An introductory statistical study of Hungarian word order

Dávid Márk Nemeskey[1,2]

[1]Department of Digital Humanities – Eötvös Loránd University
[2]National Laboratory for Digital Heritage
nemeskey.david@btk.elte.hu

Abstract Hungarian is often cited as a language with free word order. While this is not strictly true, the rules that govern the sentence structure are derived from pragmatics and are thus much more flexible than they are for analytical languages such as English. This paper presents an introductory statistical study into Hungarian word order. We report the order of verbal arguments in simple sentences in two corpora: the Hungarian Wikipedia and TrendMiner. An experimental method for ordering adjectives in noun phrases is also presented.

Keywords: word order, argument structure, order of adjectives

1 Introduction

Hungarian is frequently called a free word order language. It is true that, being an inflecting language with a rich case system, Hungarian does not need a set word order to distinguish between the main constituents of a clause (i.e. the verbal arguments). This poses a challenge for foreign learners of the language and influences the language acquisition process of Hungarian children as well (Nagyházi, 2013; Pléh, 1981).

It must be noted that most of the time the term *word order* comes up in the literature, it does not refer to the order of each individual word, but that of the *constituents* of the sentence; more specifically, the order of the verb and its arguments. The language has parts where the order of words is determined by syntactic rules, such as the structure of noun phrases (NPs; although see Section 4). These usually cause little confusion and will not be discussed in this paper.

It is not only learners of the language who are baffled by the seemingly random word order. Initially, linguists also had a hard time explaining it, and even advised forming sentences following German rules (Márton, 1805)[1]. It was Brassai (1852–53) (following earlier attempts by Táncsics (1833) and Fogarasi (1838)) who first suggested that constituents in the Hungarian sentence are ordered based on their discourse functions.

[1] Although this was probably only because the book was written to a German audience.

É. Kiss (1981, 1994) proved that the Hungarian sentence has an invariant structure, which marks constituents from a pragmatics point of view[2]. The main communicative roles, the *topic* (already known background information) and the *focus* (new information introduced by the sentence) occupy the two "slots" before the verb; the postverbal part serves no discourse function. The order of constituents within the topic and postverbal slots is free, while the focus may only host a single constituent.

The original discussion has since been extended to complex sentences (Kenesei, 1984b), exclamatives (Lipták, 2006) and discontinuous phrases (Barta et al., 2004). Behavior of constituents have been given further thought in both the preverbal (Puskás, 2000) and postverbal parts (Szalontai and Surányi, 2020). The word order of certain constituents, such as PPs, have also received attention (Dékány and Hegedűs, 2015). Competing theories also exist, such as Kenesei (1984a), but they argue against the particular grammatical framework used in É. Kiss (1981), not the general observations.

The interest in word order is not purely academic, however. Aside from the aforementioned impact on language acquisition, there are certain natural language processing (NLP) tasks where the order of sentence constituents is important. One example is dependency parsing, where determining the correct (or a valid) order of head and dependents can be a challenge. The problem also comes up in data generation or augmentation for machine learning in domains where training corpora are scarce or nonexistent.

Unfortunately, the studies above do not provide practical answers to these challenges, for two reasons. First, our NLP pipelines sorely lack any form of semantic, let alone pragmatic processing capabilities. Second, the studies are highly theoretic and are not data driven, and hence, cannot serve as bases for a machine learning system.

In this paper, we aim to conduct an initial statistical study into Hungarian word order. In the long run, we aim to create a framework for acquiring statistical data that could serve as the basis for the tasks mentioned above. Here, as a proof of concept, we focus on the methodology and examine a few simpler aspects of word order:

1. the order of verb arguments in simple sentences;
2. the tendency of oblique arguments to occupy the topic or focus position;
3. the order of adjectives in a noun phrase.

2 Methodology

2.1 Data

We run our experiments on two Hungarian corpora, which were selected to be markedly distinct in their style. We hypothesize that the word order statistics are not homogeneous across the language but depend on the style of the corpus.

