# Variation in Linguistic Complexity and its Cognitive Underpinning

**Anastasia Smirnova (smirnov@sfsu.edu)**
Department of English Language and Literature, 1600 Holloway Ave
San Francisco, CA 94132 USA

**Abstract**

Linguistic complexity – manifested in terms of hierarchical recursive structures generated by grammar – is often discussed from the perspective of cross-linguistic comparison (cf. Everett, 2005; Nevins, Pesetsky, & Rodrigues, 2009 on Pirahã). In this paper, we focus instead on the variation in complexity within a single language, English, and on the lower bound of complexity, specifically (cf. Futrell et al., 2016). We report results of two studies, a corpus study (Study 1) and a production experiment (Study 2), that investigate syntactic complexity of expressions that arise in the context of human-computer interaction and compare them to the standard language. The results of both studies show that the expressions generated in the context of human-computer interaction exhibit lesser structural complexity and often violate the norm of the language (cf. *margaret mead culture famous research*). Our results suggest that such expressions are generated by a qualitatively different type of formal grammar, Linear Grammar (Jackendoff & Wittenberg, 2017), rather than by recursive grammar (Roeper, 1999).

**Keywords:** linguistic complexity; syntax; linear grammar; human-computer interaction.

## Introduction: Syntactic Variation in Language

The expression *margaret mead culture famous research* would most likely strike native speakers of English as ungrammatical. Despite this assessment, such strings of words are generated on a regular basis by native speakers. They arise in a context of human-computer interaction, and information search, specifically. The fundamental question is how linguistic expressions that violate the core rules of the language get generated by the grammar in the first place, and what they reveal about language organization in the brain. We show that a systematic exploration of such linguistic expressions contributes to the debate about complexity in human language, and its lower bound, specifically (cf. Futrell et al., 2016).

Deviations from the syntactic norm of a language appear in a wide range of contexts and are relatively well documented. Sadock (1974) observed that product labels can omit subjects (*Contains methanol*) and objects (*Shake before using*). Yanofsky (1978) discussed utterances consisting of noun phrases only, such as *Teamwork* (in the context of a tennis doubles match win), and the problem they presented for the contemporary syntactic theory, which disallowed the generation of structures smaller than a sentence. A special volume by Kittredge and Lehrberger (1982) analyzed structural differences between standard English and language of technical manuals, stock market reports, cooking recipes, weather reports, and other specialized registers, collectively referred to as sublanguages, a term attributed to Zellig Harris. The editors observe that "specialized linguistic systems can differ quite sharply, both in *complexity* and in the particular linguistic features that set them apart from the general or standard language" (Kittredge, 1982; our emphasis).

Parallel with Kittredge's work on sublanguages, Charles Ferguson, working in the sociolinguistic tradition, studied linguistic modification in the speech of adult native speakers addressed to someone who is believed to lack full linguistic proficiency, such as children and foreigners (Ferguson, 1975). Ferguson (1982) argued that the observed modifications are best characterized in terms of simplification, thus converging with Kittredge's observation on sublanguages.

A more recent volume by Progovac et al. (2006) extends the range of linguistic phenomena that deviate from syntactic norms by focusing on telegrams, personal ads, newspaper headlines and other. Similar to previous authors working on the topic, the volume contributors emphasize that linguistic expressions that belong to specialized registers would be considered unnatural and perhaps ungrammatical from the perspective of standard English. The main puzzle is how such variation can be explained.

Take, for example, the phenomenon of subject omission, observed in diaries (*Dreamt that I picked up a New Yorker*; Haegeman, 2013), recipes (*Serves ten people*; Haegeman, 1987), telegrams (*Am ill*; Barton, 1998), sports announcers' commentaries (*Dribbles into the lane*; Reaser, 2003), medical records (*Is on folic acid*; Sager 1986), as well as in search queries (*forgot password*). To explain subject omission in a language like English, which in general does not allow subject drop, one can postulate a mechanism that would disallow subjectless sentences in the standard language, but permit them in specialized registers of the same language. The main challenge faced by such an analysis is that it would have to postulate two contradictory rules within a single grammar.

