EÖTVÖS LORÁND TUDOMÁNYEGYETEM
Bölcsészettudományi Kar

DIPLOMAMUNKA

Névszói kötőhangzók variabilitásának korpuszalapú vizsgálata

Corpus-based analysis of the variability of linking vowels in nouns and adjectives

Témavezető:
Rebrus Péter, Ph.D
tudományos főmunkatárs

Készítette:
Lévai Dániel
Elméleti nyelvészet M.A.
Számítógépes nyelvészet
és neurolingvisztika

2020

Contents

1 Introduction

Phonological hesitation of suffixes in Hungarian has been studied for a long time. Traditional linguistics treats this matter as a binary phenomenon – a suffix either vacillates or not, and traditional theories do not explain the degree of vacillation. We aim to explain the degree of hesitation based on a frequentivist approach (Bybee and Hopper, 2001).

It is a well known fact that the category of nouns and adjectives overlap in Hungarian (Elekfi, 2000; Moravcsik, 2001), and while we have a basic idea of what constitutes the former or the latter, rigorous classification of words can be challenging. There are markers which can help, such as the existence of comparative or plural forms, but it is easy to see that these kinds of explanations can lead to context-dependent classes.

There are other differentiating markers of which the speakers are not consciously aware of, such as the vowel height harmony in case marker suffixes. Nouns prefer mid vowels, adjectives prefer low vowels, but these are only tendencies, since there is no clear distinction between categories.

In the present thesis, we take an analogy and frequency-based approach to quantify these tendencies. Considering the behavior of a typical noun and adjective, we can hypothesize that if an 'adjectival' word is used as a 'nouny' word, be in a syntactic or semantic way, it shall behave as a noun in a phonological sense, and vice versa. In this manner, with the help of the tools of computational linguistics, we analyze the interaction of phonology, syntax and semantics and attempt to show the interconnectivity between these linguistic modules through this particular phenomenon.

In section 2, we introduce the vowel harmony in Hungarian and the lowering/non-lowering to be analyzed later. In section 3, we describe our methodology to extract the vacillating forms from the corpus, then in section 4, we present our hypotheses on the graduality of variation. sections 5 and 6, presents evidence for the hypothesis, and afterwards, section 7, summarizes the results.

2 Vowel harmony in Hungarian

This section will give an overview on the system of Hungarian vowels, the types of harmony, and the problems in the domain of Hungarian vowel harmony.

2.1 Hungarian vowels

There are numerous papers on the vowel harmony in Hungarian, however, this section follows the notation of Törkenczy (2011). The Hungarian vowel harmony is rich in features: there is long distance effect, neutral vowel transparency, variation of harmony, alternating and invariant suffixes, antiharmonic roots, backness and roundness harmony. The system of Hungarian vowels is given in table 1.

	front		back	
	unrounded	rounded	unrounded	rounded
high	i <i>, iː <í>	y <ü>, yː <ű>		u <u>, uː <ú>
mid	eː <é>	ø <ö>, øː <ő>		o <o>, oː <ó>
low	ɛ <e>		aː <á>	ɒ <a>

Table 1: The phonological classification of the Standard Hungarian vowels. The characters appearing in the angled brackets show the corresponding letter in the Hungarian orthography.

As shown in table 1, Hungarian has a rich vowel system: there are 7 different short-long pairs of vowels.

There is backness (palatal) and roundness (labial) harmony in Hungarian. Both are controlled by the stem, i.e. the harmonic properties of the stem determine the harmonic properties of the affixes. The direction of harmony is left-to-right, and the (last) root forms a harmony domain with the succeeding affixes, e.g.: *vas-pöröly-ök-nek* 'iron-sledgehammer-PL-DAT'. The *vas* 'iron' is back, low, unrounded, while *pöröly* 'sledgehammer' is front, mid, rounded, but the latter is the last stem, thus the affixes belong to the harmony domain of *pöröly*, taking the corresponding *-ök-nek* form, instead of *-ak-nak* or *-ok-nak*. There are many interacting phenomena in the Hungarian harmony, but due to the time and length constraints, we are only focusing on backness harmony in this thesis. On other kinds of harmony, Törkenczy (2011) gives a good overview on the matter.

Backness harmony requires that vowels should agree in backness within the harmony domain based on the following system:

Back (B)	u, uː, o, oː, ɒ, aː
Front rounded (Fr)	y, yː, ø, øː
Neutral: front unrounded (N)	i, iː, ɛ, eː

Table 2: Backness harmony domains in Hungarian. Note that neutral vowels may behave in a transparent or opaque way.

2.2 Linking vowels

In Hungarian, the suffixes are mostly consonantal, but in most cases, there are phonotactic restrictions which prohibit many types of consonant clusters. Moreover, lexical traits of word forms plays also a role in the presence and/or the quality of the linking vowels. The following examples are based on Kálmán, Rebrus, and Törkenczy (2012), while prioritizing information relevant to this thesis.

The traditional definition of linking vowels depend on morphological segmentation, i.e.: A linking vowel is a vowel that appears in certain word forms at the boundary of a stem and a suffix, and which does not appear in some other word forms containing the same suffix (but a different stem).

For example:

	stem	ACC	PL	linking vowel presence	gloss
a,	hal	hal-**a**-t	hal-**a**-k	in ACC and in PL	'fish'
b,	lap	lap-**o**-t	lap-**o**-k	in ACC and in PL	'sheet'
c,	dal	dal-t	dal-**o**-k	only in PL	'song'
d,	kocsi	kocsi-t	kocsi-k	no linking vowel	'car'

Table 3: Some examples of words containing linking vowels (written in boldface).

In section 2.2, we can see the linking vowels for some words. The linking vowels, as per the definition, occur at morpheme boundaries, like in case a, the stem is *hal*, the accusative suffix is *-t*, and the linking vowel *-a-* is between these two morphemes. The presence of linking vowels and phonotactic motivation correlates, yet there are examples for each of the 4 possibilities (Kálmán, Rebrus, and Törkenczy, 2012, p. 26).

There is a similar phenomenon concerning the quality of the linking vowels. The quality is determined mostly by the stem and the suffix, however, other factors also do affect the vowel quality, such as context, part-of-speech, dependency relation and

semantics. We will examine only the nominal stems in this thesis.

group	stem	linking vowel		gloss
	height	low-mid PL	low-mid ACC	
1.	hal	hal**a**k	hal**a**t	'fish'
1.	lassú	lassú**a**k	lassút	'slow'
1.	őz	őz**e**k	őz**e**t	'roe'
1.	fűz	fűz**e**k	fűzt	'willow'
2.	pap	pap**o**k	pap**o**t	'priest'
2.	dal	dal**o**k	dalt	'song'
2.	adó	adók	adót	'tax'
2.	tök	tök**ö**k	tök**ö**t	'pumpkin'
2.	gőz	gőz**ö**k	gőzt	'steam'
2.	tető	tetők	tetőt	'roof'

Table 4: Stems behaving as group 1. are the **lowering stems**, while group 2. is called **non-lowering stems**.

As seen in table 4, there are two groups of nominals according to the quality of the linking vowels they receive: group 1. is the lowering stems, group 2. is the non-lowering stems. The linking vowel agrees in backness with the stem and the suffix, only the vowel height changes. In group 1., the plural suffix takes the forms *-**a**k*, *-**e**k* (low), while in group 2., it takes the forms *-**o**k*, *-**ö**k*, *-k* (mid). The accusative suffix shares (almost) the same linking vowels: *-at*, *-et*, *-t* (low) in group 1., *-ot*, *-öt*, *-t* (mid) in group 2.

It can also be seen from table 4 that it is the vowel quality of the stem which defines the vowel backness, while the vowel height is seemingly controlled by some hidden lexical feature of the stem in the case of PL and ACC.

2.3 Variation

It was shown in table 4 that the backness groups are established, and it is only the vowel height which keeps on changing. The possible vowel pairs with the same backness are the following[1]:

- Plural, nominative (PL)

[1]Hungarian orthography for the respective phonemes are written in angle brackets

1. -back, low-mid: ɒ-o <a-o>, *hangos-Vk* 'loud-PL'
2. -back, low-none: ɒ-∅ <a-∅>, *alvó-Vk* 'sleeping-PL'
3. +back, low-mid: ɛ-ø <e-ö>, *pöttyös-Vk* 'polka dotted-PL'
4. +back, low-none: ɛ-∅ <e-∅>, *fekvő-Vk* 'laying-PL'

- Accusative, singular (ACC)

1. -back, low-mid: ɒ-o <a-o>, *passzív-Vt* 'passive-ACC'
2. -back, low-none: ɒ-∅ <a-∅>, *hangos-Vt* 'loud-ACC'
3. +back, low-mid: ɛ-ø <e-ö>, *pörkölt-Vt* 'meat stew-ACC'
4. +back, low-none: ɛ-∅ <e-∅>, *pöttyös-Vt* 'polka dotted-ACC'

There are 4 different kinds of lowering variation we are analyzing in this thesis, these are presented above. We can notice that the plural and accusative suffixes behave differently, but have the same vowel pairs. Hungarian phonotactics does not allow linking vowel for the accusative suffix if the stems ends with vowel, but it permits to not use linking vowel after coronal sonorant consonant *fájl-t* 'file-ACC', *var-t* 'scar-ACC', *hangos-t* 'loud-ACC'. In plural, linking vowel is always necessary after consonant, and is optional after vowel.

In the later sections, we will refer to both plural-lowering and accusative-lowering stems as lowering stems, since these are both governed by the same lowering mechanism.

2.4 Vacillating suffixes

Nouns and adjectives behave differently with respect to linking vowel lowering (Siptár and Törkenczy, 2001, p. 227). There is a fixed set of nouns which undergo lowering, the lowering is not productive, recent nouns cannot undergo lowering, for example: *hal-**a**k* 'fish-PL', *ár-**a**k* 'price-PL', *haj-**a**k* 'hair-PL', but *baj-**o**k* 'trouble-ACC', *sör-**ö**k* 'beer-PL', *fájl-**o**k* 'file-PL'.

In the case of adjectives, the tendency is reverse: most adjectives prefer lowering, although there is also a small number which are non-lowering. Some non-lowering stems are: *nagyok* 'big-PL', *gazdagok* 'rich-PL', *vakok* 'blind-PL', but the typical adjectives are lowering, such as *pirosak* 'red-PL', *finomak* 'delicious-PL', *hűvösek* 'cool-PL'.

Greek and Latinate loan adjectives, on the other hand, usually show high degree of variation: *okkult-ak*, *okkult-ok* 'occult-PL', *konkáv-ak*, *konkáv-ok* 'concave-PL', *morbid-ak*, *morbid-ok* 'morbid-PL'.

2.5 Occasional nominalization

This section presents the main hypothesis of the thesis. In the previous sections, we mentioned that adjectives and nouns behave differently when it comes to lowering the linking vowel. Adjectives tend to lower the linking vowel, nouns tend not to. But what if we have a language where these two groups are not separated by any obvious marker or suffix?

Hungarian is exactly like that - there is no distinct boundary between these two groups (Elekfi, 2000; Moravcsik, 2001). The language tends to treat adjectives as attributes and nouns as things or entities, but adjectives can be used in positions where only nouns can appear - in which they behave as a noun, referring to the (previously established) entities having that attribute. Consider the following example:

(1) Vettem két-féle almát, **pirosat** és **zöldet**
Buy-PAST-1.SG two-types apple-ACC **red-ACC** and **green-ACC**
'I bought two kind of apples: **red ones** and **green ones**.'

(2) A **pirosak** finomak voltak, a **zöldek** pedig nem.
the **red-PL** delicious-PL be-PAST-3.PL the **green-PL** but no
'The **red ones** were delicious, but the **green ones** weren't.'

In example 1, we can see that the adjectives red and green agree in case with the object of the sentence, the apple. In Hungarian, usually, there is no agreement of modifier and head in NP, but there are certain constructions in which there is. In example 2, however, the 'red' and 'green' refer to a previously mentioned group of entities (apples), thus can replace it in subject position – in the sentence 'The **red ones** were delicious', the *red ones* refer to the apples from the previous example.

