Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian

Balázs Indig and **Tímea Borbála Bajzát**
Doctoral School of Linguistics
National Laboratory for Digital Heritage
Eötvös Loránd University Department of Digital Humanities
indig.balazs@btk.elte.hu, bajzat.timi9696@gmail.com

Abstract

The quantitative turn in functional linguistics has emphasised the importance of data-oriented methods in describing linguistic patterns. However, there are significant differences between constructions and the examples they cover, which need to be properly formalised. For example, noun chains introduce significant variation in the examples, making it difficult to identify underlying patterns. The compression of noun chains into their minimal form (e.g. as they appear in abstract constructions) is a promising method for revealing linguistic patterns in corpora through their examples. This method, combined with identifying the appropriate level of abstraction for the additional elements present, allows for the systematic extraction of good construction candidates. A pilot has been developed for Hungarian infinitive structures, but is adaptable for various linguistic structures and other agglutinative languages.

1 Introduction

Functional construction grammars (Goldberg, 1995; Langacker, 2005) have recognised that human language consists of a network of symbolic form-meaning pairs (Langacker, 2008), which are influenced by frequency of use (Bybee, 1995). As a result, several methods have been developed and applied to support empirical exploration, using corpus-based and corpus-driven methods to identify linguistic patterns by considering frequency data, rather than relying on introspection, such as collocation metrics and collostructural analysis (Glynn and Robinson, 2014; Gries and Stefanowitsch, 2007; Luodonpää-Manni et al., 2017). The trend, which has led much of functional linguistic research to adopt a data-driven and quantitative approach, is referred to as the *quantitative turn* (Janda, 2013). In linguistic description, collecting datasets of sufficient quantity and quality, and processing them in an unbiased manner has become a clear challenge for linguists. However, without a well-structured technological apparatus, intuition distorts the objectivity of the query and the analysis used to identify patterns. The ability to validate patterns derived from theory using corpus-driven methods has, therefore, become a pressing issue. It is not sufficient for the results precisely match the theory, each step and the chain of reasoning must be examined to correct intuition.

Extracting constructions from corpora requires the processing of a large number of individual examples. We have found that the combinatorial explosion mainly arises from noun phrases that vary greatly in length and internal structure. However, the constructions we are looking for are not primarily concerned with the noun phrases they contain and, therefore, do not define their form precisely. To simplify entries, we compressed the noun phrases into their minimal form, which allowed us to extract shorter, more schematic patterns that better aligned with our theoretical expectations.

To test our hypothesis, we developed a rule-based method for compressing noun chains using practically POS tags, which we applied to the example clauses. The remaining elements in the constructions are replaced by either a word form, lemma or POS tag, depending on the expectations and their statistical variations, to achieve the best coverage. In this paper, we focus on the noun phrase compression method in light of the constructions found.

2 Data Sources

We used two corpora for the measurements, allowing us to compare and validate our method against overfitting. The first corpus, the *Hungarian Gigaword Corpus 2.0.5* (HGC2) (Oravecz et al., 2014), with 1.04 billion words, contains texts from six stylistic and five regional varieties. The second corpus, the *Hungarian Webcorpus 2.0*

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 96–103
November 28-29, 2024 ©2024 Association for Computational Linguistics

Type of the Auxiliary Verbs	POS	Original (HGC2)	Filtered (HGC2)	% (HGC2)	Original (HW2)	Filtered (HW2)	% (HW2)
Akar [‘want to’]	verb	610 836	419 324	68.65	650 000	518 123	79.71
Bír [‘can’, ‘endure’]	verb	22 191	15 387	69.34	179 846	112 112	62.34
Hajlandó [‘prone to’]	adj.	48 267	36 334	75.28	272 806	179 330	65.74
Képes [‘able to’]	adj.	134 843	86 833	64.40	650 000	462 001	71.08
Képtelen [‘incapable to’]	adj.	48 036	14 424	30.03	164 909	104 274	63.23
Kíván [‘wish to’]	verb	192 678	139 498	72.40	650 000	391 413	60.22
Mer [‘dare to’]	verb	63 729	39 177	61.47	473 966	278 887	58.84
Szeret(ne) [‘(would) like (to)’]	verb	484 448	278 834	57.56	650 000	448 324	68.97
Tud [‘can’]	verb	675 000	466 863	69.16	650 000	540 175	83.10

Table 1: The distribution of samples in the two corpora is as follows: From the HGC2, a total of 2 097 149 instances were collected (1 496 674 remaining after automatic filtering, constituting 71.37% of the total sample). From the HW2, 3 691 527 instances were collected (3 034 639 remaining after automatic filtering, constituting 82.21% of the total sample).