In what follows, we discuss two alternative approaches to how syntactic variation, and, specifically, variation that violates syntactic norms of a language can be explained. Both approaches avoid the postulation of contradictory rules within the grammar, but make very different assumptions about the nature of the grammar that operates in specialized registers. We then proceed to show how a systematic analysis of information requests that arise in the context of human-computer interaction can shed light on the theoretical debate presented in the next section.

## Two Approaches to Syntactic Variation

A novel position on how structural variation within a language can be explained emerges from the framework of Theoretical Bilingualism (TB), developed by Roeper (1999). Roeper (1999) rejects the idea that the grammar of a language can have contradictory rules that e.g. permit subject deletion in specialized registers, but require subjects elsewhere. He proposes that these apparently contradictory phenomena can be explained if we assume that speakers of a language have access to different types of grammars, all generated by Universal Grammar (UG). Some grammars, such as the grammar of Italian, allow subject drop on the condition of recoverability from context. Other grammars, such as the grammar of standard English, disallow subject drop. Registers within English that allow subject omission, such as diaries, telegrams, and recipes, have access to a different grammar, which, unlike the grammar of the standard English, allows the generation of subjectless sentences. The TB framework can thus be applied to explain linguistic variation and divergence from syntactic norms observed in different registers of the same language, without proposing contradictory rules.

An alternative approach to syntactic variation is proposed by Jackendoff and Wittenberg (2017). Similar to Roeper (1999), the authors are interested in the range of syntactic variation observed in language, from fully-formed sentences that obey the syntactic principles of the grammar to expressions that violate word order constraints and argument selection principles. Unlike Roeper (1999), Jackendoff and Wittenberg (2017) are interested in what is minimally needed in a language for it to be fully operational and convey meaning. Coming from the perspective of language evolution, they propose that full-fledged grammars of contemporary languages have an evolutionary predecessor, LINEAR GRAMMAR (LG). LG is a rudimentary system, which has access to words and their meanings, but lacks syntax and, as a consequence, hierarchical structure and syntactic subordination. Since there is no syntax, there are no subjects and objects, and no fixed word order. Instead, the order in which event participants are expressed is often governed by semantic and pragmatic considerations, such as agent first. Jackendoff and Wittenberg (2017) argue that many of the observed linguistic phenomena that cannot be easily explained with reference to the full-fledged grammar – the authors discuss pidgins, home signs, village signs, incomplete second language acquisition, etc. – can be accounted for with reference to LG. Additionally, the authors propose that LG is still active in the brains of contemporary speakers. Speakers fall back on LG when access to full-fledged grammar is blocked, as a result of e.g. brain damage (cf. linguistic output in aphasia). In healthy population, structures generated by LG can be observed in certain areas of the language, such as compounds. Compounds are structures like *kitchen table* and *undercurrent* that emerge from the combination of two or more lexical items. The meaning of compounds is loosely derived from the meaning of their parts with heavy reliance on context. For example, the nominal compound *apple juice seat* can mean "seat with apple juice on the table in front of it" or "seat on which apple juice was spilled" (Downing, 1977; Levi, 1978). Greater reliance on context in the absence of structure is one of the features that distinguish LG from full-fledged grammars.

To summarize, both Jackendoff and Wittenberg (2017) and Roeper (1999) propose that variation from syntactic norms can be explained if speakers have access to alternative grammars. The two approaches differ in what they assume about the nature of such alternative grammars. For Roeper (1999), whose framework is rooted in the generative grammar tradition, the available grammars do not differ in terms of complexity. The grammar that allows subject drop is structurally as complex as a grammar that prohibits subject omission – both have fully-developed syntactic components that can generate hierarchical structure. On the other hand, the alternative grammar that is available to speakers in Jackendoff and Wittenberg's (2017) framework implies lesser structural complexity: linear grammar, by definition, cannot produce hierarchical structures.