The phenomenon where adjectives can behave as nouns in certain contexts is called **occasional nominalization**. This can happen in almost any case, and nominalized adjectives behave the same way as nouns.

3 Variation in practice

3.1 Corpus

The corpus on which we conducted our measurements is the prepublished version of the Webcorpus 2 (Nemeskey, 2020). It is based on the Common Crawl webcorpus, which is a collection of pages downloaded each month from the Internet. The corpus consists of documents, i.e. web pages, with HTML tags removed, in plain text format, and it is deduplicated on document and paragraph level, thus practically every page and paragraph is unique in the corpus. The version of the corpus we are using measures around 10 billion word tokens and 27 million word types.

The corpus can be downloaded from the web page of the Human Language Technology group of SZTAKI (direct link).

3.2 Extracting vacillating stems

Due to the size of the corpus, we were restricted to using simple searches, string comparisons, and we needed to parallelize. To conduct the experiments, we used the `python`[2] programming language. We used the tokenized, morphologically annotated and disambiguated version of the corpus, and created simple descriptive statistics for each word: the number of occurrences, the different tags they have received, and the different lemmata for the word form.

We only searched for the plural forms, since the sentences containing these words had to be dependency parsed, costing a lot of processor time. As for the accusatives, we needed to find the singular accusative form based on the stem of the plural, which is an easy task.

First, we have to find pairs or triplets of words having the same morphological analysis and the same lemma but different word forms, but it has proven to be unreliable and slow due to the number of root variations produced by emMorph. We had more success with a simpler, more straightforward method. We took the list of words, and filtered those which have at least 10 occurrences in the whole corpus. We marked the vowel of the suffix for each word and constructed three rules:

[2] https://en.wikipedia.org/wiki/Python_(programming_language)

variation class	penult.	check for	example	gloss
a-o variation	a	o	*túlsúlyosVk*	overweight-PL
a-∅ variation	a	∅	*állóVk*	standing-PL
e-ö variation	e	ö	*szőrösVk*	hairy-PL
e-∅ variation	e	∅	*fekvőVk*	laying-PL

Table 5: Only four rules and we acquire the vacillating words.

As table 5 shows, we have three rules, let us see the `a-o` rule for example. We cycle through our list of word forms from the corpus, we check if the penultimate letter is `a`, and if it is, we check if there is a word in the corpus with the same letters as the first word, but the penultimate letter changed to `o`, for example, when the word is *Csokon**a**y*, we check for *Csokon**o**y* (which can appear in the corpus if someone mistyped it, however, later steps will filter these out). The a-∅ and e-∅ rules find the words ending with the present participle suffix *-ó/-ő*, and we filtered out the *-ő* ending due to the large number of pairs, halving the number of word pairs.

However, this naive method introduces a lot of false positives, like **árnyékat-árnyékot* 'shadow-ACC', *weboldalakat-*weboldalakt* 'webpage-PL-ACC' (stop consonant clusters are forbidden at the end of a word), *hallak-hallok* 'fish-dwelling/hear-1SG-2SG.ACC, hear-1SG' (only the first one could be noun, though highly improbable), and *hullámat-hullámot* 'corpse-1SG.POSS-ACC/wave-ACC' (different roots).

To reduce the number of these flaws, we introduced a number of checks to the algorithm:

1. Must be in the same case
2. Must have the same lemma
3. Must not have a stop cluster at the end of the word
4. Must be similarly frequent, that is, the rarer word of the pair must have a frequency of at least 0.01 times that of the more frequent word
5. Each word from the pair must have a frequency of at least 10
6. Must not contain symbols (`!?%_-`)

These additional checks reduced to number of pairs from 4313 to 3689. and the number of tokens from 35.366.019 to 5.596.849, saving us a lot of time when

dependency parsing the containing sentences. One additional step we added is that we excluded sentences longer than 40 words – there were some (erroneously) very long sentences, but that is usual with web-based corpora.

3.3 Grouping stems

We divided the word pairs into 3 groups based on their phonological forms and their frequency. There are derivational suffixes which induce variation in a word form. By far the two most frequent derivational suffixes are the present participle suffix -ó, -ő (*áll-ó* 'stand-ing', 'tör-ő' 'crack-ing') and the adjectivizer suffix *-s* (*rutin-os* 'experienced', *szőr-ös* 'hairy').

From the two variants of the present participle suffixes, the *-ő* does not partake in backness variation – words ending in *-ő* cannot be suffixed with -ö (*fekvő-ek*, *fekvő-k*, **fekvő-ök*, 'laying-PL'), and e is already low, thus the *e-∅* is not lowering. The *-ó*, however, does induce lowering. After *-ó*, *-o* cannot occur, but *-a* and *-∅* can (*álló-ak*, *álló-k*, **álló-ok* 'standing-PL') in case of the plural suffix. The a-∅ is a lowering pair. The accusative suffix *-t* cannot take linking vowel with this derivational suffix however, (*álló-t*, **álló-at*, **álló-ot* 'standing-ACC'). This group counts 1799 pairs and has a cumulative frequency of 2072668.

The derivational suffix *-s* behaves differently than the present participle *-ó*. Words with this suffix can receive both the mid and low variants of the plural suffix (*rutinos-ak*, *rutinos-ok*, **rutinos-k* 'experienced-PL' *szőrös-ek*, *szőrös-ök*, **szőrös-k* 'hairy-PL'), and linking vowel is mandatory for the plural suffix in this position. In accusative case, the suffix *-t* can occur without linking vowel (*rutinos-at*, *rutinos-t*, **rutinos-ot* 'experienced-ACC', *szőrös-et*, *szőrös-t*, **szőrös-öt* 'hairy-ACC'), however, the linking vowel cannot take mid height. This group counts 848 pairs and has a cumulative frequency of 594513

The third group are the leftovers. There are pairs ending with vowels -ú, -i (*hosszú* 'long', *földalatti* 'underground'), and the consonant stems are *-t* (mainly past participle), *-n, -r, -sz, -v*, (*vasalt* 'ironed', *profán* 'profane', *bíbor* 'purple', *gonosz* 'evil', *pozitív* 'positive'), but nearly every consonant can appear in word ending position. This group counts 137 pairs and has a cumulative frequency of 144121.

group	id	PL	ACC	pairs	freq
pr.part. -ó	o	a-∅	∅	1799	2072668
adj. -s	s	a-o, e-ö	a-∅, e-∅	848	594513
misc.	m	a-o, e-ö, a-∅, e-∅	a-∅, e-∅, ∅, ∅	137	144121

Table 6: The 3 distinct groups of endings, based on phonology and convenience.

4 Hypotheses

Our hypothesis revolves around the interaction of the occasional nominalization, mentioned in section 2.5 and the different lowering characteristics of the nouns and adjectives, mentioned in section 2.4. We have already seen that there is a sharp categorical difference between nouns and adjectives: the former prefers not to lower the linking vowel, the latter prefers lowering the linking vowel. However, if a true adjective undergoes occasional nominalization, these effects clash: from one side, it should lower the linking vowel, since it is an adjective, but on the other side, it should not, since it is in the typical position of a noun, and nouns do not undergo lowering. Regarding the occasional nominalization, we only consider the prototypical cases[3] of the lowering effect: the accusative suffix and the plural suffix.

We propose two different approaches to test the lowering: syntactic and semantic test.

4.1 Syntactic hypothesis

First, we discuss the syntactic hypothesis: table 7 summarizes the syntactic positions versus the cases in Hungarian.

The syntactic hypothesis is the following: If we count the ratio of lowering for a certain stem[4] in predicative position, plural, nominative case (R_1), in NP-head position, plural, nominative case (R_2), in NP-head position, singular, accusative case (R_3), the ratio of lowering will be the highest in the first, followed by the second, followed by the third case.

Table 7 shows the previously mentioned cases in table form. The syntactic

[3]Other suffixes do undergo lowering, but these behave either the same way as the accusative or the plural

[4]It is really important to aggregate only by stem, not to create cumulative statistics for a group of stems.

position	plural	accusative
predicative	low : non-low R_1	×
NP-head	low : non-low R_2	low : non-low R_3

Table 7: Different positions for nominals in Hungarian. Note that accusative forms cannot appear in predicative position, predicative is marked by the nominative case in Hungarian. R under the pairs marks the ratio of lowering.

hypothesis can also be formulated as the following if we use the information shown in table 7: in the case of vacillating stems, there is a general pattern regarding the ratio of lowering per stem: $R_1 > R_2 > R_3$ holds generally true for every stem.

To create an example, let us consider the stem *mosolygós* 'smiley', which has a variation of *-ak*/*-ok* in plural and *-at*/*-t* in accusative. The hypothesis states that in predicative position, the stem prefers lowering (*mosolygós-ak* instead of *mosolygós-ok*) more than it does in NP-head position (so the lowering form is rarer in this position), and the stem prefers lowering more in NP-head position than it does in object position (*mosolygós-at* instead of *mosolygós-t*).

To test this hypothesis, we use the built-in dependency parser (emDep) of e-magyar. We will analyze the syntactic positions of the lowering and the non-lowering forms and aggregate by stem. See more in section 5.

4.2 Semantic hypothesis

The second hypothesis refers to the semantics. We suspect that there is a distinction in meaning between the lowering and the non-lowering form of each stem.

Our hypothesis is the following:

Since the true adjectives are always lowering and the true nouns are always non-lowering, the lowering form for each stem is more similar in meaning to the true adjectives, while the non-lowering forms are more similar in meaning to the true nouns. We cannot simply measure how noun-like of adjective-like these forms are easily, hence we measure how coherent the lowering and non-lowering groups are. If the lowering forms are similar in meaning to other lowering forms more than to non-lowering forms, we can say that indeed, the lowering forms have some kind of common attribute, and if the lowering and non-lowering forms are systematically

separated in semantic similarity, we prove that there is a semantic difference between the lowering and non-lowering forms. As a result of the pairs being different in only nouniness and adjectiveness, we could conclude that the semantic difference is caused by that.

Let us give an example for what we expect to happen, let us take the stem *vallásos-ak* 'religious-PL', with plural variation of *-ak/-ok*. The form *vallásos-ak* will be close to other lowering forms, such as *babonás-ak* 'superstitious-PL', *cinikus-ak* 'cynical-PL', and the non-lowering form *vallásos-ok* will be more similar to non-lowering forms, such as *jobbos-ok* 'rightist-PL', *szkeptikus-ok* 'skeptics-PL'[5].

To test this hypothesis, we use the modern neural embedding techniques (Mikolov et al., 2013a; Mikolov et al., 2013b). We create a high-dimensional representation for each word form (type), then we conduct similarity measurements in that high-dimensional space. See more in section 6.

[5]These examples are extracted by the model we used, and are the mildest ones from the original word's neighborhood.

5 Syntax

5.1 Dependency grammar

The two most influential grammatical theories in modern linguistics are the phrase structure grammar and the dependency grammar. The phrase structure grammar (PSG) is a term coined by Chomsky (1957), and every relation in this grammar adheres to the constituency relation, thus these are also known as constituency grammars. In this approach, clause is divided into a subject (noun phrase) and predicate (verb phrase). This binary division creates one-to-one-or-more correspondence between the nodes in the tree structure and the words in the sentence. Theories developed from phrase structure grammar are government and binding (Chomsky, 1981), head-driven phrase structure grammar (Pollard and Sag, 1994) or lexical-functional grammar (Ford, Bresnan, and Kaplan, 1982).

In contrast to PSG, dependency relations are based on head and dependent relations. The correspondence is one-to-one: for every element of the sentence, there is only one node in the tree structure, thus dependency trees are always minimal and of the same size for a sentence. Traditionally, the edges in a dependency tree are marked with a predetermined inventory of primitive syntactic functions, e.g. subject, object, oblique, determiner, attribute, predicative, ...

