Radek Čech, Barbora Benešová, Ján Mačutek

Why does negation of the predicate shorten a clause?

Abstract: According to the Menzerath-Altmann law, the mean word length is greater in shorter clauses than in longer ones. In Czech, negation is mostly realized by adding the prefix ne- to the beginning of the word, which makes the word longer (and, consequently, it also increases the mean word length in the clause). Therefore, we predict that clauses in which the predicate is in the affirmative form are longer than ones with the negative predicate. We test the hypothesis on a sample of 59 pairs of affirmative and negative forms of the same verb from the Prague Dependency Treebank 3.0.

Keywords: clause length, negation, Menzerath-Altmann law

1 Introduction

Language is a complex system composed of many units of different kinds which interact with each other. The complexity of the system seems to be the cause of the difficulty connected with describing the system and with explaining its properties. However, some properties of many different complex systems are results of relatively simple "mechanisms" (Barabási & Albert 1999, Newman 2010) which have a decisive impact on the system behaviour. Capturing some of the mechanisms in the form of an empirically testable law (or a hypothesis, at least) enables a verification of its validity and opens a way towards an explanation of system properties.

In this paper, we analyse the impact of the Menzerath-Altmann law (Cramer 2005, MAL hereafter), which expresses a very general mechanism controlling a relation between the sizes of language units (for details see Section 2), on certain grammar characteristics. According to the MAL, there is a systematic relation between lengths of language units belonging to the neighbouring

Acknowledgement: J. Mačutek was supported by the grant VEGA 2/0096/21.

Radek Čech, University of Ostrava, e-mail: cechradek@gmail.com
Barbora Benešová, University of Ostrava, e-mail: benesovaba@seznam.cz
Ján Mačutek, Mathematical Institute, Slovak Academy of Sciences and Constantine the Philosopher University in Nitra, e-mail: jmacutek@yahoo.com

https://doi.org/10.1515/9783110763560-001

levels in the language unit hierarchy – in a simplified way, the longer the "higher" unit, the shorter the mean length of the "lower" unit. It means that a change in length of one unit should cause a change in length of the unit from the neighbouring levels (e.g. if words are made longer, clauses are expected to become shorter).

We focus on negation in Czech which is (not exclusively, but in the vast majority of cases) realized by adding the prefix *ne-* to the beginning of a word, e.g.

(1) *Marie přišla,*
[Mary came]

(2) *Marie* ***ne****přišla.*
[Mary did not come]

We will consider only this realization of the negation in our paper.

This prefixation means that the word becomes one syllable (and also one morpheme, but we measure word length in syllables in this paper) longer, which, in accordance with the MAL, should generally make the clauses with negated forms of the word shorter than ones containing the same word without the negative prefix. It must be emphasized that the MAL is of a stochastic character and the law represents a general tendency. Thus, some instances which do not follow the law are admissible – as an example, see clauses (1) and (2) which have the same length (counted by the number of words) despite the fact that the mean word lengths in syllables (and also in morphemes) differ. In other words, the validity of the stochastic law is manifested on a large sample and the existence of some counterexamples does not mean a violation of the law (as is the case with a deterministic law).

The aim of the study is to test the following hypothesis based on the MAL: Clauses with the negative form of the predicate contain (on average) fewer words than clauses with the affirmative form of the predicate.

The paper is organized as follows. Section 2 provides a brief description of the MAL, focusing on its realization at the level of clauses and words. The methodology we applied and the language material which was used are presented in Section 3. Section 4 brings the results, together with their interpretation for some verbs which do not behave according to the hypothesis. Finally, Section 5 concludes the paper with a short discussion.

2 Menzerath-Altmann law

The MAL was formulated for the first time by the German linguist Paul Menzerath as a relation between the duration of sounds in a syllable and length of the syllable in which the sounds occur (Menzerath 1928 – longer syllables consist of relatively shorter sounds), and later as a relation between word length and length of syllables in the word in a dictionary (Menzerath 1954 – the more syllables a word contains, the shorter its syllables are on average).