(HW2) (Nemeskey, 2020), has approximately 9 billion words and was obtained from the Web (Common Crawl). Both corpora are accessible through the *NoSketch Engine* corpus query framework (Kilgarriff et al., 2007). However, the amount of data that can be exported is limited, so the aim was to obtain as comprehensive a sample as possible within the system’s constraints.

We used the samples from an experiment conducted by Indig and Bajzát (2023) and extended them to include the most common modal auxiliaries with infinitives in Hungarian, as well as patterns of adjective + infinitive associated with modal meaning (Van linden, 2010), to get a more comprehensive picture. Table 1. summarises the sample sizes. The extracted concordances were re-analysed to ensure consistent and up-to-date annotation (Indig et al., 2019; Novák et al., 2016).

3 The Mosaic Method

The mosaic n-gram method can model linguistic data at different levels of abstraction, such as word, lemma, and POS tag, simultaneously (Indig, 2017). This concept aligns with the usage-based approach, as linguistic schemas become entrenched at various levels of sematicity, and patterns are conventionalised based on their frequency of use (Bybee, 1995). The mosaic n-gram method efficiently generates and ranks all possible abstractions of linguistic data, thereby reducing the reliance on linguistic intuition to identify statistically sound construction candidates that have the appropriate level of abstraction for each element to maximise coverage.

The method includes a classification step to identify *inferior mosaic n-grams* (i.e. subset relation): any less frequent mosaic n-gram that generalises from a set of examples that are a subset of those covered by a more frequent mosaic n-gram. Conversely, the more frequent n-gram is labelled as major compared to the less frequent one if the latter only covers examples that are also covered by the former. Among entries of equal frequency covering the same set of examples, all but the least abstract are deleted as redundant. In addition, by setting a frequency threshold to discard entries that are rare despite their abstraction, the processing time and the number of entries to be manually checked can be further reduced. This approach allows for a high degree of customisation, and the generated mosaic n-gram patterns can be easily converted into a query expression (e.g. the CQL in Sketch Engine (Kilgarriff et al., 2007) to check them in the corpus and explore linguistic data matching the pattern). In an abstraction of the method to additionally handle free word order n-grams are substituted for bag of words (Indig and Bajzát, 2023).

4 Compressing Noun Phrases

The formal definition of constructions typically includes only the bare minimum of noun phrases, as the focus is on the whole structure. When clauses that may contain elements at various levels of abstraction are automatically compared (e.g. element by element), it becomes difficult to match patterns of different lengths without introducing additional measures. However, noun phrases with modifiers can separate functionally and structurally similar samples (see the first three rows of Table 2).

N	Frequency	Example				
3	1688			[/N][Acc]	lemma:akar	[/V][Inf]
4	1103		[/Adj][Nom]	[/N][Acc]	lemma:akar	[/V][Inf]
5	1665	[/Adj][Nom]	[/Adj][Nom]	[/N][Acc]	lemma:akar	[/V][Inf]
5	1365	[/Det|Art.Def]	[/Adj][Nom]	[/N][Acc]	lemma:akar	[/V][Inf]
5	997	[/Det|Art.Def]	[/N][Nom]	[/N][Poss][Acc]	lemma:akar	[/V][Inf]

Table 2: All five entries could be reduced to '[/N][Acc] lemma [/V][Inf]' without violating syntax. (*akar* 'want to')

On the other hand, sequences of the same length with different modifiers in the noun phrases present a different issue. The last three rows of Table 2. show that, although the modifiers differ, the head noun appears in the same grammatical case, making the sequences practically analogous from our perspective. In summary, based on the examples presented above, it can be assumed that the majority of the found examples originate from such partially abstracted (i.e. not simplified) sequences.

