West Germanic noun-noun compounds and the morphology-syntax trade-off

Pablo Mosteiro
Utrecht University
Utrecht, the Netherlands
p.mosteiro@uu.nl

Damián Blasi
Pompeu Fabra University
Barcelona, Spain
dblasi@fas.harvard.edu

Denis Paperno
Utrecht University
Utrecht, the Netherlands
d.paperno@uu.nl

Abstract

This paper examines the linguistic distinction between syntax and morphology, focusing on noun-noun compounds in three West Germanic languages (English, Dutch, and German). Previous studies using the Parallel Bible Corpus have found a trade-off between word order (syntax) and word structure (morphology), with languages optimizing information conveyance through these systems. Our research question is whether manipulating English noun-noun compounds to resemble Dutch and German constructions can reproduce the observed distance between these languages in the order-structure plane. We extend a word-pasting procedure to merge increasingly common noun-noun pairs in English Bible translations. After each merge, we estimate the information contained in word order and word structure using entropy calculations. Our results show that pasting noun-noun pairs reduces the difference between English and the other languages, suggesting that orthographic conventions defining word boundaries play a role in this distinction. However, the effect is not pronounced, and results are statistically inconclusive.

1 Introduction

The linguistic distinction between *syntax* and *morphology* is well-known and contentious (Tallman and Auderset, 2023; Crystal, 2010). Syntax is often understood as the study of word combinations into phrases and sentences, while morphology focuses on internal word processes. However, the boundary between these domains is blurred, and attempts to distinguish them often hinge on the complex notion of *wordhood* (Haspelmath, 2023). Some patterns in language, nonetheless, seem to support a morphology-syntax divide, with languages relying more on one or the other system.

Previous research using the Parallel Bible Corpus (PBC) (Mayer and Cysouw, 2014) found this trade-off across many languages, suggesting that languages optimize information conveyance through these systems (Koplenig et al., 2017). More recently, Mosteiro and Blasi (2025) postulated that the statistical order-structure trade-off observed could be ascribed to the conventional word boundaries used in the curation of the dataset.

We focus on a phenomenon straddling the border between morphology and syntax: noun-noun compounds in West Germanic languages. While Dutch and German form noun-noun compounds by creating a single orthographic word, English conventionally writes noun-noun compounds as separate words. For example, the expression *winter garden* is two words in English, but its equivalents *wintertuin* and *Wintergarten* in Dutch and German, respectively, are composed of one word. At the surface level, the English construction seems syntactic, while the Dutch and German constructions seems morphological. Noun-noun compounds have been the subject of extensive linguistic study (Sun and Harald Baayen, 2021; Gast, 2008; Berg, 2006).

Our research question is: **can we reproduce the distance between English on the one hand and Dutch and German on the other in the study by Koplenig et al. (2017) by manipulating English noun-noun compounds, so that they stick together as in Dutch and German?**

To answer this question, we reproduce the word-pasting experiment of Mosteiro and Blasi (2025), but this time we only allow word pairs to be merged if both of the words involved are nouns. Thus, we investigate whether the observed distance between English and Dutch/German in the order-structure plane is merely an artifact of orthographic conventions defining word boundaries.

2 Materials and Methods

We use a multilingual parallel corpus, on which we apply a word-pasting methodology (Mosteiro and Blasi, 2025) to paste common word-pairs to-

Proceedings of the The 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 15–22

gether, and then we compute the amount of information contained in word-order and word-structure using an entropy estimator based on Shannon's entropy (Shannon, 1948; Koplenig et al., 2017). [1]

