Morphological ambiguity of Lithuanian uninflected parts of speech

Prepared by: Miglė Žemrietė
Supervisor: Assoc. Prof. Dr. Erika Rimkutė
Institute of the Lithuanian Language

Morphological ambiguity

- Morphological ambiguity is a phenomenon where the same word form can have multiple meanings depending on context. Both inflected and uninflected word forms may coincide.
- For example: *light* can be an adjective meaning "not heavy," a noun meaning "illumination," or a verb meaning "to ignite."
- Different forms of the same word may coincide. For example, *mama* in Lithuanian can be the nominative singular, the instrumental, or the vocative form.

Morphological taggers

- **Morphological taggers** – programs that, once given a word or a text, display the grammatical tags for each word – such as part of speech, gender, number, case, etc.
- In Lithuania, there is three morphological taggers:

Lemuoklis

Semantika.lt

Morfuoklis

Lemuoklis

Kompiuterinė lingvistika – taikomosios kalbotyros šaka, nukreipta į natūralios kalbos apdorojimą ir analizę įvairiomis kompiuterinėmis technologijomis.

◉ **Pateikti vieną tikėtiniausią variantą** ○ **Pateikti visus galimus variantus**

◉ Lema + gramatinės pažymos ○ Tik lema ○ Tik gramatinės pažymos

Rezultatas puslapyje | Rezultatas faile

<word="Kompiuterinė" lemma="kompiuterinis" type="bdv., teig, nelygin. l., neįvardž., mot. g., vns., V."/>
<space/>
<word="lingvistika" lemma="lingvistika" type="dkt., mot. g., vns., V."/>
<space/>
<word="taikomosios" lemma="taikyti(-o,-ė)" type="dlv., teig., nesngr., neveik. r, es. l., įvardž., mot. g., vns., K."/>
<space/>
<word="kalbotyros" lemma="kalbotyra" type="dkt., mot. g., vns., K."/>
<space/>
<word="šaka" lemma="šaka" type="dkt., mot. g., vns., V."/>
<sep=","/>
<space/>
<word="nukreipta" lemma="nukreipti(-ia,-ė)" type="dlv., teig., nesngr., neveik. r, būt. l., neįvardž., mot. g., vns., V."/>
<space/>
<word="į" lemma="į" type="prl."/>
<space/>
<word="natūralios" lemma="natūralus" type="bdv., teig, nelygin. l., neįvardž., mot. g., vns., K."/>
<space/>
<word="kalbos" lemma="kalba" type="dkt., mot. g., vns., K."/>
<space/>
<word="apdorojimą" lemma="apdorojimas" type="dkt., vyr. g., vns., G."/>
<space/>
<word="ir" lemma="ir" type="jng."/>
<space/>
<word="analizę" lemma="analizė" type="dkt., mot. g., vns., G."/>
<space/>
<word="įvairiomis" lemma="įvairus" type="bdv., teig, nelygin. l., neįvardž., mot. g., dgs., Įn."/>
<space/>
<word="kompiuterinėmis" lemma="kompiuterinis" type="bdv., teig, nelygin. l., neįvardž., mot. g., dgs., Įn."/>
<space/>
<word="technologijomis" lemma="technologija" type="dkt., mot. g., dgs., Įn."/>
<sep="."/>
<p/>

Semantika.lt

Automatinis tikrinimas | Analizuojamas tekstas | Rašybos klaidos | **Morfologija**

Tekstas:

Kompiuterinė lingvistika – taikomosios kalbotyros šaka, nukreipta į natūralios kalbos apdorojimą ir analizę įvairiomis kompiuterinėmis technologijomis.

Pasirinktas teksto segmentas:

Kompiuterinė

Ankstesnis | Kitas

Segmento morfologinė analizė:

Ankstesnis	Kitas
Pagrindinė forma (1)	*kompiuterinis*
Kalbos dalis	*Būdvardis*
Pobūdis	*Bendras*
Laipsnis	*Nelyginamasis*
Giminė	*Moteriškoji giminė*
Skaičius	*Vienaskaita*
Linksnis	*Vardininkas*
Apibrėžtumas	*Ne*

Morfuoklis

Kompiuterinė lingvistika – taikomosios kalbotyros šaka, nukreipta į natūralios kalbos apdorojimą ir analizę įvairiomis kompiuterinėmis technologijomis.