[2] We do not aim at presenting the full timeline of the field in this paper. For a historical overview of theories on word order, the reader is referred to Nagyházi (2013)

Wikipedia (WP) The Hungarian Wikipedia[3] is a semi-edited corpus, created by volunteers. As an encyclopedia, most of the articles use formal, written language that aims to convey factual information in a neutral tone. For our study, we used the version available in the Hungarian Webcorpus 2.0 (Nemeskey, 2020).

TrendMiner (TM) The TrendMiner corpus is a collection of 1.9 million political Facebook comments, harvested over the course of 3 months in 2013 and 2014 (Miháltz et al., 2015). As expected of such material, the language is highly informal, often rude, and as close to spoken Hungarian as the medium allows. Compared to Wikipedia, it contains a high amount of questions, exclamatives and imperatives.

TrendMiner contains about 46M tokens in 4M sentences; Wikipedia is about 3 times larger at 175M tokens in 13.8M sentences.

Both corpora were processed with `emtsv` (Indig et al., 2019) using the morphological analyzer, lemmatizer and dependency parser modules. For Wikipedia, we only needed to run the latter as the rest of the annotations is already available in Webcorpus 2.0. TrendMiner also comes fully annotated, but it uses different tagsets and so had to be re-processed to make the two sources compatible.

2.2 Tooling

In order to quickly find sentences with a certain argument structure, we loaded the data into the treebank search tool `dep_search`[4] (Luotolahti et al., 2015). `dep_search` allows the user to index and query data in the CoNLL-U format[5] using a custom query language[6] . The language enables the specification of a dependency subgraph complete with morphological constraints that is matched against the indexed treebank. Matching sentences are returned in their original CoNLL-U format, with the single target token of the query marked as such.

We indexed and queried our two corpora separately to allow for comparison between the two. Since `dep_search` returns sentences as-is, we developed our own scripts to extract the argument structure from the query results. Both our code[7] and a patched version of `dep_search`[8] that fixes a few issues with the original are available on GitHub.

3 Argument structure

3.1 Preprocessing

In this initial study, we concentrate on the simplest sentences: namely,

[3] https://hu.wikipedia.org/wiki

[4] https://github.com/fginter/dep_search

[5] https://universaldependencies.org/format.html

[6] http://bionlp.utu.fi/searchexpressions-new.html

[7] https://github.com/DavidNemeskey/word_order

[8] https://github.com/DavidNemeskey/dep_search

1. the sentence is declarative;
2. it has a single finite verb as its ROOT;
3. the verb is not negated;
4. it has exactly one subject and object, each noun(phrase)s.

As Hungarian is a pro-drop language (certainly as far as subjects are concerned), we repeated our measurements also for the case when the SUBJ relation is not realized on the surface.

To satisfy the first condition, we simply dropped all sentences that did not end with a single period "." token. The second to fourth conditions were implemented by the dep_search query[9]

```
VERB&Mood=Ind&VerbForm=Fin
!< _ >SUBJ NOUN >OBJ NOUN !>OBL _ !>COORD _ !>NEG _
```

Due to a limitation of the dep_search query language, we could not narrow the query down for it to discard sentences containing multiple subjects and/or objects. Instead, these were filtered later from the final statistics. As a result, the percentage values in Tables 2, 3 and 5 do *not* sum to 100%; the remaining mass is accounted for by these more complicated sentences.

Corpus	No obliques (see 3.2)		With obliques (see 3.3)		Retained
	SUBJ-OBJ	OBJ	SUBJ-OBJ	OBJ	
Wikipedia	276 788	126 794	398 541	350 805	8.37%
TrendMiner	20 487	34 004	14 890	15 399	2.07%

Table 1. Number of sentences in the filtered collections

Table 1 reports the sizes of the filtered corpora. As shown in the last column, only a small fraction of the original corpora passed our filters; in fact, the number of sentences we ended up from TrendMiner is so low that the feasibility of a more in-depth study than the one presented below is questionable. The stylistic differences between the two corpora are also reflected in that the sentences in WP tend to be more complex (including a subject and/or obliques most of the time), while subject dropping is more frequent in TM.