2.1 Data

We use the Parallel Bible Corpus (Mayer and Cysouw, 2014), which comprises 2000 translations of the Bible in 1460 languages, covering over 40 language families worldwide[2]. Each translation is preprocessed with tokenization, Unicode normalization, and insertion of spaces between words and punctuation. As in Koplenig et al. (2017), we split translations into individual books, and focus on six New Testament books (Matthew, Mark, Luke, John, Acts, and Revelation). We do not lowercase the texts in the preprocessing step, and instead do this after part-of-speech tagging. We only consider bibles in Dutch, German, or English, as indicated by file names starting in `nld`, `deu`, or `eng`, respectively. Because contemporary German only became fully standardized in the 19$^{\text{th}}$ century, we exclude bibles translated before 1800. We only included bibles that contained at least 90% of all the indexed verses in each of the six books considered. Our final dataset consists of 23 bible translations in German, 29 in English, and 4 in Dutch. The list of bible translations can be found in Appendix A.

2.2 Part-of-speech tagging

We employ part-of-speech (POS) tagging to identify nouns in our datasets, utilizing SpaCy's `en_core_web_lg` model, version 3.8.0, for its large size and CPU-optimization. POS tagging is applied to English data only, while German and Dutch data are excluded due to the subsequent omission of the word-pasting algorithm on those languages. For each token, we extract the Universal POS (UPOS) tag to classify words as nouns (`NOUN`) or otherwise. After POS tagging, we lowercase all tokens as in Koplenig et al. (2017).

2.3 Entropy calculator

Following the work of Koplenig et al. (2017), we aim to estimate the amount of information carried by word order and word structure, which are proxies for syntax and morphology. To do this, we create three versions of each bible translation. One is the original text, named `orig`. The second version, named *shuffled*, is obtained by shuffling all tokens within each verse in the text. This effectively destroys word order. The third version, named *masked*, is obtained by replacing each word type in a book by a unique randomly generated character sequence of the same length. This effectively destroys word structure. An example is shown on Table 1. After applying these operations at the verse level, all verses in a book are concatenated in a shuffled manner, thus creating an original, a shuffled, and a masked version of each book. Each of these versions is fed into an entropy calculator (Koplenig et al., 2017), which returns the amount of information contained in each of these versions, in bits per unit character. This results in the quantities H^b_{original}, H^b_{order}, and $H^b_{\text{structure}}$, corresponding to the `orig`, `shuffled`, and `masked` versions of book b, respectively. We then compute the information contained in word order and word structure as:

$$D^b_{\text{order}} = H^b_{\text{order}} - H^b_{\text{orig}} \quad (1)$$

$$D^b_{\text{structure}} = H^b_{\text{structure}} - H^b_{\text{orig}} \quad (2)$$

2.4 Word pasting

We replicate and extend the word-pasting experiment from Mosteiro and Blasi (2025) on English Bible translations. For each book in each translation, we iteratively generate new versions by merging the most frequent noun-noun word pair into a compound. Following each merge, we create original, shuffled, and masked versions of each book and compute estimates of D^b_{order} and $D^b_{\text{structure}}$ using the entropy calculator introduced in Section 2.3. We do not paste proper nouns, as exploratory analysis showed limited pasting of proper nouns in German and Dutch[3]. This splitting methodology is applied solely to English translations.

2.5 Final pipeline

We take the 56 bible translations described in Section 2.1. Following previous work (Koplenig et al., 2017), we consider only six books of the New Testament. We thus arrive at 342 book-translation pairs. We split each of these into verses and create an original, a masked, and a shuffled version of it. We then paste the verses back together to obtain an original, a masked, and a shuffled version of each book-translation. We feed each of these into an entropy calculator to obtain the information in bits

[1] All our code can be found at `https://github.com/PabloMosUU/WordOrderBibles`.

[2] We use commit `9e66cf47f`. Newer versions contain even more translations.

[3] For example, English *Jesus Christ* is *Jezus Christus* in Dutch and *Jesus Christus* in German.