We chose a rule-based approach because, in Hungarian, apart from a few well-separated cases, noun phrases can be trivially compressed using word order, morphology, and POS tags with simple regular expressions, and we can retain the property of converting the resulting patterns to CQL. A challenging case is possessives, where there are two ways of expressing the genitive function, both of which are homonymous. The first is marked by the nominative case, followed immediately by its property (e.g. *a* [/Det][Art|Def] *kutya* [/N][Nom] *háza* [/N][Poss.3Sg][Nom] 'the dog's house'). The second is expressed with the dative case (e.g. *a* [/Det][Art|Def] kutyának [/N][Nom] *a* [/Det][Art|Def] *háza* [/N][Poss.3Sg][Nom] 'the house of the dog'), which allows for flexible word order and even interruption (e.g. a *kutyának* [/N][Dat] *lefestette* [/V][Pst.Def.3Sg] *a* [/Det][Art|Def] *házát* [/N][Poss.3Sg][Acc] 'he/she painted the house of the dog'). While nominative homonymy is easy to handle because of its word order (Ligeti-Nagy et al., 2019) (see the last row in Table 2.), examples with the dative variant were excluded from our sample. This concerns 0.36% of instances (across 11 types) in HGC2, whereas 0.59% of instances (across 27 types) in HW2.

The rules are iterated in two steps, as running them simultaneously would produce incorrect results due to the aforementioned ambiguity of the nominative case. The first iteration specifically compresses cases where nouns with possessive suffixes are preceded by unmarked genitive cases (e.g. ***János*** **[/N][Nom]** ***könyvét*** **[/N][Poss.3Sg][Acc]** *olvasták* [/V][Pst.Def.3Pl] '(They) read the book of John' compared to *János* [/N][Nom] ***a*** **[/Det][Art|Def]** ***könyvét*** **[/N][Poss.3Sg][Acc]** *olvasta* [/V][Pst.Def.3Sg]) 'John read his book'). Then the remaining cases are processed to avoid overlaps. Finally, the aggregated frequencies of identical entries are calculated and classified according to their new lengths.

The procedure could be trivially refined and applied similarly to other agglutinative languages (e.g. Uralic languages). However, such rule-based transformations rely heavily on morphological markedness, and the results depend on both the language analysis tool and the quality of the corpus used (e.g. the amount of noise present).

5 Evaluation

We evaluated the results in two ways. First, we examined the changes in the distribution of frequencies and the total number of resulting entries. Next, we analysed how the most frequent and general patterns are reflected in the usage patterns found, which enables further analysis.

5.1 Changes in the top candidates

Our first concern was to see how the number of different patterns (types) associated with each length had changed. First, the longest patterns (N = 7 and above) have disappeared, likely because they contain noun phrases with two or more elements. There is also a noticeable decrease in the number of types for the shorter constructions, but the rate flattens out as the length decreases (e.g. N = 7 to 5). However, the opposite trend can be seen for N = 3 and 4 as a result of the compression. The number of types increases significantly for 4 grams, while the increase for 3 grams is more limited. Most auxiliaries exhibit these trends (see Table 3.).

In cases where significant differences were observed between the two corpora (e.g. *Hajlandó* N = 4), the discrepancy arises from the fact that the sam-

	N = 3		N = 4		N = 5		N = 6		N = 7	
	H	W	H	W	H	W	H	W	H	W
Akar	103.26	103.06	117.22	109.87	52.79	49.05	33.33	34.67	0	0.99
Bír	100	100	115.38	122.78	57.14	63.73	0	41.67	0	65.28
Hajlandó	100	112.50	200	163.41	87.10	60.55	85.71	76.47	0	5.00
Képes	108.33	103.92	268.42	229.63	63.33	47.64	5.26	21.95	0	3.26
Képtelen	125.00	107.14	150	116.67	25.00	52.94	0	5.88	-	0
Kíván	148.15	116.95	225.49	217.56	23.76	29.64	29.41	37.82	0	0.83
Mer	100	110.94	107.41	107.83	74.42	68.07	0	30.77	-	0
Szeret	111.32	103.40	119.21	119.29	28.92	34.34	6.25	10.80	0	0
Tud	111.25	109.52	137.65	137.46	53.92	50.47	40.25	39.34	2.04	0.88

Table 3: Change in samples compared to the initial type distribution (%) from HGC2 (H) and HW2 (W)

ples drawn from HGC2 were considerably smaller than those from HW2. This resulted in a proportionally greater increase in the number of entries as they were compressed. Excluding the three adjective samples, the average difference between the two corpora is 2.56%, indicating relatively similar values across both corpora.