Which theory is better suited to explain syntactic structure of information requests that emerge in the context of computer-human interactions? If such requests can be shown to exhibit lesser structural complexity, then they can be viewed as a product of LG, as predicted by Jackendoff and Wittenberg (2017). If, on the other hand, information requests exhibit the same level of structural complexity as standard English, then their generation cannot be explained in terms of linear grammar and should be viewed as a product of an alternative full-fledged grammar, as predicted by Roeper (1999).

To understand better the predictions of the two approaches, in what follows we discuss manifestation of structural complexity in language and ways to evaluate it.

## Manifestation of Structural Complexity

Complexity and its cognitive underpinnings are among the most dividing theoretical issues in linguistics and cognitive science. Hauser, Chomsky, and Fitch (2002) argue that recursion – the ability of a linguistic pattern to reproduce itself – is the core property of human language, which makes it distinct from animal communication systems. All human languages are expected to show recursion. This position has been most notably challenged by Everett (2005), who argues that Pirahã, an indigenous Amazonian language, does not have recursion. While there is a disagreement on whether recursion should be considered a defining property of the human language (Jackendoff & Pinker, 2005), most of the researchers agree on how complexity is manifested in language. This position is summarized by Givón (2008) as follows: "What makes the syntactic structure of human language complex […] is the embedding of clauses in a subordinate – hierarchically lower – position inside other clauses, yielding recursive structure" (Givón, 2008, cited in Bickerton, 2008).

One manifestation of syntactic complexity is subordinate relative clauses. The sentence [$_{S1}$ *The boy,* [$_{S2}$ *who lives next door*]*, is friendly*] constitutes the case of structural subordination, because one structure of type S embeds another structure of type S. In this case, the relative clause [*who lives next door*] is embedded within the subject of the matrix clause, *the boy*, and functions as its modifier. The relative pronoun *who*, which introduces the relative clause, stands in anaphoric relation to the head noun that it modifies.

The relative clause construction can be generated by a full-fledged grammar that allows for structural embedding, but not by linear grammar, which can only generate flat structures. The option available for a language with linear grammar would be to package information into two syntactically independent structures, such as *A boy lives next door. He is friendly*. Such strategy is observed in child language (cf. *He met 'Toothless'. That was this big lion* vs. the targeted sentence with the relative clause *He met 'Toothless', who was a big lion* (Romaine, 1984)). The lack of relative clauses does not impose constrains on what kind of information can be communicated, since the same content can be presented in two independent clauses.

Everett (2005) argues that Pirahã does not have relative clauses and any other type of syntactic embedding for that matter, which he takes as evidence for lesser structural complexity. Everett's analysis was challenged by Nevins at al. (2009), but supported by recent corpus work on Pirahã reported in Futrell et al. (2016).

Following the line of argumentation developed in the previous literature, in what follows, we examine manifestation of syntactic complexity in information requests that arise in the context of human-computer interaction. We focus specifically on relative clauses, a typical case of syntactic subordination. If the proportion of relative clauses in queries is comparable to that in the standard language, then we will have to reject the LG approach proposed by Jackendoff and Wittenberg (2017) and assume that search queries are generated by a different type of grammar, which is as complex as the grammar of standard English, as proposed by Roeper (1999). On the other hand, if queries show a significantly smaller proportion of relative clauses, compared to the standard language, and consistently utilize linguistic means to avoid relative clauses, then such a finding can be taken as evidence in support of the LG model.

We side with Jackendoff and Wittenberg (2017) and predict that information requests in the contexts of human-computer interaction will display lesser structural complexity compared to their counterparts in the human-human condition. One crucial difference between our study of complexity in a specialized register of the English language and the study of complexity in e.g. aphasiacs or child language is that healthy adult speakers of English, unlike patients with brain damage and small children, always have access to full-fledged grammar. Previous studies show that in the context of search, older and less experienced users, as well as experienced users who are presented with complex search tasks might prefer to use standard English (Aula, Khan, & Guan, 2010). As a consequence, we do not expect to find categorical differences in the distribution of relative clauses in the context of search vs. in the standard language. What we expect, instead, is that there will be a significantly smaller proportion of relative clauses in the context of human-computer interaction compared to human-human interaction.