Version	Text
orig	immediately **they** left the boat and **their** father and followed him .
shuffled	followed boat him and the and father **their they** left . immediately
masked	aihuraovaha **phun** fafa luh avnn wso **octaa** otstsh wso tehreaed fed e

Table 1: Three versions of each verse of the bible are created before computing the entropy in bits per unit character of each book. The **boldface** merely highlights the effect, by showing that two words that are related in both their form and their meaning in the original text are mapped to completely different words in the masked text.

per character, then we compute two differences to obtain the information contained in word order and word structure for each book-translation. In the case of English, we expand this analysis by pasting noun-noun pairs iteratively from the most common to the least common and recomputing the word-order and word-structure information at each step. For each language, book, and number of merges[4], we average the values of D^b_{order} and $D^b_{\text{structure}}$, as in Koplenig et al. (2017).

3 Results

Figure 1 shows our main result. The red squares and green stars are the word-order and word-structure information for Dutch (nld) and German (deu), respectively. The blue dot labeled "eng-orig" corresponds to the average of the original English bible translations. The blue dot labeled "eng-nn-pasted" is the average value of word-order and word-structure information across English translations after all noun-noun pairs have been pasted together. For comparison, the cyan triangles are reproductions of Mosteiro and Blasi (2025), in which the first 100 and 200 most common word pairs have been pasted together, regardless of POS tag. The fit line shown is the one found by Koplenig et al. (2017) by fitting an inverse proportionality line on word-order and word-structure information across all languages in the PBC. The number of noun-noun merges required for each book to reach the point when no more noun-noun pairs can be pasted, averaged over translations, is shown on Table 2. In Appendix B we report the results of repeating the study using SpaCy's en_core_web_trf, a transformer model. Qualitatively we observed similar results.

[4]For German and Dutch, we do not do any merging, so we average values of D^b_{order} and $D^b_{\text{structure}}$ for each language and book.

Book	Max verses	Max NN merges
Acts	1007	34.5
John	879	23.7
Luke	1151	48.9
Mark	678	28.8
Matthew	1071	43.5
Revelation	404	28.4

Table 2: For each book considered, the maximum number of verses found across the available English translations, and the maximum number of noun-noun merges, averaged over translations.

4 Discussion

Figure 1 indicates that pasting noun-noun pairs together either leaves the English data point unchanged brings it closer to the Dutch and German data points. The effect is much smaller than when we paste all words regardless of POS tag. For Acts and John, no effect is observable altogether. Table 2 shows the maximum number of verses for each book in English. Note that Acts and John are neither the longest nor the shortest books, so there a priori no reason to believe that the effect should be smaller or negligible for those books.

Let $\Delta(D_{\text{order}})$ and $\Delta(D_{\text{structure}})$ be the differences between D_{order} and $D_{\text{structure}}$, respectively, before and after the noun-pasting procedure. Table 3 shows the values of $\Delta(D_{\text{order}})$ and $\Delta(D_{\text{structure}})$ across the various translations in English, together with the p-value for a paired permutation test to discard a null hypothesis in which both Δ is 0. From this table, we can conclude that only $\Delta(D_{\text{structure}})$ in Revelation is significantly different from 0.

It would be desirable to study longer corpora, not only because it would allow more noun-noun pairs to be pasted, but also because the entropy estimator we used converges to the entropy for long texts (Kontoyiannis et al., 1998). Convergence was checked in a previous study (Koplenig et al., 2017). Still, in future work we plan to evaluate