Overall, the number of types decreased to 54.43% (2593 types) for HGC2, while HW2 showed a reduction to 60.83% (5573 types). This indicates that, on average, between 289 and 622 different types per auxiliary/adjective remain. The resulting entries are construction candidates that can be easily validated manually due to their limited quantity (cf. the large number of examples they cover). Thus, the expected reduction in the number of individual entries, aimed at enhancing the analysis, has been successfully achieved, leaving only the qualitative evaluation.

5.2 Constructional similarities

The quality of the remaining patterns can best be observed by looking at common types and their variations across several auxiliaries. To illustrate this, we selected the ten most frequent entries of the 3- and 4-gram types (see Table 4 in the Appendix.) as they cover 64.69% of the instances for 3-grams and 34.37% for 4-grams in HGC2, while 65.97% for 3-grams and 29.44% for 4-grams in HW2. Longer n-grams ($\geq$ 5, which account for only 25.92% and 27.38% of the examples respectively) would require a different approach due to data sparsity, as we cannot be sure why they are missing for individual auxiliaries.

The present study (Table 4.) validates the previous findings on Hungarian auxiliary verb structures (Indig and Bajzát, 2023; Bajzát, 2022; Kálmán et al., 1989) using a corpus-driven approach and extends the scope of the investigation to predicative adjectives with modal meanings. The verbs *akar* (‘want’), *tud* (‘can’) and *szeret(ne)* (‘(would) like (to)’) were omitted from Table 4., because they were found for all the patterns listed (i.e. they do indeed show frequent prototypical patterns for auxiliaries as expected). Beyond that, one can identify patterns in Table 4 that are specific to auxiliaries (e.g. insertion [/Prev] L [/V][Inf], which indicates a greater semantic integration between the infinitive and the auxiliary (Imrényi, 2013; Modrián Horváth, 2020; Bajzát, 2020) and serves as a key criterion for auxiliary (Kálmán et al., 1989)). While the insertion of certain predicative adjectives (*be* [/Prev] *képes* [/Adj][Nom] *menni* [/V][Inf] ’he/she is able to go into’) is theoretically possible and could represent the next step in their grammaticalisation process (Langacker, 2006; Heine and Narrgog, 2012), we found no evidence of such behaviour. However, frequent patterns (e.g. [/N][Acc] L [/V][Inf]) do not exclusively specify auxiliary structures, as they also often occur with other verbs as well. A more sophisticated and systematic analysis of construction patterns (e.g. using this method) could reveal auxiliaries and predicative adjectives with similar modal functions functions but belonging to different word classes. The characterisation of such relationships makes it possible to draw a network of auxiliaries, verbs and compatible predicates through their usage patterns.

The most frequent word order pattern in Table 4. covers instances of grammatical focus (e.g. [/N][Acc] L [/V][Inf]), where the speaker emphasises a particular component of the event rather than the event as a whole. Negative contexts are also typical for all modal auxiliaries and predicative adjectives, except for *képtelen*, which already contains the negation (the *-tAlAn* suffix, e.g. *Mari képes/képtelen aludni* 'Mary is able/unable to sleep'). classification (Imrényi, 2013; Kálmán et al., 1989). For three auxiliary verbs (*kíván*, *mer*, and *bír*), not all of the top 10 patterns were identified. For *kíván*, only one pattern was missing from the HGC2 sample, likely due to limited occurrence of special structures. A similar issue arose with *mer*, although all relevant patterns were found in the HW2 sample. From this we can conclude that the auxiliary verb *bír* is the one that actually shows distinct patterns. The verb *bír* typically occurs in contexts with negative polarity (e.g. *Nem bírja felemelni azt* 'He/She cannot lift it'), which inherently increases the number of words in the patterns. Additionally, *bír* often appears in structures with restrictive adverbs (e.g. *Alig bírja felemelni azt* 'He/She can hardly lift it'). Furthermore, this phenomenon is also evident in the fact that when certain grammatical cases of a noun are placed in the 3-grams, these patterns do not appear with the *bír*. Based on its specific patterns, it can be assumed that it is less advanced on the grammaticalisation path to a typical auxiliary verb than, for example, the more abstract auxiliary *tud* ('can'). Among the predicative adjectives (*képes*, *hajlandó* and *képtelen)*), *képes* ('able to') most closely matches the patterns of general modal auxiliary structures (i.e. it has instances for most of the patterns listed). This may be functionally because it is quasi-synonymous with the auxiliary *tud* ('can') in contexts expressing ability.

