The Mental Lexicon of Communicative Fragments and Contours: The Remix N-gram Method

Emese K. Molnár, Andrea Dömötör
National Laboratory for Digital Heritage (DH-LAB)
Budapest, Hungary
emesekmolnar@gmail.com, domotor.andrea2@btk.elte.hu

Abstract

The classical mental lexicon models represent the lexicon as a list of words. Usage-based models describe the mental lexicon more dynamically, but they do not capture the real-time operation of speech production. In the linguistic model of Boris Gasparov, the notions of communicative fragment and contour can provide a comprehensive description of the diversity of linguistic experience. Fragments and contours form larger linguistic structures than words and they are recognized as a whole unit by speakers through their communicative profile. Fragments are prefabricated units that can be added to or merged with each other during speech production. The contours serve as templates for the utterances by combining linguistic elements on specific and abstract level. Based on this theoretical framework, our tool applies remix n-grams (combination of word forms, lemmas and POS tags) to identify similar linguistic structures in different texts that form the basic units of the mental lexicon.

Keywords: remix n-gram, communicative fragment, communicative contour, text reuse

1. Introduction

Models of the mental lexicon have long been central to linguistics and several very different approaches have been developed over time. Our paper seeks to answer the following questions: (i) What kind of mental lexicon model can capture the everyday linguistic experience of the language user and describe the language production at the discourse level? (ii) How can we operationalise the theoretical model, or in other words which (NLP) method can fit the theory?

After the structuralist concept of Saussure, the Chomskyan generative grammar (Chomsky, 1965) has dominated the field of linguistics. Based on this theory the basic unit of language is the sentence which can be built using elements and rules. The building blocks are the words that are contained in the lexicon as a list of elements. In contrast to this formal concept, the functional and usage-based approach (Langacker, 1987; Croft, 2001; Goldberg, 2006; Bybee, 2010) focuses on not just the language itself, but on the speaker as well, since these two cannot be separated. Just as the knowledge of language cannot be separated from general knowledge of the world, neither can the lexicon be separated from the rest of the language. They describe linguistic units as form and meaning pairs, holistic and emergent.

Although these lexicon models are no longer simple collections of words, as they change from static to continuum concept, they are still not dynamic enough to describe the online, i.e. real-time process of language production. Even if they focus on the speaker, these approaches still concentrate on the abstract form of the language, and fail to capture the everyday linguistic experiences of language users. Accordingly, this paper presents a theoretical model in which lexical items and grammar are emergently linked and proposes an NLP-based tool inspired by this model. We tested our method on a Hungarian poetry corpus as a first step in the development of the tool. By the end of the project, the aim is to develop a tool that can detect the typical communicative fragments and contours in different texts.

2. Communicative Fragments and Communicative Contours

Boris Gasparov built his intertextual concept of language on Bakhtinian theory. Gasparov criticizes construction grammars and Cognitive Grammar because, although they are not rule-based descriptions, their concepts are still too abstract and rigid. His central concept is the communicative fragment (CF), which is „a concrete segment of speech of any shape, meaning, and stylistic provenance that speakers are able to recognize spontaneously and to use as a conventional expression that fits certain communicative purposes" (Gasparov, 2010: 38). CFs can be more varied than constructions, since their boundaries are defined by the linguistic experience of the language user. Thus, fragments - unlike lexical units in a traditional lexicon model - are not listable but are constantly changing; they do not necessarily have either a compact or a fixed syntactic structure; they have the ability to evoke, allude to, and merge with each other (Gasparov, 2010: 50-55). Besides, CFs are prefabricated and

The Workshop on Cognitive Aspects of the Lexicon, pages 133–139

ready-made pieces as well that are embedded in context. Every CF has a texture that is an imprint of the specific situations in which it is used. This texture determines the expectations of discourse, i.e. the communicative profile that marks the genre, style, potential topic and conversing parties associated with the fragment (Gasparov, 2010: 55-58).