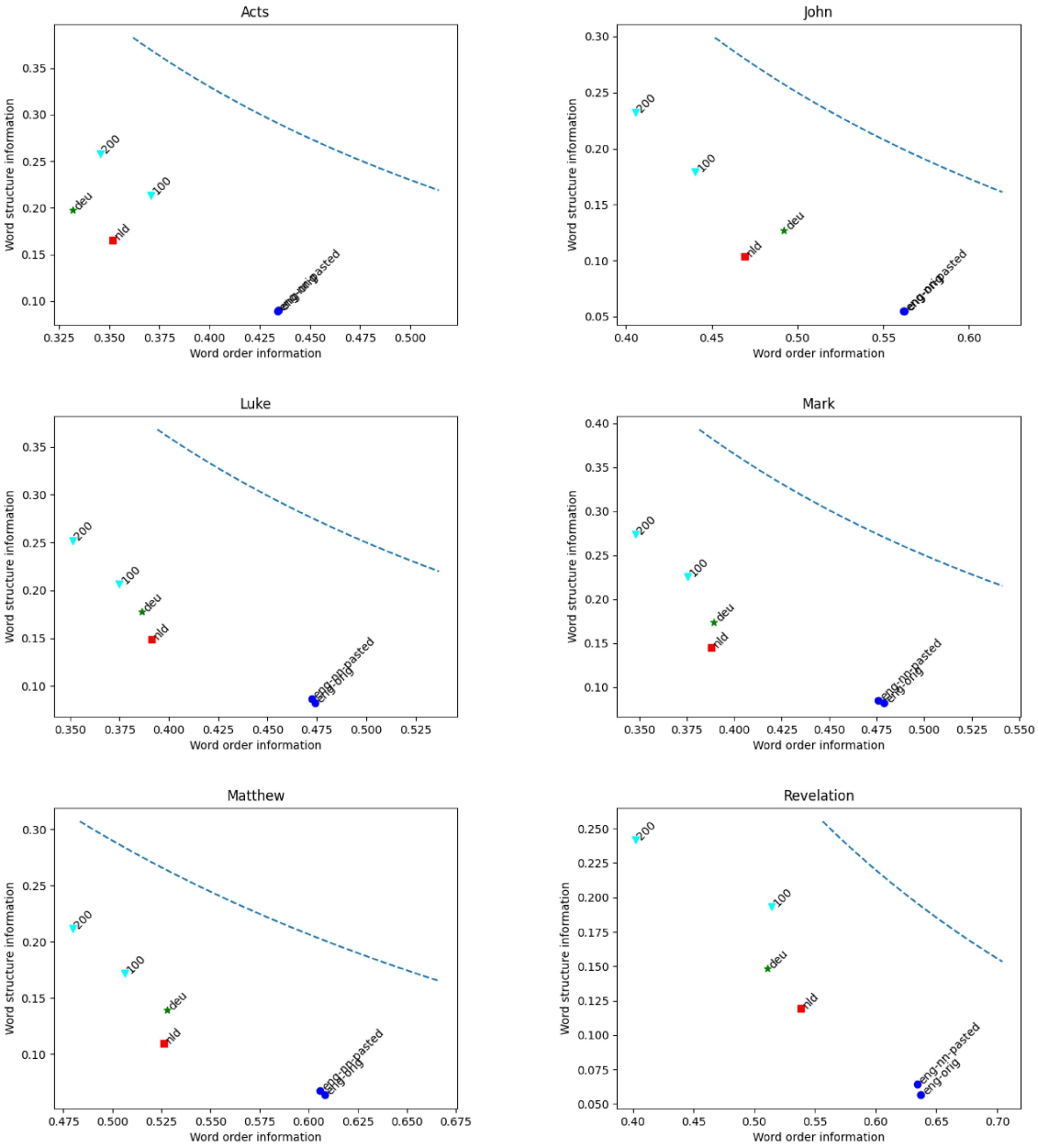

Figure 1: Word-structure versus word-order information for English, Dutch, and German. The green stars and red squares are German (deu) and Dutch (nld), respectively. The blue dots are English data, before (eng-orig) and after (eng-nn-pasted) pasting all noun-noun pairs in the book together. The cyan triangles are obtained by pasting the 100 and 200 most common word pairs regardless of POS tag. The dashed line is the best fit found by Koplenig et al. (2017) using all languages in the Parallel Bible Corpus.

these quantities over multiple books combined.

Although the evidence is not conclusive, there seems to be an indication that pasting noun-noun pairs together, which implicitly turns English noun-noun compounds from two words into one, brings the English word-order and word-structure information closer to the values for Dutch and German.