The main advantage to dependency grammar is the easier handling of languages with free word order (such as Hungarian), eliminating the need of transformations, moving, or commanding. (Jurafsky and Martin, 2009)[6][7]

Let us take the sentence *Megettem a burgeremet pénteken.* 'I ate my burger on Friday'. This sentence consists of 3 elements: *megettem* 'PART-eat-PAST-1SG', *a burgeremet* 'burger-POSS.1SG-ACC', *pénteken* 'friday-ON'. But due to the free word order in Hungarian, these 3 elements can appear in any[8] order[9]: *megettem a burgeremet pénteken*, *megettem pénteken a burgeremet*, *a burgeremet megettem pénteken*, *a burgeremet pénteken megettem*, *pénteken a burgeremet megettem*, *pénteken megettem a burgeremet*. A phrase structure grammar would need two rules for the place of the location (temporal) adjectival: one for when the object is before a such adver-

[6] Since the 2009 edition, a new one is being written by the same authors, featuring a section on dependency relation, previously not present in the book

[7] https://web.stanford.edu/~jurafsky/slp3/

[8] note that there is a slight difference in topic and focus between these sentences

[9] n variable elements can appear in $n!$ order, that is n times n-1 times n-2 times ... times 2 times 1, thus a sentence with 3 variable elements can have 6 different orderings

bial, and one for after, whereas a dependency grammar encodes this information on the edges of the syntactic tree, eliminating the need for word order-dependent rules. An additional advantage of the dependency grammars is that the head-dependent relations provide direct information on the semantic relationship between the predicates and their arguments, while constituent-based approaches have to be further parsed to distill semantic information.

To illustrate the differences on a specific case, we will take a simple English sentence from Jurafsky and Martin (2009): *I prefer the morning flight through Denver*, shown in fig. 1. As we can see, the dependency tree is more compact, however, it really does only encode the head-dependent relation, while the phrase structure tree is more abundant with information. To mark the type of relation on dependency trees, we can name the edges. Moreover, it is a common practice to illustrate the relations with arrows over the words, shown in fig. 2, allowing us space-efficient dependency trees.

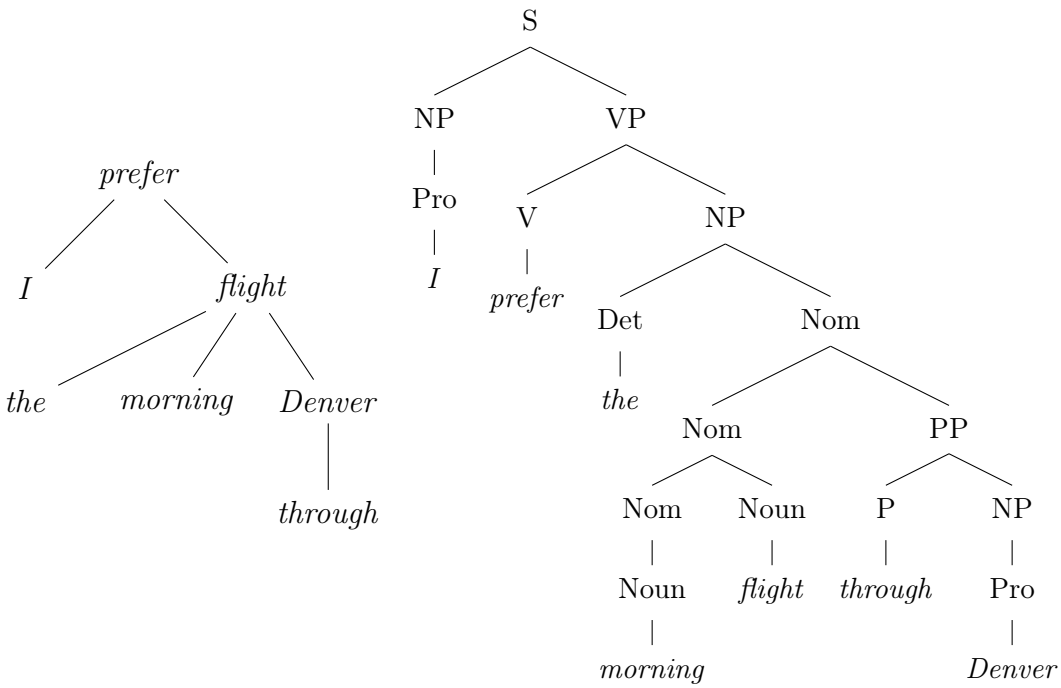

Figure 1: Dependency grammar on the left, phrase structure grammar on the right

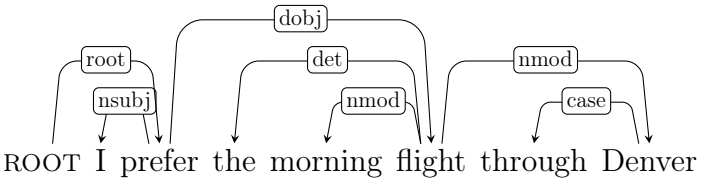

Figure 2: The dependency relations of the example

5.2 Eisner algorithm

There are 4 main algorithmic approaches to dependency parsing: dynamic programming (Eisner, 1996), graph algorithms (McDonald et al., 2005)[10], constraint satisfaction (Karlsson, 1990), and greedy, deterministic parsing (Nivre, 2008)[11].

The first approach is the dynamic programming-based methods. The idea is similar to that of the constituent parsing (CKY) – some rules establish a tree during the use of language, we only have to search for the steps to reproduce the same structure. Graph algorithms are a natural choice to dependency parse a sentence since a dependency tree is a directed, acyclic graph (tree, in graph theoretical sense). The basic idea is that we can score arcs in a fully connected graph, and once they are scored, the Chu–Liu/Edmond's (Chu and Liu, 1965; Edmonds, 1967) algorithm can be used to find the maximum scoring spanning tree in the graph. Another, nowadays less popular method is constraint satisfaction. In this approach, we prune the arcs in a fully connected graph by using constraints, i.e. eliminating the ones that do not satisfy the hard constraints. The third main approach is greedy, deterministic parsing. Here, the main idea is that the algorithms takes the words of the sentence left-to-right, and at each step, the oracle decides between a few simple operations: put the word in the stack, draw an edge in some direction, or pop the stack. The oracle is a machine learning algorithm, pretrained on treebanks. This family of algorithms is also called shift-reduce algorithms.

The approach that e-magyar uses by the Bohnet parser is a combination of the maximum spanning tree algorithm and Eisner algorithm.

The Eisner algorithm searches for the highest scoring set of arcs. The algorithm is recursive, it accumulates the sub-problems and solves the overall problem by composing them. Using dynamic programming, the algorithm does not reparse the sub-trees, it looks them up, and it looks up all possible sub-parses.

[10]MSTParser
[11]MaltParser

Definition 5.2.1. Scoring function

Let $G = (V, A)$ be an arbitrary graph over some sentence S, let $\{w_i \in S\}$ be the words of the sentence, let r be the relation between two words, let ψ be an oracle, telling the score of a single arc. The score of G is the sum of the scores of the relations in the graph, that is, the sum of individual (head, relation, modifier) triplets' scores.

$$\text{score}(G) = \sum_{(w_i, r, w_j) \in A} \psi(w_i, r, w_j)$$

We are not detailing the inner workings of the algorithm, but there great educational youtube videos which show the steps of the algorithm.

5.3 Methodology

To process the corpus, we used the state-of-the-art `e-magyar`[12] (Indig et al., 2019; Váradi et al., 2018) text processing system. The `e-magyar` is a modular toolchain, with separate modules for tokenizing, morphological analysis, lemmatizing, part-of-speech tagging, constituency and dependency parsing, NP-chunking and named entity recognition. The newer version of the `e-magyar` toolchain is named emtsv[13]. It can be used as a Python library or as a Docker image[14], and it features an easy-to-use command line interface and an even easier-to-use REST API web frontend.

The dependency parser we used is the built-in parser of `e-magyar`, called `magyarlanc` (Zsibrita, Vincze, and Farkas, 2013), which is based on the Mate (Bohnet, 2010) parser. The `magyarlanc`, after being built into an `e-magyar` module, received the `emDep` name.

The output of `emDep` can look intimidating at first, but it is easy to understand. Note that we are not looking at the entirety of the output, only at the important columns. The sample input is the sentence: *A **pirosak** elvitték a **magasakat** autóval Abonyba.* 'The **red ones** took the **tall ones** with car to Abony'. On fig. 3, we can see the dependency relations in visual form, and table 8 represents the relations in table form.

In Farkas, Vincze, and Schmid (2012), they elaborate on the methodology of detecting the empty copulae. The Szeged Dependency Treebank (Vincze et al., 2010) on which they have trained the algorithm introduces a virtual node for the copula. The algorithm lacked compatibility with these virtual nodes, thus they

[12] https://e-magyar.hu/en

[13] https://github.com/dlt-rilmta/emtsv

[14] https://hub.docker.com/r/mtaril/emtsv

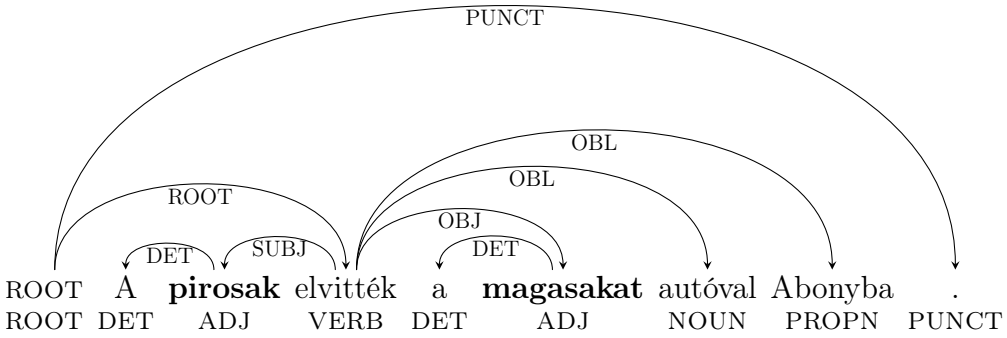

Figure 3: The dependency relations of the example

form	lemma	xpostag	upostag	id	deprel	head
A	a	[/Det|Art.Def]	DET	1	DET	2
pirosak	piros	[/Adj][Pl][Nom]	ADJ	2	SUBJ	3
elvitték	elvisz	[/V][Pst.Def.3Pl]	VERB	3	ROOT	0
a	a	[/Det|Art.Def]	DET	4	DET	5
magasakat	magas	[/Adj][Pl][Acc]	ADJ	5	OBJ	3
autóval	autó	[/N][Ins]	NOUN	6	OBL	3
Abonyba	Abony	[/N][Subl]	PROPN	7	OBL	3
.	.	[Punct]	PUNCT	8	PUNCT	0

Table 8: The dependency relations of the example as a table

have removed these nodes and all of their dependents were attached to the head of the original virtual node. This can be the reason why there is a large number of uncertainty when predicting the dependency relation in the not-subject cases, although they achieved 87.2% when predicting labels and relations simultaneously. The sentence demonstrating this example is: *A piros autók* ***szépek*** 'the red cars are **pretty**', shown on fig. 4, table 9. Note that the ROOT node points to an adjective, indicating empty copula.

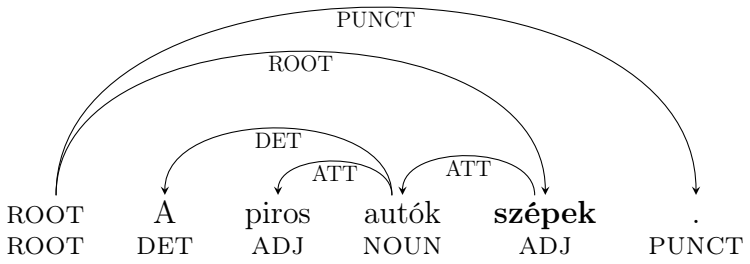

Figure 4: While the dependency parser is not perfect, it can show us whether the target word is the subject of the sentence or not.

form	lemma	xpostag	upostag	id	deprel	head
A	a	[/Det|Art.Def]	DET	1	DET	3
piros	piros	[/Adj][Nom]	ADJ	2	ATT	3
autók	autó	[/N][Pl][Nom]	NOUN	3	ATT	4
szépek	szép	[/Adj][Pl][Nom]	ADJ	4	ROOT	0
.	.	[Punct]	PUNCT	5	PUNCT	0

Table 9: In this case, it did manage to find that the adjective is indeed in the place of the empty copula

5.4 Results

We calculated the ratios R_1, R_2 for every group, and also R_3 for the s-stem group, as explained in section 4.1. In the next paragraphs, we analyze the groups, starting with the o-stems. In figs. 5 to 7, a point marks a stem, ρ marks the Spearman's rank correlation coefficient between the axes, and u marks the ratio of points where $x \geq y$, that is, the ratio of points under the dashed line.

