BRIEF REPORT

Syntax-semantics interactions – seeking evidence from a synchronic analysis of 38 languages [version 1; peer review: awaiting peer review]

Tom S Juzek[1], Yuri Bizzoni[2]

[1]Kauz Linguistic Technologies, Düsseldorf, 40223, Germany
[2]Saarland University, Saarbrücken, 66123, Germany

v1 **First published:** 01 Apr 2021, **10**:265
https://doi.org/10.12688/f1000research.50988.1
Latest published: 01 Apr 2021, **10**:265
https://doi.org/10.12688/f1000research.50988.1

Abstract

The notion that, to facilitate processing, as semantic complexity increases, syntactic complexity decreases, follows from various linguistic theories. This brief report presents the results of testing that notion, by analysing synchronic data from 38languages and correlating canonical measures of semantic and syntactic difficulty. We expected an overall positive tendency. However, the results came out mixed to negative. There is a notable degree of variation and there are no clear tendencies within language families. After detailing the theoretic and cognitive reasons that support the original hypothesis, we conclude with a short discussion about the potential causes and implications of our findings. A possible interpretation is that the interaction we are looking for is more subtle than one might have assumed.

Keywords

cognitive linguistics, corpus analysis, syntax, semantics, syntax-semantics interactions, efficient communication, dependency grammars, Universal Dependencies, null results

Open Peer Review

Reviewer Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Tom S Juzek (tsjuzek@posteo.be), Yuri Bizzoni (yuri.bizzoni@gmail.com)

Author roles: Juzek TS: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Bizzoni Y**: Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: At the time of the project, both authors were employedthrough the SFB1102 research grant by the DFG.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Juzek TS and Bizzoni Y. **Syntax-semantics interactions – seeking evidence from a synchronic analysis of 38 languages [version 1; peer review: awaiting peer review]** F1000Research 2021, **10**:265
https://doi.org/10.12688/f1000research.50988.1

First published: 01 Apr 2021, **10**:265 https://doi.org/10.12688/f1000research.50988.1

Introduction

The syntactic complexity and semantic difficulty of a sentence are two different aspects of language, each requiring an independent amount of processing effort.[1] Many linguists also assume that the interplay between different linguistic levels can facilitate efficient communication.[2]

A relatively popular assumption is that syntax facilitates semantics.[1,3–5] Several linguists and philosophers of language assume that grammar was optimized for rational and efficient communication.[6] Various recent studies that have looked into the question from an evolutionary and typological viewpoint also support the idea that the drive for communicative efficiency has shaped syntactic features such as subject-verb-object order[7] or average dependency length,[8] and semantic domains such as numeral systems[9] and colour terms.[10] Concerning grammar, it seems safe to suppose that denser semantic content tends to reduce syntactic complexity. This notion is an explicit part of some linguistic theories, e.g., systemic functional grammar,[11] and is in harmony with other theories, e.g., the uniform information density hypothesis.[12–14] However, despite an increasing body of research, a large-scale study of synchronic data, comparing a semantic with a syntactic analysis, is still missing.

Corpus study

Under the view that syntax facilitates semantics, it is plausible to assume that *on average* writers and speakers will deliver difficult semantics via simpler syntax and simplify semantics in the presence of difficult syntax in order to ensure a healthy level of communicative complexity - price the risk of misunderstanding. That is, *we expect a tendentially inverse relationship between semantic and syntactic complexity.* Whether real data corroborates this expectation is the research question framing this report. We elaborate it through two competing hypotheses.

H0: There is no systematic interaction between semantic complexity and syntactic complexity.

H1: There is an inverse relationship between semantic complexity and syntactic complexity: as lexical semantics complexity increases, syntactic complexity decreases, and vice versa.

A repository containing all files relevant for replication of this research can be found online.[33,34]

Methods

Corpora and languages

To test the hypotheses, we used the corpora made available through the Universal Dependencies (UD) project[a], accessed in June 2020. The UD corpora are, especially for large languages, a mix of different registers and domains, ranging from blog posts to legal texts, the vast majority of which comes from contemporary sources. Critically, the corpora contain *manually* annotated syntactic information in the form of dependency relations. The fact that syntactic annotation was done manually and not automatically is beneficial for our goals, since it removes the risk of systematic mistakes introduced by automatic syntactic annotation. For any language with more than one (sub)corpus, we merged the corpora and only included languages for which we have more than 0.1m tokens in total. Further, we only considered languages that have their lemmas specified, which leaves us with the languages we list in Table 1.