Future work will use word splitting (Mosteiro and Blasi, 2025) to split Dutch and German noun-noun compounds and evaluate whether the data points move closer to the English data points.

5 Conclusions

In conclusion, our study aimed to investigate whether the observed distance between English and Dutch/German in the order-structure plane, as reported by Koplenig et al. (2017), is merely an

Book	Metric	Δ	p
Acts	D_{order}	-0.00074	0.470
	$D_{\text{structure}}$	-0.00164	0.648
John	D_{order}	-0.00071	0.473
	$D_{\text{structure}}$	0.00002	0.497
Luke	D_{order}	-0.00155	0.438
	$D_{\text{structure}}$	0.00445	0.146
Mark	D_{order}	-0.00291	0.391
	$D_{\text{structure}}$	0.00334	0.234
Matthew	D_{order}	-0.00233	0.426
	$D_{\text{structure}}$	0.00358	0.139
Revelation	D_{order}	-0.00310	0.356
	$D_{\text{structure}}$	0.00816	0.006

Table 3: $\Delta(D_{\text{order}})$ and $\Delta(D_{\text{structure}})$ for each book as estimated by a paired permutation test in which each paired sample consists of D_{order} and $D_{\text{structure}}$ for a single English translation, before and after merging noun-noun pairs, for a total of 29 paired datapoints. We used 100000 resamples. Only the p-value for $D_{\text{structure}}$ for Revelation is less than 0.05, meaning that noun-noun merges caused a statistically significant increase in $D_{\text{structure}}$ for Revelation.

artifact of orthographic conventions defining word boundaries. By replicating and extending the word-pasting experiment from Mosteiro and Blasi (2025) on English Bible translations, we found that pasting increasingly common noun-noun pairs together reduces the difference between English and the other languages, suggesting that the distinction is at least partially due to this factor.

However, the effect was not as pronounced as observed in the original study due to a small number of noun-noun pairs present in the corpus (see Table 2), and the shift was not statistically significant. This small effect could also be caused by the fact that the words we pasted are less frequent than those pasted by Mosteiro and Blasi (2025), because we selected a subset of their words. Future work will check this effect on a bigger corpus with more noun-noun pairs.

Limitations

In this study we applied POS tagging at the individual verse level. Future work could check whether tagging entire books of the bible would increase POS-tagging performance.

We used commit `9e66cf47f` of the PBC for consistency with prior work. There might be additional bible translations in our languages of interest in more recent versions of the PBC.

Not all noun-noun clusters in English are compounded in their German or Dutch translations. But in our study we pasted all occurring pairs of nouns. A refinement of this work would check that all noun-noun pairs pasted are linguistically accurate, in the sense that their counterparts in German or Dutch would be compounds.

As for the linguistic question, we only considered one phenomenon in one language family, namely noun-noun compounds in West Germanic languages. It would be interesting to find another phenomenon occurring in another language family, to validate our methodology.

References

Thomas Berg. 2006. The internal structure of four-noun compounds in english and german. *Corpus Linguistics and Linguistic Theory*, 2(2):197–231.

David Crystal. 2010. *The Cambridge encyclopedia of language*. Cambridge University Press Cambridge.

Volker Gast. 2008. Verb-noun compounds in english and german. *Zeitschrift für Anglistik und Amerikanistik*, 56(3):269–282.

Martin Haspelmath. 2023. Defining the word. *Word*, 69(3):283–297.

Ioannis Kontoyiannis, Paul Algoet, Yuri Suhov, and Abraham Wyner. 1998. Nonparametric Entropy Estimation for Stationary Processesand Random Fields, with Applications to English Text. *Information Theory, IEEE Transactions on*, 44:1319 – 1327.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLOS ONE*, 12(3):1–25. Publisher: Public Library of Science.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pablo Mosteiro and Damián Blasi. 2025. Word boundaries and the morphology-syntax trade-off. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 86–93, Abu Dhabi, UAE. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Kun Sun and R. Harald Baayen. 2021. Hyphenation as a compounding technique in english. *Language Sciences*, 83:101326.