Rezultatų formatas:

- CoNLL-U
- Sketch Engine
- NoSketch Engine

Anotavimo standartas:

- "Jablonskis"
- MSD

Analizuoti | Naudojant samplaikas

Kompiuterinė	kompiuterinis	bdv.nelygin.mot.vns.V.	kompiuterinis-a
lingvistika	lingvistika	dkt.mot.vns.V.	lingvistika-n
–	–	skyr.	–-t
taikomosios	taikomasis	bdv.nelygin.įvardž.mot.vns.K.	taikomasis-a
kalbotyros	kalbotyra	dkt.mot.vns.K.	kalbotyra-n
šaka	šaka	dkt.mot.vns.V.	šaka-n
<g/>			
,	,	skyr.	,-t
nukreipta	nukreipti	vksm.dlv.neveik.būt.mot.vns.V.	nukreipti-v
į	į	prl.G.	į-s
natūralios	natūralus	bdv.nelygin.mot.vns.K.	natūralus-a
kalbos	kalba	dkt.mot.vns.K.	kalba-n
apdorojimą	apdorojimas	dkt.vyr.vns.G.	apdorojimas-n
ir	ir	jng.	ir-c
analizę	analizė	dkt.mot.vns.G.	analizė-n
įvairiomis	įvairus	bdv.nelygin.mot.dgs.Įn.	įvairus-a
kompiuterinėmis	kompiuterinis	bdv.nelygin.mot.dgs.Įn.	kompiuterinis-a
technologijomis	technologija	dkt.mot.dgs.Įn.	technologija-n

Morphological taggers

Taggers do not always annotate correctly – they often fail to recognize foreign words, slang, abbreviations.

Morphologically ambiguous words are particularly problematic: taggers struggle to choose the correct grammatical tag when a word has several possible ones.

Morphological ambiguity in other languages

- In Czech, morphologically ambiguous forms make up 46%
- English – 38.65%
- Hungarian – 21.58%
- Estonian – 50%
- Latvian – 50%

Hajič 2004, p. 173; Puolakainen 2012, p. 194; Paikens, Rituma, Pretkalnina 2013, p. 270.

Morphological taggers in other languages

- Latvian taggers identify lemmas with 97% accuracy and grammatical tags with 93%
- Polish taggers work with about 94% accuracy
- Czech – 97%
- Estonian taggers – 97% (*Estmorf*, *Vabamorf*, *Neural network-based taggers*).

Paikens, Rituma, Pretkalnina 2013, p. 270; Wróbel 2017, p. 389; Jelínek et al. 2021, p. 55; Tkachenko & Sirts 2018, p. 167.

Morphological ambiguity in Lithuanian

Ambiguity may occur between:

- **Two (or more) inflected words:**

e.g., *būdą* – meaning both "character" [masculine] and "doghouse" [feminine];

- **Inflected and uninflected words:**

e.g., *dėlei* meaning both "a slug" and a preposition "for";

- **Two (or more) uninflected words:**

e.g., *kaip* can be an adverb, a particle, or a conjunction.

Morphological ambiguity of uninflected parts of speech

About 25% of all morphologically ambiguous words are uniflectice.

These overlaps can be divided into nine categories:

1. Conjunctions and particles (*ar*, *būtent*, *ir*, *ne*, *nors*, *nei*, *kad*);
2. Adverbs, conjunctions, and particles (*tik*, *kaip*, *čia*);
3. Adverbs and particles (*jau*, *nelyg*, *taip*, *tarsi*, *kur*, *ten*, *vis*);
4. Particles, conjunctions, and interjections (*o*);
5. **Adverbs and prepositions** (*aplinkui*, *anapus*, *prieš(ais)*, *arti*, *aukščiau*, *išilgal*, *netoli*, *paskui*, *skradžiai*);
6. Particles and interjections (*na*, *oi*);
7. Interjections and onomatopoeias (*šast*, *cha*);
8. Prepositions and particles (*ant*);
9. Conjunctions and prepositions (*iki*).

Morphologically ambiguous adverbs and prepositions

- Focus on adverbs overlapping with prepositions;
- e.g. **arti** → *namas stovi **arti*** ("the house is **near**") – adverb; *namas stovi **arti** mišk**o*** ("the house is **near** the forest") – preposition;
- Taggers struggle to assign correct part of speech in such cases;
- Word order rules (e.g., preposition before noun) are not always reliable;
- Dictionary analysis shows borderline cases with unclear part of speech.
- Prepositions and adverbs can form clusters, such as ***iš anapus** pasaulio* ("**from beyond** the world"). In Lithuanian, identifying the part of speech in such cases is complicated.