(1) Open the door to the west veranda!
[open the door]
[the door to]
[the door to the veranda]
[the west veranda]

Example (1) shows that CFs can vary in length and overlap. As Gasparov (2010) describes: „Due to their doubleedged connection to fluid mental processes on the one hand and to linguistic hardware on the other, CFs constitute the crucial link between the cognitive and the operational aspects of language – between creative efforts of the mind and the concrete material that allows those efforts to emerge as tangible facts of speech" (Gasparov, 2010: 64).

According to Gasparov (2010) CFs constitute a primary vocabulary that is at least as important for speakers' knowledge of their language as the vocabulary of lexical units. Speakers compose and interpret speech primarily based on CFs rather than words. Although a CF can be divided into smaller meaningful components like words and morphemes, they remain a single unit for the speaker. Although they can recognize the components and complex structure of the unit this analytic process is not reflected during the speech production like the subprocesses are not reflected either in the habitual operations of their everyday life (Gasparov, 2010: 39).

In the Gasparovian model, CFs are considered the basic unit of language, but for speech production a speech prototype (SP) is needed that leads to the realization of the speech artifact (SA) (Gasparov, 2010: 117-121). The starting point for the SA is the fragments, which are linked together through communicative profiles. The fragments are connected to each other on the basis of both structural and semantic similarities, and the resulting networks are the speech prototypes. In addition to the SP, the realisation of the SA requires to define the specific contextual framework. While the fragment carries its specific context in its texture, the speech artifact always needs motivation based on the actual context. The essential aspect of the speech process is the way in which the prefabricated fragments are organised into prototypes, as possible variants of which the speech artifacts are created according to the specific context.

1st segment	2nd segment	3rd segment	4th segment
Actually I'm	[surprised] [amazed] [glad] [so glad] [so happy]	you/he/she/they/John	[VP].

Figure 1: Communicatice Contour

Besides, the communicative fragments, communicative contours (CC) are available for the speakers to create utterances. Similarly to a CF, CC is recognized comprehensively by the speaker and it has an imprint of the situation where it occurred. The difference between the CF and CC is the character of their shape. Contours can be seen as a template rather than a blueprint. CC is a semiconcrete design with some prefabriceted pieces and gaps between the structural elements. The gaps are flexible during the speech process, they can be contracted, expanded or reshaped in order to complete the utterance. While CFs are incomplete and fragmentary, its borders are often vague, so it can be easily modified or fused with other CFs in speech. In contrast, a CC has to be structurally complete and it has a sharply outlined frame. Its flexibility comes from within, filling of the gaps can be varied, while the structural elements retain the specific character of the CC (Gasparov, 2010: 151-159).

Accordingly, the CC is built up of three constituents: a lexical-structural template (Gasparov, 2010: 158), a prosodic template (Gasparov, 2010: 162) and lacunae (Gasparov, 2010: 166). The lexical-structural template containing morphemes, words or word combinations is considered to be the most sharpened constituent with the most concrete elements. This template includes the elements that function as signposts, which help to select and place other possible elements in key positions in the structure. In comparison, a prosodic template is less a concrete form that is not identical with the actual pronunciation of the utterance, but rather a comprehensive sound shape in the inner perception of speakers. It determines, among others, the intonational contours of pronunciation and pauses. Together, the pitch curve, rhythmical texture, accent and timbre of voice as prosodic signposts guiding the vocalisation of the CC to complete syntactic patterns as well as to select specific elements of the lexicon. The least specific constituent of the contour is the lacunae between the lexical and prosodic markers. However, these gaps cannot be considered as blank space, as they consistently fit into the overall structure of the contour. The main lexical items, the rhythmic and intonational contour and the general communicative profile delimit the set of items that can be inserted into the lacunae.

Speakers keep in memory a large number of CCs of different shapes, lengths, and styles. Similar to the vocabulary of CFs, the vocabulary of CCs does not form a coherent system. It can rather be described as a shapeless agglomeration of templates in speakers' memory. Different CCs are linked to different communicative situations and speech experiences, each activated in a speaker's mind opportunely. They are package of knowledge whose relation to each other, and to a presumable overall system, is simply irrelevant to speakers (Gasparov, 2010: 157).

