Decreasing Entropy: How Wide to Open the Window?

Balázs Indig[1,2], Noémi Vadász[1,3], and Ágnes Kalivoda[3]

[1]MTA–PPKE Hungarian Language Technology Research Group, Budapest, Hungary
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
[3]Pázmány Péter Catholic University, Faculty of Humanities and Social Sciences
{indig.balazs,vadasz.noemi,kalivoda.agnes}@itk.ppke.hu

Abstract. On the basis of the literature about human sentence processing we examined the parsing process from two aspects. With the help of a sentence-completion experiment we show that there is a strong relationship between the entropy of the words in the sentence and the look-ahead window of a two-phase sentence processing model. The result of our experiment showed that people intend to close the verbal complex and the noun phrase as soon as possible and our corpus-measurements support that it happens in a trigram window.

Keywords: natural language processing, psycholinguistics, entropy

1 Introduction

Natural language processing (NLP) is the task of handling human languages with the aid of computers. Unfortunately, in complex tasks, such as machine translation, computers are far behind the human alternative even though to train a system, scientists use more text than one would see in a life. We think that the difference in performance is based on the main principles of how the two parsers or translators (machine and human) work.

Our parsing model, AnaGramma [6, 12] is a performance-based, psycholinguistically motivated system following the patterns of human language processing as much as possible. The model utilizes a strictly left-to-right approach for processing the input word by word imitating the language input.

In this paper we examine human sentence processing. We explore the first phase of a two-phase sentence processing model, on one hand in production – with a sentence-completion experiment based on entropy –, and on the other hand in perception – with measurements on multiple corpora using a look-ahead window.

As the theory of entropy is widely-used in other disciplines, its application introduced in the paper in conjunction with the look-ahead window can be directly applied in other fields as well[1]. After a short theoretical background we present our results from the experiment and the corpus measurements.

[1] For example in music, because it has essential relationship with natural language. Similarly to language, in music perceptually discrete events are structured into

2 Background

2.1 Sausage Machine

The AnaGramma system aims to model the human language processing based on the *Sausage Machine* where the parsing process consists of two main phases. The first phase is – as Frazier and Fodor [5] calls it – the *Preliminary Phrase Packager* which assigns the lexical and phrasal nodes to groups of words within the string input. In the second phase, these packaged phrases get their roles in the sentence by adding non-terminal nodes, this phase is called the *Sentence Structure Supervisor.*

Frazier and Fodor [5] set a window of roughly six words which is used in the first phase of the sentence processing for preparing the packages for the second phase. In Section 4 we prove that for Hungarian a trigram window is enough due to its agglutinative nature. As human parsers try to bind the arguments of the verb as soon as possible [7], they sometimes fail and therefore need to backtrack. The most extreme manifestations of reanalysis are garden path sentences as in Example 1.

(1) The horse raced past the barn fell.

During the reading of these garden path sentences word by word we need to backtrack which increases the time required to understand them. It is related to Kimball's *principle of Fixed Structure* [7], which claims that *'recalling a shunted phrase out of memory to restructure it is costly'.*

2.2 Entropy

Traditionally, entropy is a quantitative measure of the *randomness* of a system. For example a brand new deck of cards has low entropy since it is ordered, and a shuffled one has high entropy. Shannon and Weaver introduced entropy into information theory [13, 14]. Miller was trying to show that statistical approximations to English have a predictive value for sentence recognition [10].

According to Shannon and Weaver *information* is the measure of one's freedom of choice when one selects a message [14]. Natural language that produces a sequence of symbols (letters and phonemes) according to certain probabilities is a *stochastic* process, and when the probabilities depend on the previous events it is called a *Markov chain.*

On the level of words this probabilistic behaviour of natural language works as well: *'...If we are concerned with English speech, and if the last symbol chosen is 'the', then the probability that the next word be an article, or a verb form other than a verbal, is very small. This probabilistic influence stretches over more than two words, in fact. After the three words 'in the event' the probability for 'that' as the next word is fairly high, and for 'elephant' as the next word is very low.'* [14].

hierarchical sequences according to syntactic principles [9]. According to this, music can also be observed from this aspect.

