist native speaker. A judgment is stamped for the time and date of judgment and the user that provided it. Judgments can not be altered and can only be deleted.

Researchers can choose whether they want an example to be open or closed. For open examples, other linguists can also provide judgments. For each example, the platform provides plots and summary statistics on the judgments provided so far. Summary of judgments are provided as bar plots for categorical judgments and as histograms for scalar (Likert) judgments. Each example can be accessed using a shareable URL with a unique identifier. Users can search for examples using keywords in an example's transcription, gloss, translation, or tags.

# 4 An Example Workflow

To illustrate how we imagine the platform to be used, we provide an example workflow for a project. Consider the problem of verb agreement in Farsi, discussed at the beginning of this chapter, and suppose we want to write a chapter on this topic. Since, we also have an account on Open Linguistics, we can go to our notebook and start a project titled "Subject Verb Agreement in Farsi" with the following description: "Does Farsi show subject verb agreement? (Research question) Hypotheses include: Yes, and No. If yes, then we expect the morphological form of the verb to vary systematically based on features of the subject, for example, person, number, or gender. If no, we expect no such pattern."

Then we can think of examples to add to this project. For example, we might add examples like (1) as well as similar ones that vary the verb or the subject noun phrase to control for specific lexical effects (Ferreira 2005). To increase our confidence in the reliability of our examples and judgments, we can post examples for judgments from other Farsi speaking linguists (Philips 2009, Gibson & Fedorenko 2013, Linzen & Oseki 2018). We can write up the generalizations we have drawn from the examples and our analyses, exporting examples that illuminate the discussion from the project page and including them in the manuscript. Examples exported from Open Linguistics will include a short URL to the example online, allowing readers to access to original data and judgments. When the manuscript is finished, we include the project URL as a footnote on the first page and upload the manuscript to our project page. This way, readers will have access to all the data we have consulted, in addition to those included in the paper. If not public yet, we can make the project public and share it with other linguists so that they can see our manuscript and all its examples. Finally, we can submit our manuscript to a journal for publication.

In summary, the steps of the example workflow discussed in this section are:

1. Create a project on the platform. Add the research questions and hypotheses in the description.

2. Add all tested examples to the project. The examples may be constructed, from previous literature, from informants, or existing examples on Open Linguistics, contributed by other linguists.
3. Collect judgments from other linguists who speak the language to increase the reliability of the examples.
4. Write up the analysis and include examples that best illuminate the issues in the paper. It is possible to export them from Open Linguistics to LaTeX or MS Word. The export will include a link to the example online.
5. When done with writing manuscript, include the link to its Open Linguistics project (for example as a footnote). This way reviewers and readers will have access to all the data.
6. Upload the manuscript to its Open Linguistics project.
7. If not public until now, make the project public so that other linguists can read the manuscript and see its data.

# 5 Discussion

Informal judgments provided by linguists have been a major source of insight for theoretical linguistics. They are easy and convenient to obtain and allow for rapid theory building and testing. However, as many researchers have argued, collection of informal judgments is unsystematic and can result in unreliable linguistic data and subsequently theories. Critics have raised many important issues with informal judgments. We summarized them in Table 1. However, we argued that informal methods can be improved, without losing their advantages such as ease and convenience of data collection. In fact, it is possible to make informal collection of acceptability judgments easier, more open, and more systematic at the same time. We introduced an online platform, Open Linguistics, that aims to accomplish this task.