3. Linguistic Remix

Gasparov built his theory on the basis of intertextualty. "The prevalent mode of speakers' linguistic activity can be called "intertextual," in the sense that speakers always build something new by infusing it with their recollection of textual fragments drawn from previous instances of speech" (Gasparov, 2010: 3). His thoughts are very close to those of another author, who describe not just the language, but the whole culture in the spirit of reuse. Lessig (2008) is credited with the concept of remix culture which questioned and renegotiated not only the term of authorship, but also the understanding of creativity and culture. The concept of remix can be adapted successfully to the examination of cultural practices because – as Lessig points out – while the phenomenon of remixing may seem novel, its core mechanism has long been a part of human culture: remixing with (digital) media is identical to the fundamental process of language use. Evoking and incorporating the words of others into written works or conversations is so natural that we do not even notice the borrowing.

Popular text generation tools which are based on the Large Language Models (LLMs) owe their success to the fact that they exploit this fundamental linguistic mechanism. These tools learn from large amount of textual data sets to determine the probabilistic values of linguistic patterns for generating textual content. In other words, they remix by recognising and regenerating existing linguistic patterns. However, the output of LLMs is not completely transparent from this point of view, so it is also worth developing tools that highlight and and reveal the basic patterns of language production.

Remix can be seen as a concept that can capture both theoretical and methodological approaches based on the everyday linguistic experiences of language users. In contrast to intertextuality, it is not only suitable for describing prototypical, lexical repetitions, but can also be applied to the investigation of linguistic similarities on the structural level. The remix can serve as a framework and can link the Gasparovian language model and an NLP solution that fits the theory.

4. Related Work

There are many examples of computational methods for intertextuality and text reuse detection. These studies usually focus on the relationships between texts in different text types. For example, one line of research focuses on texts that are reused in academic work, with a particular focus on plagiarism (Citron and Ginsparg, 2015; Anson and Moskovitz, 2020, Gienapp et al., 2023). Large corpora of newspapers are available for the study of text repetition as well (Smith et al., 2013; Vesanto et al., 2017; Rosson et al., 2023). Intertextuality in literary texts is a long-established and widely researched phenomenon, which has been further enhanced by the increasing availability of large corpora and the emergence of computational methods. (Kahane and Mueller, 2001; Lee, 2007; Coffee et al., 2013; Büchler et al., 2014; Gladstone and Cooney, 2020).

The closest in spirit to our own project were those tools that make possible to detect text reuse within literary corpora. Both the Chicago Homer (Kahane and Mueller, 2001), the Tesserae (Coffee et al., 2013) and the Commonplace Cultures (Gladstone and Cooney, 2020) have a query interface with search and comparison function. Among these the Chicago Homer is a bilingual database of Early Greek epic. The corpus is tokenized, lemmatized and annotated with morphological and narratological tagging. The tool makes possible to find repetitions (sequence of two or more words) in the corpus and to filter them by various criteria. The Tesserae Project provides an online tool that allows users to compare two texts in ancient Greek, Latin, or English. The basic Tesserae search finds sentences or poetic verse lines in two different texts that share two or more similar lemmata based on an n-gram method. Experimental search options were added for sound similarity, for semantic relatedness of Greek to Latin and for context similarity using a topic modeling approach. The Commonplace Cultures project aims to detect text reuses in the Eighteenth Century Collection Online (ECCO). For the comparison of the texts PhiloLine, a sequence alignment tool was developed. The model is based on shingles of n-grams to find shared passage according to the number of common contiguous n-grams between two textual sequences.

5. Method

While text reuse research usually focuses on a single text type, our goal is to develop a method to detect typical patterns in several text type. Most of the tools are designed for English texts, so our

further goal is to have a suitable tool for examining another type of language, a morphologically rich language.