## Assessing Structural Complexity in Human-Computer Interaction

We present two studies that aim to address the question about structural complexity of information requests that emerge in the context of human-computer interaction. We focus specifically on the distribution of relative clauses, which we treat, following previous work, as a possible measure of structural complexity. Study 1 is a corpus study, in which we compare the distribution of relative clauses in a corpus of naturally occurring queries to the distribution of relative clauses in standard English. Study 2 reports results of a production experiment, in which participants are primed to produce relative clauses. We compare the distribution of relative clauses in two conditions, information requests addressed to a human (human-human condition) and information requests addressed to a computer (human-computer condition).

### Study 1: Corpus Analysis

**Design and Procedure** A corpus of ~80,000 naturally occurring queries was analyzed with a custom computer program in Python. The queries were entered into the California State University (CSU) search system by users between July 1, 2016 and October 20, 2016. Our primary interest was syntactic embedding and relative clauses, specifically. Since search queries are structurally different from the standard language, we were not able to utilize standard parsing libraries to extract relative clauses automatically (cf. Barr, Jones, & Regelson, 2008). We also considered searching for queries with overt relative clause pronouns, such as *who*, *which*, etc., an approach suggested in Szmrecsanyi (2004). However, there is a problem with this method: the program returns syntactically simple queries where *wh*-words function as question words rather than relative pronouns, as in *who takes ept*. To address the limitation of the automatic method, we randomly selected a subset of 1,200 queries, consisting of two-word queries and longer expressions, and examined them manually for the presence of relative clauses. We specifically focused on clauses with overt relative clause pronouns, such as *who*, *when*, *where*, etc.

**Results** we found a very small number of relative clauses in our set: 3. This is 5 times less than the frequency of relative clauses reported for English conversations by Biber (1991).

Notably, the examples with relative clauses in our dataset, such as *can I retake the class and replace what I received from it*, adhere to the syntactic norms and retain the key grammatical (subject, object, predicate) and functional elements that are usually absent in search queries.

**Discussion** Despite the fact that our data show very few instances of relative clauses, we are hesitant to take it as an accurate assessment of complexity. One possible reason for the small number of relative clauses in our dataset might be that queries differ from the standard language in terms of their lengths. Our analysis shows that the average length of queries in our corpus was 2.6 words (cf. Aula et al., 2010). For comparison, the average sentence length in the Brown corpus is 20 words. The distribution of query lengths is shown in Table 1.

Table 1: Query length in the corpus.

| Query length type | Proportion in corpus |
|---|---|
| 1-word queries | 14.4% |
| 2-word queries | 43.8% |
| 3-word queries | 23.4% |
| 4-word queries | 10.4% |
| 5-word queries | 4.3% |
| 6 & up | 3.6% |

While length is not in general indicative of syntactic complexity, it might be a contributing factor responsible for the low number of observed relative clauses in the corpus. To address this concern, we conducted a production experiment, which primed participants to use relative clauses, and compared their distribution in human-computer and human-human conditions.

## Study 2: Production Experiment

**Participants** Eighty participants took part in the experiment. They were recruited from the Amazon Mechanical Turk (AMT) platform, and redirected to a survey hosted on Qualtrics. All participants indicated that they were native speakers of English. The average age was 40 years old (the youngest participant was 23, and the oldest 69 years old). 62% were male, and 38% female. The participants were compensated for their participation.

**Stimuli** The stimuli consisted of 16 unique scenarios, designed to elicit an information request on the part of the user. Thematically, the scenarios focused on popular topics chosen to spike participants' curiosity, thus approximating the experimental setting to natural situations in which users look for information. Structurally, each scenario contained a relative clause introduced by an overt *wh*-pronoun. The priming conditions belonged to three groups: (i) subject relative clauses (6 total); (ii) relative clauses introduced by *when* (5 total); and (iii) relative clauses introduced by *where* (5 total). (See Appendix for the list of stimuli).

**Design and Procedure** In the beginning of the experiment, participants were introduced to the main protagonist, Maria. Maria was looking for some information. She had access to a computer and could type her information requests into a text box that appeared on computer screen. In the human-computer condition, participants learned that Maria's information requests will be answered by Google. In the human-human condition, participants learned that Maria's requests for information will be read and answered by a knowledgeable person. In both conditions participants were instructed to help Maria formulate her information requests.