In fig. 6, we can see the plot of the ratio of lowering in predicative position versus in subject position. The colors are consistent between the three little plots, meaning that the rightmost and center plot are filtered plots of the leftmost one.

The leftmost plot shows us a tendency that stems prefer to lower in predicative position more than to lower in subject position, although with a rather weak evidence of 0.633 ratio. Spearman's ρ is 0.631, showing a strong evidence of rank correlation between R_1 and R_2 – meaning that the more a stem tends to lower in predicative position, the more it tends to lower in subject position. The center plot filters the *-ható* '-ble' adjectives, which proves to be its own subcategory inside the o-stems.

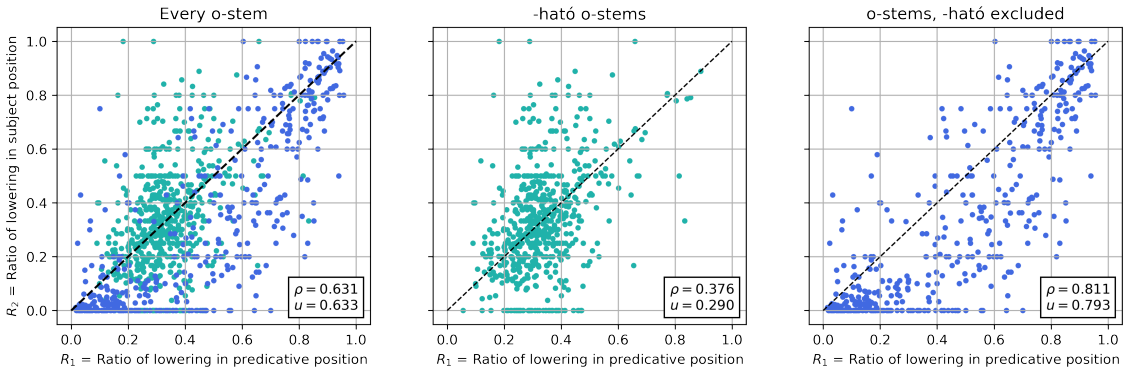

Figure 5: Plots of o-stems. Due to the sharp difference between *-ható* and the rest of the forms, we divided this group into two subcategories.

This group shows a weaker correlation between R_1 and R_2, but these stems prefer to lower in subject position rather than in predicative position.

If we exclude the *-ható* adjectives, we get a more coherent adjectival group, with very high correlation and high ratio of stems being lowering in attributive position than in subject position. In this group, we can see a dense group of words in the lower part of the plot, these are the (almost purely) nominal stems, with occasional (but clear) adjectival use, such as *látó-a-k* 'seer-PL', *zaklató-a-k* 'harasser-PL', *buzgó-a-k* 'zealous-PL', *támogató-a-k* 'supporter-PL'. Overall, this group has a ρ of 0.811 and almost 80% of the points prefer to lower in predicative position rather than in subject position.

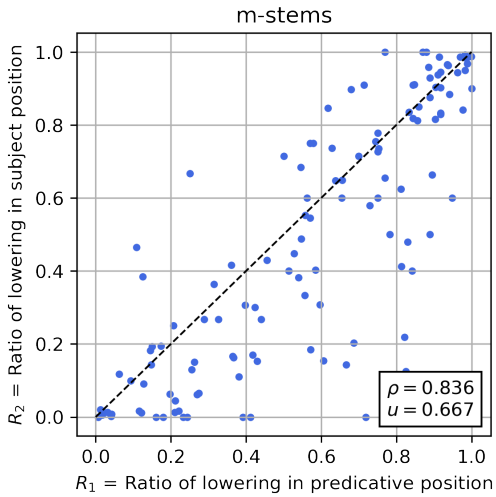

Figure 6:
Plot of m-stems.

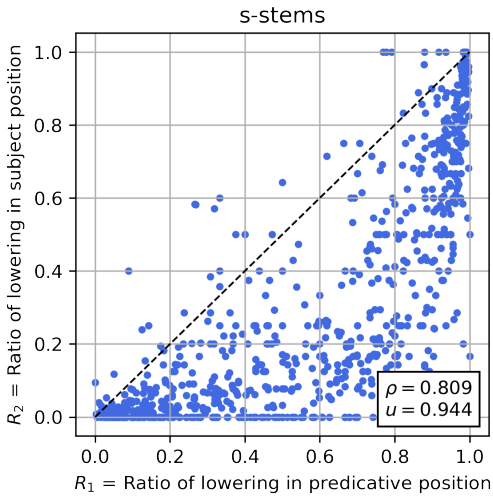

Figure 7:
Plot of PRED vs SUBJ in s-stems.

In fig. 6, we can see the plurals of the miscellaneous stems. This group counts only a few elements. We can see the usual, strong correlation between R_1 and R_2 and that two thirds of the points are under the dashed line.

In fig. 7, we can see the plurals of the s-stems. Here, 94% of the points prefer lowering in predicative position to in subject position, and the correlation is very high, 0.809.

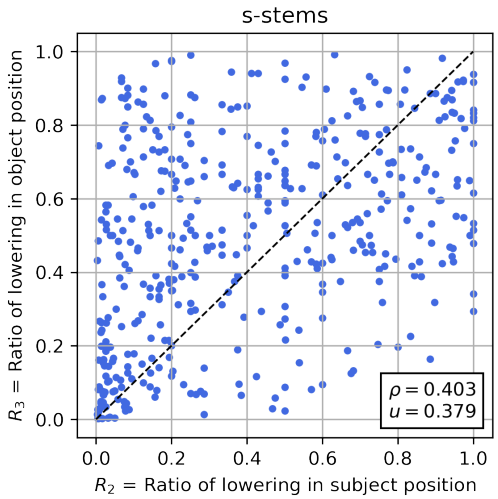

Figure 8:
Plot of SUBJ vs OBJ in s-stems.

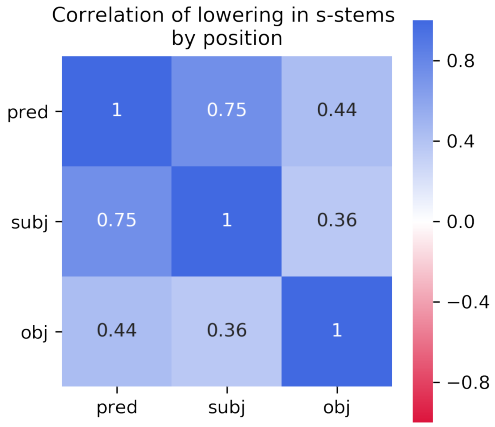

Figure 9: Correlation of lowering in different cases.

In fig. 8, we can see the ratio of lowering in subject position versus in object position in s-stems. This is the first major contradiction to our hypothesis, since there are significantly more points over the dashed line, meaning the majority of the stems prefer lowering in subject position to lowering in object position. Despite of this, the correlation is significant ($\rho = 0.403$, meaning the higher the lowering in object position of the stem, the higher the lowering in subject position.

Figure 9 summarizes the correlation of lowering in the s-stems group. Note that the correlations are different than in fig. 7 – we had to exclude some forms which did not appear in both lowering and non-lowering forms in the corpus. The lowering in predicative is the richest in information - it has the highest correlation with the others, indicating that if a stem prefers lowering in predicative position, it strongly prefers lowering in subject position ($\rho = 0.75$), and weaker, but prefers lowering in object position ($\rho = 0.44$). The other forms present lower correlation coefficients, indicating lesser information.

6 Semantics

There are many ways to capture the meaning of words, but in this thesis, we are using vector space models. Vector space models (VSM) extract knowledge automatically from a corpus, requiring much less labor than manually creating onthologies and knowledge bases. A vector space model is a model which assigns a vector (or multiple vectors) for each word from the vocabulary of the corpus. Once a vector space model is established, the powerful tools of mathematics can be used to explore and manipulate the vector space. We can decompose vectors, search for similar vectors, solve analogies using vector addition and subtraction, and even utilize them to represent more complex structures, such as sentences, documents, texts.

In the following subsections, we are introducing the notion of 'word embeddings', demonstrate how the embedding algorithm works, give a quick overview of the basic mathematical definitions needed for our hypothesis, and at the end of the section, we present the results of our analysis.

6.1 The basic idea of word embeddings

The idea of transforming text into vectors dates back to 1975 (Salton, Wong, and Yang, 1975), when Salton decided to create a large occurrence dictionary from multiple documents, then characterizing each document by the occurrences of the words from the vocabulary. The 'statistical semantics hypothesis' assumes that the word frequencies describe the meaning of a document. One could then measure the similarity of these document vectors by using their euclidean, cosine, Manhattan, or Jacquard distance. A document about cats must contain the word 'cat' with high frequency, another document about domestic animals also contains the word 'cat' with high frequency, so the respective coordinates for the two document vectors should be similar. One can even organize these vectors into clusters, creating groups of documents covering specific topics. Having document vectors enables querying a document database, we can create queries for specific phrases, thus allowing indexing and searching through large libraries. The so-called 'bag-of-words' hypothesis states exactly this phenomenon, a document's word frequencies show the relevance of the document to a query of the words. We can also calculate the 'term frequency – inverse document frequency' (tf-idf) product for each word. If a word appears frequently in one document, it has high 'term frequency' in that document, and if that word appears only in a few documents, it has high 'inverse document frequency', the tf-idf product is high, so that word is descriptive of that document in

a document collection (Turney and Pantel, 2010). On the other hand, if a word is present and frequent in every document (such as the determiner 'the'), it has high term frequency and very low inverse document frequency, thus is not descriptive of any document, does not convey any information of that particular document.

Taking this one step further, we could imagine that the surrounding words for each word is the document itself. For each word, we could count how many times the other words from the vocabulary appear in a nearby window (in a certain length to each direction, usually chosen between 2 and 5). That way, we create a list of words and corresponding frequencies (context vectors) for every word in the vocabulary, and as the quote says, the context of a word describes the word itself. Furthermore, the 'distributional hypothesis' states that words in similar contexts have similar meanings (Deerwester, Dumais, and Harshman, 1990; Harris, 1954). Spotting this among the word vectors is quite easy – if two vectors are similar by some similarity measure, the corresponding words have similar meaning.

While the counting of the surrounding words is easy and straightforward, there is a newer method to create vectors: not simply counting the surrounding words, but using neural networks to predict the words based on the context, then use the error of the prediction to improve the algorithm. (Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington, Socher, and Manning, 2014; Bojanowski et al., 2016)

6.2 Mathematical background

Definition 6.2.1. Vector

A vector of dimension n over the set of real numbers $\mathbb{R}$ is a finite ordered list of n real numbers.

$$\mathbf{v} \in \mathbb{R}^n = (v_1, v_2, \ldots, v_n), \text{ where } \forall v_i \in \mathbf{v} : v_i \in \mathbb{R}$$

Definition 6.2.2. Norm

The norm $\|\cdot\|$ is a function that shows the length of a vector $\mathbf{v}$ in a vector space.

$$\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}, \text{ where } \mathbf{v} \in \mathbb{R}^n$$

Definition 6.2.3. Dot product

The dot product $\langle\cdot\rangle$ of vectors $\mathbf{u}$,$\mathbf{v}$ is the sum of the element-wise product of $\mathbf{u}and\mathbf{v}$.

$$\mathbf{u} \cdot \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{n} u_i v_i, \text{ where } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$$

Definition 6.2.4. Cosine similarity

The cosine similarity shows the cosine of the angle of two vectors, mapping the angle in $[-1, 1]$.

$$\text{sim}_{cos}(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}, \text{ where } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$$

Definition 6.2.5. Euclidean distance

The Euclidean distance is a distance metric used in Euclidean spaces.

$$\text{d}(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|, \text{ where } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$$

Definition 6.2.6. Cosine distance

The cosine distance is a distance metric based on cosine similarity.

$$\text{d}_{cos}(\mathbf{u}, \mathbf{v}) = 1 - \text{sim}_{cos}(\mathbf{u}, \mathbf{v}), \text{ where } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$$

Definition 6.2.7. Softmax function

The softmax or exponential normalized function is a logistic function that enables to interpret a series of values as a probabilistic variable. Let $\mathbf{x} \in \mathbb{R}^n$ be a sample, $x_i \in \mathbf{x}$ a numerical observation. We can define the σ softmax function:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \quad \text{for } i \in 1, \ldots, n$$

6.3 Neural network-based word embedding

The idea of neural networks dates back to the 40s (McCulloch and Pitts, 1943), when McCulloch created a computational network. The main idea is that our brain is composed of neurons (nodes) and synapses (edges). We, as humans, learn and memorize by creating and strengthening synapses, and an artificial neural network – by analogy – should learn by strengthening and weakening weights on the edges based on the sample it receives. Constructing and training a neural network is a difficult task, because we do not have a strong idea how to interpret the weights of the edges or the nodes themselves – a neural network is a black box, and we do not always know how the architecture of the network should look like, or how we should train a network.