Pléh et al. attempted to demonstrate the relevance of information theoretical accounts to understanding word recognition and morphological processing in Hungarian [11]. Their work is based on that of Antal [1, 2] who used the entropy notion developed by Shannon and Weaver [14] for equal probability outcomes where entropy is a function of the number of possible outcomes. See these papers for details on statistical complexity and entropy of language.

Morphological boundaries influence the degree of this monotonous decrease and intuitively correspond to slowing declining (locally increasing entropy values). Figure 1 shows the entropy values over the graphemes of a morphologically complex word [2], the entropy value gradually decreases over the stem, and then a suddenly increasing entropy indicates a morpheme boundary.

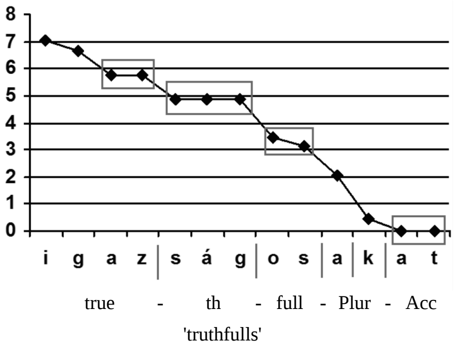

Fig. 1. Entropy value changes over a multiple suffixed Hungarian noun

Entropy can be captured between not only morpheme boundaries but bigger units like words as well. In the following section we show an experiment in which the effect of entropy fluctuation is measured on subsequent words in order to show empirically the inner-workings of the first phase of the Sausage Machine [5] for Hungarian to shed light on this detail of human parsing.

2.3 Corpora

We have made our measurements in two corpora:

The InfoRádió Corpus contains short Hungarian political and economical news. Each utterance here consists of a title and a body containing 2-3 sentences that describe a single political or economic event. The corpus of 54.996 leads containing 135.587 sentences and 1.953.419 is tokens taken from a news portal's RSS feed (www.inforadio.hu).

[2] The example and the figure is from Pléh et al. [11]

The Pázmány Corpus [4] consists of Hungarian texts collected from the internet. The downloaded texts form the basis of the Pázmány Corpus with 1.2 billion tokens from more than 30 000 domains.

3 An entropy experiment

We performed an on-line test which focused on the entropy in processing a sentence word by word. We were curious to see how some words constrain the possible continuation of a sentence. In order to achieve this, the participants could see a Hungarian sentence with each successive word revealed one after another; each time they had to guess the next word in no more than 20 seconds. After their guess the solution has appeared and they had another 20 seconds to guess the next word, seeing its right context.

This test simulates how the human parser processes a sentence left-to-right and word-by-word skipping uninformative words and making predictions to speed up reading which is to be modeled in AnaGramma. Our prediction regarding to the meaning of a sentence is not the most important factor of language processing, however, 'there is good evidence[3] that expectancy generation plays a role in language comprehension' [3].

60 participants were involved in our test: 45 women and 15 men. The average age was 32 years, more than half of the participants had a university degree. Their reaction times (the time they needed to type their guess) were measured as well[4]. In the following section we present two sentences from the experiment and discuss the results.

3.1 Results

Two sentences from our test are presented below. Figure 2 and 3 show the reaction times needed to guess the next word by every segment of the sentences. The figures are followed by a detailed description of the results.

The first word (Figure 2) was a verb in imperative, *Térjünk* 'Turn+Imp+p2'. Most of the participants thought that a detached preverb in post-position could be the next word. 40% guessed the preverb *vissza* 'back', thus implying the meaning *visszatér* 'to return', with a reaction time of 5.3 seconds which is really fast. In case of *Térjünk még* (még 'still' is just a filler-word) the guessed word was *vissza* 'back' again, now with 50%. Other preverbs appeared as well (e.g. *ki* 'out', *be* 'in', *le* 'down').

[3] EEG experiments show that the N400, a component of EEG signals is sensitive to semantic and structural priming. This amplitude is high when an already introduced fragment of a sentence is followed by a word that is not related to it [8].