It can be seen that the ten most frequent n-grams from HGC2 and HW2 differ slightly at the same length. This could be due to differences in the size of the two corpus, but it also raises further questions about the correlation between the patterns used and the text types or genres. Additionally, the still high number of variations (see Table 3.) is partly due to the flexibility of Hungarian word order and partly to the presence of idiomatic patterns as separate entries. However, this aligns with our aim of preserving structural diversity, reflecting both the variety and similarities among auxiliaries.

6 Conclusion and Future Work

We have presented the steps[1] for reducing examples in the corpus to a form a form that is nearly identical to theoretical constructions in a corpus-driven manner through a case study. The semi-automatic clause extraction and feature reduction are covered by Indig and Bajzát (2023), from which we used the samples, while the selection of the optimal level of abstraction is obtained automatically with Mosaic methods (Indig, 2017). This work focuses on the precise compression of the noun phrases (cf. NP chunking), which halved the number of the remaining candidates, yielding manually comparable results that still need further evaluation and reduction. The method is currently under development, so the latter steps are not yet in their final form.

The limitations of the compression include the handling of interrupted possessive structures. However, we have shown that their small number is more of a usage pattern than an oddity, opening up new directions for research. Free argument order poses another challenge, which we plan to address by using bag of words instead of n-grams after the compression, allowing for the classification of otherwise similar neutral and focused structures. The idiomatic structures identified (e.g. entries with specific elements kept) are not included in this evaluation and should be examined separately.

We have used rule-based components (cf. LLMs) to maintain maximum control over the workflow and to be able to examine and validate the chain of reasoning at each step in order to develop a more correct intuition from which theories can benefit. For example, the need for a rigorous formalisation of the constructions sought and the development of tools to achieve such reduction of examples. This approach has revealed unexpected behaviours (e.g. the absence of interrupted possessive structures) that would otherwise remain hidden. If studied further, they could provide more insight into the reasoning of the language users through their usage patterns and would shape theoretical thinking. On the other hand, the empirical validation of the revealed theoretical results based on human cognition is essential to support the cognitive aspect of the approach. Therefore, in the next step of the research we plan to validate the identified patterns with cloze tests in the form of a language game (Indig and Lévai, 2023)), as this approach also supports the testing of LLMs for the same task.

The proposed steps are loosely language and task dependent as they can be easily adapted to other languages and phenomena. They are tested on Hungarian, so they are particularly suitable for other Uralic and morphologically rich languages.

[1] The full source code is available under copyleft licence at https://github.com/bajzattimi/Research-of-infinitive-structures-related-to-the-modal-semantic-domain

Acknowledgments

Supported by the EKÖP-24-3 University Research Fellowship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

References

Tímea Bajzát. 2020. de akarám kdnek is tudtára adni' - az akar, a tud és a mer segédige + főnévi igenév kompozitumszerkezetek szintaktikai vizsgálata az időbeliség perspektívájában. In Gábor Simon and Gábor Tolcsvai Nagy, editors, *Nyelvtan, diskurzus, megismerés*, pages 119–149. ELTE Eötvös Kiadó, Budapest.

Tímea Bajzát. 2022. A premodális tartományokkal összekapcsolódó segédige/melléknév + főnévi igenév konstrukciók mondatszintű szemantikai vizsgálata. In Szilárd Tátrai and Gábor Tolcsvai Nagy, editors, *A magyar mondat és kontextuális környezete*, pages 141–194. ELTE Eötvös Kiadó, Budapest.

Joan Bybee. 1995. *Frequency of Use and the Organization of Language*. Oxford University Press, Oxford.

Dylan Glynn and Justyna A. Robinson, editors. 2014. *Corpus Methods for Semantics. Quantitive studies in polysemy and synonymy*. John Benjamins, Amsterdam–Philadelphia.

Adele E. Goldberg. 1995. *Constructions: A construction grammar approach to argument*. University of Chicago Press, Chicago.

Stefan Th. Gries and Anatol A. Stefanowitsch, editors. 2007. *Corpora in Cognitive Linguistics*. Mouton de Gruyter, Berlin.

Bernd Heine and Heiko Narrgog. 2012. *The Oxford Handbook of Grammaticalization*. Oxford University Press, Oxford.