In most cases, analysis of text reuse apply n-gram-based methods to find text similarities. In the case of word n-grams, the text is divided into sequences of n adjacent words in particular order. The result is similar to the example about CFs that was shown in example (1). This suggest that n-gram based methods can be suitable for detecting potential CFs within a text. By comparing the n-grams of different texts, we can identify the fragments that are usually characteristic of a discourse based on their repetition. Based on this, our method started by comparing trigrams of texts.

5.1. Remix n-gram

As we have seen in the case of CCs in figure (1), the production of utterances requires not only the combination of concrete words or fragments, but a template that are consist of linguistic units on the different levels of abstraction. Our method, called remix n-gram, is based on the concept of mosaic n-gram of (Indig and Bajzát, 2023). Mosaic n-grams are combinations of words, lemmas and POS tags representing different levels of language both specific and abstract. The comparison of such n-grams can capture structural similarities of texts besides the textual ones.

The first challenge was the sentence segmentation as we used a poetry corpus (see in detail in section 5.2). In the case of poems that contain sentence punctuation marks we segmented the text according to these. Many poems, however, does not contain sentence punctuations, therefore in these cases we considered each stanza a sentence.

As a next step we removed determiners (*a, az* "the", *egy* "a, an"), the conjunction word *és* ("and") and all the words tagged as "other" (X) by the morphological analyzer. We extracted the trigrams of the remaining words of each sentence. The trigrams contained all the information of the word: the word form, the lemma and the POS tag. For POS tag we used the UD tagset enhanced with some language-specific characteristics. This was necessary because Hungarian is a morphologically rich language therefore a UD POS tag and lemma cover many different word forms that are not really grasped as similar words by language users. Nominal tags (NOUN, ADJ, PROPN, PRON, NUM) were enhanced with case (2a) and a "Poss" feature if they were possessive (2b). In the case of adjectives we also marked the different degrees (2c). Verbs were divided into finite and non-finite (infinitive) groups (2d and 2e) and the tags of finite verbs were enhanced with the mood feature (2e).

(2) a. *erdőbe* 'to the woods'
`NOUN` → `NOUN.Ill`

b. *lelke* 'his/her soul'
`NOUN` → `NOUN.Poss`

c. *kisebb* 'smaller'
`ADJ` → `ADJ.Cmp`

d. *látni* 'to see'
`VERB` → `VERB.Inf`

e. *mennék* 'I would go'
`VERB` → `VERB.Cnd`

We compared the trigrams of each poem, taking into account word forms, lemmas and POS tags, and ranked the degree of similarity between them. The matching word forms got the highest score, followed by matching lemmas and POS tags. This means that a matching word form is worth 3 points for each token, a matching lemma is worth 2 points and a matching POS tag is worth 1 point. So, if three word forms match, the trigram is worth 3x3 points, that is 9 points, if two word forms and one lemma match, the trigram is worth 2x3+2, that is 8 points in total and so on. The minimal requirement of similarity was having all three POS tags and at least one lemma matched.

5.2. Corpus

To test our method, we chose a corpus that was available with the annotation layers to extract the remix n-grams. For this reason, we used The ELTE Poetry Corpus (Horváth et al., 2022a) of 3,441,864 tokens contains the complete poems of 50 Hungarian canonical poets. Besides, tokenization, lemmatization, as well as the part-of-speech and morphological analysis, the automatic annotation of the structural elements (title, stanzas, lines) and the sound devices (rhyme scheme, rhyme pairs, rhythm, alliteration, phonological structure of words) of the poems were completed in XML format.

6. Results

As a pilot study, we chose a poet from the poetry corpus and compared the remix trigrams of the poems among each other. This allowed us to identify potential candidates of CFs and fragments of CCs specifically, instead of the prototypical, literal cases of intertextuality. We compared 514 poems, in which we found more than 200,000 matching trigrams.

Examples (3)-(7) show different degrees of matching in the examined subcorpus. (3) is an exact match, (4) has two matching words and one matching POS tag, (5) and (6) have one matching word and two matching POS tags, and (7) has one matching lemma and two matching POS tags.