[4] The test was performed under uncontrolled circumstances. Something may have diverted the participant's attention or he/she may have typed slowly. Nevertheless, thanks to a sufficient number of participants, these time frames provide valuable information regarding to the sentence processing tendencies.

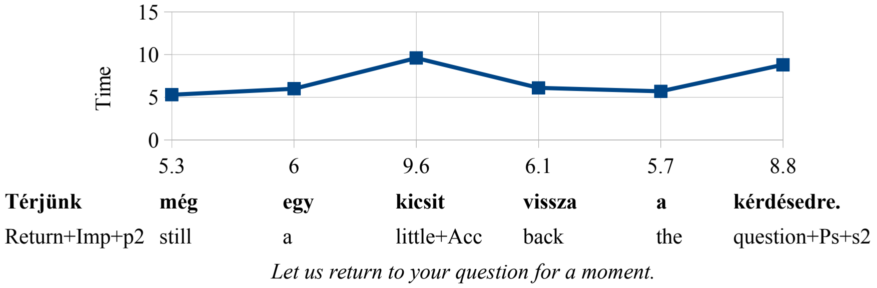

Fig. 2. Average reaction time influenced by the newly appearing word (The translated sentence and separately the translation of each individual word according to the full sentence i.e. after the part-of-speech and word-sense disambiguation are displayed.)

After *egy* ‘a(n)’ indefinite article appeared, the reaction time increased with another 3.6 seconds: the participants thought of a collocation (*egy kicsit* ‘a little’) or wrote a noun which indicates a direction (where to turn to). After the word *kicsit*, only preverbs were guessed: *vissza* ‘back’ had 60% at this point. This happened because the resulting sentence is not typical, it would be more natural if the preverb would follow the verb immediately. This tendency is quite obvious, because some participants reported that they became irritated when the newly appearing word was not a preverb, even though their guess was a preverb for the second time already, because they wanted to close the verbal complex as soon as possible.

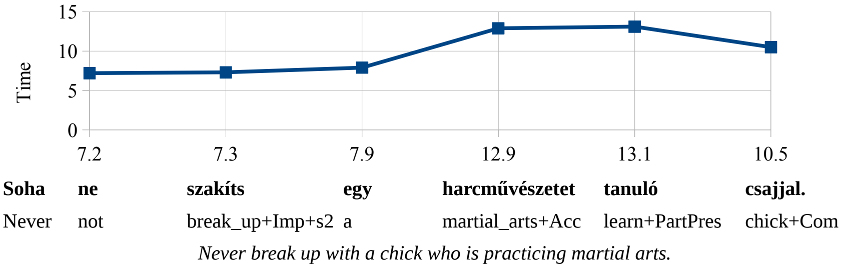

Fig. 3. Average reaction time influenced by the newly appearing word (The translated sentence and separately the translation of each individual word according to the full sentence i.e. after the part-of-speech and word-sense disambiguation are displayed.)

In step 1. of the second sentence (Figure 3 and Table 3.1) the starting word was *Soha* ‘Never’. Most of the participants tried to continue it by extending the negation with the following words: *nem/ne* ‘not’, *sem/se* ‘not even’, *többé/többet*

Table 1. The second sentence (Figure 3) as the participants saw it step-by-step (2nd column). The words that have just appeared are in **boldface**. The 3rd column is the translation of the last words according to their left-context.

1.	**Soha**	‘Never’
2.	Soha **ne**	‘not’
3.	Soha ne **szakíts**	for the meaning see Table 3.1
4.	Soha ne szakíts **egy**	‘a’
5.	Soha ne szakíts egy **harcművészetet**	‘martial_arts+Acc’
6.	Soha ne szakíts egy harművészetet **tanuló**	‘learn+PartPres’
7.	Soha ne szakíts egy harművészetet tanuló **csajjal**	‘chick+Com’

‘no more’, *sehol* ‘nowhere’, *napján* ‘never ever’. They needed 7.2 seconds on average to make this decision and type in the answer, this could be regarded as a fast reaction. This can be explained with the frequent co-occurrence of *soha* ‘never’ and these words.

