VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII

Towards the Automatic Extraction of Term-defining Context in Lithuanian

Agnė BIELINSKIENĖ, Loïc BOIZOU, Jolanta KOVALEVSKAITĖ and Andrius UTKA

Centre of Computational Linguistics,

Vytautas Magnus University (Kaunas, Lithuania)

Framework of the research

- project “Automatic Identification of Education and Science Terms” (ŠIMTAI 2) that is funded by a grant (No. LIT-2-44) from the Research Council of Lithuania

- The project aims to apply a corpus-based descriptive approach, in order to obtain contextual information that would help to define automatically extracted terms for the terminology dictionary of Education and Science.

Set of extracted terms

- The amount of terms used in the extraction of contextual information from corpus – 300;
- the length of terms varies from 1 to 5 words.

Corpus-based descriptive approach

- important conceptual characteristics are expressed by semantic relations (hypernymy, hyponymy, meronymy etc.) which can be identified by patterns (e. g. knowledge-rich contexts, Meyer 2001);

- it is possible to use this terminography-oriented knowledge from corpora as a starting point for dictionary definitions;

- to our knowledge the methodology has not yet been applied to Baltic languages.

Term-Defining Contexts (TDCs)

- One type of knowledge-rich contexts typically presented through semantic relations of hypernymy and hyponymy.

- TDCs present the most useful information for dictionary definitions, as it is nearest to a classic form of the definition:

  T *(term) = X (hypernym) + differentiating characteristics*.

Research method

The stages of **pattern-based approach** include

- analysis of constituent elements in definitional patterns;
- formalization of definitional patterns;
- automatic extraction of term-defining contexts, using an extraction tool, specifically designed for the task;
- evaluation of results.

Constituent elements in definitional patterns

- verb lexical items, such as *būti* (to be)*, sudaryti* (to consist of), *apimti* (to include), *laikyti* (to be considered as), which are possible markers of relevant semantic relations, among them the hypernymy being the most typical;

- some punctuation marks (dashes, quotation marks, colons) which often indicate a definitional structure;

- particular grammatical features, such as case, which express significant syntactic relations with lexical elements of the patterns.

VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII

Definitional patterns

1. Tn … *sudaryti* … Na (Tn … constitutes … Na)
2. Ta … *sudaryti* … Nn (Ta … constitutes … Nn)
3. Tn … *apimti* … Na (Tn … includes … Na)
4. Ta … *apimti* … Nn (Ta … includes … Nn)
5. Tn … *laikomas* … Ni (Tn … is considered … as Ni)
6. Ti … *laikomas* … Nn (Ti … constitutes … Nn)
7. Tn – … Nn (Tn – … Nn)
8. Nn … – … Tn (Nn … – … Tn)
9. Tn: … Nn (Tn: … Nn)
10. Nn: … Tn (Nn: … Tn)
11. Tn … *būti* … N (Tn … is … N)
12. *N … būti … Tn (N … is … Tn)*
13. *terminas “Tn” (the term "Tn")*
14. *terminas Tn (the term Tn)*
15. *Tg terminas (the Tg term)*
16. *apibrėžimas “Tn” (the definition "Tn")*
17. *apibrėžimas Tn (the definition Tn)*
18. *T apibrėžimas (the T definition)*

Qafe, a pattern concordancer

- *A specific tool was designed to automatically extract concordances for lexico-grammatical patterns from morphologically annotated corpus.*
- Qafe combines a pattern matcher with a flexible representation of lexical items based on a Haskell module called Tefirt developed at the CCL-VDU.

Pattern structure

- Patterns are strings of abstract lexical items (punctuation symbols included), possibly with inserted gaps.
- The maximum gap size (expressed in term of number of inserted word forms) can be stated.

Lexical items

- The lexical items are made of a word form, a lemma and a grammatical information. Each part is optional.
- It allows searching for (for example):
  - a definite word form,
  - a definite lexeme,
  - a lexeme in a given case,
  - an entirely undefined word,
  - an undefined word with some grammatical features (for example a genitive plural substantive).