3.2 The simple sentence

Table 2 lists the relative frequencies (in percentages) for all possible orderings of the VERB and its SUBJ and OBJ dependencies in our simple sentences. Looking at the top, we can see that the most frequent orderings are SVO and SOV, although TrendMiner clearly prefers the former and Wikipedia the latter. While the order of the rest of the orderings is the same for both corpora, the distributions

Structure			Counts	%
SUBJ	OBJ	VERB	77 673	28.06
SUBJ	VERB	OBJ	64 841	23.43
OBJ	SUBJ	VERB	58 851	21.26
OBJ	VERB	SUBJ	15 950	5.76
VERB	OBJ	SUBJ	1328	0.48
VERB	SUBJ	OBJ	992	0.36

Structure			Counts	%
SUBJ	VERB	OBJ	6901	33.68
SUBJ	OBJ	VERB	4571	22.31
OBJ	SUBJ	VERB	1837	8.97
OBJ	VERB	SUBJ	1636	7.99
VERB	OBJ	SUBJ	474	2.31
VERB	SUBJ	OBJ	375	1.83

Table 2. SUBJ-VERB-OBJ word order variations in WP (left) and TM (right)

are decidedly different: TM's has a "fatter" tail, whereas in WP, `VERB`-initial sentences are virtually nonexistent, and `OSV` is almost as prevalent as `SVO`.

Table 3 shows the results for the `SUBJ`-drop sentences. Again, the two corpora differ in their preferred orderings: Wikipedia prefers `OV` by a large margin, while TrendMiner contains the two possible ordering of `VERB` and `OBJ` in almost equal measure, slightly preferring `VO`. While the order of the first two rows are the same as in Table 2 with respect to `VERB` and `OBJ`, their distribution is not: the difference between `SVO` and `SOV` was bigger in TM than in WP, while here it is the opposite. We are going to show, however, that the disparity is only skin-deep.

Structure			Counts	%
OBJ	VERB		82 905	65.39
VERB	OBJ		35 807	28.24
OBJ	OBJ	VERB	3525	2.78
OBJ	VERB	OBJ	3135	2.47
VERB	OBJ	OBJ	1191	0.94

Structure			Counts	%
VERB	OBJ		17 309	50.90
OBJ	VERB		15 376	45.22
OBJ	VERB	OBJ	763	2.24
OBJ	OBJ	VERB	356	1.05
VERB	OBJ	OBJ	183	0.54

Table 3. VERB-OBJ word order variations in WP (left) and TM (right)

We can interpret all sentences with no overt subject as having a virtual pronoun subject that was dropped from the realized sentence. (Although explicitly realizing the pronoun changes the pragmatics of the sentence. Example (1) from Wikipedia illustrates this: the two sentences are semantically equivalent, but (1b) emphasizes the overt subject, while the covert one in (1a) is neutral.) If we drop the `SUBJ` relation from all structures in Table 2 and merge the (now) identical rows, we end up with 55.08% `OBJ VERB` to 24.27% `VERB OBJ` in Wikipedia and 39.27% to 37.82% in TrendMiner, which is much closer to the relative ratios in Table 3.

[9] For an explanation of the query, see the documentation linked above.

(1) a. *Bevallottan Robert Bresson stílusát követi.*
admittedly Robert Bresson 's style.ACC follow.3SG

b. *Ő bevallottan Robert Bresson stílusát követi.*
he admittedly Robert Bresson 's style.ACC follow.3SG

'He admits to following Robert Bresson's style.'