The architecture in a neural word embedding consists of only a single hidden layer of neurons and an output layer with a softmax classifier, shown in fig. 10, with an example iteration. The task of the model is, for context in the corpus, to learn the probabilities of vocabulary words being in that context.

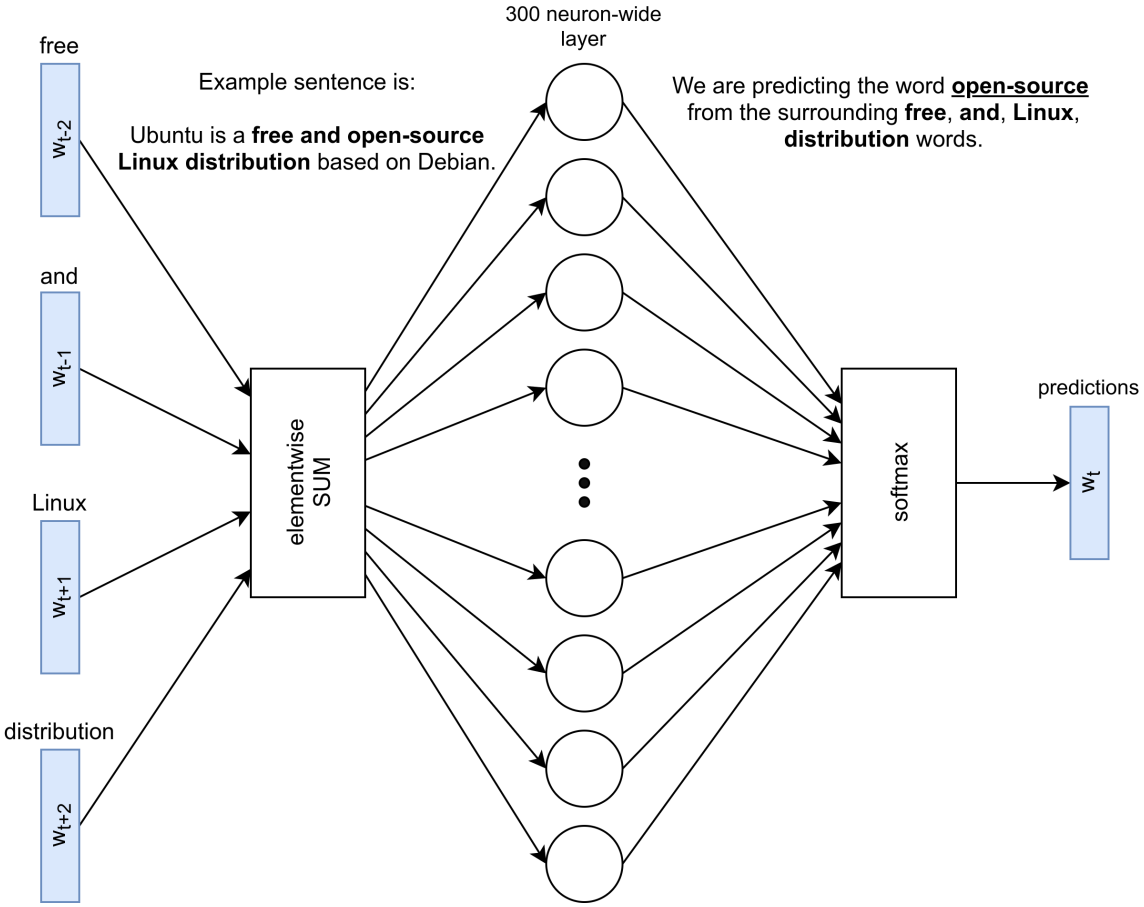

Figure 10: Network architecture

The example sentence is 'Ubuntu is a free and open-source Linux distribution based on Debian.', we are using a windows size of 2 in each direction, and we are guessing what word is in the context of the words 'free', 'and', 'Linux', 'distribution'. In the example, the target word is 'open-source', and the model will map probabilities to the words of the vocabulary. Using these probabilities, we can backpropagate the error and teach the model. In the example, the size of the vocabulary is N, and the size of the hidden layer is D.

The input is a binary vector representing the context, and the output is a probability vector. First, we are creating a context vector by summing up the index vectors for each word, marked $w_{t\pm i}$ in fig. 10. The context vector is classified by the neurons, and then the softmax function creates an easily explainable probability distribution for the target word.

To summarize, we create a model to do a fake task (predicting words based on the contexts) only to learn the input weights which will be used as vectors. The same way, the model also learns the output weights and the classification problem

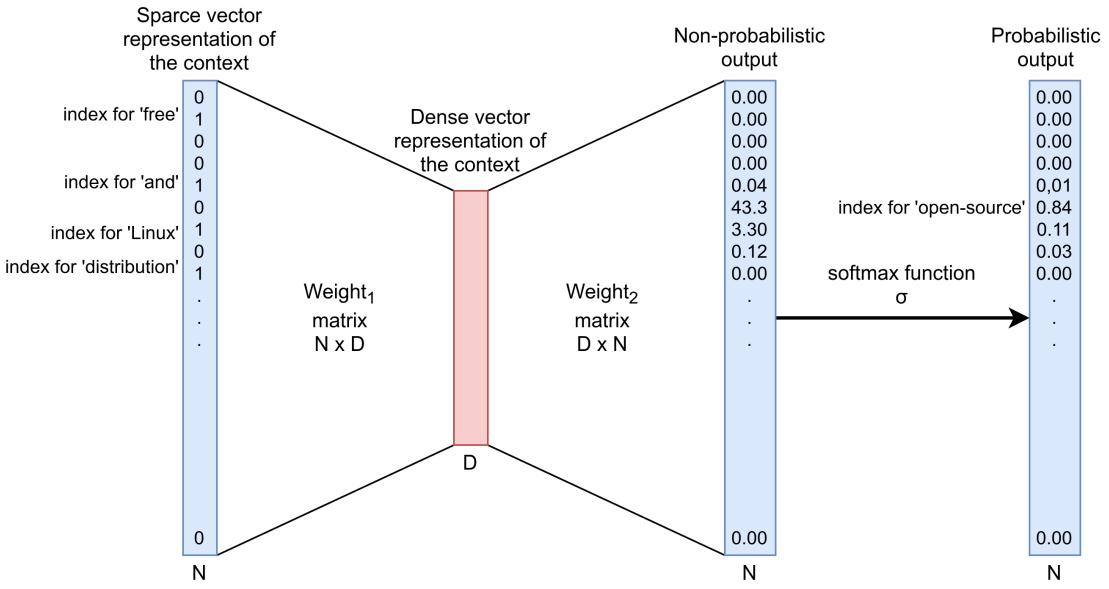

Figure 11: Classifying text with CBOW. The dimension are written below the objects.

reduces to a matrix dot product and to a softmax classification problem.

In fig. 11, we can see the dimensionality of the mapping functions and the importance of the softmax function.

A word's index vector is a one-hot[15] vector encoding its position in the vocabulary. When we sum the index vectors up, we lose the order of the context, but we get a context index vector, encoding the position of the words in context in the vocabulary. After the context is represented by a sparse[16] vector of dimension N, it is mapped onto the hidden layer of dimension D by a linear transformation (Weight_1 matrix) of dimension $N \times D$, creating a D-dimensional dense[17] representation vector. The dense representation vector then is processed by another linear transformation (Weight_2 matrix) of dimension $D \times N$, creating a vector of vocabulary dimension N. The softmax function is then applied on this vector, providing us with a probabilistic vector, where each coordinate means the probability of that word being in the given context.

For adjusting the weights of the matrix, the model uses backpropagation, which is a method for refreshing the weights by calculating the partial differentials of the error caused by each weight.

As stated previously, predicting the target word based on the context is actually

[15] A vector, that has exactly one 1 value, and the rest are 0 values.

[16] Sparse, as in having lots of 0 values.

[17] Dense, as in not having 0 values.

a fake task. We are not interested in guessing missing words from contexts, we are interested in representing word forms in a compact and dense manner. That is why the mapping function (matrices) are important. The first mapping function is of dimension $N \times D$, that is, a vocabulary-tall and neuron layer-wide matrix. For the word of index i in the vocabulary, this matrix encodes a dense representation in its row i, and that is what we were after. The corresponding row of the Weight_1 matrix for a word is the **word vector**.

The main difference from the traditional methods are the subsampling, negative sampling, and lower dimensionality. Subsampling is a technique to reduce the importance of the common words in the corpus. The intention behind the subsampling is that the words like 'the', 'a', 'have' occur very frequently, yet encode little semantic information about the context. The subsampling is probability based and uses the following function to determine the probability that the word w_i should be taken into account when updating the weights of the neural network, where z is the relative frequency function.

$$\mathbb{P}(w_i) = \left(\sqrt{\frac{z(w_i)}{0.00001}} + 1 \right) \cdot \frac{0.00001}{z(w_i)}$$

Negative sampling is a technique to reduce calculation time. Without negative sampling, for each word, we would need to increase the weight of the edges of the correct guess (reward the specific edges for guessing the word right), and we would need to decrease the weights which predicted the context wrong, so we would update every weight for every item in the vocabulary each training step. With negative sampling, we select a few noise words (5, in our case), and we update the network only by the error produced by these noise words. Thus, for each training step, we would update the network 1+5 times.

6.4 Obtaining word vectors

The software we are using to create word embeddings is Radim Řehůřek's `gensim`. (Řehůřek and Sojka, 2010) It is a Python package created in 2010 to ease text processing, but ended up being one of the most robust, efficient and hassle-free softwares to process plain text. While the software offers the possibility of fine-tuning every hyperparameter, by default it reduces noise, smooths vectors, and even removes words with low frequency and low semantic distinguishing value. We used the followed hyperparameters in the creation of our models: method of training is 'continuous-bag-of-words with negative sampling', the dimension of the word embedding is 300,

the window used is 5 words in both directions, and negative sampling is set to 5, meaning that for every training iteration, 5 'noise words' are drawn. The minimal samples were set to 10, meaning that the model automatically cropped rare words. The training consisted of 5 epochs, which was really fast, around 600000 words/second with an Intel Core i9-9920XE, it took around 20 hours for the whole corpus.

6.5 Information encoded in word vectors

Word vectors mainly encode semantic information, but some syntactic information is also present. One method to evaluate word embeddings are inspecting some words' closest neighbors. We will take our model trained in the previous section and see what words are the closest in cosine distance to the target word.

In table 10, we can see 7 examples for the relations extracted from the text by the word embedding. In example 1 and 3, we can see hyponymy-hypernymy relation, as 'man', 'child, 'woman' are hyponyms of 'human', and 'block of flats' and 'cottage' are hyponyms of 'house'. In example 2, we can see that the words belong to the same category. The 'dog', 'cat', 'doggy', 'kitten' belong to the same category, while the word 'animal' is a hypernym of 'dog'. Example 4 shows that word vector similarity also encodes antonymy relation. This is not surprising, since the model learns from contexts – and antonyms usually share the same contexts, that is, if a thing can be 'good', it can also be 'bad'. Example 5 and 6 shows that similarly spelled words can have entirely different meanings, and that it can be seen trivially from the surrounding words. In fact, it is a very rare occurrence that a language has word pairs with similar meaning and spelling (Blevins, Milin, and Ramscar, 2017, p. 15). Example 7 shows an example of synonymy relation. *Hatos* 'number six', written in letters, has a high degree of similarity with *6-os*, written with a suffixed number. The same relation can be seen with unit measures, abbreviations, and words written in caps or with bad orthography.