Book	Max verses	Max NN merges
Acts	1007	40.1
John	879	28.0
Luke	1151	54.8
Mark	678	35.0
Matthew	1071	50.3
Revelation	404	28.4

Table 4: Maximum number of noun-noun merges for each book, averaged across English translations. Compared to Table 2, this one was generated using SpaCy's transformer model `en_core_web_trf` as a POS tagger instead of `en_core_web_lg`. The **Max verses** column is unchanged because that is a property of the input texts.

Adam JR Tallman and Sandra Auderset. 2023. Measuring and assessing indeterminacy and variation in the morphology-syntax distinction. *Linguistic Typology*, 27(1):113–156.

A List of bibles used

Table 5 shows the file names of the bible translations used for this study.

B Transformer-based POS tagger

We repeated the entire analysis of this paper using SpaCy's `en_core_web_trf`, a transformer model, instead of `en_core_web_lg`. The results are shown on Figure 2 and Table 4. We note that the average number of noun-noun pairs is higher than with `en_core_web_lg`. This means that either the `en_core_web_lg` model incorrectly classifies nouns as other parts of speech, or that `en_core_web_trf` incorrectly selects non-nouns as nouns. However, the figure shows that the downstream results are slightly less significant than those found with `en_core_web_lg` (Figure 1).

Translation name	
deu-x-bible-albrecht.txt	deu-x-bible-elberfelder1871.txt
deu-x-bible-elberfelder1905.txt	deu-x-bible-freebible.txt
deu-x-bible-genfer2011.txt	deu-x-bible-greber.txt
deu-x-bible-gruenewalder.txt	deu-x-bible-interlinear.txt
deu-x-bible-konkordant.txt	deu-x-bible-lebendig.txt
deu-x-bible-luther1912.txt	deu-x-bible-luther2017.txt
deu-x-bible-meister.txt	deu-x-bible-menge.txt
deu-x-bible-neue.txt	deu-x-bible-newworld.txt
deu-x-bible-pattloch.txt	deu-x-bible-schlachter.txt
deu-x-bible-schlachter2000.txt	deu-x-bible-tafelbibel.txt
deu-x-bible-textbibel.txt	deu-x-bible-volxbibel.txt
deu-x-bible-zuercher.txt	eng-x-bible-amplified.txt
eng-x-bible-basic.txt	eng-x-bible-catholic.txt
eng-x-bible-clontz.txt	eng-x-bible-common.txt
eng-x-bible-darby.txt	eng-x-bible-diaglot.txt
eng-x-bible-easytoread.txt	eng-x-bible-etheridge.txt
eng-x-bible-godsword.txt	eng-x-bible-goodnews.txt
eng-x-bible-lexham.txt	eng-x-bible-literal.txt
eng-x-bible-majority.txt	eng-x-bible-modern.txt
eng-x-bible-montgomery.txt	eng-x-bible-new2007.txt
eng-x-bible-newcentury.txt	eng-x-bible-newinternational.txt
eng-x-bible-newliving.txt	eng-x-bible-newreaders.txt
eng-x-bible-newsimplified.txt	eng-x-bible-newworld1984.txt
eng-x-bible-newworld2013.txt	eng-x-bible-passion.txt
eng-x-bible-riverside.txt	eng-x-bible-treeoflife.txt
eng-x-bible-world.txt	eng-x-bible-worldwide.txt
nld-x-bible-1951.txt	nld-x-bible-2004.txt
nld-x-bible-2007.txt	nld-x-bible-newworld.txt

Table 5: Bible translations from the PBC that were used in the present study.