After reading the introduction, participants saw 16 different scenarios, and were asked to type an information request in the text box. The order of scenarios was randomized. Each participant saw one condition only (between-subject design). The assignment of conditions to participants, human-computer vs. human-human, was randomized.

Based on the literature on syntactic priming (Bock, 1986), we expected that participants will tend to use relative clause constructions in their information requests. If the grammar responsible for generating information requests in the context of human-computer interaction does not differ in structural complexity from the grammar that operates in the human-human condition, there should be no difference in the proportion of relative clauses produced by participants in the two conditions. Such a finding would be compatible with the assumption of the TB framework proposed by Roeper (1999). If, on the other hand, we find a smaller proportion of relative clauses in the human-computer condition compared to the human-human condition, then such a finding can indicate that the grammar that operates in the context of human-computer interaction is structurally less complex, which will be compatible with the predictions of Jackendoff and Wittenberg (2017).

**Results** To analyze the proportion of relative clauses in the human-human vs. human-computer condition, each information request generated by experimental participants was manually coded. Expressions with relative clauses were coded as 1, and expressions lacking relative clauses were coded as 0.

Collapsing across the scenarios, there was a strong preference to use more relative clause constructions in the human-human condition ($m = 0.62$, $SD = 0.28$) compared to the human-computer condition ($m = 0.33$, $SD = 0.32$). The differences between the two conditions were statistically significant on a two-sided t-test ($t(78) = 4.4$, $p < 0.001$).

A closer look at the three types of priming scenarios, (i) scenarios with subject relatives, (ii) scenarios with relative clauses introduced by *when*, and (iii) scenarios with relative clauses introduced by *where*, reveals strong preference for relative clauses in the human-human condition compared to the human-computer condition across all three types. The results are presented in Figure 1.

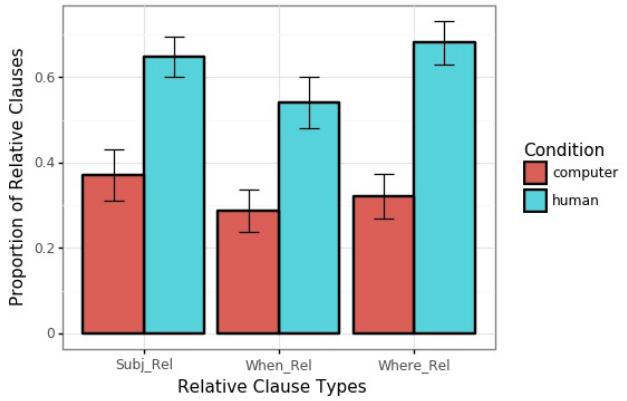


Figure 1: Proportion of relative clauses produced by participants across different types of priming scenarios. The computer condition represents human-computer interaction and the human condition the human-human interaction. Error bars indicate +/-1 SE.

**Discussion** The analysis of the data reveals that relative clauses are less likely to be used when participants formulate information requests in the human-computer condition compared to the human-human condition. Since constructions without relative clauses are considered to be syntactically less complex compared to constructions with relative clauses, we conclude that our results support the hypothesis that linguistic expressions generated in the context of human-computer interaction exhibit structural simplification relative to the norm.

The differences between information requests generated in the two conditions can be summarized along the following lines. In the human-human condition information requests usually have a form of well-formed questions, such as *How old was the player who set the record for the most home runs in 1975?* (Baseball scenario). Rarely do we observe any deviations from syntactic norms, such as missing arguments or lack of function words. When participants choose not to use relative clauses in the human-human scenarios, they often employ syntactic co-ordination, another syntactically complex construction that shows embedding, packaging information into two clauses connected by the conjunct *and*, as in *Which carnivorous indoor plants are safe for pets, and which nursery carries such plants?* (Nursery scenario).