So how does Open Linguistics do the job? The more detailed answer is provided in Table 1. But in short, since Open Linguistics is an online platform, it can automatically perform tasks that are not possible with current methods. First, linguistic examples have a specific structure (transcription, gloss, translation, judgment, etc.) that makes them hard to handle by commonly used software for data storage, visualization, and analysis such as Excel. Open Linguistics, on the other hand, is designed for handling linguistic examples and judgment data. It provides an interface to users that makes example entry intuitive and easy, removing any concerns over how the example should be formatted. All formatting and visualization is done by the platform. Since examples can also be exported in the right format for LaTeX or MS Word, the platform reduces the amount of time and effort spent by linguists on storing and formatting linguistic examples. Second, using previous methods, getting judgments from a large number of linguists or informants and keeping track of them was a laborious task. Open Linguistics makes this task easy because it automatically keeps track of judgments and provides summary of them. It also makes examples searchable and retrievable based on topics that they address or theories that they support. Finally, Open Linguistics allows for storing and sharing all the data that was consulted for hypothesis formation or testing. This is currently not possible, because papers have limited space and there is no appropriate platform for storing linguistic examples. Open Linguistics fills this gap.

We hope that Open Linguistics can bridge the gap between informal and formal methods. Theoretical linguists can collect judgments on their examples from other linguists and informally test the reliability of the judgments. If judgments are divided or unclear, the distribution of judgments would let them know. In such cases, they can set the example aside as an unclear case, or if too important for their theory, they can use formal methods to further investigate the judgment patterns. Therefore, Open Linguistics increases the rigor and reliability of informal judgments, without taking away the advantages that have made them so valuable for linguists and linguistics.

# 6 References

Adger, David (2003). *Core syntax: A minimalist approach* (Vol. 33). Oxford: Oxford University Press.

Chomsky, Noam. (1969). Linguistics and philosophy, in *Language and Philosophy: A Symposium*, Sidney Hook, (ed.), New York: New York University Press, 51–94.

Culicover, Peter, & Jackendoff, Ray (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, 14(6), 234-235. DOI: `https://doi.org/10.1016/j.tics.2010.03.012`

Dabrowska, Ewa (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The linguistic review*, 27(1), 1-23. DOI: `https://doi.org/10.1515/tlir.2010.001`

Ferreira, Fernanda (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review.* 22, 365–380. DOI: `https://doi.org/10.1515/tlir.2005.22.2-4.365`

Gibson, Edward. & Fedorenko, Evelina. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14, 233–234. DOI: `https://doi.org/10.1016/j.tics.2010.03.005`

Gibson, Edward & Fedorenko, Evelina (2013): The need for quantitative methods in syntax and semantics research, *Language and Cognitive Processes*, 28:1-2, 88-124. DOI: `https://doi.org/10.1080/01690965.2010.515080`

Gibson, Edward, Piantadosi, Steven, & Fedorenko, Evelina. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, 28(3), 229-240. DOI: `https://doi.org/10.1080/01690965.2012.704385`

Linzen, Tal, & Oseki, Y. (2018). The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1). DOI: `http://doi.org/10.5334/gjgl.528`

Phillips, Colin (2009). Should We Impeach Armchair Linguists?. In Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy, and Sung-Ock Sohn, *Japanese/Korean Linguistics.* 17. Stanford, CA: CSLI Publications.

Scholz, Barbara C., Pelletier, Francis Jeffry and Pullum, Geoffrey K., "Philosophy of Linguistics", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.),
URL = <https://plato.stanford.edu/archives/win2016/entries/linguistics/>.

Schütze, Carson (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Berlin: Language Science Press. DOI: `10.26530/OAPEN_603356`

Schütze, Carson and Sprouse, Jon. (2014). Judgment Data. In R. J. Podesva and D. Sharma, *Research Methods in Linguistics* (pp. 27-50) Eds. Cambridge, UK: Cambridge University Press. DOI: `https://doi.org/10.1017/CBO9781139013734`

Sprouse, Jon, & Almeida, Diego (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, 48(3), 609-652. DOI: `https://doi.org/10.1017/S0022226712000011`

Sprouse, Jon, Schütze, Carson, & Almeida, Diego (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134, 219-248. DOI: `https://doi.org/10.1016/j.lingua.2013.07.002`