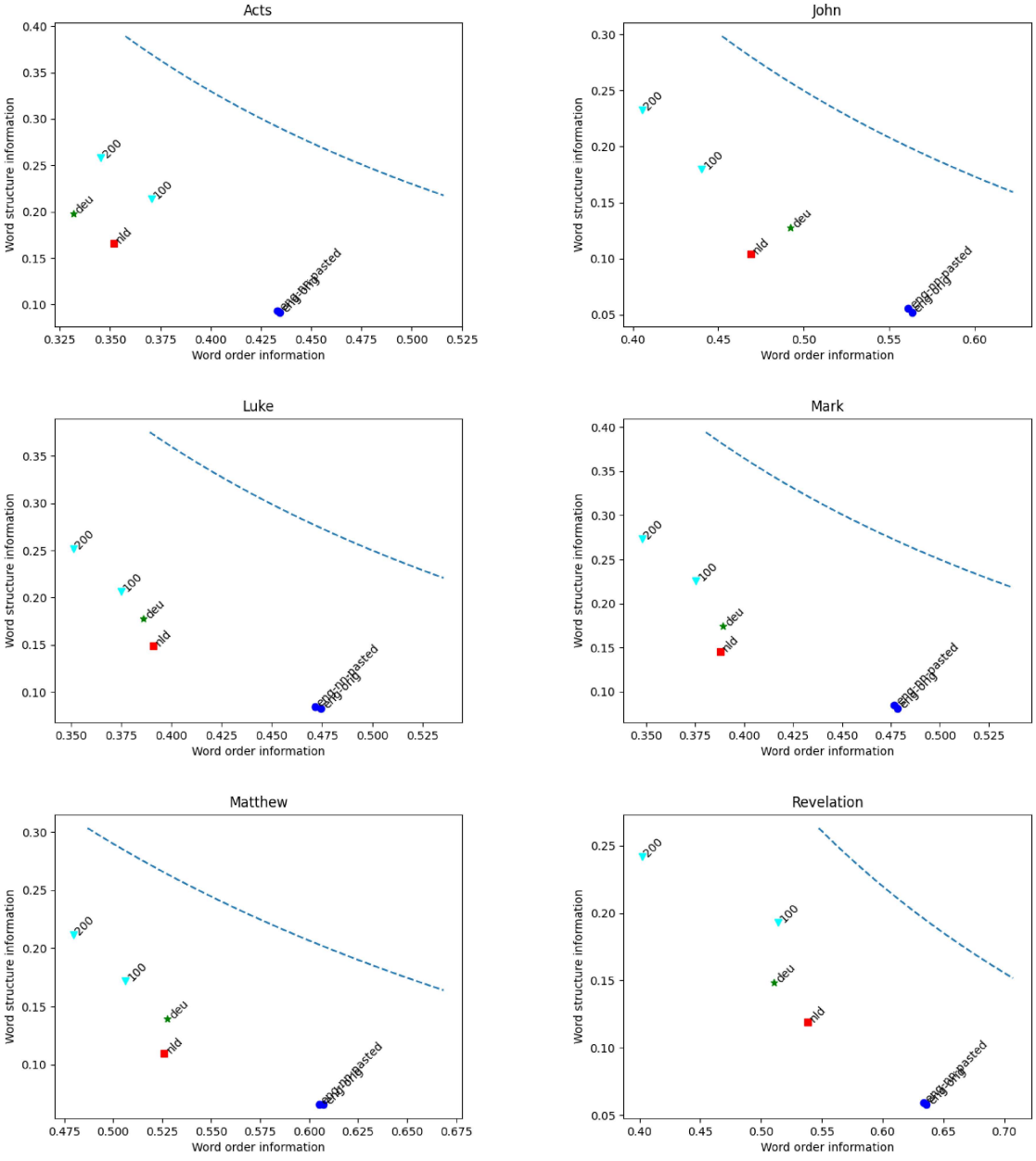

Figure 2: The same analysis as presented in Figure 1, this time using SpaCy's transformer model `en_core_web_trf` instead of `en_core_web_lg`.