It is, however, valid much more generally, presumably for all immediate neighbours in the language unit hierarchy (such as e.g. phoneme – morpheme/syllable – word – clause – sentence), see examples provided by Altmann (1980) and by Cramer (2005). Mačutek et al. (2019) suggested a very general formulation of the law as follows: *The mean size of constituents is a function of the size of the construct*, where a construct is a higher language unit (e.g. word) composed of constituents, i.e. lower-level units (e.g. syllables). The usual mathematical expression of this general form of the MAL is

$$y(x) = ax^b e^{-cx}, \tag{3}$$

with $y(x)$ being the mean size of constituents if the size of the construct is x; a, b, and c are parameters. Very often, a more simple formula is sufficient, namely

$$y(x) = ax^b, \tag{4}$$

with parameter b attaining a negative value. It is a special case of formula (3) for $c = 0$. The function (4) is decreasing, and the MAL can be reworded as *the longer the construct, the shorter its constituents* (which is true also for the original Menzerath's observations).

The MAL is valid, among others, also for the relation between sentence length (measured in the number of its clauses) and clause length (measured in words), see Köhler (1982) for English, Heups (1983) for German, and Teupenhayn and Altmann (1984) for eight languages (German, English, French, Swedish, Hungarian, Slovak, Czech, and Indonesian). For all texts under study, the mean clause length decreases with the increasing sentence length, i.e. formula (4) can be used to model the relation.

The MAL can be interpreted in a context of both Zipf's principle of least effort (Zipf 1949) and the synergetic linguistic theory (Köhler 2005). Specifically, the MAL expresses the mechanism which controls proportions of lengths of units of different linguistic levels. These proportions can be seen as a result of the speaker's and hearer's communication requirements which are determined

by a strategy to achieve a communication goal(s) with the least effort. In other words, the co-existence of these requirements leads to a dynamic equilibrium among lengths of linguistic units.

Specifically, negation makes the predicate verb one syllable longer. The equilibrium is thus shifted and the speaker is "forced" – by the least effort requirements – to "find" it by making shorter the clause in which the negative predicate occurs. On the contrary, the use of affirmative predicate verbs allows for longer clauses. Thus, if the hypothesis is not falsified, we can state that there is a systematic relation between negation of the predicate and length of the clause, the fact which has not been observed yet, to our knowledge.

3 Methodology and language material

A clause is considered as the construct that consists of units of a lower degree that constitute it, i.e. words. Within the MAL context, the word is traditionally regarded as the closest lower unit to the clause (see Section 2), however, some approaches have accepted the syntactic phrase as the neighbouring level (cf. Mačutek et al. 2017). Although there are many definitions of the clause, they do not usually differ essentially and they share some crucial features. For instance, according to Crystal (2003), the clause is "a unit of grammatical organization smaller than the sentence, but larger than phrases, words or morphemes". Similarly, in the Prague Dependency Treebank 3.0 (Bejček et al. 2013, PDT 3.0 hereafter), which is used for this analysis (see below), clauses are defined as "grammatical units out of which complex sentences are built. A clause typically corresponds to a single proposition expressed by a finite verb and all its arguments and modifiers (unless they constitute clauses of their own)" (Mikulová et al. 2013). This annotation of clause is used in our analysis. As for the lower unit, i.e. the word, we determine it in accordance with the annotation of the PDT 3.0, which means that a word is identified as a sequence of letters between spaces.

In this study, only predicates were chosen for the analysis, mainly because the predicate constitutes a root of a clause structure, according to dependency grammar formalism (Meľčuk 1988, Hudson 2010, Osborne 2019), and, consequently, it has a decisive impact on the clause structure, including its length. Further, we decide to compare average lengths of clauses containing different forms (affirmative vs. negative) of the same verb. Specifically, clauses containing either affirmative or negative form of certain verbs were chosen, creating pairs that enabled measuring and testing the differences in size of clauses containing the same verb, but either with or without negation. Fifty-nine pairs of

verbs were tested altogether, with the minimal number of occurrences of each form of the verb in the PDT 3.0 being twenty. We distinguish among different word forms (for verbs in Czech they can differ for different grammatical categories such as person, gender, number, tense, etc.), i.e. we did not lemmatize the sentences.