In step 2. the starting word was extended with the next word resulting in the sequence *Soha ne*, ‘Never ever’ (literally ‘Never not’). Every participant thought of a verb in imperative, second person singular, mostly *mondd* ‘say’, *tedd* ‘do’, *gondold* ’think’ – these tend to be frequently used warnings and requests. The average reaction time was nearly the same as before (7.3 seconds).

Step 3. presented the sequence *Soha ne szakíts.* The resulting sequence is quite complex, because the meaning of the verb *szakít* depends on the particle used with it, furthermore, it can stand without a particle or form collocations like *szakít időt* ‘to find the time to do sth’[5] (see Table 3.1). Almost half of the participants (27 people) thought of *szakíts félbe* (‘interrupt’+Imp+s2), 12 people gave an answer that suggests the meaning ‘to break up with sb’ where the Hungarian verb doesn’t have a particle but an argument in instrumental case. Some of the answers imply the meaning ‘to tear apart sth’ (*szakíts szét, szakíts ketté*) and ‘to tear off sth’ (*szakíts le*). Due to this large amount of options we can see a slightly increasing reaction time (7.9 seconds). It is caused by the predicted – however, different – verb modifier as a consequence of the intention to close the verbal complex as soon as possible.

In step 4. the appearance of *egy* (‘a’, indefinite article) caused a sudden increase in reaction time (+5 seconds), 6 participants didn’t write a new guess at all. Those who were waiting for a verb modifier are now forced to backtrack to the beginning of the sentence to correct the path of the parse which needs time. Furthermore, the indefinite article indicates the beginning of a noun phrase which can continue in many ways. This high entropy regarding to the possible continuation is the cause of increased reaction times and 43 different answers[6].

[5] The more common word order would be *időt szakít* or *félbeszakít* (in this case, *félbe* functions as a prefix).

[6] The most common answers were *virágot* (‘flower’+Acc, resulting in ’Never tear off a flower’, guessed by 7 people) and *nővel* (‘woman’+Com , resulting in ‘Never break up with a woman’, guessed by 3 people).

Table 2. The argument structures of the Hungarian *szakít* verb and some of its possible verb modifiers

(verb modifier) + verb	meaning	arguments		
szakít$_1$	to break up with	Nom	Com	
szakít$_2$	to tear	Nom	Acc	
meg + szakít	to cut off	Nom	Acc	
félbe + szakít	to interrupt	Nom	Acc	
ketté + szakít	to tear in two	Nom	Acc	
szét + szakít	to tear apart	Nom	Acc	
le + szakít	to tear off	Nom	Acc	Del
ki + szakít	to pluck	Nom	Acc	Ela
el + szakít	to tear apart	Nom	Acc	(Abl)

The task became even more difficult when the next word of the sentence appeared in the 5. step. 17 people wrote *félbe*, even if this answer doesn't have a reasonable explanation. It would indicate the meaning 'Never interrupt [a] martial arts' which has a semantic incompatibility. The explanation behind the results is that two phenomena are opposed to each other: the content does not match the semantic expectations[7] and the urge to place the verb modifier, which has been proven stronger. 11 people didn't guess. Only less than half of the participants recognized this as the beginning of a complex noun phrase. They wrote present participles, e.g. *tanuló* 'someone who studies sth', *ismerő* 'someone who knows sth', etc. The highest reaction time (13.1 seconds) can be seen at this point. It is caused caused both by the aforementioned opposing phenomena and the complex noun phrase containing a participle.

In step 6. (final step), participants guessed a noun in comitative or accusative case. The latter can be explained as the participant's plan to add the verb modifier *félbe* to the end of the whole sentence (as it was unclear when the sentence will come to an end)[8], thus he or she chose the meaning 'Never interrupt [someone] who is practicing martial arts!'. The former case (words in comitative case (58%)) imply the meaning 'Never break up with a [man/girl/friend/person] who is practicing martial arts!'. Both solutions make sense, however, the decision depends on whether the participant is influenced by his or her answers in the earlier steps. The average reaction time decreased (10.5 seconds) due to the semantic and structural constraints of the context. Still, the time needed is high, because of the urge to complete the complex NP and identify it as an argument of main verb.