Data

- corpus files, where patterns are to be searched for;
- the list of searched patterns;
- a list of terms (optional):

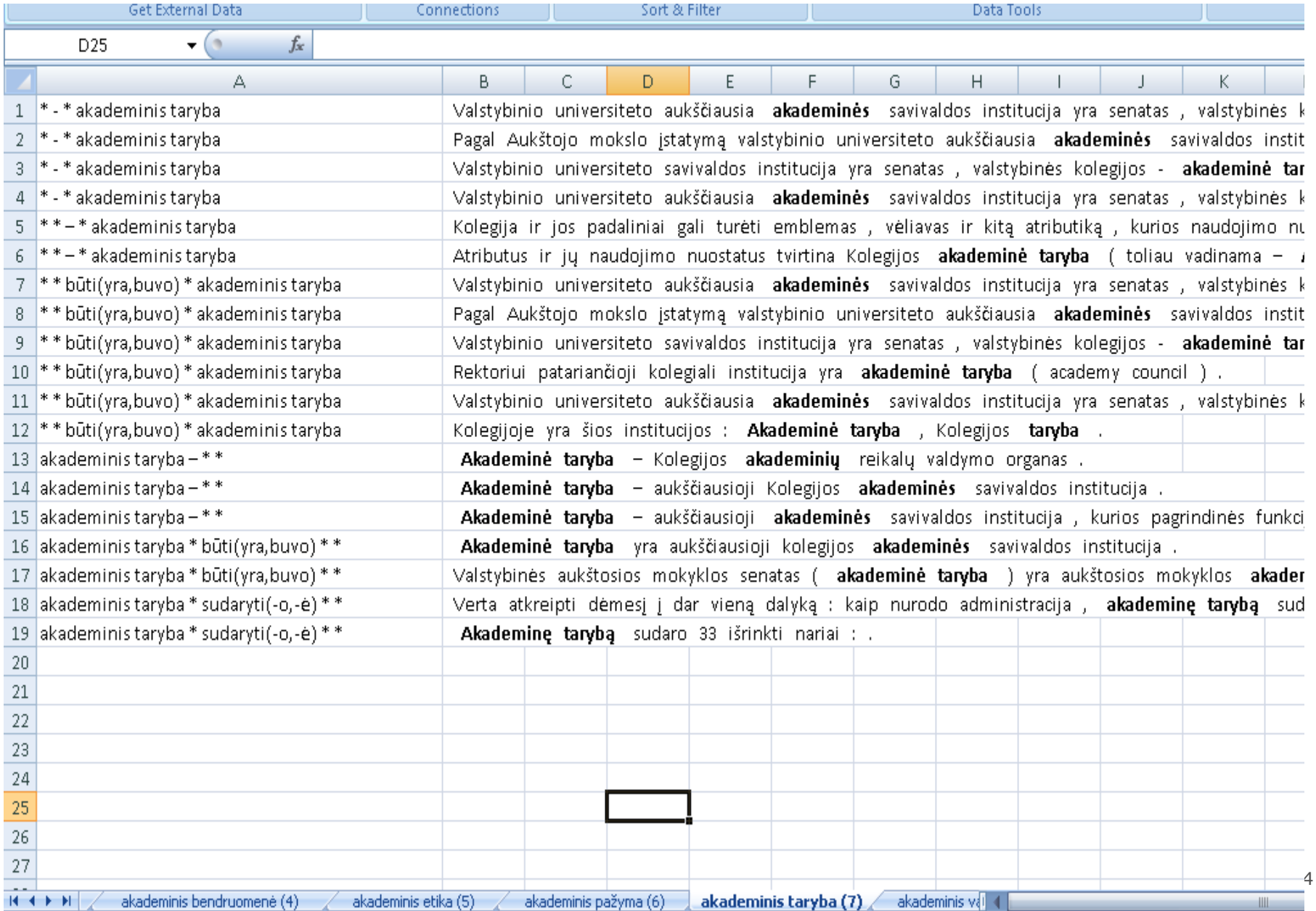
→ allows separating terms and patterns, so that the same pattern can be searched for in combination with different terms,

→ it avoids explicit repetition of the same pattern for each term.

Results

- Results are given as a list of concordances. The dimension of concordances is the sentence.
- Results can be generated in plain text or as XML MS Excel files. In Excel files, each term appears in a different table and each concordance appears as a line preceded by an indication of the extracted pattern.
- In concordances, the extracted term is in bold.

VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII


Get External Data | Connections | Sort & Filter | Data Tools

D25

	A	B
1	* - * akademinis taryba	Valstybinio universiteto aukščiausia **akademinės** savivaldos institucija yra senatas , valstybinės k
2	* - * akademinis taryba	Pagal Aukštojo mokslo įstatymą valstybinio universiteto aukščiausia **akademinės** savivaldos instit
3	* - * akademinis taryba	Valstybinio universiteto savivaldos institucija yra senatas , valstybinės kolegijos - **akademinė tar**
4	* - * akademinis taryba	Valstybinio universiteto aukščiausia **akademinės** savivaldos institucija yra senatas , valstybinės k
5	* * – * akademinis taryba	Kolegija ir jos padaliniai gali turėti emblemas , vėliavas ir kitą atributiką , kurios naudojimo nu
6	* * – * akademinis taryba	Atributus ir jų naudojimo nuostatus tvirtina Kolegijos **akademinė taryba** (toliau vadinama –
7	* * būti(yra,buvo) * akademinis taryba	Valstybinio universiteto aukščiausia **akademinės** savivaldos institucija yra senatas , valstybinės k
8	* * būti(yra,buvo) * akademinis taryba	Pagal Aukštojo mokslo įstatymą valstybinio universiteto aukščiausia **akademinės** savivaldos instit
9	* * būti(yra,buvo) * akademinis taryba	Valstybinio universiteto savivaldos institucija yra senatas , valstybinės kolegijos - **akademinė tar**
10	* * būti(yra,buvo) * akademinis taryba	Rektoriui patariančioji kolegiali institucija yra **akademinė taryba** (academy council) .
11	* * būti(yra,buvo) * akademinis taryba	Valstybinio universiteto aukščiausia **akademinės** savivaldos institucija yra senatas , valstybinės k
12	* * būti(yra,buvo) * akademinis taryba	Kolegijoje yra šios institucijos : **Akademinė taryba** , Kolegijos **taryba** .
13	akademinis taryba – * *	**Akademinė taryba** – Kolegijos **akademinių** reikalų valdymo organas .
14	akademinis taryba – * *	**Akademinė taryba** – aukščiausioji Kolegijos **akademinės** savivaldos institucija .
15	akademinis taryba – * *	**Akademinė taryba** – aukščiausioji **akademinės** savivaldos institucija , kurios pagrindinės funkci
16	akademinis taryba * būti(yra,buvo) * *	**Akademinė taryba** yra aukščiausioji kolegijos **akademinės** savivaldos institucija .
17	akademinis taryba * būti(yra,buvo) * *	Valstybinės aukštosios mokyklos senatas (**akademinė taryba**) yra aukštosios mokyklos **akader**
18	akademinis taryba * sudaryti(-o,-ė) * *	Verta atkreipti dėmesį į dar vieną dalyką : kaip nurodo administracija , **akademinę tarybą** sud
19	akademinis taryba * sudaryti(-o,-ė) * *	**Akademinę tarybą** sudaro 33 išrinkti nariai : .
20		
21		
22		
23		
24		
25		
26		
27		

akademinis bendruomenė (4) | akademinis etika (5) | akademinis pažyma (6) | **akademinis taryba (7)** | akademinis v

Further improvements

- Some bugs and shortcomings have to be addressed.
  - The sentences with an asterisk are not handled properly.
  - Extracted terms are properly emphasized as bold in concordances, but words which constitute subparts of the term (if the term is a multiword expression) are also in bold everywhere in the concordance (even when they are not included in the multiword term).
  - The way the extracted pattern is indicated for each concordance has to be improved.

Evaluation of Automatically Extracted TDCs

- Evaluation set: 57 two-word terms.
- We have established the number of patterns used and their frequency;
- we have manually evaluated automatically extracted TDCs as relevant or irrelevant;
- we tried to establish limitation criteria for the patterns, in order to improve the relevance among automatically extracted patterns.