4 Results

The mean lengths of the clauses which contain the affirmative and negative forms were enumerated and compared for each of the 59 verbs from our sample (see Section 3). The significance of the differences between them were statistically tested by the Mann-Whitney-Wilcoxon test (the Student t-test cannot be applied because the data are not distributed normally, as was shown by the Shapiro-Wilk test). Results are presented in Tab. 1, where AFF is the affirmative verb form in Czech, ENG its English equivalent,[1] f(AFF) is the frequency of the affirmative form in the PDT 3.0, NEG is the negative verb form in Czech, f(NEG) the frequency of the negative form in the PDT 3.0, CL(AFF) and CL(NEG) are the mean lengths of clauses which contain the affirmative and the negative verb form, respectively, and p is the p-value of the test. Verbs which do not behave according to our hypothesis (i.e. the mean clause length is higher for their negative forms) are highlighted in bold.

For 48 out of 59 verbs, the hypothesis from Section 1 is corroborated, as the clauses with the affirmative forms of these verbs are on average longer (the p-value of the Mann-Whitney-Wilcoxon test is below 0.05 for 24 verbs, below 0.01 for 15 verbs).

The remaining 11 verbs contradict the hypothesis. However, this behaviour can be explained for the majority of them. For instance, the only case when the clauses containing the negated predicate are significantly longer than clauses with the affirmative predicate is represented by pair of verbs *myslím* [I think] and *nemyslím* [I do not think]. Formally, the word *myslím* [I think] is a finite verb which has the function of the predicate. However, a closer observation of particular clauses reveals that the word *myslím* has two different grammatical functions in the sample. In many sentences it is used as an adverb or a particle which expresses uncertainty of the statement (actually, it is a parenthesis). Grepl and Nekula (2017) provide as an example e.g. the sentence *Bude myslím pršet* (which,

1 Naturally, the English translations of the verbs depend on the context. We present here the most obvious “dictionary translations”.

Tab. 1: Clause length for affirmative and negative forms of verbs.

AFF	ENG	f(AFF)	NEG	f(NEG)	CL(AFF)	CL(NEG).	p
jsou	they are	2641	nejsou	302	9.18	8.05	< 0.01
bude	he/she/it will	2199	nebude	309	10.08	8.74	< 0.01
byl	he was	1785	nebyl	202	9.54	8.50	< 0.01
má	he/she/it has	1783	nemá	272	9.45	8.08	< 0.01
bylo	it was	1390	nebylo	185	9.13	8.62	0.12
jsem	I am	1264	nejsem	41	7.07	4.95	0.12
jsme	we are	1250	nejsme	34	8.00	6.74	0.17
byla	she was	1200	nebyla	134	10.04	8.02	< 0.01
může	he/she/it can	1013	nemůže	203	10.15	9.09	< 0.01
budou	they will	896	nebudou	127	10.56	8.63	< 0.01
měl	he had/should	810	neměl	112	10.72	8.47	< 0.01
mají	they have	753	nemají	144	8.94	7.99	0.11
musí	he/she/it/they must	701	nemusí	120	9.12	8.81	0.51
byly	they were	679	nebyly	77	10.47	8.43	< 0.01
jde	**he/she/it goes**	**623**	**nejde**	**119**	**8.34**	**8.76**	**0.10**
měla	she had/should	563	neměla	65	11.30	8.57	0.01
mohou	they can	466	nemohou	88	11.09	8.57	< 0.01
měli	they (masc. anim.) had/ should	373	neměli	57	10.12	7.93	< 0.01
patří	he/she/it/they belong(s)	366	nepatří	24	10.97	6.79	< 0.01
byli	**they (masc. anim.) were**	**352**	**nebyli**	**29**	**9.22**	**9.34**	**0.96**
mělo	it had/should	300	nemělo	54	11.08	11.00	0.71
mohl	he could	289	nemohl	70	10.27	8.80	0.02
chce	he/she/it wants	287	nechce	54	9.69	7.87	0.02
měly	they had/should	268	neměly	29	11.91	9.93	0.06
znamená	he/she/it means	230	neznamená	32	7.28	4.63	0.06
platí	he/she/it pays/is true	219	neplatí	41	8.20	6.44	0.07
máme	**we have**	**212**	**nemáme**	**51**	**7.19**	**7.47**	**0.80**
stojí	he/she/it stands/costs	210	nestojí	21	9.23	6.33	< 0.01
podařilo	it succeeded	194	nepodařilo	44	12.55	10.89	0.02
došlo	it came	186	nedošlo	39	9.81	8.00	0.03
mohli	they (masc. anim.) could	184	nemohli	35	10.18	7.69	< 0.01
mohla	she could	180	nemohla	20	11.31	9.60	0.14
mám	I have	168	nemám	45	6.13	6.02	0.64
můžeme	we can	155	nemůžeme	50	8.99	6.82	0.02
chtějí	they want	152	nechtějí	31	8.78	7.45	0.19
budeme	we will	139	nebudeme	24	7.94	7.63	0.84
dá	he/she/it will give	139	nedá	63	7.77	7.19	0.43
chtěl	he wanted	131	nechtěl	35	8.37	8.06	0.87
existuje	he/she/it exists	122	neexistuje	63	7.75	6.24	0.17
myslím	**I think**	**122**	**nemyslím**	**28**	**1.81**	**2.82**	**<0.01**
stalo	he/she/it became	117	nestalo	24	8.58	4.79	< 0.01