1.	*ember*	*ember-nek, gyerek, férfi-ember, nő, férfi*
gloss	human[a]	man-DAT, child, male-human, woman, man
2.	*kutya*	*macska, kutyus, cica, kiskutya, állat*
gloss	dog	cat, doggy, kitty, puppy, animal
3.	*ház*	*lakó-ház, ház-nak, ház-ikó, tömb-ház, panel-ház*
gloss	house	apartment, house-DAT, cottage, block of flats, panel house
4.	*jó*	*rossz, remek, szuper, jo*[b]*, kellemes*
gloss	good	bad, great, super, good, pleasant
5.	*vörös-ek*[c]	*barná-k, piros-ak, sárgá-k, kék-ek, fehér-ek*
gloss	red-PL	brown-PL, red-PL, yellow-PL, blue-PL, white-PL
6.	*vörös-ök*[d]	*német-ek, franciá-k, bajor-ok, muszk-ák*[e]*, spanyol-ok*
gloss	red-PL	German-PL, French-PL, Bavarian-PL, Russian-PL, Spanish-PL
7.	*hat-os*	*kilenc-es, öt-ös, 8-as, tizenkett-es, 6-os*
gloss	number six	number nine, number five, number 8, number twelve, number 6

[a] *Ember* means both man and human in Hungarian.

[b] With bad orthography.

[c] In Hungarian, there is different word for the red of the flag or the blood, but nowadays, generally used as deeper red.

[d] As a noun-lowering noun stem, red-PL means communist, Soviet.

[e] Pejorative, derived from Moscow.

Table 10: Examples for the closest neighbor relation in our word embedding. The gloss gives the translation in order.

The second important piece of information that vectors encode are analogy. Word vectors allow us to do analogies, like France : Paris :: Germany : _____, where we can fill in the blank using simple vector addition:

$$\text{nearest neighbor}(\mathbf{v}_{\text{Paris}} - \mathbf{v}_{\text{France}} + \mathbf{v}_{\text{Austria}}) = \mathbf{v}_{\text{Wien}}$$

	$word_1$: $word_2$ :: $word_3$: _____			
1.	*férfi*	*fiú*	*nő*	***lány***
gloss	man	boy	woman	girl
2.	*Franciaország*	*Párizs*	*Ausztria*	***Bécs***
gloss	France	Paris	Austria	Wien
3.	*Magyarország*	*pörkölt*	*India*	***curry***
gloss	Hungary	meat stew	India	curry
4.	*jó*	*rossz*	*magas*	***alacsony***
gloss	good	bad	tall	small

Table 11: Solving analogies using word embeddings. The last column is the answer given by the embedding.

As you can see in table 11, we can extract general semantic relations (examples 1.-3.) and antonym relations (example 4.).

6.6 Methodology

Visualizing and processing 300 dimensions is a difficult task. The phenomenon of the curse of dimensionality (Bellman, Corporation, and Collection, 1957) states that many problems arise in high-dimensional space that are not present low dimensional spaces. In our case, traditional distance functions between vectors is hard to interpret since there are large gaps between points, thus it is very hard to densely populate high-dimensional spaces. However, there are lots of methods to reduce dimensionality while preserving local features of groups of points. The two most commonly used modern methods are t-distributed stochastic neighbor embedding (t-SNE, Maaten and Hinton (2008)) and uniform manifold approximation and projection (UMAP, McInnes, Healy, and Melville (2018)). These methods predominantly preserve local structure and are essential when visualizing data where dimensions are equally important and hard to interpret.

The word vectors were normalized, since word vector length strongly correlates with the relative frequency of the corresponding word (Arora et al., 2016). Both UMAP and t-SNE have adjustable hyperparameters, thus for each plot in the next section, these were tuned based on Wattenberg, Viégas, and Johnson (2016).

Our hypothesis in section 4.2 stated that lowering forms are generally more simi-

lar to adjectives than to nouns and non-lowering forms are generally more similar to nouns than to adjectives. Measuring the adjectival and nominal components of word vectors is incredibly difficult. The most promising paper in this matter is Rothe, Ebert, and Schütze (2016), but reproducing their results, even when reimplementing their algorithm has proven difficult. Nevertheless, we do not have to resort to orthogonal transformations, word vector distillation and ultradense reembedding to prove this hypothesis. We know that the vacillating pairs of words are nearly identical in meaning (except for polysemy), and that if we separate the lowering forms from the non-lowering forms, the systematic difference should be based on the preferred category of the word form. Thus, we are trying to show that the lowering and non-lowering forms separately form coherent groups. To see the coherence in 300 dimensions, we are using a simple method based on the nearest neighbors in cosine distance. For each word, we take some (depending on the stem's type) odd number of nearest neighbors. Among the nearest neighbors, some will be lowering and some will be non-lowering, but one group will have a higher count than the other one. We will call the ratio of the words of the same height to the size of the neighborhood neighborhood score, marked with S. And if the original word has the same height as most of its neighborhood ($S > 0.5$), then its neighborhood is descriptive of the word. If most of the words have correct neighborhood, we can conclude that the lowering and non-lowering words form coherent clusters. To measure this coherency, we take the very simple accuracy measure:

Definition 6.6.1. Accuracy

Accuracy is the ratio of correctly classified items to the number of items.

$$\text{accuracy} = \frac{\text{number of correctly classified points}}{\text{total number of points}}$$

If the accuracy is around 0.5, we can say that the points are randomly distributed. The higher the accuracy is, the more descriptive neighborhoods are, therefore the points form more coherent groups.

We do have to note that we cannot compare more inflected words with less inflected words, thus we cannot use for example comparative suffixed words to enrich our dataset. More inflection adds more syntactic information to the word vectors, thus increasing the vector similarity between similarly inflected words (Lévai and Kornai, 2019).

6.7 Results

All of the plots are available in HTML format in this project's thesis folder, annotated, in which you can zoom into after using the buttons in the upper right corner.

We will first see the least coherent group, which is the group of plural miscellaneous stems, then the plural o-group, then the plurals of s-group, and lastly, the accusatives, which are mainly s-stemmed[18]. On the plots, the blue points mark the mid forms with similar neighborhood, the green ones mark the mid forms with dissimilar neighborhood, the red ones mark the low forms with similar neighborhood and the orange points mark the low forms with dissimilar neighborhood. The legend is the same for all of the figures in this subsection, and the axes are abstract, due to the nature of dense word embeddings and dimension reduction.

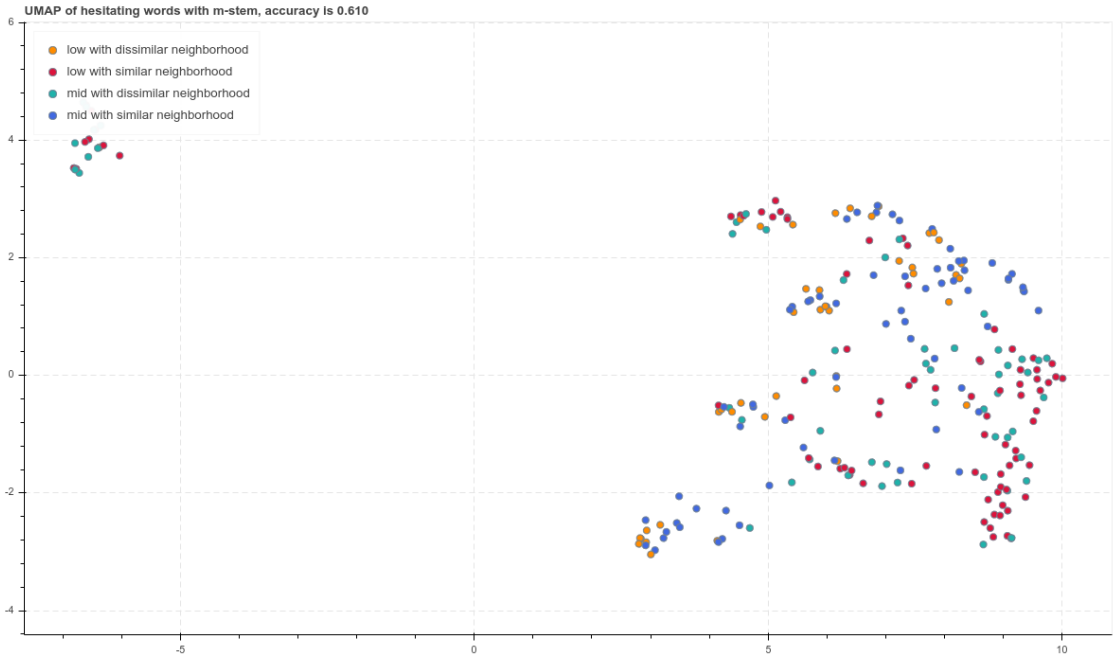

Figure 12: UMAP of the word forms in the m-group. Note that the neighborhood relation is based on the 300-dimensional neighborhood, not on the 2-dimensional, projected neighborhoods.

In fig. 12, we can see a UMAP reduction of the m-group word forms.

On the right hand side, we can see a dense core of red points with a few green points surrounding. The words do not share an apparent common meaning (examples include *szinonim-ak* 'synonym-PL', *hazug-ak* 'liar-PL', *tudatalatti-ak*

[18] o-stemmed cannot be lowering in accusative, seen in section 3.2

'subconscious-PL', *árnyalt-ak* 'shaded-PL', *bumfordi-ak* 'clumsy-PL', ...). On the left side, the little cluster is the suffixated numbers, like *harmadik-ak/harmadik-ok* 'third-PL', *hányadik-ak/hányadik-ok* 'which number-PL', *sokadik-ok/sokadik-ak* 'umpteenth-PL'. Both the lowering and non-lowering forms are in the same cluster. In the top side, there are the words derived from place names, i.e. *csány-i-ak* 'Csány-FROM-PL', *kanári-ak* 'Canary-FROM-PL', *somló-i-ak* 'Somló-FROM-PL. The other dense areas show little system.

Generally speaking, we can see that the colors are not entirely randomly distributed – there is a red cluster on the right, and it is mainly surrounded by blue points. The accuracy reflects this tendency – it is 0.610, not entirely convincing, but higher than 0.5, showing us that there is a slight coherence in high dimension.

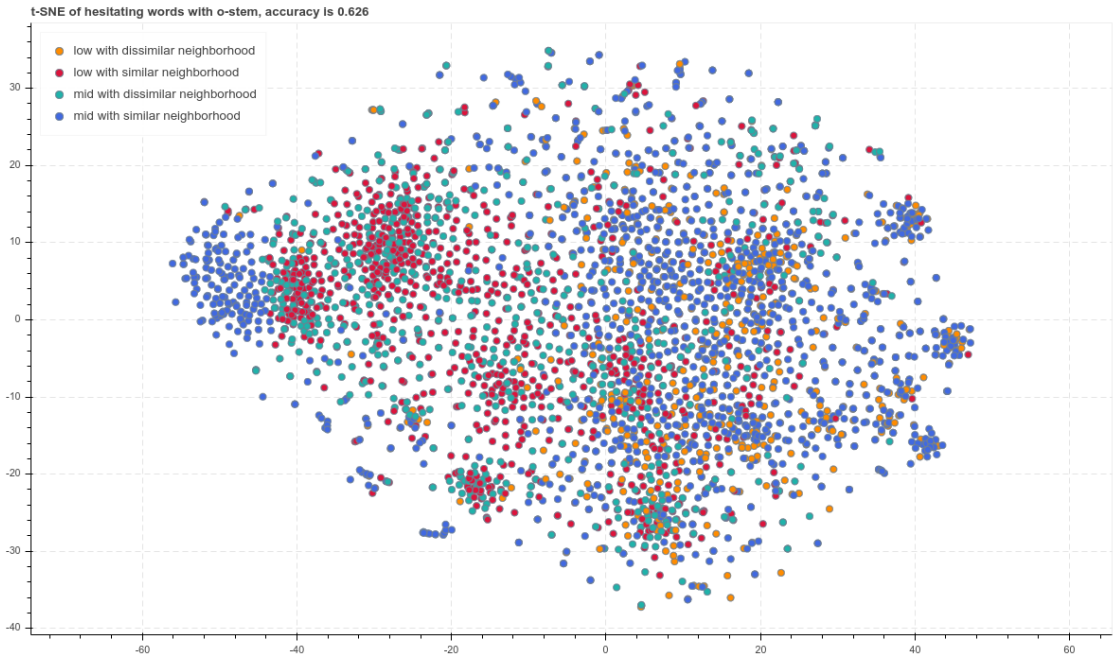

Figure 13: t-SNE of the o-stem group. One can immediately notice the eye-catching circular shapes, that is an effect of the t-SNE.