Semantic measure

We measured semantic complexity with a canonical and straightforward measure: the sum of the log frequencies of a sentence's lemmas[b]. The inverse correlation between word frequency and its processing difficulty is widely known,[15,16] while lemmas and inflected forms do not appear to behave very differently in terms of distribution.[17] If readers encounter a particular word several times, they become more effective in processing it both at the orthographic and semantic levels; and the more they encounter the word, the more effective they become. Thus, all other things being equal, a sentence's sum of log lemma frequency should be a proxy for its processing cost. The reason why we did not opt for sequence surprisal or perplexity, for example, is that sequences longer than one word necessarily reflect syntactic behaviours. That is, at an n-gram level, surprisal and perplexity contain both semantic and syntactic information. They are functionally equivalent to the frequency measure at a unigram level, so we use the latter as our semantic measure.

Syntactic measure

We used the UD framework to measure syntactic complexity. The UD framework expresses syntactic relations through dependencies: each element depends on another element, its head.[18] In UD, the head is the semantically most salient element, and the dependent modifies the head. The top-level head is the root of a sequence, typically the main verb of the

[a]https://universaldependencies.org/

[b]We log the absolute frequencies. Since we apply the measure only to the corpus from which we take it, this does not lead to distortions.

Table 1. The correlations between the syntactic and the semantic measures of different languages, given in the "r" columns and ordered accordingly. An asterisk indicates a p<=0.05. The median sentence length for any given language is in the "med. len." column. "N" is the number of sentences analyzed. "r long" is the correlation coefficient for twice the median sentence length.

Language	med. len.	N	r	r long
Finnish	5	379	0.20*	−0.04
Estonian	6	849	0.18*	0.15*
Slovak	5	338	0.15*	0.11
Latvian	9	309	0.13*	−0.07
Slovenian	10	183	0.12	0.10
Swedish	8	234	0.10	−0.10
Polish	6	1330	0.09*	0.07
Arabic	11	92	0.09	−0.06
Norweg.	11	765	0.09*	−0.10*
German	13	5134	0.08*	0.03
Cl. Chin.	4	6319	0.07*	0.03
Roman.	14	631	0.05	0.04
Czech	10	2342	0.05*	−0.03
Hindi	14	565	0.05	−0.11
Naija	5	168	0.05	−0.07
Croatian	15	207	0.04	0.03
Spanish	17	408	0.04	−0.10*
Chinese	14	296	0.02	−0.22*
Basque	8	357	0.02	−0.10
French	10	549	0.01	0.11*
Portug.	10	141	0.01	−0.16*
Russian	10	1516	−0.01	−0.03
Persian	13	89	−0.02	0.16
Icelandic	15	27	−0.02	−0.38
English	12	527	−0.03	0.00
Urdu	17	123	−0.05	−0.38*
Catalan	23	280	−0.06	−0.37*
Ukrainian	8	142	−0.06	−0.26*
Old Russ.	5	896	−0.06	−0.05
Korean	9	1421	−0.07*	−0.10*
Dutch	6	476	−0.08	0.03
Galician	26	144	−0.09	−0.73*
Japanese	13	214	−0.13	−0.07
Latin	8	2256	−0.14*	−0.14*
Danish	11	105	−0.15	−0.01
Italian	14	670	−0.28*	−0.07
Hebrew	13	137	−0.29*	−0.22*
Bulgarian	6	368	−0.48*	0.01

matrix clause. Critical to our syntactic measure is the notion of dependency length. For each element, we can calculate the distance to its head, where each intervening element increases the distance by one, and for each sentence, we can add up the distances for a sentence's elements, which results in the *sum of dependency lengths*. Figure 1 illustrates a dependency analysis; the sum of distances in that example is 6 (2+1+2+1). The measure is taken from the literature.[2,19–21] We consider this an intuitive measure: all other things being equal, sentences with shorter dependency lengths will be easier to process.[2,22]

Controlling for confounds

Since the relationship between sentence length and summed dependency distances is not necessarily linear,[2] there is no simple way to normalise for sentence length. Thus, to remove the risk of spurious sentence length effects, we computed our correlation only at the median sentence length per language, excluding punctuation.[c] For example, the median sentence length for English is 12 tokens, while for Dutch, it is 6. Another aspect we controlled for is word length. Word length is known to roughly correlate with information content,[23] so for any median token length, we further extracted the median character length and considered a range of +/-10%. For instance, the median character length for English sentences of 12 tokens is 47 characters, so we only considered sentences with 12 tokens that are within a range of 42 to 52 characters. Since the median sentence length is relatively short for some languages, we also analysed longer sentences, viz. twice the median sentence length, which is the 'long' condition.