Information requests in human-computer condition look syntactically very different. They can be characterized as strings of words that mention the main information points, as in *india tanneries water pollution* (Tanneries scenario). Information requests in this condition usually lack the main building blocks of a sentence. One prominent tendency we observed was the packaging of words into compounds, such as *pet safe carnivorous plants*. Some of the examples are presented in Table 2. Compounding in the context of human-computer interaction seems to provide alternative means to convey information that would otherwise be expressed with a relative clause (cf. *carnivorous plants [that are safe for pets]*). Well-formed compounds are frequently followed by modifiers that provide additional information about location or time. Unlike in the standard language, there are no prepositions to link this material to the rest of the construction: *[home run record] [1975] [age]*. Lack of prepositions is expected and predicted in structures generated by LG.

Table 2: Compounds in human-computer condition.

| Name of Scenario | Compounds |
|---|---|
| Margaret Mead | margaret mead research culture |
| LA restaurant | LA deaf restaurant |
| Chicago stadium | vegan meal sport stadium owner |
| Tectonic plates | tectonic plates meet properties |

Importantly, information requests that contain relative clauses in the human-computer condition, such as *What's the composition of soil in forests that are above 300 feet* (Trees scenario), retain all features of the standard language. This finding aligns with the corpus data from Study 1. In cases like this, it looks like language users are utilizing their dominant, full-fledged grammar, even within a register that can potentially trigger access to LG. This is not unexpected. While the LG model does not discuss the motivation for switching to the default grammar in the context of specialized registers, it does not preclude such a possibility. Previous studies have shown that certain factors, such as complexity of a search task and the domain expertise of the user, are likely to trigger full-length grammatical questions rather than strings of words (Aula et al., 2010). Our analysis of the queries generated in Study 2 shows that there are no scenarios that would consistently trigger switch to the full-fledged grammar, which suggests that within-user variation might be accounted for by other factors, such as familiarity with particular domains or experience with search engines.

## General Discussion

If healthy adult speakers can choose to use full-fledged grammar, even in the contexts of simplified registers, what is the utility of LG? Ferguson (1982) argues that simplification facilitates communication between proficient and less proficient speakers, using baby talk and foreigner talk as an example. At the same time, the process of simplification in telegrams, headlines, sports announcers' talk and academic note-taking is attributed to economy considerations, the need to use language efficiently under the time pressure, space limitation and/or under increased cognitive load (Janda, 1985).[1] The switch to LG in the context of information search can be driven by economy

[1] While the question about the cognitive and neurological foundations of simpler vs. complex grammars and their processing cost is beyond the scope of this paper, a promising direction for this type of work is presented in Rodriguez and Granger (2016), who argue that the same mechanism, if powerful enough, can process both complex and simpler structures.

considerations, such as space limitation. Another possibility, suggested by a reviewer, is that simplification is driven by speakers' recursive reasoning about the type of grammar computers can understand (cf. Frank & Goodman, 2012 on Rational Speech Act).

This discussion raises the question of whether information requests should be considered a form of natural language? Strong argument in support of this position is that search queries exhibit systematic properties, albeit different from that of the standard language. For example, Li (2010) argues that search queries have a clearly identifiable semantic structure, consisting of intent head (IH) and intent modifiers (IM), as in [IM *alice in wonderland*] [IM *2010*] [IH *cast*]. Moreover, properties of information requests align with the core properties of other specialized registers that are indisputably considered to be forms of natural language. Of particular notice is the effect of context on word order. In the absence of syntax, word order in specialized registers is governed by semantic and pragmatic considerations, seen in the tendency to mention agent before action (Jackendoff & Wittenberg, 2017). While agents and actions are extremely rare in information requests, pragmatic considerations nevertheless affect word order. Specifically, Smirnova, Lenarsky, and Romero Sanchez (2019) show that in search queries with multiple adjectival modifiers, the attribute that is more important to the speaker tends to be mentioned first, even if the resulting word order (cf. *aluminum red bike*) violates the default adjective ordering in the standard language (*red aluminum bike*). These findings challenge the position that information requests that arise in the context of human-computer interaction are "unstructured collections of terms" (Barr et al., 2008) rather than a form of natural language.