(3) *este van már – este van már*
evening is already – evening is already

'It's evening already'

(4) *soha nem látott – soha nem hallott*
never not seen – never not heard

'never seen' – 'never heard'

(5) *mint rossz madár – mint jámbor állatok*
like bad bird – like pious animals

'like a bad bird' – 'like pious animals'

(6) *mint rossz madár – mint puszta rom*
like bad bird – like mere ruin

'like a bad bird' – 'like a mere ruin'

(7) *pogányok lelke volt – emberek*
pagans soul.Poss was – people
emléke van
memory.Poss is

'It was pagans' souls' – 'there is a memory of people'

As the examples show the remix n-gram method is suitable for capturing structural similarities, however many of the matches on more abstract levels were not matching in terms of semantic similarity. It can be seen from the (6) example that one word and two POS tag matching does not always fulfill our expectations. While in example (7) the bird and the animal are still semantically close to each other, in the case of example (6) the bird and the ruin are not elements of similar semantic categories.

7. Discussion and future work

The aim of our method was to constract a model of the mental lexicon that correlates with the everyday linguistic experience of language users. Instead of a formal and list-like description of the lexicon, a functional and usage-based approach was adapted for this purpose. The communicative fragment and contour of the Gasparovian concept were suitable to serve as the basis for a dynamic model. The proposed remix n-gram method is effective in identifying the potentional text passages of more specific word-level CFs and sequences similar to CCs with more abstract structures. Since we extract typical language structures by comparing texts, we get different results from the comparisons of different texts, just as different language users have different language experiences. Thus, an NLP tool was provided that is theoretically grounded and the concepts of the theoretical framework became operationalized in practice as well.

We are able to find potential CFs and fragments of CCs using the remix n-gram method, but the large number of hits is still difficult to manage, meaning that the method needs to be refined in order to increase precision. We can reduce the number of hits by extending the stop word list or by further specifying the POS tags. Currently, only mood is specified as a criterion for verbs, but by adding number, person and tense could give more accurate hits. Furthermore, the (5) and (6) examples show that filtering words that are semantically closer to each other could also lead to better results. To find semantic similarities, it would be worth using word embedding to rank matches such as *bird* and *animal* over *bird* and *ruin*.

The next stage of the project will be to test the method on different corpora. Firstly, the ELTE Poetry Corpus has a folk song subcorpus (Horváth et al., 2022b), which is closer to the oral culture, and besides this the ELTE Novel Corpus (Bajzát et al., 2021) and the ELTE Drama Corpus (Szemes et al., 2022) are also available. The ongoing Lyrical Poetry Corpus project (Horváth et al., 2021) of ELTE DiAGram Research Centre for Functional Linguistics will contains not only canonical poems, but also song lyrics and slam texts in Hungarian. The Hungarian gold standard corpus project (K. Molnár and Dömötör, 2023) of DH-Lab will provide the opportunity to test less artistic text types, closer to everyday discourse, such as texts from blogs, educational and cultural web sites, in addition to novels.

In addition to testing on different corpora, we also aim to compare with other methods. For example, PhiloLine, which is used in the Commonplace Culture project, is an open source tool, so it can be used on different corpora. As it is also an n-gram based method, it offers the possibility to test the effectiveness of remix n-grams for languages with rich morphology. This may also pave the way for further developments aiming at applying the remix n-gram method to other types of languages.

These methods of identifying language structures are also useful because they bring us closer to understanding how language works in general. Unlike the tools based on LLMs, the results are more transparent and easier to interpret. In the long term, they can therefore contribute to the development of LLMs.

Acknowledgement

The research was supported by the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the

National Research, Development and Innovation Fund.

References

Ian G. Anson and Cary Moskovitz. 2020. Text recycling in stem: a text-analytic study of recently published research articles. In *Accountability in Research 28*, pages 349–371.

Tímea Borbála Bajzát, Botond Szemes, and Eszter Szlávich. 2021. Az elte dh regénykorpusz és lehetőségei. In *Online térben – az online térért. Networkshop 30: országos online konferencia*, pages 63–72, Budapest. HUNGARNET Egyesület.