Relevance of the research

- Ambiguity between adverbs and prepositions is one of the reasons for annotation inaccuracies (Piečytė 2021, p. 37).
- The issues of overlapping prepositions and adverbs are relevant to both theoretical and practical linguistics.

Object

- For this study, all morphologically ambiguous words were automatically extracted from the morphologically annotated corpus. From that list, adverbs coinciding with prepositions were selected. These are the **object** of this study.
- A total of 46 such words were found:

abigaliai, *abipus(iai)*, *anapus*, *antrapus*, *aplink(ui)*, *apsukui*, *arčiau, arčiausiai, arti*, *artyn*, *aukščiau, greta*, *iki*, *įkandin, įkypai, įstrižai, išilgai, kiaurai*, *kitapus*, *netoli*, *paraleliai*, *paskiau*, *paskui*, *paskum*, *perdėm, piečiau, pirm(a)*, *prieš(ais)*, *pusiau*, *skersai*, *skradžiai, statmenai*, *šalia, šiapus, šiauriau, toliau*, *už, viduj(e)*, *vidur(y)(je)*, *vienapus*, *vietoj(e)*, *virš, viršuj, viršum, žemėliau, žemiau.*

Goal and sources

- The **goal** of this study is to examine how the distinction between adverbs and prepositions is reflected in three main Lithuanian dictionaries as well as in the main grammars. It also examines how frequently and in what ways these word types are used in the corpora, and provides proposals on how to describe adverbs and prepositions more clearly in lexicographic resources.

- Three **dictionaries** were analyzed: Dictionary of Contemporary Lithuanian (DCL), Dictionary of the Lithuanian Language (DLL), and Dictionary of Standard Lithuanian (DSL);
- four **grammars**: "Grammar of Contemporary Lithuanian" (GCL), "Lithuanian Grammar" (LG), "Practical Grammar of Standard Lithuanian" (PGSL), and J. Šukys's book "Lithuanian Cases and Prepositions";
- two **corpora**: the morphologically annotated corpus "MATAS" and "the Corpus of Contemporary Lithuanian Language" (CCLL).

Netoli (“not far”)

- DCL (Dictionary of Contemporary Lithuanian):

Preposition: 1) “near (indicating distance)”; 2)“near (indicating time)”; 3) “about (indicating approximate quantity or number)”.

Adverb: 1) “nearby, close (in terms of location)”; 2) “near (in terms of time)”.

- DLL (Dictionary of the Lithuanian Language):

Preposition: 1) “near, at a short distance”; 2. “near (indicating time)”; 3. “about (to express approximation in quantity)”;

Adverb: 1) “near, close by (in terms of location)”: maudėmės **netoli nuo** kranto (“we bathed **not far from** the shore”; 2) “soon, in a short time”; 3) “not much, slightly, not very”; 4) “almost, nearly”.

Netoli nuo (“not far from”) – an adverb and a preposition? Two prepositions? A prepositional compound?

LKG (Lithuanian Grammar, p. 581): A compound is not formed, since in such cases the adverb can be moved after the prepositional phrase or another word can be inserted between them.

BUT:

J. Šukys (1998, p. 418): “with the adverbs toli, netoli [...] the preposition nuo tends to form compound prepositions.”

PGSL (Practical Grammar of Standard Lithuanian): In the expression netoli nuo, the word netoli is a preposition: “Secondary elements preceding a primary preposition may be interpreted as adverbs with independent syntactic function [...]. Their interpretation as two prepositions may result from uniform intonation.”

Arčiau, arčiausiai (“closer”, “the closest”)

- **PGSL:** Even comparative and superlative forms of adverbs can undergo prepositionalization.
- **J. Šukys** questions whether the use of prepositions formed from comparative or superlative adverbs is always correct: even “standardized prepositions such as *arčiau* (‘closer’), *aukščiau* (‘higher’), *žemiau* (‘lower’) are used only in the locative sense and are not very common in quantity-related constructions with nouns denoting units of measurement” (Šukys 1998, p. 564).
- **DLL:**

Arčiau – a preposition used with the genitive: *arčiau namų* (“closer to the house”); *eikš eikš, antele, arčiau manęs* (“come, come, little duck, closer to me”).

However, *arčiausiai* is not included in the dictionaries.