Analysis

We calculated the Pearson correlation coefficient between the summed logged frequencies and the summed dependency distances for each language, using R.[24] For our two measures, *positive* correlation coefficients are evidence for H1: We expect low frequencies (high semantic complexity) to correlate with shortened dependency lengths (low syntactic complexity).

Results and discussion

The results are given in Table 1 and are in parts illustrated in Figure 2. Due to space limitations, we omitted the N for the long condition. In total, we analyzed about 31.6k sentences in the normal condition and about 16.1k sentences in the long condition.

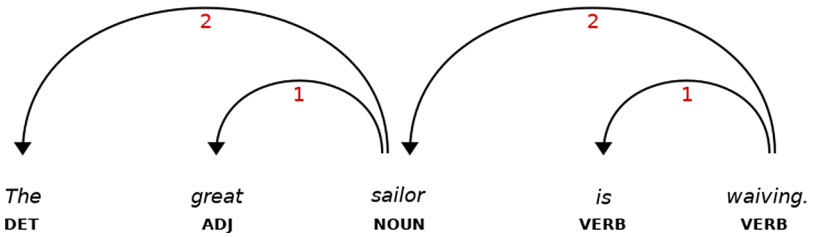

Figure 1. Visualising a simple sequence in the Universal Dependencies framework, including dependency distances (red numerals below a dependency arch/edge). (Created with spaCy: https://spacy.io/.)

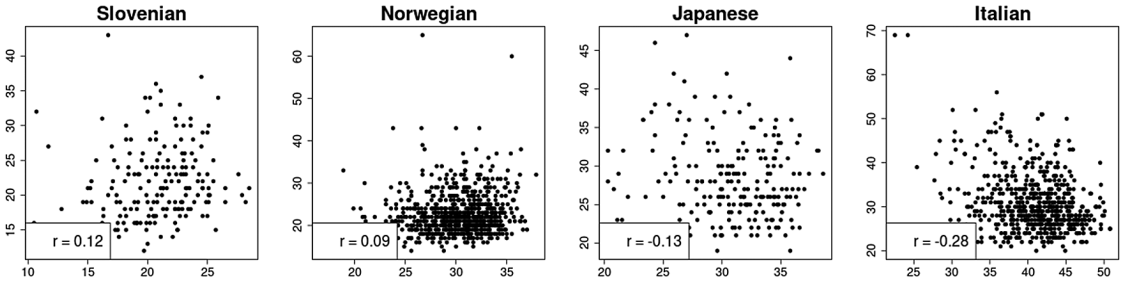

Figure 2. An illustration of the results of the semantic analysis (x-axis), and the syntactic analysis (y-axis) including Pearson's r for a sample of languages. Each dot represents the results for a certain sentence. Within a language, all sentences are of the same token length. (Created with R.[24])

[c]As a test, we computed this correlation for all sentence lengths in English, finding consistent results.

Our results are mixed-to-negative and seem to be language-dependent. There is a rather high variance centred around a mean near zero: 21 languages display the expected positive correlations, 9 of which are significant, but 17 languages display negative correlations, 5 of which are significant. We cannot interpret these findings as evidence for H1, and while it is difficult to affirm H0, we cannot reject it.

Taken at face value, this would mean that in some languages, an increase in semantic complexity implies an increase in syntactic complexity, which seems implausible, also since there is quite some variation within language groups, e.g., Slovenian and Bulgarian or Swedish and Dutch. There could be various other causes for the observed variation and the lack of a systematic correlation as discussed below.

Semantic measure weakness

We used a canonical, easy to compute, and corpus-dependent semantic measure, amply verified in linguistics, information theory, and cognitive science. While it is a *proxy* for semantic complexity, as such it leaves room for improvement, we find it unlikely that the mixed results we observe are entirely due to our measure of complexity.

Syntactic measure weakness

Our syntactic measure depends on the UD framework's validity, which aims to be language universal.[18] We cannot exclude that there are inconsistencies in theoretical choices across languages, e.g., how prepositions or relativizers are attached.[25] An in-depth qualitative analysis is needed to follow up on this, and those choices might have contributed to some variance. The extensive amount of work gone into the formulation of the UD framework, together with the efforts made to ensure the consistency of UD syntactic annotations across different languages, makes us think that internal inconsistencies should not shift the mean to a degree such that it conceals a real effect.

Domain/register effects

Specific registers, or even domains, might be characterised by an unusual complexity of both the semantic and the syntactic side, or have a technical vocabulary that becomes "simple" to their usual readers, leaving room for more syntactic complexity. However, the fact that we draw our semantic measures from the corpora themselves should mitigate such an effect: if a rare semantic unit frequently appears in one of our corpora, its frequency increases, and its complexity level decreases. While differences in the corpora's makeup could cause parts of the observed variance, there is no apparent reason why they should have caused the mean to shift towards zero.