András Imrényi. 2013. A beférkőző segédigés szerkezetek függőségi nyelvtani elemzéséhez. *Magyar Nyelv*, 109(3):291–308.

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai. 2019. One format to rule them all – the emtsv pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Balázs Indig. 2017. Mosaic n-grams: Avoiding combinatorial explosion in corpus pattern mining for agglutinative languages. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 147–151, Poznań. Adam Mickiewicz University.

Balázs Indig and Tímea Borbála Bajzát. 2023. Bags and mosaics: Semi-automatic identification of auxiliary verbal constructions for agglutinative languages. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 111–116, Poznań. Adam Mickiewicz University.

Balázs Indig and Dániel Lévai. 2023. I'm smarter than the average bert! – testing language models against humans in a word guessing game. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 106–111, Poznań. Adam Mickiewicz University.

Laura A. Janda, editor. 2013. *The Quantitive Turn: The Essential Reader*. De Gruyter Muoton, Berlin.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2007. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

C. György Kálmán, Lászkó Kálmán, Ádám Nádasdy, and Gábor Prószéky. 1989. A magyar segédigék rendszere. In Zsigmond Telegdi and Ferenc Kiefer, editors, *Általános Nyelvészeti Tanulmányok XVII. Tanulmányok a magyar mondattan köréből*, pages 49–103. Akadémiai Kiadó, Budapest.

Ronald W. Langacker. 2005. Construction grammars: cognitive, radical, and less so. In Francisco J. Ruiz de Mendoza Ibáñez and Sandra M. Peña-Cervel, editors, *Construction Grammars: cognitive, radical, and less so*, 32, pages 101–159. Mouton de Gruyter, Berlin–New York.

Ronald W Langacker. 2006. Subjectification, grammaticalization, and conceptual archetypes. In Angeliki Athanasiadou, Costas Canakis, and Bert Cornillie, editors, *Subjectification. Various paths to subjectivity*, pages 49–103. Mouton de Gruyter, Berlin–New York.

Ronald W. Langacker. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press, Oxford.

Noémi Ligeti-Nagy, Andrea Dömötör, and Noémi Vadász. 2019. What does the nom say? an algorithm for case disambiguation in Hungarian. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 27–41, Tartu, Estonia. Association for Computational Linguistics.

Milla Luodonpää-Manni, Esa Penttilä, and Johanna Viimaranta, editors. 2017. *Empirical Approaches to Cognitive Linguistics*. Cambridge Scholars Publishing, Cambridge.

Bernadett Modrián Horváth. 2020. Beférkőzés és keretképzés a magyar nyelvben. In Géza Balázs, András Imrényi, and Gábor Simon, editors, *Hálózatok a nyelvben*, pages 281–296. Magyar szemiotikai társaság, Budapest.

Dávid Márk Nemeskey. 2020. *Natural language processing methods for language modeling*. Ph.D. thesis, Eötvös Loránd University.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1315–1322, Portorož, Slovenia. European Language Resources Association (ELRA).

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723, Reykjavik, Iceland. European Language Resources Association (ELRA).

An Van linden. 2010. From premodal to modal meaning: Adjectival pathways in english. *Cognitive Linguistics*, 21:537–571.