Tab. 1 (continued)

AFF	ENG	f(AFF)	NEG	f(NEG)	CL(AFF)	CL(NEG).	p
šlo	**it went**	**116**	**nešlo**	**30**	**8.12**	**9.10**	**0.14**
chceme	we want	84	nechceme	25	7.35	5.56	0.12
chtěli	they (masc. anim.) wanted	84	nechtěli	25	8.19	6.24	0.06
ví	**he/she/it knows**	**82**	**neví**	**62**	**4.41**	**4.69**	**0.63**
vidí	he/she/it/they see(s)	67	nevidí	26	9.28	7.81	0.31
hrozí	he/she/it/they threaten(s)	61	nehrozí	33	8.43	6.48	0.04
víme	**we know**	**61**	**nevíme**	**27**	**3.18**	**3.56**	**0.77**
dokáže	he/she/it can do	54	nedokáže	20	9.09	8.60	0.67
mění	he/she/it/they change(s)	53	nemění	33	9.45	7.48	0.03
vím	I know	52	nevím	75	2.50	2.27	0.04
mohu	I can	47	nemohu	33	7.40	6.06	0.35
brání	**he/she/it/they defend(s)**	**44**	**nebrání**	**22**	**8.55**	**9.14**	**0.78**
chci	I want	40	nechci	34	5.30	4.76	0.53
zbývá	**he/she/it remains**	**38**	**nezbývá**	**24**	**7.55**	**9.71**	**0.13**
vědí	**they know**	**37**	**nevědí**	**22**	**5.32**	**5.59**	**0.47**
věděl	**he knew**	**34**	**nevěděl**	**20**	**4.59**	**5.10**	**0.48**
zná	he/she/it knows	29	nezná	22	7.31	6.36	0.63
souhlasí	he/she/it/they agree(s)	24	nesouhlasí	31	8.42	6.65	0.53

according to the PDT 3.0 syntactic formalism, consists of two clauses, *Bude pršet* and *myslím*) with the literal translation *I think it will rain*, but its meaning is *Probably it will rain* or *Supposedly it will rain* etc. In these sentences, *myslím* is annotated as a one-word clause in the PDT 3.0. Consequently, the mean length of clauses with the affirmative predicate decreases substantially.

Further, the biggest difference between clause lengths among verbs which contradict our hypothesis is observed for *zbývá* [he/she/it remains] and *nezbývá* [he/she/it does not remain], one can find. Here, for 21 out of 23 instances of the negative form, the verb occurs in the syntactic phrase *nezbývá než* [there is no other way but to . . .]. However, this syntactic structure (i.e., verb + *než*) is not attested in the affirmative form in the sample. Obviously, *nezbývá než* cannot be considered a "pure" negation of affirmative form *zbývá*.

A brief examination of all verbs which contradict the hypothesis suggests that considering only the presence or absence of the prefix *ne-* can be too rough a criterion to extract the affirmative and negative forms of verbs from a corpus, and that both semantics and phraseology play an important role. A finer-grained approach which would take them (and possibly other factors) into account can shed more light on the problem and thus enhance results achieved in this pilot study.