Wasow, Thomas and Arnold, Jennifer (2005). Intuitions in linguistic argumentation, *Lingua*, 115: 1481–1496. DOI: `https://doi.org/10.1016/j.lingua.2004.07.001`

| **Issue** | **Open Linguistics Solution** |
|---|---|
| **Small-N**: few speakers, few items (Gibson & Fedorenko 2013; Gibson, Piantadosi, Fedorenko 2013) | The platform allows collection of judgments from a large number of linguists and using as many items as needed. |
| **Between-Spaker Variation**: Detecting unclear, inconclusive, or "divided judgments" (Arnold & Wasow 2005), detecting between speaker variation (possibly due to idiolectal or dialectal variation), providing a measure of agreement among linguists | The platform allows linguists to provide judgments to an example and provides summary statistics on collected judgments. Unclear or inconclusive cases will show judgment distributions that are almost uniform for all responses. "Divided judgments" should show a bimodal distribution for judgments; conclusive cases should show peaks on particular response options. |
| **Within-Speaker Variation**: Measuring judgment variation within speakers | judgments are time-stamped and cannot be changed. Linguists can add as many judgments as they want. The history of judgments provided by the same linguist can provide estimate of within-speaker variation. |
| **Item Effects**: Detecting effects specific to the example used such as idiosyncrasies of lexical items or the context of the example (Ferreira 2005) | The platform allows for reporting and collecting a large number of examples that vary by the lexical items used. The platform's systematic collection of judgments allows for tracking such item effects. |
| **Marginal Cases**: Detecting cases were judgments are uncertain (Arnold & Wasow 2005) | In addition to judgments, the platform allows linguists and informants to provide their certainty on a judgment. |
| **Measurement**: Current informal methods have limited response options not allowing for graded judgments, or estimating aggregate measures and acceptability effect sizes (Philips 2009; Gibson, Piantadosi, Fedorenko 2013) | Linguists can provide both categorical and graded (scalar) judgments. Recorded judgment data is used to estimate aggregate measures and informal effect sizes. |
| **Linguist Bias**: Detecting differences in judgments of linguists vs. non-linguists, or naive vs. informed informants (Dabrowska 2010; Scholz, Pelletier, & Pullum, 2016) | Judgments contain information about whether they are provided by linguists or non-linguists, and whether they were informed or uninformed with respect to the hypotheses. |
| **Tracking Empirical Evidence**: Data that have supported a particular hypothesis, analysis, or theory are sometimes forgotten (Philips 2009) | Example tags and project tags make it possible to search and retrieve examples that relate to a particular phenomenon or support a hypothesis. |
| **Selective Reporting**: Linguists often informally test many sentences before finding an interesting contrast. However, only such interesting cases are reported which allow for unintentional confirmation bias and possibly some type of data-dredging (p-hacking) | Linguists can provide all examples they considered on the platform and reviewers can verify that a reasonable number of examples was considered and that all examples provide evidence for the hypothesis. |
| **Cognitive biases** influencing linguists' or informants' judgments; collecting/reporting examples or judgments from colleagues and acquaintances ("Hey Sally", "Personal Communication") (Ferreira 2005; Gibson & Fedorenko 2010) | The platform allows collection of data from a large number of linguists, reducing the possibility of bias in a particular direction |
| **Access** to crosslinguistic and cross-domain data for generalization and theory building (Philips 2009) | Examples are searchable by tags and glosses. Linguists can easily find a collection of cross-linguistic examples for a specific phenomenon (indicated by a tag). |
| **Openness**: Open sharing of all data, hypotheses, and analyses | Each example and project has a shareable permanent URL. |
| **Peer-review of data**: vetting reliable examples and weeding out unreliable ones (Philips 2009, Linzen & Oseki 2018) | The platform provides summary statistics on the examples judgments. Unreliable examples can be detected using these summaries. |

Table 1: Overview of previously reported issues with informal judgments and how Open Linguistics addresses these issues.