Further factors

Other factors that we could not account for could be at play. This concerns pragmatic aspects[26] or even morpho-phonetic aspects of the lexicon.[27–30] Another source of noise in our data could be that syntactic structures can increase in complexity to a good extent without requiring any semantic simplification or even favoring semantic complexity.[31]

We have identified some factors that could cause the observed variance. However, if the interaction between syntax and semantics was as strong as we assumed, then there should have been a measurable shift towards a positive mean in most languages. That means that there should have been *some* signal from the data at hand "shining through" the noise. Nonetheless, the results are not evidence of our H0: the absence of evidence is not the evidence of absence.[32] A practical takeaway could be that the interaction between syntax and semantics is more complex and more subtle than we sometimes conceived it to be, certainly at a synchronic level.

Conclusions

This brief report explored in how far syntax and semantics interact in synchronic, large-scale data from 38 languages. To compute syntactic and semantic complexity, we relied on two common measures: dependency length and unigram frequency. These measures are only one way to approach the problem, but their simplicity and empirical robustness make them powerful and unassuming proxies, widely employed in computational and cognitive linguistics. Against our expectation, the analysis did not produce a widespread positive correlation between the two measures that would express an inverse correlation between syntactic and semantic complexity. In many of the languages represented in our data, even when the semantic processing load is high, its syntactic complexity does not appear systematically lowered. Since we used the most straightforward method we conceived of to test our hypothesis, we think that if the effect we hypothesised was obvious and consistent across languages, it should have been visible with the methodology we used, i.e., the mean of all correlations should have had a clear positive tendency. Since this is not the case, our results hint at the possibility that the interaction we are looking for is more subtle than one could have assumed. We find our negative results surprising, and we hope that our report stimulates both discussion and further research.

Data availability

Underlying data

Zenodo: Syntax-semantics interactions – seeking evidence froma synchronic analysis of 38 languages: dataset repository. https://doi.org/10.5281/zenodo.4542536.

This project contains the following underlying data:

- A subset of the Universal Dependencies Corpora v2.6 used in this study.

Apache 2.0 license, with the Universal Dependency Consortium holding the copyright.

Extended data

Zenodo: Syntax-semantics interactions – seeking evidence froma synchronic analysis of 38 languages: scripts repository. https://zenodo.org/record/4643152.

This project contains the following extended data:

- The scripts used for data preparation and analysis.

The scripts are available under a CC0 license ("No rights reserved"). While the scripts modify the data sets, the modified data sets continue to be subject to the same Apache 2.0 license as the original data sets.

Author contributions

YB and TSJ co-designed the study. TSJ coded the analysis.

Competing interests

There are no competing interests.