In fig. 13, we can see a t-SNE reduction of the o-group word forms. We can immediately see a dense blue core on the left, followed by mainly red clusters a bit right, then a large, fuzzy area of blue-green-red-orange. The lackluster fuzzy area is mostly populated by the *-ható* '-ble' forms, with many word pairs being being almost in the same position, indicating the lack of difference in meaning between the lowering and non-lowering forms. Examples include: *irányítható* 'controllable', *lebontható* 'deconstructible', *bevállalható* 'bearable', *szabályozható* 'regularizable'. The blue cluster on the left side are mainly human attributes, occupations, groups: *gyógyulók* 'healing-

PL', *vágyakozók* 'longing-PL', *provokálók* 'provoking-PL', *fontoskodók* 'officious-PL', *ordítók* 'shouting', *mozgók* 'moving'. It is really interesting to see how coherent this cluster is while not having high degree of semantic relatedness intracluster. The two red clusters on the left side are also very hard to characterize. The words in those two clusters are very similar in meaning to those in the blue cluster – they generally mean human attributes, occupations, groups: *provokálóak* 'provoking', *vendégcsalogató* 'inviting'.

Prediction based on the neighborhood gives us a 0.626 accuracy in this case, which is a bit better and means a bit more coherent clusters than the m-group.

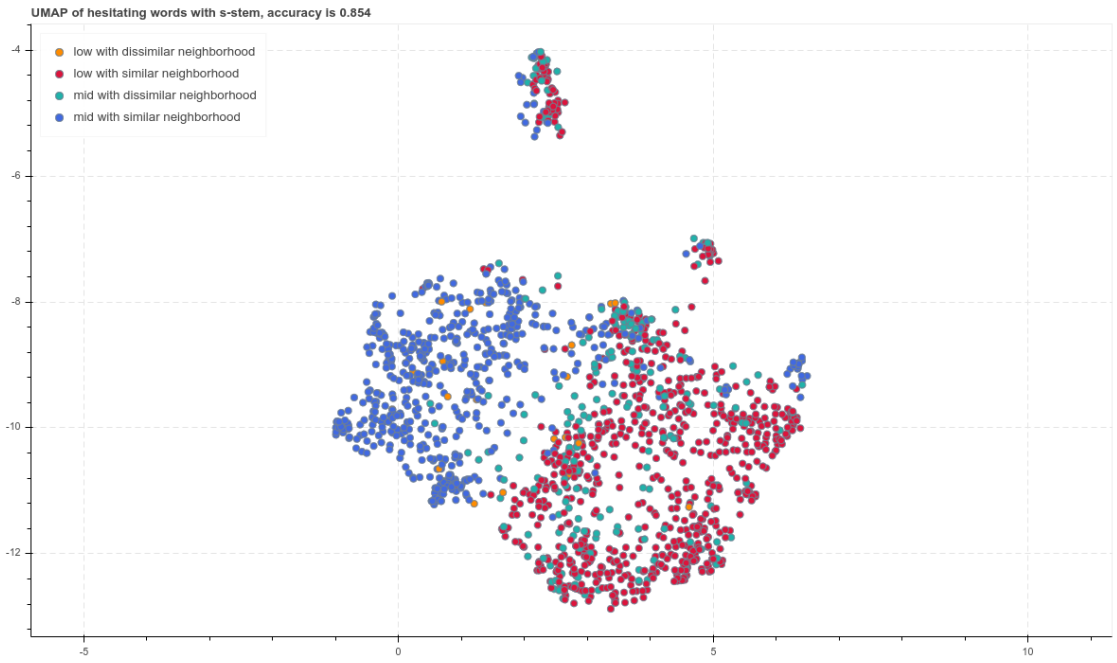

Figure 14: UMAP of s-group.

In fig. 14, we can see the most coherent group, the group of the s-stemmed forms. The non-lowering forms are on the left hand side, and the lowering forms are on the right hand side. There are three little clusters on the top, we start by describing those. The blue cluster in (6.5, -9) contains plant-related and geographical words: *diós-ok* 'walnuty-PL', *hínáros-ok* 'seaweedy-PL', *pozsgás-ok* 'succulent-PL', *mocsaras-ok* 'marshy-PL' *szoros-ok* 'gorges-PL'. Their respective lowering pairs are not so far away, they are in the lower-left direction from this group. This is the most coherent small cluster, and we can definitely see the semantic relatedness here, and this small cluster is also showing us that there is a distinct difference in the meaning of lowering and non-lowering forms.

The second cluster is located in (5, -7). These words are used to describe attributes of buildings and apartments: konyhás**a**k 'kitchened-PL', *teraszos**a**k* 'terraced-PL', *egyágyas**a**k* 'one-bedded-PL', *parkettás**a**k* 'hardwood-floored-PL', *betonos**a**k* 'concreted-PL'. Most forms are lowering, with a few exceptions, like *egyszobás**o**k* 'one-roomed-PL', *egyágyas**o**k* 'one-bedded-PL'. It seems that these words do not form distinct pairs in meaning.

The upper, big cluster is also multi-colored. These are mainly derived from units, such as *szavasak* 'worded-long-PL', *kilogrammosok* 'kilogramm-heavy-PL', *órásak* 'hour-long-PL', *lóerősök* 'horse-power-strong-PL', *forintosak* 'forint-cost-PL'. Here, the semantic relatedness is so strong that the lowering and non-lowering forms barely separate.

Then there is the big cluster, with a dense blue core on the left, and a dense red core on the right. In the case of this big cluster, we cannot really tell the reason of the semantic similarity, but that is exactly why we are using the word embeddings. The two clusters are coherent and we can indeed see that intragroup similarity is way higher than the intergroup similarity.

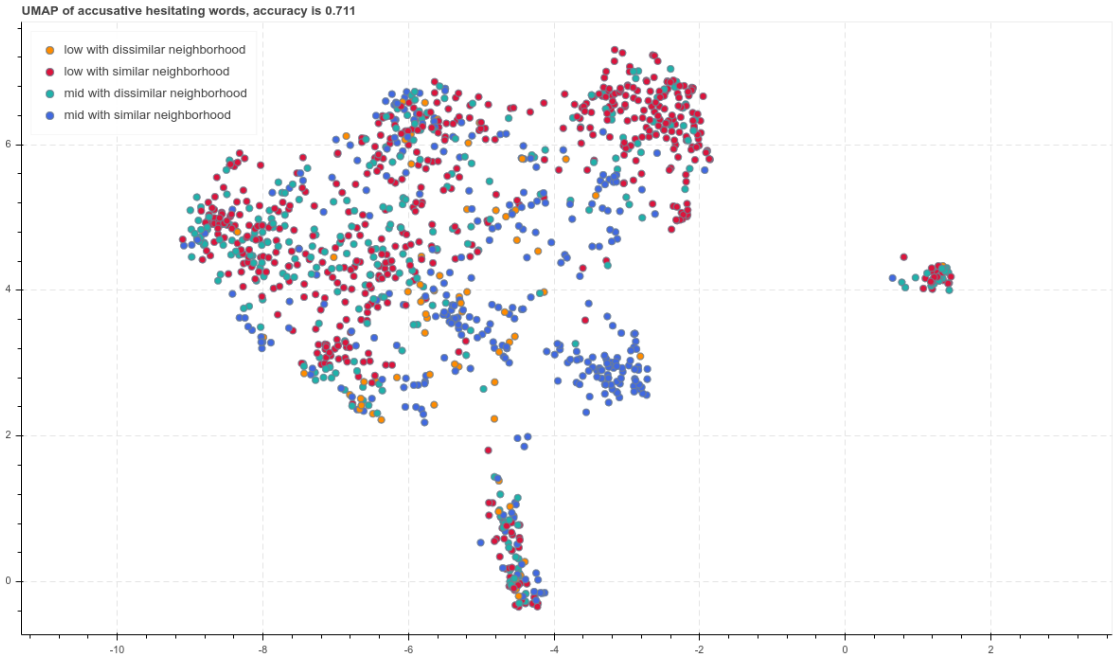

Figure 15: UMAP of accusative words.

In fig. 15, we can see the UMAP of the accusative forms. We are not dividing the accusatives into subcategories, since the o-stems cannot take accusatives, and the m-stems are very low in number.

On the right side, there is an incoherent cluster very far from the dense area.

These words are generally children's games: *doktorosat* 'playing doctor-ACC', *papás-mamást* 'playing house-ACC'. These forms appear with their pairs close by, indicating the lack of semantic difference.

The upper right red cluster contains mainly attributional adjectives, *aranyosat* 'cute-ACC', *tudományosat* 'scientific-ACC', *erőszakosat* 'violent-ACC'. The blue cluster at (-3, 3) contains words meaning people having that attribute: *lovast* 'horseman-ACC', *kártyást* 'gambler-ACC', *vallásost* 'religious-ACC'. The colorful bottom cluster are made up of words derived from unit measures, *tonnást* 'ton.heavy-ACC', *szavasat* 'word.long-ACC', *órást* 'hour.long-ACC'. Other than these 4 clusters, there is a dense and fuzzy core.

The accuracy in this case is 0.711, showing that the words generally form coherent neighborhoods, thus the lowering and non-lowering forms are usually distinct in meaning.

7 Overall results

In this section, we are joining together the results shown in section 5.4 and in section 6.7. In the following plots, S_i marks the neighborhood score, defined in section 6.6, that is, for each word, it shows ratio of points having the same suffix height in its neighborhood, and S_1 marks the score of the plural, lowering form, S_2 marks the plural, non-lowering form. In case of the s-stems, S_3 marks the accusative, lowering form and S_4 marks the accusative, non-lowering form. R_1 marks the ratio of lowering in predicative position, R_2 marks the ratio of lowering in subject position, and R_3 marks the lowering in object position. We will analyze the correlation between the ratio of lowering (R_i) and the neighborhood score (S_i) per stem.

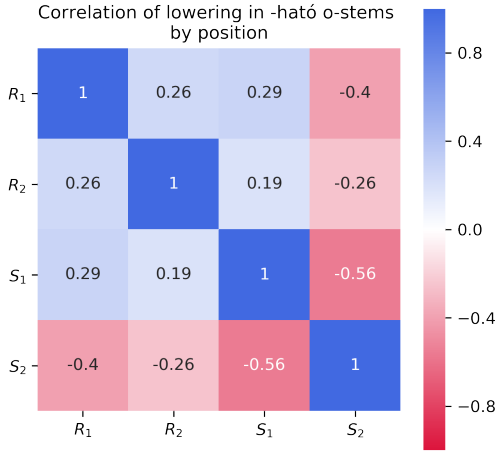

Figure 16: Plot of *-ható* o-stems.

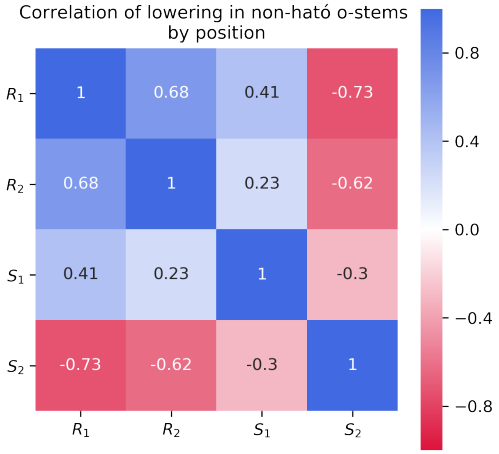

Figure 17: Plot of non-*ható* o-stems.