The bottom half of Table 3, which lists statistics for sentences with two objects, paints the same picture. In TrendMiner, the "neutral" (although rather unusual) `OBJ VERB OBJ` is the most frequent ordering, with the other two being much rarer, while in Wikipedia, `OBJ OBJ VERB` not only outnumbers `VERB OBJ OBJ` 3 to 1, but is also the most numerous variant overall.

These results confirm our earlier hypothesis that the two corpora would differ in their word orders. Yet there might be other factors at play that may explain the differences. The two corpora might simply contain different sets of verbs. In addition to that, the verb distributions might also exhibit a bias not present in the language in general. In the following, we are investigating these hypotheses.

`VERB, OBJ`				`SUBJ, VERB, OBJ`			
Wikipedia		TrendMiner		Wikipedia		TrendMiner	
Verb	Percent	Verb	Percent	Verb	Percent	Verb	Percent
nevez	5.77	*kíván*	7.44	*tartalmaz*	3.15	ad	2.53
kap	2.00	*köszön*	4.00	ad	2.36	kap	2.22
ír	1.89	lát	2.21	jelent	2.15	jelent	1.41
ad	1.78	kap	1.89	kap	2.02	tesz	1.21
tart	1.61	*kiván*	1.78	alkot	1.56	lát	1.19
végez	1.50	kér	1.74	mutat	1.56	hoz	1.13
használ	1.38	vár	1.63	vesz	1.12	jár	0.84
talál	1.20	ad	1.59	okoz	1.07	okoz	0.79
készít	1.13	hoz	1.41	hoz	1.03	fizet	0.71
vezet	1.11	szeret	1.18	képez	0.99	vesz	0.69
épít	1.09	olvas	1.16	biztosít	0.99	mutat	0.68
hív	1.09	tesz	1.11	használ	0.99	tűr	0.67
tekint	0.96	ismer	1.08	tart	0.95	megszáll	0.66

Table 4. Relative frequency of the top words in both corpora for the `VERB OBJ` and `SUBJ VERB OBJ` frames

Table 4 lists the most frequent verbs found in the two corpora for the `VERB OBJ` and `SUBJ VERB OBJ` frames. What is clear at first glance is that the two frames have widely different verb usage characteristics. `VERB OBJ` is dominated by a few words in both corpora: "*nevez*" (*call*) in WP and "*kíván*" (*wish*, also misspelled as "*kiván*")[10] and "*köszön*" (*greet*) in TM. These verbs seem to be

[10] With many occurrences in expressions like "*Jó reggelt kívánok.*" (*Good morning.*)

idiosyncratic of each corpus, as they are outliers to the linear functions that can be fitted to the frequencies of the rest of the words on the lists (with a R^2 of 0.94 in both cases). Consequently, removing them from the data might improve the accuracy of our statistics.

The updated statistics are listed in Table 5. Comparing them to Table 3, we can see that removing *nevez* from WP has left the relative frequencies of the structures unchanged. On the other hand, the frame distribution in TM has become closer to that WP, proving the existance of the bias hypothesized above and demonstrating the importance of filtering idiosyncratic word (usage)s from the data.

Structure	Counts	%
OBJ VERB	76 881	60.64
VERB OBJ	34 517	27.22

Structure	Counts	%
VERB OBJ	15 836	46.56
OBJ VERB	12 355	36.34

Table 5. VERB-OBJ word order variations in WP (left) and TM (right) without corpus-specific words

With the bias (largely) out of the way, we can turn our attention to the vocabularies of the two datasets and see if they explain the differences in the argument structure distributions. Table 5 shows a mixed picture. In the top 13 verbs for VERB OBJ, there are only two words ("*kap*" (*receive*) and "*ad*" (*give*)) common to both list; however, for SUBJ VERB OBJ, about half of the words (6) fall in the intersection. In the longer tail (up to the top 50 verbs), the ratio of words common to both corpora are around 33% for VERB OBJ and 50% for SUBJ VERB OBJ. How much of the difference can be explained by this is left for future study.