[7] The existence of this phenomenon has been verified using EEG experiments by [8].

[8] This form is unlikely in edited texts, and has minimal occurrences in unedited ones, however, some participants were desperately stuck at the form *félbeszakít.* This phenomena is responsible for these rare forms found in corpora.

3.2 Discussion

The diversity of the answers and the length of reaction times show how easily and accurately the next item (the word itself or at least its part of speech without its adjuncts) can be predicted based on the context already known matching the concept of entropy [14]. In the results two trends can be observed: Firstly, when the lexical elements are predominant because of the strong lexical collocations like in *Soha ne* and *Térjünk vissza/rá.* We also have found that when a collocation can not be ruled out, the participants' decision strategy was risky and fast by choosing a collocation to speed up parsing like in *Soha ne **[imperative]**.*

Secondly, in view of the results we can state that the participants tried (1) to choose a verb modifier for the verb as soon as possible (*Térjünk **[verb modifier]***), and (2) to close the NPs quickly with a case marker corresponding to the requirements of the verb (*Soha ne szakíts **[commitative/accusative]***) even though it is less constrained lexically, because the appropriate category is more important. The aforementioned intention of closing the verbal complex and the NP is related to the Preliminary Packaging Phase [5].

As a side-effect, we empirically have found a way how the rare occurrences of far-strolled verb modifiers (see Section 4.1) are created. This rare case happens when the intentions to close the verbal complex and not to distract the more meaningful constituents are conflicting. This problem is usually solved by putting the verb modifier right after the verb, but when this phase is missed for some reasons in production, the other constituents become more important and the verb modifier is held back till the end of the sentence. In the following section we will show how the aforementioned phenomena are manifested in real word texts. To achieve this, we use written corpora.

4 Window in parsing

In the previous section we presented an experiment which helped us to capture the phenomena of word-level entropy. To verify Fodor and Frazier's statement [5] that the size of the window used in the first phase of the Sausage Machine is 'approximately six words' we used corpora and measured the right detached verb modifier and the nominative posessor–possessum distance for Hungarian. As Hungarian is an agglutinative language (and most of the information is stored in morpheme suffixes of the content words in contrast to the many function words and fixed word-order in English) we show that a narrower window is enough. Based on our preliminary experiments we set the size to 'three content words'. In those cases where this window is not enough we show that it is likely that another strategy is used for parsing.

4.1 The verb and its preverb

Figure 4 shows the distribution of the Hungarian verbs that can bear preverb, grouped by the number of their possible preverbs. A little more than half of these

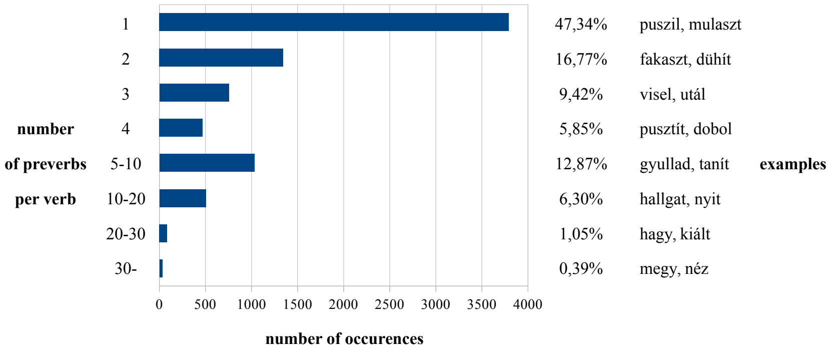

Fig. 4. Number of possible preverbs per verb. A little more than half of the Hungarian verbs that have preverbs can bear various preverbs in the sentence.

verbs can take various preverbs in the sentence, which is crucial because the verb itself can have more than one possible – but contradicting – argument structures (see Table 3.1 for example in the previous section) at this point of processing. The appearance of a preverb after the verb itself can filter impossible argument structures and prune false branches of the analysis resulting in faster parsing.