Joan Bybee. 2010. *Language, usage and cognition.* Cambridge University Press, Cambridge.

Marco Büchler, Philip R. Burns, Marco Müller, Emily Franzini, and Greta Franzini. 2014. Towards a historical text re-use detection. In *Text Mining: From Ontology Learning to Automated Text Processing Applications*, pages 221–238. Springer.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

Daniel T. Citron and Paul Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences*, 112(1):25–30.

Neil Coffee, Jean-Pierre Koenig, Poornima Shakti, Roelant Ossewaarde, Chris Forstall, and Sarah Jacobson. 2013. The tesserae project: Intertextual analysis of latin poetry. *Literary and Linguistic Computing*, 28:221–228.

William Croft. 2001. *Radical Construction Grammar. Syntactic theory in typological perspective*. Oxford University Press, Oxford.

Boris Gasparov. 2010. *Speech, memory, and meaning: Intertextuality in everyday language*. De Gruyter, Berlin.

Lukas Gienapp, Wolfgang Kircheis, Bjarna Sievers, Benno Stein, and Martin Potthast. 2023. A large dataset of scientific text reuse in open-access publications. *Scientific Data*, 10(58).

Clovis. Gladstone and Charles Cooney. 2020. Opening new paths for scholarship: Algorithms to track text reuse in ecco. In *Digitizing Enlightenment: Digital Humanities and the transformation of Eighteenth-Century Studies*, pages 353–374. Voltaire Foundation in association with Liverpool University Press.

Adele Goldberg. 2006. *Constructions at work. The nature of generalization in language*. Oxford University Press, Oxford.

Péter Horváth, Péter Kundráth, Balázs Indig, Zsófia Fellegi, Eszter Szlávich, Tímea Borbála Bajzát, Zsófia Sárközi-Lindner, Bence Vida, Aslihan Karabulut, Mária Timári, and Gábor Palkó. 2022a. Elte poetry corpus: A machine annotated database of canonical hungarian poetry. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3471–3478, Paris. European Language Resources Association (ELRA).

Péter Horváth, Péter Kundráth, and Gábor Palkó. 2022b. Elte népdalkorpusz – magyar népdalok gépileg annotált adatbázisa. In *Valós térben – Az online térért: Networkshop 31: országos konferencia*, pages 276–283, Budapest. HUNGARNET Egyesület.

Péter Horváth, Gábor Simon, and Tátrai Szilárd. 2021. A lírai személyjelölés konstrukcióinak annotálási elveiről. In *Líra, poétika, diskurzus*, pages 133–166, Budapest. ELTE Eötvös Collegium.

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.

Emese K. Molnár and Andrea Dömötör. 2023. Experiments on error detection in morphological annotation. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 186–190, Poznan. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza.

Ahuvia Kahane and Martin Mueller. 2001. *The Chicago Homer*. University of Chicago Press/Northwestern University Library.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar*. Stanford University Press, Stanford.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic. Association for Computational Linguistics.

Lawrence Lessig. 2008. *Remix: Making Art and Commerce Thrive in a Hybrid Economy*. Penguin Press.

David Rosson, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. Reception reader: Exploring text reuse in early modern british publications. *Journal of Open Humanities Data*, 9(1).

David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*, pages 86–94.

Botond Szemes, Tímea Bajzát, Zsófia Fellegi, Péter Kundráth, Péter Horváth, Balázs Indig, Anna Dióssy, Fanni Hegedüs, Natali Pantyelejev, Sarolta Sziráki, Bence Vida, Balázs Kalmár, and Palkó Gábor. 2022. Az elte drámakorpuszának létrehozása és lehetőségei. In *Valós térben – Az online térért: Networkshop 31: országos konferencia*, pages 170–178, Budapest. HUNGARNET Egyesület.

Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Nivala, and Tapio Salakoski. 2017. A system for identifying and exploring text repetition in large historical document corpora. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 330–333, Gothenburg, Sweden.