3.3 Obliques

For oblique arguments, a similar study could be conducted as for SUBJ and OBJ. However, the resulting table would be hard to interpret and would tell very little, as OBL is an umbrella relation that covers any of 15 noun cases (nominative and accusative excluded). Instead, we opted to compare oblique argument types based on their inclination to move to a preverbal position.

As explained in Section 1, the preverbal positions in a Hungarian sentence are taken by the pragmatically most important parts of the information structure: the topic and the focus. Unfortunately, without pragmatical analysis, we cannot determine which role the argument takes, so we simply check if the argument precedes the verb.

Looking at the results in Table 6, we can see that half of the cases (namely All, Cau, Ill, Ins, Sub, Ter, Tra) have very similar (within 3%) relative frequencies in both corpora, while some of them (Abl, Del, Ela) differ by more

Case	Suffix	Wikipedia		TrendMiner	
		Frequency	Preverbal	Frequency	Preverbal
Abl	-tŐl	5029	55.72%	419	38.90%
Ade	-nÁl	2414	70.30%	114	64.91%
All	-hOz	4333	46.83%	188	44.68%
Cau	-ért	1355	43.03%	244	45.08%
Dat	-nAk	6513	55.43%	253	42.59%
Del	-rŐl	6227	62.15%	294	36.39%
Ela	-bŐl	6642	64.57%	652	46.17%
Ess	-ként	7225	73.74%	221	82.35%
Ill	-bA	7019	48.17%	690	44.93%
Ine	-bAn	65 659	71.54%	2234	62.67%
Ins	-vAl	25 504	63.30%	1421	62.14%
Sub	-rA	21 226	53.24%	1715	56.15%
Sup	-n	44 772	75.19%	1457	65.41%
Tem	-kor	1404	85.04%	49	71.43%
Ter	-ig	2757	77.73%	142	74.65%
Tra	-vÁ	1178	63.67%	41	60.98%
Nom		193 458	89.91%	9218	87.31%
Acc	-t	125 399	58.28%	4561	43.20%

Table 6. The total number of oblique arguments with specific cases and the percentage they occur in a preverbal position. Nominative and accusative included for reference. Only sentences with a single SUBJ, OBJ and OBL were taken into account.

than 15%. On the whole, Wikipedia seems to employ more preverbal obliques; the difference is only 4% on the type level, but significantly larger (67.15% vs 57.65%) on the token level. The difference makes sense intuitively, as the goal of Wikipedia is to convey factual information and new information enters the sentence in the focus position, which, as we have seen, immediately precedes the verb in Hungarian.

Why certain oblique arguments prefer the preverbal position more than others is an interesting topic for future work. Looking at the directional cases, TO-type cases (All, Ill, Sub) are around 50% in both corpura, while AT-types (Ade, Ine, Sup) are the “most preverbal” with 75-75% in WP and 60–65% in TM. FROM types (Abl, Ela, Del) tend to be between the two in WP, but have the lowest percentages of all cases in TM. Again, the explanation of these tendencies requires further research.

4 Adjective order

It has long been theoreticized that the (neutral) order of attributive adjectives in an NP is subject to restrictions based on their semantic category. Dékány (2021) mentions the variations below:

a. value > dimensions > physical property > speed > human propensity > age > colour (Dixon, 1982)

b. cardinal > ordinal > quality > size > shape > colour > nationality (Cinque, 1994)
c. ordinal > cardinal > size > length > height > speed > width > weight > temperature > wetness > age > shape > colour > origin > material (Scott, 2002)
d. subjective comment > evidential > size > length > height > speed > depth > width > weight > temperature > ?wetness > age > shape > colour > nationality/origin > material (Laenzlinger, 2005)

, and posits that these restrictions are also in place for Hungarian.