We measured the distances between verbs and their right detached preverbs[9] on the InfoRádió Corpus. Table 4.1 summarizes our findings.

Table 3. Positions of post-verbal detached preverbs – In edited texts 99% of the detached preverbs appear immediately after the verb, even in unedited texts the maximum two tokens after the verb contain the 99% of preverbs

FIN	+1	+2	+3	+4	+5	+6	+7
SUM	23.552	220	-	-	-	-	-
%	99.999%	0.001%	-	-	-	-	-

In the InfoRádió corpus there is no example of a preverb following the verb at a distance larger than two positions[10].

Our results show that a trigram window is enough for connecting the verb and its right detached preverb. It means that in the first phase of the Sausage Machine

[9] In Hungarian the preverb can take various places: (1) on the verb as a prefix, (2) detached to the left, (3) detached to the right.

[10] There is a low number of examples of more than two positions distance, but all of them are caused by tagging errors which were not counted.

the verbal complex is completed and its argument structure is disambiguated, so the speed-up effect of the window can prevail.

4.2 Possessive structure

In Hungarian there are two syntactic possessive constructions. When the possessor is in nominative case the possessum can be modified by numerals and/or adjectives, but it cannot take an article. It means that the possessor and the possessum form an NP, their order is fixed and the verb can not intervene. In possessive contructions the possessum agrees with its possessor in person and number.

(2) Peti kutyája
Peter+N+Nom dog+N+s3
Peter's dog

When the possessor is in dative case there are two individual NPs for it and its possessum. This means that their order is not fixed[11] and the verb can intervene.

(3) Petinek elveszett a kutyája
Peter+N+Dat lose+V+Past the+Article dog+N+s3
Peter's dog is lost.

The most neutral realization of the iterated possessive construction is when the first possessor is in nominative case, and the second one is in dative case.

(4) Peti kutyájának a nyakörve
Peter+N+Nom dog+N+s3 the+Article collar+N+s3
Peter's dog's collar

We measured the distance between the possessor and the possessum in nominative case using the Pázmány Corpus. We were looking for structures that start with a word in nominative case and end with the closest word having a possessive affix. The possessor's position was marked with 0 and the position of the possessed was determined automatically, compared to the possessor. With this method more than 7.700.000 phrases were matched. Figure 5 shows the positions of possessed, given in percent.

The +1 position of the possessum covers 52.46% of the cases. It means that the possessor is followed immediately by the possessum (e.g. *Chopin művei* '[the] works of Chopin'). The +2 position has 27.45%. The intervening word is usually an adjective (e.g. *Bozsik legfontosabb tulajdonsága* 'Bozsik's most important property') or a numeral (e.g. *ingatlanok 10 százaléka* '10 percent of [the] real estates'). 10.66% goes to the +3 position. The two intervening words are mostly an adverb and an adjective/numeral (e.g. *népesség csaknem 60 százaléka* 'almost 60 percent of the population'), sometimes a complex substantive derived from a verb (e.g. *döntések hatályon kívül helyezése* 'repeal of decisions').

[11] The most neutral is the possessor-possessum order, the reverse is still grammatical but somehow marked

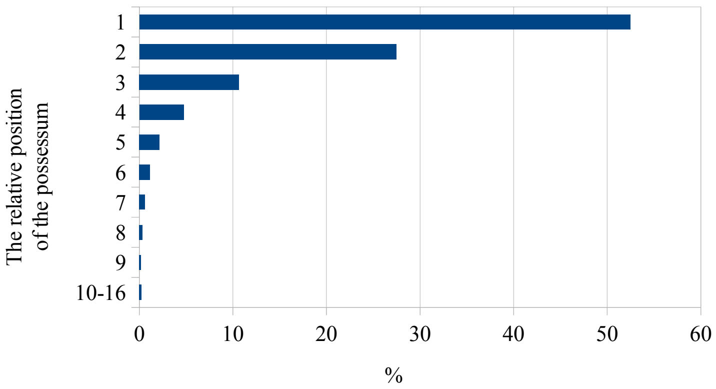

Fig. 5. The distance of the nominative possessor and its possessum (in nominative)

The summarized frequency of the latter positions is less than 10%. These phrases usually include enumerations or a participle where the derived verb keeps its arguments and adjuncts to its left (see Example 5 found in Pázmány Corpus (+16 position of the possessum[12]).