A Appendix

TOP	C	N	l	kíván	mer	bír	*hajl*	képes	*képt*
[/N][Acc] L [/V][Inf]	H	35273	3	H W	H W	0 W	H W	H W	0 0
nem L [/V][Inf]	H	8683	3	H W	H W	H W	H W	H W	0 0
[/Cnj] [/V][Inf] L	H	6972	3	H W	H W	H W	0 0	0 W	0 0
[/N][Nom] L [/V][Inf]	H	6052	3	H W	H W	H W	0 W	H W	0 W
[/Prev] L [/V][Inf]	H	5197	3	H W	H W	H W	0 0	0 0	0 0
[/N][Ins] L [/V][Inf]	H	5181	3	H W	H W	0 0	0 W	H W	0 0
[/N][Subl] L [/V][Inf]	H	3936	3	H W	H W	0 0	0 0	0 W	0 0
L [/V][Inf] [/N][Acc]	H	3468	3	H W	H W	0 0	0 W	H W	0 0
[/N][Ine] L [/V][Inf]	H	2721	3	H W	0 W	0 0	0 W	H W	0 0
[/N][Nom] [/V][Inf] L	H	1768	3	H W	H W	0 0	0 0	0 0	0 0
[/N][Acc] L [/V][Inf]	W	66688	3	H W	H W	0 W	H W	H W	0 0
nem L [/V][Inf]	W	12775	3	H W	H W	H W	H W	H W	0 0
[/N][Ins] L [/V][Inf]	W	12134	3	H W	H W	0 0	0 W	H W	0 0
[/N][Nom] L [/V][Inf]	W	10435	3	H W	H W	H W	0 W	H W	0 W
[/Cnj] L [/V][Inf]	W	10202	3	H W	H W	H W	H W	H W	0 W
L [/V][Inf] [/N][Acc]	W	9349	3	H W	H W	0 0	0 W	H W	0 0
[/Cnj] [/V][Inf] L	W	8093	3	H W	H W	H W	0 0	0 W	0 0
[/N][Ine] L [/V][Inf]	W	6639	3	H W	0 W	0 0	0 W	H W	0 0
[/N][Subl] L [/V][Inf]	W	6520	3	H W	H W	0 0	0 0	0 W	0 0
[/Prev] L [/V][Inf]	W	5721	3	H W	H W	H W	0 0	0 0	0 0
[/N][Nom] nem L [/V][Inf]	H	9930	4	H W	H W	H W	H W	H W	0 0
[/Cnj] nem L [/V][Inf]	H	8323	4	H W	H W	H W	H W	H W	0 0
[/Cnj] [/N][Acc] L [/V][Inf]	H	8274	4	H W	H W	0 W	H W	H W	0 W
[/N][Acc] nem L [/V][Inf]	H	8107	4	H W	H W	H W	H W	H W	0 0
[/N][Nom] [/N][Acc] L [/V][Inf]	H	7984	4	H W	0 W	0 W	0 W	H W	0 0
[/N][Acc] [/Prev] L [/V][Inf]	H	7028	4	H W	0 W	0 W	0 0	0 0	0 0
[/Cnj] [/Prev] L [/V][Inf]	H	5742	4	H W	H W	H W	0 0	0 0	0 0
[/N][Nom] [/Prev] L [/V][Inf]	H	4410	4	H W	H W	0 W	0 0	0 0	0 0
[/Prev] L [/V][Inf] [/N][Acc]	H	4130	4	H W	H W	0 0	0 0	0 0	0 0
[/N][Nom] [/Post] L [/V][Inf]	H	3343	4	H W	0 W	0 W	0 W	H W	0 W
[/Cnj] [/N][Acc] L [/V][Inf]	W	18950	4	H W	H W	0 W	H W	H W	0 W
[/N][Nom] nem L [/V][Inf]	W	16290	4	H W	H W	H W	H W	H W	0 0
[/N][Nom] [/N][Acc] L [/V][Inf]	W	16130	4	H W	0 W	0 W	0 W	H W	0 0
[/Cnj] nem L [/V][Inf]	W	15768	4	H W	H W	H W	H W	H W	0 0
[/N][Acc] nem L [/V][Inf]	W	13869	4	H W	H W	H W	H W	H W	0 0
[/N][Acc] [/Prev] L [/V][Inf]	W	9242	4	H W	0 W	0 W	0 0	0 0	0 0
[/Cnj] [/Prev] L [/V][Inf]	W	8663	4	H W	H W	H W	0 0	0 0	0 0
nem L [/V][Inf] [/N][Acc]	W	7127	4	H W	H W	H W	H W	H W	0 0
[/Cnj] L [/V][Inf] [/N][Acc]	W	6738	4	0 W	H W	0 W	H W	H W	H W
[/N][Nom] [/Prev] L [/V][Inf]	W	5100	4	H W	H W	0 W	0 0	0 0	0 0

Table 4: The most frequent mosaic 3-gram and 4-gram types found in the samples of the two corpora. The auxiliary verb lemmas *akar*, *tud* and *szeret(ne)* were detected in all the patterns in the table and are therefore not shown individually. Abbreviations: C = the original corpus which the pattern was derived, N = Number of occurences, l = length of the pattern, L = lemma (of the auxiliary); hajl = *hajlandó*, képt = *képtelen*; H = Hungarian Gigaword Corpus 2.0., W = Hungarian Webcorpus 2.0., 0 = Not present. Glossary: *nem* ‘not’;)