Similarly to prior theoretical work on word order in general, the category lists above make sense intuitively, but are not readily usable for NLP. One reason is that there is no mapping defined to actual words. The lists are also incomplete, so even if the mappings existed, many adjectives would be left without a category. In this section, we investigate if similar, but more complete, ordering rules can be extracted from a corpus in a data-driven fashion.

4.1 Categories

We started out by collecting all noun phrases with at least two adjectives in them:

```
NOUN >ATT@L ADJ&Case=Nom >ATT@L ADJ&Case=Nom
```

A custom script was then used to retain only those where the adjectives immediately precede the noun. This allowed us to concentrate on the core issue without having to deal with numerals (`NUM`) and coordinating conjunctions (`CCONJ`) for now. We then deleted all tokens other than the adjectives and added the dummy tokens `DET` and `NOUN` to the beginning and the end of each adjective group, respectively. This gave us a list of 1.6M virtual NPs from Wikipedia; we decided against using TrendMiner in this experiment so as to have a cleaner dataset. The total number of adjective types is 132 795.

A word2vec (Mikolov et al., 2013) embedding was trained on the concatenated virtual NPs using gensim[11] (Řehůřek and Sojka, 2010). We used a CBOW model of 25 dimensions with a window size of only 2, so that adjectives that belong to separate NPs do not see each other over the dummy tokens. To allow us to actually evaluate our results, we only trained embeddings for adjectives that occurred in at least 4000 NPs; 113 in total.

As a last step, we ran the k-means algorithm (MacQueen et al., 1967) in Scikit-learn[12] (Pedregosa et al., 2011) on the normalized vectors to obtain our candidate categories. Table 7 shows the results of a clustering with $k = 8$.

As can be seen, the clusters are remarkably consistent, aside from the few odd words (*in italics*). It is also not difficult to assign a name to the clusters, apart from the last one, which plays the role of the "kitchen sink". Some of the

[11] `https://github.com/RaRe-Technologies/gensim`

[12] `https://github.com/scikit-learn/scikit-learn`

Cluster	**Words**
Nationality	magyar amerikai német francia brit angol *saját* olasz japán orosz
Importance	nagy legnagyobb jelentős teljes ismert fontos kisebb önálló fő
Position	első egyik *című* egyes különböző második utolsó további elleni másik
Cultural	katolikus katonai politikai gazdasági zenei televíziós tudományos
Affiliation	nemzetközi nemzeti válogatott állami egyetemi városi műszaki
Age	új római királyi kis régi *jános* helyi erdélyi századi egykori modern
Sports	évi olimpiai nyári női legjobb országos európai budapesti téli bajnoki
Misc.	*nevű álló magyarországi lévő hagyományos található egész ún.*

Table 7. Adjective categories generated by an 8-way k-means clustering

clusters, such as Nationality and Position, clearly correspond to well-established categories. Given the capacity of the model and the very limited input data, we conjecture that what the embedding learned are indeed the adjective's category (i.e. role or place in the NP) and not its full semantic representation. However, this requires formal validation.

Curiously, most categories do not appear at all, which is due to the frequency limit we employed to make our data interpretable. When no such limit is enforced, various other categories manifest, such as color, year, ordinals, etc. These usually cluster together really well, but we have yet to find the right cluster number or algorithm to make sure that each cluster is meaningful.

4.2 Ordering

To discover if an ordering exists for our clusters, we converted our virtual NPs to a graph. Each token was mapped to a vertex, and a directed edge $e : v_1 \rightarrow v_2$ was put between vertices v_1 and v_2 iff the token corresponding to v_1 directly precedes the one corresponding to v_2 in any of the NPs. The weight of e equals to the number of NPs in which the connection was found.