- Prepositions are not subject to comparison or any other morphological change, so including only the preposition *arti* is not sufficient – after all, one cannot derive the preposition *arčiau* from *arti*, as is possible in the case of adverbs.
- The inclusion of the form *arčiau* in dictionaries is also supported by actual usage – in the MATAS corpus, *arčiau* is used prepositionally in 62% of all occurrences; in the CCLL – 47%. *Arčiausiai* is used prepositionally in 50% of cases.

Paraleliai (“parallel”)

- **DLL:**

1) adverb: *With the establishment of Christianity in our land, the establishment of Polish culture also took place in **parallel**.*

- *Paraleliai* – not a preposition?

BUT:

In the MATAS corpus, there are 6 occurrences of *paraleliai*, 1 of which is a preposition (16%). In the CCLL, there are 197 instances of this form, including 22 prepositional uses (11%).

The preposition *paraleliai* is used **ONLY** with the dative case.

GCL (1994, pp. 586–649): prepositions are used only with the accusative, instrumental, and genitive cases. From the examples, it is clear that this is not dialectal or archaic language. It seems likely that an atypical use of the preposition with the dative is emerging. However, questions of grammatical correctness arise. Therefore, its inclusion in dictionaries could be complicated.

Conclusions and Recommendations

The analysis of adverbs and prepositions revealed that their definitions and usage differ across dictionaries and in *Morfuoklis*, which is relevant for automatic morphological analysis.

1. A key challenge is distinguishing adverbs from prepositions due to combinations (e.g., *netoli nuo*), which should be treated as separate words.
2. There are doubts about including prepositions from comparative and superlative adverbs in dictionaries, but it is important to include them in databases for accuracy.
3. Frequent usage patterns, such as *paraleliai* as a preposition, are not well represented in dictionaries. It should be included in Morfuoklis for better annotation.

The results of the study show that to improve Lithuanian lexicographical sources, a more thorough and systematic analysis of adverbs and prepositions is needed, along with clearer presentation of these parts of speech in dictionaries. Such an analysis will help clarify the rules of uninflected words in the Lithuanian language, as Lithuanian annotators rely on quantitative data and linguistic rules.

References

- Ambrazas, V. (Ed.). (1994). *Grammar of Contemporary Lithuanian Language*. Vilnius: Mokslo ir enciklopedijų leidykla.
- Drukteinis, A., Kazlauskaitė, R., Balčiūnienė, A., & Vaskelienė, J. (2024). *Practical Grammar of the Lithuanian Language*. Vilnius: Vilnius University Press. Available at: https://www.knygynas.vu.lt/praktine-bendrines-lietuviu-kalbos-gramatika [Accessed: 20 Feb. 2025].
- Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., & Šindlerová, J. (2021). SYN2020: A new corpus of Czech with an innovated annotation. In *Text, Speech, and Dialogue: 24th International Conference, Proceedings* (No. 24, pp. 48–59). Available at: https://link.springer.com/chapter/10.1007/978-3-030-83527-9_4 [Accessed: 12 Mar. 2024].
- Piečytė, M. (2021). *Morphologically Ambiguous and Unrecognized Words in the "Educational Corpus of the Lithuanian Language"* (Bachelor's thesis). Vytautas Magnus University.
- Paikens, P., Rutuma, L., & Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics*, 267–277. Available at: https://aclanthology.org/W13-5624.pdf [Accessed: 12 Mar. 2024].
- Puolakainen, T. (2012). How Does the Choice of Morphological Analyser Influence the Quality of Syntactical Analysis? In *Human Language Technologies – The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012* (pp. 193–200). IOS Press. Available at: https://dblp.org/rec/conf/hlt/Puolakainen12.html [Accessed: 12 Mar. 2024].
- Šukys, J. (1998). *Lithuanian Cases and Prepositions: Usage and Norms*. Kaunas: Šviesa.
- Tkachenko, A., & Sirts, K. (2018). Neural Morphological Tagging for Estonian. In *Human Language Technologies – The Baltic Perspective* (Vol. 307, pp. 166–174). Available at: https://ebooks.iospress.nl/volumearticle/50318 [Accessed: 27 Aug. 2024].
- Wróbel, K. (2017). KRNNT: Polish recurrent neural network tagger. *Proceedings of the 8th Language & Technology Conference*, 386–391. Available at: http://ltc.amu.edu.pl/book2017/papers/PolEval1-6.pdf [Accessed: 12 Mar. 2024].

Morphological ambiguity of Lithuanian uninflected parts of speech

Prepared by: Miglė Žemrietė
Supervisor: Assoc. Prof. Dr. Erika Rimkutė
Lithuanian Language institute