BRIEF REPORT

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

Tom S Juzek1, Yuri Bizzoni2

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

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 38 languages 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

Invited Reviewers

1. Chunshan Xu, Anhui Jianzhu University, Hefei, China
2. Diego Frassinelli, University of Konstanz, Konstanz, Germany

Any reports and responses or comments on the article can be found at the end of the article.
**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.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 or average dependency length, and semantic domains such as numeral systems and colour terms.⁷⁻⁸ 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.⁹⁻¹² 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, 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. The inverse correlation between word frequency and its processing difficulty is widely known, while lemmas and inflected forms do not appear to behave very differently in terms of distribution. 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. 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

---

¹https://universaldependencies.org/

²We 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.

<table>
<thead>
<tr>
<th>Language</th>
<th>med. len.</th>
<th>N</th>
<th>r</th>
<th>r long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finnish</td>
<td>5</td>
<td>379</td>
<td>0.20*</td>
<td>−0.04</td>
</tr>
<tr>
<td>Estonian</td>
<td>6</td>
<td>849</td>
<td>0.18*</td>
<td>0.15*</td>
</tr>
<tr>
<td>Slovak</td>
<td>5</td>
<td>338</td>
<td>0.15*</td>
<td>0.11</td>
</tr>
<tr>
<td>Latvian</td>
<td>9</td>
<td>309</td>
<td>0.13*</td>
<td>−0.07</td>
</tr>
<tr>
<td>Slovenian</td>
<td>10</td>
<td>183</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>Swedish</td>
<td>8</td>
<td>234</td>
<td>0.10</td>
<td>−0.10</td>
</tr>
<tr>
<td>Polish</td>
<td>6</td>
<td>1330</td>
<td>0.09*</td>
<td>0.07</td>
</tr>
<tr>
<td>Arabic</td>
<td>11</td>
<td>92</td>
<td>0.09</td>
<td>−0.06</td>
</tr>
<tr>
<td>Norweg.</td>
<td>11</td>
<td>765</td>
<td>0.09*</td>
<td>−0.10*</td>
</tr>
<tr>
<td>German</td>
<td>13</td>
<td>5134</td>
<td>0.08*</td>
<td>0.03</td>
</tr>
<tr>
<td>Cl. Chin.</td>
<td>4</td>
<td>6319</td>
<td>0.07*</td>
<td>0.03</td>
</tr>
<tr>
<td>Roman.</td>
<td>14</td>
<td>631</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>Czech</td>
<td>10</td>
<td>2342</td>
<td>0.05*</td>
<td>−0.03</td>
</tr>
<tr>
<td>Hindi</td>
<td>14</td>
<td>565</td>
<td>0.05</td>
<td>−0.11</td>
</tr>
<tr>
<td>Naija</td>
<td>5</td>
<td>168</td>
<td>0.05</td>
<td>−0.07</td>
</tr>
<tr>
<td>Croatian</td>
<td>15</td>
<td>207</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Spanish</td>
<td>17</td>
<td>408</td>
<td>0.04</td>
<td>−0.10*</td>
</tr>
<tr>
<td>Chinese</td>
<td>14</td>
<td>296</td>
<td>0.02</td>
<td>−0.22*</td>
</tr>
<tr>
<td>Basque</td>
<td>8</td>
<td>357</td>
<td>0.02</td>
<td>−0.10</td>
</tr>
<tr>
<td>French</td>
<td>10</td>
<td>549</td>
<td>0.01</td>
<td>0.11*</td>
</tr>
<tr>
<td>Portug.</td>
<td>10</td>
<td>141</td>
<td>0.01</td>
<td>−0.16*</td>
</tr>
<tr>
<td>Russian</td>
<td>10</td>
<td>1516</td>
<td>−0.01</td>
<td>−0.03</td>
</tr>
<tr>
<td>Persian</td>
<td>13</td>
<td>89</td>
<td>−0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>Icelandic</td>
<td>15</td>
<td>27</td>
<td>−0.02</td>
<td>−0.38</td>
</tr>
<tr>
<td>English</td>
<td>12</td>
<td>527</td>
<td>−0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Urdu</td>
<td>17</td>
<td>123</td>
<td>−0.05</td>
<td>−0.38*</td>
</tr>
<tr>
<td>Catalan</td>
<td>23</td>
<td>280</td>
<td>−0.06</td>
<td>−0.37*</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>8</td>
<td>142</td>
<td>−0.06</td>
<td>−0.26*</td>
</tr>
<tr>
<td>Old Russ.</td>
<td>5</td>
<td>896</td>
<td>−0.06</td>
<td>−0.05</td>
</tr>
<tr>
<td>Korean</td>
<td>9</td>
<td>1421</td>
<td>−0.07*</td>
<td>−0.10*</td>
</tr>
<tr>
<td>Dutch</td>
<td>6</td>
<td>476</td>
<td>−0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>Galician</td>
<td>26</td>
<td>144</td>
<td>−0.09</td>
<td>−0.73*</td>
</tr>
<tr>
<td>Japanese</td>
<td>13</td>
<td>214</td>
<td>−0.13</td>
<td>−0.07</td>
</tr>
<tr>
<td>Latin</td>
<td>8</td>
<td>2256</td>
<td>−0.14*</td>
<td>−0.14*</td>
</tr>
<tr>
<td>Danish</td>
<td>11</td>
<td>105</td>
<td>−0.15</td>
<td>−0.01</td>
</tr>
<tr>
<td>Italian</td>
<td>14</td>
<td>670</td>
<td>−0.28*</td>
<td>−0.07</td>
</tr>
<tr>
<td>Hebrew</td>
<td>13</td>
<td>137</td>
<td>−0.29*</td>
<td>−0.22*</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>6</td>
<td>368</td>
<td>−0.48*</td>
<td>0.01</td>
</tr>
</tbody>
</table>
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.\textsuperscript{2,19-21} We consider this an intuitive measure: all other things being equal, sentences with shorter dependency lengths will be easier to process.\textsuperscript{2,22}