Ideally, this graph would be a directed acyclic graph (DAG) with `DET` as the source and `NOUN` as the sink. In reality, the graph is full of cycles: for 132 797 vertices it has about 15k cycles of length 2 and 610k of length 3. An example of a 2-cycle is *idegen írású* , from the NPs[13]

(2) a. *latin betűs írású idegen nyelveknek*
Latin letter writing foreign language.PLUR.DAT

‘of foreign languages written in Latin script’

[13] As one reviewer rightly pointed out, having both “*idegen*” and “*írású*” on the same level in (2b) is a parsing error. Unfortunately, the dependency parser is far from perfect and it affects our results as well. Another class of errors we discovered with it is that sometimes the morphological and the dependency labels contradict, e.g. a `[/N][Acc]` word gets the `SUBJ` label.

b. *idegen* *írású* *alakok*
foreign writing form.PLUR

‘forms written in foreign writing (style)’

We took this “raw” graph and replaced each vertex with the cluster to which the associated token belongs; vertices not part of any of the 8 clusters above were removed, along with the edges connected to them. This resulted in a much simpler graph with 8 nodes. Unfortunately, this graph still contains cycles, so we iterated through the cycles and dropped the edge with the least weight from each until we were left with a DAG (all graph processing steps were done in NetworkX[14] (Hagberg et al., 2008)).

Topologically sorting the final graph (and disregarding *Misc.*) yielded the order *Position > Importance > Age > Nationality > Sports > Affiliation > Cultural.* Where the categories are comparable, our list matches the theories above. For instance, *Position > Importance > Age > Nationality* roughly corresponds to *ordinal > size > age > origin* in (Scott, 2002).

Hungarian	**English**
harmadik európai irodalmi	third European ... of literature
újabb híres	another famous
jelentős római	significant Roman
másik olasz állami	another Italian state
saját legjobb	own best
japán olimpiai központi	*Japanese Olympic central*
kisebb lengyel téli kereskedelmi	smaller Polish winter commercial

Table 8. Adjective sequences generated randomly from our categories

Table 8 presents a few adjective combinations generated randomly according to our ordering. All but one of the examples feel valid. The error comes from the *Sports > Affiliation* pair in our ordering, which is probably not entirely consistent. For instance, the sequence in question should be *központi olimpiai*, but the ordering is correct for *országos egyetemi.* We leave the task of finding out whether better or more fine-grained clustering could solve such problems for future work.

5 Conclusion and future work

In this paper, we have conducted an initial, “proof of concept” statistical study into various aspects of Hungarian word order. We suggested a basic machinery for acquiring frequency data on word order variations from a corpus. We have

[14] `https://networkx.org/`

shown that word order is dependent on the type of text (formal or informal) and the corpus. Our results show that the study of word order should not be confined to the field of theoretical linguistics. Our method could be used to collect statistics for e.g. text generation in a domain with no preexisting textual data, or data augmentation for machine learning.

We have devised an experimental method based on clustering of word embeddings for determining the order of adjectives in noun phrases. The method shows promising results, but the coverage and the quality of the clustering needs improvement.

There are several open avenues for further research. In this paper, we only considered very simple sentences with a single verbal predicate; future work should broaden the focus on one hand and produce more fine-grained statistics on the other. In particular, individual properties of verbs (such as the presence of preverbs, definiteness, or its semantics, i.e. position in an embedding space) might have a significant effect on the word order. Another possible future direction is linking our data to other fields of NLP such as language modeling or verb frame databases (e.g. Mazsola (Sass, 2018)), which would enable us to evaluate the impact of verbal constructions on word order.

We also intend to try and bridge the gap between our statistical approach and the theoretical work on the field. This would allow us to experimentally validate some of the theories or even to incorporate some of their predictions to our NLP toolchains.

Acknowledgments

The author would like to thank the reviewers for their insightful comments.