5 Conclusion

Our results indicate that the MAL captures a "mechanism" which has indeed a very strong impact on properties of language units. A slight increase in the mean word length (we remind reader that we focus solely on the predicate and its negation, which makes it one syllable or one morpheme longer) is reflected in shorter clauses in 48 out of 59 verbs (i.e. in more than 80% of them). For one half of them (24 out of 48), the difference is statistically significant if the significance level is set to be 0.05. The remaining 11 verbs display the opposite tendency (i.e. negative clauses are longer), but this fact can be explained by different functions which the affirmative and negative form of those verbs have in sentences. This is true especially for the verb *myslím* [I think] (see Section 4), the only one for which the negative clauses are even significantly longer.

The paper opens also several other problems which can be solved only after larger corpora from several languages are analyzed. First, verbs seem to have a special position in clauses, one could say that they are more important than other words (see e.g. Čech et al. 2011). It would be interesting to check whether negation of e.g. adjectives (which is in Czech realized mostly in the same way, i.e. by adding the prefix *ne-*) has the same effect. Second, negation of the predicate is realized by different means in different languages. In Czech, a one-syllable prefix is added, while e.g. in English or in French two one-syllable words are often needed (e.g. *I understand* vs. *I do not understand*, or *je comprends* vs. *je ne comprends pas*). The question is whether the "reaction" of clause length is "stronger" in such cases.

References

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn (ed.), *Glottometrika* 2, 1–10. Bochum: Brockmeyer.

Barabási, Albert-László & Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286(5439). 509–512.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek & Šárka Zikánová. 2013. „Prague Dependency Treebank 3.0." http://ufal.mff.cuni.cz/pdt3.0/ (accessed 18 October 2021)

Cramer, Irene M. 2005. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, 659–688. Berlin/New York: de Gruyter.

Crystal, David. 2003. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Čech, Radek, Ján Mačutek & Zdeněk Žabokrtský. 2011. The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A: Statistical Mechanics and its Applications* 390(20). 3614–3623.

Fowler, Henry W., Francis G. Fowler & Della Thompson. 1995. *The Concise Oxford Dictionary of Current English*, 9th edn. Oxford: Clarendon Press.

Grepl, Miroslav & Marek Nekula. 2017. Postojová částice. In Petr Karlík, Marek Nekula & Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [New encyclopedic dictionary of Czech]. https://www.czechency.org/slovnik/POSTOJOVÁČÁSTICE (accessed 18 October 2021)

Heups, Gabriela. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In Reinhard Köhler & Joachim Boy (eds.), *Glottometrika* 5, 113–133. Bochum: Brockmeyer.

Hudson, Richard. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Köhler, Reinhard. 1982. Das Menzeratsche Gesetz auf Satzebene. In Werner Lehfeldt & Udo Strauss (eds.), *Glottometrika* 4, 103–113. Bochum: Brockmeyer.

Köhler, Reinhard. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, 760–774. Berlin/New York: de Gruyter.

Mačutek, Ján, Jan Chromý & Michaela Koščová. 2019. Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics* 26(1). 66–80.

Mačutek, Ján, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni & Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, *Pisa, Italy, 2017*, 100-107. Linköping: Linköping University Electronic Press.

Meľčuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. Albany (NY): State University of New York Press.

Menzerath, Paul. 1928. Über einige phonetische Probleme. In *Actes du premier Congrès international de linguistes*, 104–105. Leiden: Sijthoff.

Menzerath, Paul. 1954. *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Mikulová, Marie, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková & Zdeněk Žabokrtský. 2013. „From PDT 2.0 to PDT 3.0 (Modifications and Complements)." https://ufal.mff.cuni.cz/pdt3.0/doc/tr54.pdf (accessed 18 October 2021)

Newman Mark E. J. 2010. *Networks. An introduction*. Oxford: Oxford University Press.

Osborne, Timothy. 2019. *A dependency Grammar of English: An Introduction and Beyond*. Amsterdam/Philadelphia: John Benjamins.

Teupenhayn, Regina & Gabriel Altmann. 1984. Clause length and Menzerath's law. In Joachim Boy & Reinhard Köhler (eds.), *Glottometrika* 6, 127–138. Bochum: Brockmeyer.

Zipf, George K. 1949. *Human Behavior and the Principle of the Least Effort. An Introduction to Human ecology*. Cambridge (MA): Addison-Wesley.