## Conclusion

In this paper we investigated a range of variation in linguistic complexity, focusing specifically on constructions that emerge in the context of human-computer interaction, such as *margaret mead culture famous research*. Such constructions present a puzzle for the formal theory of language, as they differ substantially from the baseline.

The results of the production study reported here suggest that linguistic output generated in the context of human-computer interaction exhibits structural simplification compared to the standard language. Focusing on the distribution of relative clauses, we show that native speakers tend to produce fewer relative clauses in human-computer condition compared to the human-human condition, despite syntactic priming.

Our study has direct theoretical implications. Structural simplification observed in the context of human-computer interaction can be explained if we assume that in such contexts speakers utilize linear grammar, a system that allows to communicate meaning but lacks the syntactic component responsible for structural complexity in the standard language. Our work, thus, provides support for the model of grammar developed by Jackendoff and Wittenberg (2017). On the empirical side, our research contributes to the discussion on what constitutes a lower bound of complexity in language (Futrell et al., 2016) by bringing to the table an overlooked linguistic phenomenon in the English language.

## Acknowledgments

I thank Peter Culicover, Ray Jackendoff, Brian Joseph, and Maria Piñango for productive conversations on language complexity. I am also grateful to the participants of the LSA 2019 for their feedback.

## References

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult? *Proceedings of the 10th SIGCHI conference on human factors in computing systems* (pp. 35-44). New York, NY: Association for Computing Machinery.

Barr, C., Jones, R., & Regelson, M. (2008). The linguistic structure of English web-search queries. *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 1021-1030). Honolulu, HI: Association for Computational Linguistics.

Biber, D. (1991). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bickerton, D. (2008). Recursion: core of complexity or artifact of analysis? In T. Givón & M. Shibatani (Eds.), *Syntactic complexity: Diachrony, acquisition, neuro-cognition, evolution*. Amsterdam: John Benjamins.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355-387.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, *53*, 810-842.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology*, *4*, 621-646.

Ferguson, C. A. (1975). Toward a characterization of English foreigner talk. *Anthropological Linguistics*, *17*, 1-14.

Ferguson, C. A. (1982). Simplified registers and linguistic theory. In L. Obler & L. Menn (Eds.), *Exceptional language and linguistics*. New York, NY: Academic Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998-998.

Futrell, R., Stearns, L., Everett, D. L., Piantadosi, S. T., & Gibson, E. (2016). A corpus investigation of syntactic embedding in Pirahã. *PloS one*, *11*: e0145289

Givón, T. (2008). *The genesis of syntactic complexity*. Amsterdam: John Benjamins.

Haegeman, L. (1987). Register variation in English: Some theoretical observations. *Journal of English Linguistics*, *20*, 230-248.

Haegeman, L. (2013). The syntax of registers: Diary subject omission and the privilege of the root. *Lingua*, *130*, 88-110.

Hauser, M., Chomsky, N., & Fitch, T. (2002). The faculty of language: what is it, who has it, and how did it evolve. *Science*, *198*, 1569-79.

Jackendoff, R., & Pinker, S. (2005). The Nature of the language faculty and its implications for evolution of language. *Cognition*, *97*, 211-225.

Jackendoff, R., & Wittenberg, E. (2017). Linear grammar as a possible stepping-stone in the evolution of language. *Psychonomic Bulletin & Review*, *24*, 219-224.

Janda, R. D. (1985). Note-taking English as a simplified register. *Discourse Processes*, *8*, 437-454.

Kittredge, R., & Lehrberger, J. (Eds.) (1982). *Sublanguage: Studies of language in restricted semantic domains*. New York, NY: Walter de Gruyter.

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York, NY: Academic Press.

Li, X. (2010). Understanding the semantic structure of noun phrase queries. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1337-1345). Uppsala: Association for Computational Linguistics.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. *Language*, *85*, 355-404.

Progovac, L., Paesani, K., Casielles, E., & Barton, E. (Eds.). (2006). *The syntax of nonsententials: Multidisciplinary perspectives*. Amsterdam: John Benjamins.

Reaser, J. (2003). A quantitative approach to (sub) registers: the case of 'Sports Announcer Talk'. *Discourse Studies*, *5*, 303-321.