Bibliography

Barta, Cs., Dormeyer, R., Fischer, I.: Word order and discontinuities in a dependency grammar for Hungarian. In: Proceedings of the 2nd Conf. on Hungarian Computational Linguistics (MSZNY), Szeged Hungary, Juhasz Nyomda. pp. 19–27 (2004)

Brassai, S.: Tapogatódzások a magyar nyelv körül. Pesti Napló (1852–53)

Cinque, G.: On the evidence for partial N-movement in the romance DP. In: Cinque, G., Koster, J., Pollock, J.Y., Rizzi, L., Zanuttini, R. (eds.) Paths towards universal grammar. Studies in honor of Richard S. Kayne, pp. 85–110. Georgetown University Press, Washington, DC (1994)

Dékány, É.: The Hungarian Nominal Functional Sequence. Springer Nature (2021)

Dékány, É., Hegedűs, V.: Word order variation in Hungarian PPs. Approaches to Hungarian 14, 95–120 (2015)

Dixon, R.M.: Where have All the Adjectives Gone?: And Other Essays in Semantics and Syntax. De Gruyter Mouton (1982)

É. Kiss, K.: Structural relations in Hungarian a ‘free’, word order language. Linguistic Inquiry 12, 185–213 (1981)

É. Kiss, K.: Sentence structure and word order. In: The syntactic structure of Hungarian, pp. 1–90. Brill (1994)

Fogarasi, J.: Euréka! Atheneum II 13, 193–198; 16, 241–249; 19, 289–297 (1838)

Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using networkx. In: Varoquaux, G., Vaught, T., Millman, J. (eds.) Proceedings of the 7th Python in Science Conference. pp. 11–15 (2008)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., Makrai, M.: One format to rule them all – the `emtsv` pipeline for Hungarian. In: The 13th Linguistic Annotation Workshop (8 2019)

Kenesei, I.: On the logic of word order in Hungarian. In: Abraham, W., de Mey, S. (eds.) Topic, Focus, and Configurationality. Benjamins (1984a)

Kenesei, I.: Word order in Hungarian complex sentences. Linguistic Inquiry 15(2), 328–342 (1984b)

Laenzlinger, C.: French adjective ordering: Perspectives on DP-internal movement types. Lingua 115(5), 645–689 (2005)

Lipták, A.: Word order in Hungarian exclamatives. Acta Linguistica Hungarica 53(4), 343–391 (2006), http://www.jstor.org/stable/26190105

Luotolahti, J., Kanerva, J., Pyysalo, S., Ginter, F.: SETS: Scalable and efficient tree search in dependency graphs. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 51–55. Denver, Colorado (2015), https://aclanthology.org/N15-3011

MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)

Miháltz, M., Váradi, T., Csertő, I., Fülöp, É., Pólya, T.: Beyond sentiment: Social psychological analysis of political facebook comments in hungary. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015). ACL (2015)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013), https://bit.ly/39HikH8

Márton, J.: Ungarische Grammatik, wodurch der Deutsche die ungarische Sprache richtig erlernen kann. Wien (1805)

Nagyházi, B.: Az egyszerű mondat szórendjének egy lehetséges tanítási modellje a magyar mint idegen nyelv oktatásában. Ph.D. thesis, Pécsi Tudományegyetem (2013)

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

Pléh, Cs.: The role of word order in the sentence interpretation of Hungarian children (1981)

Puskás, G.: Word Order in Hungarian: The syntax of Ā-positions. Linguistik Aktuell/Linguistics Today, John Benjamins Publishing Company (2000), https://books.google.hu/books?id=aZY9AAAAQBAJ

Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (5 2010), http://is.muni.cz/publication/884893/en

Sass, B.: Mazsola-mindenkinek. In: Vincze, V. (ed.) XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). Szegedi Tudományegyetem Informatikai Tanszékcsoport (2018)

Scott, G.J.: Stacked adjectival modification and the structure of nominal phrases. Functional structure in DP and IP: The cartography of syntactic structures 1, 91–120 (2002)

Szalontai, Á., Surányi, B.: Word order effects of givenness in Hungarian. In: Hegedűs, V., Vogel, I. (eds.) Approaches to Hungarian: Volume 16: Papers from the 2017 Budapest Conference, pp. 138–163 (2020)

Táncsics, M.: Nyelvészet. Pest (1833)