**Controlling for confounds**

Since the relationship between sentence length and summed dependency distances is not necessarily linear,\textsuperscript{7} 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.\textsuperscript{5} 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,\textsuperscript{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.\textsuperscript{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/.)](image)

![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.\textsuperscript{25})](image)

\footnote{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. We cannot exclude that there are inconsistencies in theoretical choices across languages, e.g., how prepositions or relativizers are attached. 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 or even morphophonetic aspects of the lexicon. 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.

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. 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.
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.

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

19. Futrell R, Mahowald K, Gibson E: Large-Scale Evidence of Dependency Length Minimization in 37 Languages. Proc Natl...


33. Juzeck TS, Bizzoni Y: Syntax-semantics interactions – seeking evidence from synchronic analysis of 38 languages: dataset repository (Version 1.0) [Data Set]. *Zenodo*. 2021, February 16. [Publisher Full Text]

In this paper, the authors test a very straightforward (at least in theory) and interesting assumption that there should be an inverse correlation between the complexity of syntax vs. semantics in language. They perform a systematic analysis of text in 38 different languages. They quantify semantic complexity as uni-gram frequency and syntactic complexity as dependency length (distance from the head). The results are neither confirming nor completely negating the predictions based on the literature and are very language-dependent.

What I liked:

All in all, this is a simple interesting research question that can (and should) be addressed from a corpus analysis perspective. It is nice to see the attempt of analyzing very different languages and trying to identify relationships between them. Moreover, the code used in this study is available and allows, as far as I can say, full replicability.

What could be improved:

I think this article needs a bit of extra work in order to be indexed. Unfortunately, at the moment, it is not really clear if we can rely on the results reported or if what we see is simply due to other strong confounding elements. Below are some comments:

- I find the statement that complex syntax calls for simple semantics and vice-versa a bit of an overgeneralization. In many domains (in particular web data), I would expect a simple-simple interaction instead.

- The authors correctly indicate that frequency is a good indicator of processing cost. However, processing cost does not exactly correspond to semantic complexity: such complexity is way more multifaceted and, I feel, not really captured by frequency alone.

- Moreover, focusing on the semantics of single words (treated as out of context entities) is probably an oversimplification: I understand that the inclusion of contextualized measures
would indirectly capture syntactic information too, but not including this information makes this measure probably too simple to capture semantic complexity.

- One possible confounding variable that, in my opinion, needs to be directly addressed in the experiment is the effect of genre/domain of the text used. At the moment the effect is only briefly described as a possible cause, but it should really be included in the statistical analysis. Very specific domains (e.g., legal text) follow specific styles that differ from all the rest and this can add a lot of "noise" to the data. This becomes even more of an issue once the study tries to compare different languages based on corpora that, as far as I understand, are not balanced in any systematic way.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computational linguistics, corpus analysis, cognitive science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
School of Foreign Studies, Anhui Jianzhu University, Hefei, China

This is an interesting and original work concerning the relation between syntax and semantics. But, there are some issues worthy of further discussion or clarification.

First of all, syntax facilitates semantics. This is true. But this does not mean that "writers and speakers will deliver difficult semantics via simpler syntax and simplify semantics in the presence of difficult syntax". In fact, the complex syntax is probably a result of pressure for communication of complex semantic message.

Secondly, it is very difficulty to quantitatively define semantic difficulty. The authors have used the sum of the log frequencies of a sentence's lemmas. This is more a measure of expectation, or, information. In fact, in conveying complex semantic message, it is important to keep the information content at a proper level at any time: it can neither be too high nor too low (uniform information density hypothesis). The relatively even distribution of information might be one of tasks of syntax. Zipf's law of word frequency distribution is probably one piece of evidence.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cognitive linguistics and dependency grammar

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com