In figs. 16 and 17, we can see the correlation of o-stems, divided as seen in section 5.4. In case of the *-ható* words the correlation between the syntactic and semantic data is weak $(-0.40, -0.26, 0.19, 0.29)$, and R_1 correlates slightly more with S_1 and S_2 than R_2. The meaning of this correlation is that if the stem prefers prefers lowering in predicative position, the lowering form tends to be similar in meaning to other lowering forms, while the non-lowering form of the same stem tends not to be similar in meaning to other non-lowering forms. In case of the non-*ható* o-stems, the tendency is similar to that of the *-ható* stems, but stronger. $(-0.73, -0.62, 0.23, 0.41)$. The high negative correlation between syntactic ratios and S_2 mark the abundance of words which are always in subject position (things, occupations), such as: *látó* 'seer', *zaklató* 'harasser', *vádlók* 'prosecutor, accuser'.

These words are very rarely lowering, but have high neighborhood score due to being similar to other prototypical nouns.

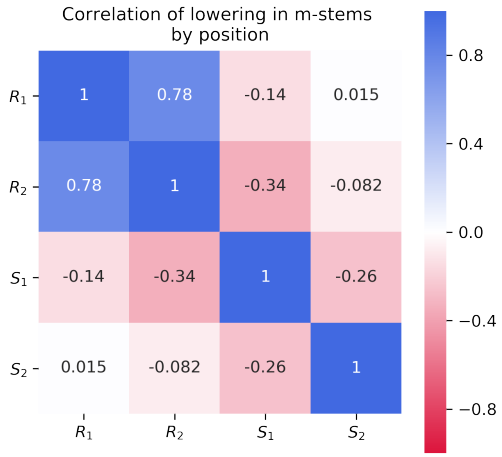

Figure 18: Plot of m-stems.

In fig. 18, we can see the correlations of the miscellaneous-stems. The correlation between the syntactic ratios and semantic scores are low, but this can be due to the low quantity of the data or the highly varying stems.

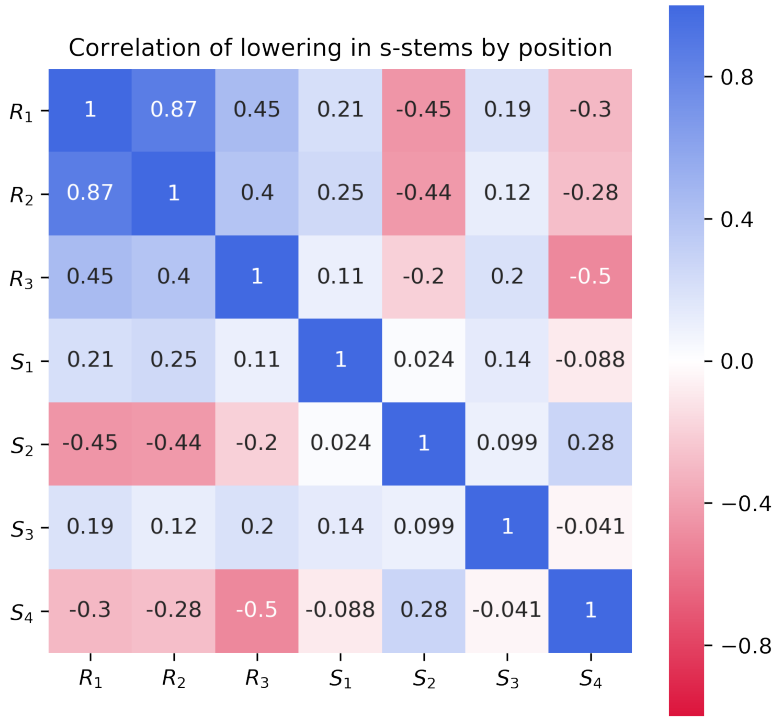

Figure 19: Plot of s-stems. Note that there are accusative forms in this figure.

In fig. 19, we can see the s-stems, which are unique by having vacillation in object position. This follows the general tendency shown in fig. 16, which is that if a stem prefers lowering in predicative or subject position, the meaning of the lowering form is in a lowering neighborhood (weakly, $\rho = 0.21, 0.25$), and the non-lowering form is in a dissimilar neighborhood. The unique part in this figure is the relation between R_3 and S_3, S_4.

Seemingly, R_3 can weakly ($\rho = 0.2$) explain S_3, that is, the ratio of lowering accusatives does not tell much about the semantics of the lowering form of the stem. However, it has high negative correlation with S_4, meaning that high R_3 shows low neighborhood score for the non-lowering form and low R_3 shows high neighborhood score for the non-lowering form.

8 Conclusion

We have analyzed the interaction of phonology, syntax and semantics in lowering nominal stems in this thesis. Our hypothesis stated that the degree of lowering in nouns and adjectives, a seemingly morphophonological phenomenon, can be attested in higher levels of the language. Our analysis has shown that there is a strong, systematical difference between the lowering and non-lowering forms, and even though the tendencies are probabilistic, they are statistically significant. This difference can be seen in a word's syntactic and semantic attributes, and strongly correlates with our intuition.

In syntax, the syntactic position of the word contributes to the quality of linking vowel. The predicative position has a bigger influence on the quality of the linking vowel, meanwhile the other positions are less influential. According to our analysis, the form in the predicative position is the richest in information, although that is an expected result since that form is the most frequent.

In semantics, the meaning influences the quality of the linking vowel. In case of the majority of the stems, the difference between the lowering and non-lowering variant is predictable, and correlates with our intuition – the non-lowering form is 'nouny', the non-lowering form is 'adjectival'.

We have also seen that the syntactic and semantic attributes are tied together, thus a word's meaning and syntactic position affect the quality of the linking vowel.

This phenomenon would be difficult to explain if we treat the language as a modular structure, with one-directional connections between the linguistic modules. Alternatively, our hypothesis fits well into usage-based theories like Bybee and Hopper (2001), where frequency is a very important factor in the usage of the language.

References

Arora, Sanjeev et al. (2016). "Linear Algebraic Structure of Word Senses, with Applications to Polysemy". In: *arXiv:1601.03764v1*.

Bellman, R., Rand Corporation, and Karreman Mathematics Research Collection (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press. ISBN: 9780691079516. URL: https://books.google.it/books?id=wdtoPwAACAAJ.

Blevins, James, Petar Milin, and Michael Ramscar (Jan. 2017). "The Zipfian Paradigm Cell Filling Problem". In: DOI: 10.1163/9789004342934_008.

Bohnet, Bernd (Aug. 2010). "Top Accuracy and Fast Dependency Parsing is not a Contradiction". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, pp. 89–97. URL: https://www.aclweb.org/anthology/C10-1011.

Bojanowski, Piotr et al. (2016). "Enriching Word Vectors with Subword Information". In: *arXiv preprint arXiv:1607.04606*.

Bybee, Joan L. and Paul J. Hopper (2001). *Frequency and the Emergence of Linguistic Structure*. ISBN: 9789027298034. DOI: 10.1075/tsl.45.

Chomsky, Noam (1957). *Syntactic Structures*. The Hague: Mouton and Co.

— (1981). *Lectures on Government and Binding*. Foris.

Chu, Y. J. and T. H. Liu (1965). "On the Shortest Arborescence of a Directed Graph". In: *Science Sinica* 14, pp. 1396–1400.

Deerwester, Scott C., Susan T Dumais, and Richard A. Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Edmonds, J. (1967). "Optimum Branchings". In: *Journal of Research of the National Bureau of Standards, Section B, 71B (4)* 14, pp. 233–240. DOI: 10.6028/jres.071b.032.

Eisner, Jason M. (1996). "Three New Probabilistic Models for Dependency Parsing: An Exploration". In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C96-1058.

Elekfi, László (2000). "Semantic differences of sufficial alternates in Hungarian". In: *Acta Linguistica Hungarica 47*, pp. 145–177. ISSN: 1216-8076.

Farkas, Richárd, Veronika Vincze, and Helmut Schmid (Apr. 2012). "Dependency Parsing of Hungarian: Baseline Results and Challenges". In: *Proceedings of the*

13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, pp. 55–65. URL: https://www.aclweb.org/anthology/E12-1007.

Ford, Marilyn, Joan Bresnan, and Ronald M. Kaplan (1982). "A competence-based theory of syntactic closure". In: *The Mental Representation of Grammatical Relations*. Ed. by Joan Bresnan. Cambridge, MA: MIT Press, pp. 727–796.

Harris, Zellig S. (1954). "Distributional structure". In: *Word* 10.23, pp. 146–162.

Indig, Balázs et al. (Aug. 2019). "One format to rule them all – The emtsv pipeline for Hungarian". In: *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 155–165. DOI: 10.18653/v1/W19-4018. URL: https://www.aclweb.org/anthology/W19-4018.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc. ISBN: 0131873210.

Kálmán, László, Péter Rebrus, and Miklós Törkenczy (Jan. 2012). "Possible and impossible variation in Hungarian". In: pp. 23–50. DOI: 10.1075/cilt.322.02kal.

Karlsson, Fred (1990). "Constraint Grammar as a Framework for Parsing Running Text". In: *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C90-3030.

Lévai, Dániel and András Kornai (Jan. 2019). "The impact of inflection on word vectors". In: *XV. Magyar Számítógépes Nyelvészeti Konferencia*.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9, pp. 2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

McCulloch, W.S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of mathematical biophysics* 5, pp. 115–133.

McDonald, Ryan et al. (Oct. 2005). "Non-Projective Dependency Parsing using Spanning Tree Algorithms". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 523–530. URL: https://www.aclweb.org/anthology/H05-1066.

McInnes, Leland, John Healy, and James Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [stat.ML].

Mikolov, Tomas et al. (2013a). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mikolov, Tomas et al. (2013b). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL].

Moravcsik, Edith A. (2001). "On the nouniness of Hungarian adjectives". In: *Naturally! Linguistic studies in honor of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*, pp. 337–346.

Nemeskey, Dávid (May 2020). "Natural Language Processing methods for Language Modeling". PhD thesis. Doctoral School of Informatics, ELTE.

Nivre, Joakim (2008). "Algorithms for Deterministic Incremental Dependency Parsing". In: *Computational Linguistics* 34.4, pp. 513–553. DOI: 10.1162/coli.07-056-R1-07-027. URL: https://www.aclweb.org/anthology/J08-4003.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: http://www.aclweb.org/anthology/D14-1162.

Pollard, Carl and Ivan Sag (Jan. 1994). "Head-Drive Phrase Structure Grammar". In: *Bibliovault OAI Repository, the University of Chicago Press*. DOI: 10.3115/981210.981231.

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50. URL: http://is.muni.cz/publication/884893/en.

Rothe, Sascha, Sebastian Ebert, and Hinrich Schütze (June 2016). "Ultradense Word Embeddings by Orthogonal Transformation". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 767–777. arXiv: 1602.07572 [cs.CL]. URL: http://www.aclweb.org/anthology/N16-1091.

Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975). "A vector space model for automatic indexing". In: *Communications of the ACM* 18.11, pp. 613–620.

Siptár, Péter and Miklós Törkenczy (2001). "The phonology of Hungarian. Oxford: Oxford University Press". In: *Phonology* 18.2. DOI: 10.1017/S0952675701004080.

Törkenczy, Miklós (Feb. 2011). "Hungarian Vowel Harmony". In: *The Blackwell companion to phonology*, pp. 2963–2990. ISBN: 978-1405184236. DOI: 10.13140/RG.2.1.4010.5368.

Turney, Peter D. and Patrick Pantel (2010). "From Frequency to Meaning: Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37, pp. 141–188.

Váradi, Tamás et al. (May 2018). "E-magyar – A Digital Language Processing System". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Vincze, V. et al. (Sept. 2010). "Hungarian Dependency Treebank". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pp. 1855–1862.

Wattenberg, Martin, Fernanda Viégas, and Ian Johnson (2016). "How to Use t-SNE Effectively". In: *Distill*. DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.

Zsibrita, János, Veronika Vincze, and Richárd Farkas (Sept. 2013). "magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, pp. 763–771. URL: https://www.aclweb.org/anthology/R13-1099.