References

1. Ostrin RK, Tyler LK: **Dissociations of lexical function: Semantics, syntax, and morphology.** *Cogn Neuropsychol.* 1995; **12**(4): 345–389.
Publisher Full Text
2. Gibson E, Futrell R, Piantadosi S, *et al.*: **How Efficiency Shapes Human Language.** *Trends Cogn Sci.* 2019; **23**(5): 389–407.
PubMed Abstract | Publisher Full Text
3. Boland JE: **The relationship between syntactic and semantic processes in sentence comprehension.** *Lang Cogn Processes.* 1997; **12**(4): 423–484.
Publisher Full Text
4. Ainsworth-Darnell K, Shulman HG, Boland JE: **Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials.** *J Mem Lang.* 1998; **38**(1): 112–130.
Publisher Full Text
5. Johns CL, Campanelli L, Kush D, *et al.*: **Individual differences in combinatorial semantic processing: Skilled comprehension facilitates complement coercion during sentence comprehension.** 2019.
Publisher Full Text
6. Frank AF, Jaeger TF: **Speaking rationally: Uniform information density as an optimal strategy for language production.** *In Proceedings of the annual meeting of the cognitive science society* 2008; volume **30**.
7. Hahn M, Jurafsky D, Futrell R: **Universals of word order reflect optimization of grammars for efficient communication.** *Proc Natl Acad Sci U S A.* 2020; **117**(5): 2347–2353.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Futrell R, Levy RP, Gibson E: **Dependency locality as an explanatory principle for word order.** *Language.* 2020; **96**(2): 371–412.
Publisher Full Text
9. Xu Y, Liu E, Regier T: **Numeral systems across languages support efficient communication: From approximate numerosity to recursion.** *Open Mind.* 2014: pages 1–14.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Steinert-Threlkeld S, Szymanik J: **Ease of learning explains semantic universals.** *Cognition.* 2020; **195**: 104076.
PubMed Abstract | Publisher Full Text
11. Halliday MAK: **Introduction: On the "architecture" of human language**. In: Webster J, editor, *On lang ling.* London, New York: Continuum; 2003; **chapter 1**: pages 1–30.
12. Aylett M, Turk A: **The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech.** *Lang Speech.* 2004; **47**(1): 31–56.
PubMed Abstract | Publisher Full Text
13. Jaeger TF, Levy RP: **Speakers optimize information density through syntactic reduction**. In: Schölkopf B, Platt JC, Hoffman T, editors, *Advances in Neural Information Processing Systems 19.* 2007: pages 849–856. MIT Press.
14. Jaeger TF: **Redundancy and reduction: Speakers manage syntactic information density.** *Cogn psychol.* 2010; **61**(1): 23–62.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Kuperman V, Van Dyke JA: **Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers.** *J Exp Psychol Hum Percept Perform.* 2013; **39**(3): 802.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Piantadosi ST: **Zipf's word frequency law in natural language: A critical review and future directions.** *Psychon Bull Rev.* 2014; **21**(5): 1112–1130.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Corral A, Boleda G, Cancho RF: **Zipf's law for word frequencies: Word forms versus lemmas in long texts.** *PloS One.* 2015; **10**(7): e0129031.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Nivre J, Abrams M, Agić Z, *et al.*: **Universal Dependencies 2.4.** 2019.
Reference Source
19. Futrell R, Mahowald K, Gibson E: **Large-Scale Evidence of Dependency Length Minimization in 37 Languages.** *Proc Natl*

Acad Sci U S A. 2015; volume **112**: pages 10336–10341.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Gibson E: **The dependency locality theory: A distance-based theory of linguistic complexity.** *Image, language, brain.* 2000: 95–126.

21. Liu H: **Dependency Distance as a Metric of Language Comprehension Difficulty.** *J Cogn Sci.* 2008; **9**: 159–191.
Publisher Full Text

22. Hawkins JA: *Efficiency and Complexity in Grammars.* Oxford: Oxford University Press; 2004.

23. Lewis ML, Frank MC: **The length of words reflects their conceptual complexity.** *Cognition.* 2016; **153**: 182–195.
PubMed Abstract | Publisher Full Text

24. R Core Team:*R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020.
Reference Source

25. Osborne T, Gerdes K: **The status of function words in dependency grammar: A critique of universal dependencies (ud).** *Glossa.* 2019.
Publisher Full Text

26. Kuperberg GR, McGuire PK, Bullmore ET, *et al.*: **Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fmri study.** *J Cogn Neurosci.* 2000; **12**(2): 321–341.
PubMed Abstract | Publisher Full Text

27. Graves WW, Grabowski TJ, Mehta S, *et al.*: **A neural signature of phonological access: distinguishing the effects of word frequency from familiarity and length in overt picture naming.** *J Cogn Neurosci.* 2007; **19**(4): 617–631.
PubMed Abstract | Publisher Full Text

28. Rothermich K, Schmidt-Kassow M, Kotz SA: **Rhythm's gonna get you: Regular meter facilitates semantic sentence processing.** *Neuropsychologia.* 2012; **50**(2): 232–244.
PubMed Abstract | Publisher Full Text

29. Canette L-H, Lalitte P, Bedoin N, *et al.*: **Rhythmic and textural musical sequences differently influence syntax and semantic processing in children.** *J Exp Child Psychol.* 2020; **191**: 104711.
PubMed Abstract | Publisher Full Text

30. King A, Wedel A: **Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing.** *Open Mind.* 2020; **4**: 1–12.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Schoenemann PT: **Syntax as an emergent characteristic of the evolution of semantic complexity.** *Minds Machines.* 1999; **9**(3): 309–346.
Publisher Full Text

32. Altman DG, Bland JM: **Statistics notes: Absence of evidence is not evidence of absence.** *Bmj.* 1995; **311**(7003): 485.
Publisher Full Text

33. Juzek TS, Bizzoni Y: **Syntax-semantics interactions – seeking evidence froma synchronic analysis of 38 languages: dataset repository (Version 1.0) [Data set].** *Zenodo.* 2021.
Publisher Full Text

34. Juzek TS, Bizzoni Y: **Syntax-semantics interactions – seeking evidence froma synchronic analysis of 38 languages: scripts repository (Version v1.1).** *Zenodo.* 2021, February 16.
Publisher Full Text