(5) Krisztina különleges , Swarovski kristályokból és minőségi
Kristina+N special , Swarovski crystal and high_quality
japán gyöngyökből készült , egyedi tervezésű romantikus , nőies
Japanese pearl made , unique designed romantic , feminine
nyaklánca
necklace+Ps+s3
'Kristina's special, romantic, feminine necklace with unique design, made of Swarovski crystals and high quality Japanese pearls'

As we can see, a trigram-window is sufficient in 80% of the possible cases. Even if the possessor stands without a suffix that could indicate the grammatical case (so its analysed as nominative), the parser is able to make a decision whether the word is a possessor or not. If a word having a possessive affix can be found in the window of the word originally marked as nominative, it is highly possible that the word without grammatical case is actually the possessor.

In 20% of the measured cases, there is more than one intervening word within the possessive structure. More than half of these cases include co-ordinations (enumeration of the possessum's attributes), and the presence of embedded participles is frequent as well. In case of these complex NPs, a decision about the possessor's role can not be made with the help of a trigram-window but in a latter phase of processing. Even so, we have to emphasize that this problem does not occur in four fifths of the possessive structures. The large distance between

[12] Punctuation marks are counted as separate tokens.

the possessor and the possessum occurs rather in – mostly formal – written texts. We assume that there is an other parsing strategy for handling these long-term dependencies which is a topic of an other research.

5 Conclusion

The data extracted from the corpus are consistent with the results of our entropy experiment. The human processor tries to close the different phrases as soon as possible, so they will appear in a trigram-window. Therefore the first phase of the Sausage Machine can be observed both in production and perception.

References

1. Antal, L.: A megnyilatkozások tagolása morfémák szerint. Magyar nyelvőr 86(2), 189–202 (1962)
2. Antal, L.: A formális nyelvi elemzés. Gondolat, Budapest (1964)
3. Elman, J.L.: An alternative view of the mental lexicon. Trends in Cognitive Sciences 8(7), 301 – 306 (2004), http://www.sciencedirect.com/science/article/pii/S1364661304001366
4. Endrédy, I.: Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz (Language Tehcnology Algorithms for Automatic Corpus Building and More Precise Data Processing). Ph.D. thesis, Pázmány Péter Catholic University Faculty of Information Technology, Budapest (6 2016), in Hungarian
5. Frazier, L., Fodor, J.D.: The sausage machine: A new two-stage parsing model. Cognition 6(4), 291–325 (1978)
6. Indig, B., Prószéky, G.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott nyelvtudomány 15(1-2), 29–44 (2015)
7. Kimball, J.: Seven principles of surface structure parsing in natural language. Cognition 2(1), 15–47 (1973)
8. Kutas, M., Hillyard, S.A.: Brain potentials during reading reflect word expectancy and semantic association. Nature 307(January 12), 161–163 (1984)
9. Lerdahl, F., Jackendoff, R.: A generative theory of tonal music. The MIT Press, Cambridge. MA (1983)
10. Miller, G.A.: Language and Communication. McGraw Hill, New York (1951)
11. Pléh, C., Németh, K., Varga, D., Fazekas, J., Várhelyi, K.: Entropy measures and predictive recognition as mirrored in gating and lexical decision over multimorphemic hungarian noun forms. Psihologija 46(4), 397–420 (2013)
12. Prószéky, G., Indig, B., Vadász, N.: Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In: Bence, K. (ed.) "Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére, pp. 223–232. MTA Nyelvtudományi Intézet, Budapest (2016)
13. Shannon, C.E.: A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5(1), 3–55 (Jan 2001), http://doi.acm.org/10.1145/584091.584093
14. Shannon, C.E., Weaver, W.: A Mathematical Theory of Communication. University of Illinois Press, Champaign, IL, USA (1963)