Rodriguez, A., & Granger, R. (2016). The grammar of mammalian brain capacity. *Theoretical Computer Science*, *633*, 100-111.

Roeper, T. (1999). Universal bilingualism. *Bilingualism: Language and Cognition*, *2*, 169-186.

Romaine, S. (1984). Relative clauses in child language, pidgins and creoles. *Australian Journal of Linguistics*, *4*, 257-281.

Sadock, J. M. (1974). Read at your own risk: syntactic and semantic horrors you can find in your medicine chest. *Proceedings of CLS 10* (pp. 599-607). Chicago, IL: Chicago Linguistic Society.

Sager, Naomi (1986). Sublanguage: Linguistic phenomenon, computational tool. In R. Grishman & R. Kittredge (Eds.), *Analyzing language in restricted domains*. New York, NY: Lawrence Erlbaum.

Smirnova, A., Lenarsky, A., & Sanchez, R. R. (2019). Contextual determinants of adjective order: Beyond itsy bitsy teeny weeny yellow polka dot bikini. *Proceedings of the 41st annual conference of the cognitive science society* (pp. 2825-2831). Montreal, QB: Cognitive Science Society.

Szmrecsanyi, B. (2004). On operationalizing syntactic complexity. *Proceedings of the 7th international conference on textual data and statistical analysis* (pp. 1032-1039). Louvain-la-Neuve: Presses universitaires de Louvain.

Yanofsky, N. M. (1978). NP utterances. *Proceedings of CLS 14* (pp. 491-502). Chicago, IL: Chicago Linguistic Society.

## Appendix: Experimental Stimuli

**Stimuli priming subject relative clauses**

1. [**Bitcoin bus**] Maria wants to know if there is a bus that only accepts bitcoins. What should she type into the text box? 2. [**LA restaurant**] Maria knows that there is a restaurant in LA that only hires deaf people, and wants to know the location of the restaurant. What should she type into the text box? 3. [**Baseball**] Maria wants to know how old the player who set the record for the most home runs in 1975 was at the time. What should she type into the text box? 4. [**Philanthropy**] Maria wants to know the names of super rich people in the Bay Area who pledged to give away most of their wealth. What should she type into the text box? 5. [**Carnivorous plants**] Maria wants to get carnivorous indoor plants that are safe for pets, and wants to know which nursery carries such plants. What should she type into the text box? 6. [**Chicago stadium**] Maria knows that there is a sport stadium in the Greater Chicago Area which serves vegan meals, and wants to know who is the owner of the stadium. What should she type into the text box?

**Stimuli priming *when* relative clauses**

1. [**Bees**] Maria wants to know what bees do when animals try to take their honey. What should she type into the text box? 2. [**Coastal cities**] Maria wants to know what will happen to coastal cities when sea-level rise reaches 1 inch. What should she type into the text box? 3. [**Einstein's brain**] Maria is trying to remember a popular quote attributed to Thomas Stoltz Harvey when it was discovered that he stole Albert Einstein's brain. What should she type into the text box? 4. [**Kasparov**] Maria wants to know what Garry Kasparov said when he was defeated by Deep Blue, an IBM supercomputer. What should she type into the text box? 5. [**Tesla**] Maria wants to know how Nikola Tesla reacted when he learned that Guglielmo Marconi won the Nobel Prize for the development of radio technology. What should she type into the text box?

**Stimuli priming *where* relative clauses**

1.[**Tectonic plates**] Maria wants to know about the geological properties of places where two tectonic plates meet. What should she type into the text box? 2. [**Trees**] Maria wants to know the composition of the soil in forests where trees grow above 300 feet. What should she type into the text box? 3. [**Margaret Mead**] Maria wants to know what the culture where Margaret Mead did her research is most famous for. What should she type into the text box? 4. [**Indian tanneries**] Maria is interested in the economic status of the states in India where tanneries are the main source of water pollution. What should she type into the text box? 5.[**Vaping**] Maria is interested in whether people who live in areas where vaping is illegal are healthier compared to the general population. What should she type into the text box?