Accuracy and productivity of definitional patterns

Pattern	Translation	Occurrences	Relevant context	Number of different occurring terms
T ... būti ... Nn	(T … is … Nn)	**88**	40	**30**
Nn ... – ... Tn	(Tn … – … Nn)	**44**	27	**16**
N ... būti ... Tn	(N … is … Tn)	**32**	13	**20**
Nn: ... Tn	(Nn: … Tn)	**19**	6	**11**
Tn – ... Nn	(Tn – … Nn)	**10**	1	**6**
Tn: ... Nn	(T: … N)	**9**	0	**8**
Tn – Nn	(Tn – Nn)	**8**	3	**7**
T apimti N	(T includes N)	**6**	6	**2**
Nn – Tn	(Nn – Tn)	**6**	3	**4**
Tn sudaro Na	(Tn constitutes Na)	**2**	2	**1**
T laikomas N	(T is considered as N)	**1**	1	**1**
T terminas	the term T	**1**	0	**1**
T apibrėžimas	the definition T	**1**	0	**1**

45 per cent (102 cases) of all identified contexts (227 cases) are relevant.

Pattern	Translation	Occurrences	Relevant context	Number of different occurring terms
T ... būti ... Nn	(T … is … Nn)	88	**40**	30
Nn ... – ... Tn	(Tn … – … Nn)	44	**27**	16
N ... būti ... Tn	(N … is … Tn)	32	**13**	20
Nn: ... Tn	(Nn: … Tn)	19	**6**	11
Tn – ... Nn	(Tn – … Nn)	10	**1**	6
Tn: ... Nn	(T: … N)	9	**0**	8
Tn – Nn	(Tn – Nn)	8	**3**	7
T apimti N	(T includes N)	6	**6**	2
Nn – Tn	(Nn – Tn)	6	**3**	4
Tn sudaro Na	(Tn constitutes Na)	2	**2**	1
T laikomas N	(T is considered as N)	1	**1**	1
T terminas	the term T	1	**0**	1
T apibrėžimas	the definition T	1	**0**	1

VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII

Examples of the relevant TDCs

(1) [Tn … *būti* … Nn]

Mokslo institutas *yra akademijos padalinys, kuriame atliekami ilgalaikiai atskirų mokslo krypčių ar šakų fundamentiniai ir taikomieji moksliniai tyrimai...*

Scientific Institute *is an academic institution, where fundamental and applied research is being conducted.*

Mokslų akademija *yra juridinis asmuo.*

The academy of science *is a legal entity*

(2) [N … būti … Tn]

Čia, kaip ir humanitariniuose moksluose, svarbiausias tyrimų rezultatas dažniausiai yra ***mokslinė publikacija***.

Here, as in Humanities, the most important research result is **a scientific publication**.

(3) [Tn … *apimti* … Na] and [Tn … *sudaryti* … Na]

Mokslo institutą *sudaro mokslo laboratorijos, skyriai ir kiti padaliniai bei mokslininkų grupės.*

Scientific institute *is made of scientific laboratories, departments and other subdivisions, as well as groups of scientists.*

Examples of the relevant TDCs

- There are cases when a given term becomes a hyponym and occurs in a row of hyponyms, e.g.:

[N … būti … Tn] *Fakultetuose yra katedros , **mokslo centrai** , laboratorijos ir kiti padaliniai*.

[N … to be … Tn] *In faculties there are departments, **scientific centres**, laboratories and other subdivisions.*

[Nn: … Tn] *Tyrimui buvo pasirinktos septynios dalykinės sritys: verslas , edukologija , **gamtos mokslai** , istorija , matematika , fizika.*

[Nn: … Tn] *For the research seven fields have been chosen: business, educology, **natural sciences**, history, mathematics, physics.*

Examples of the irrelevant contexts

- It has been determined that 55 per cent of all cases have been irrelevant (127 contexts from 227).
- The pattern [Tn … būti … Nn] has produced the biggest number of irrelevant contexts: the verb *būti* (to be) is often used as an auxiliary, which constitutes a sentence predicate in combination with a participle, e.g.

Aukštasis mokslas *yra dalinai mokamas – studentai moka mokestį už mokslą.*
Higher education *is partially paid, as students pay a fee for their education.*

- However, some specific lexical items owe special attention, since certain participles (*būti + žinomas 'is known'*, *įvardijamas 'is named'*, *laikomas 'is considered'*) may form TDCs.

Other sources of inaccuracies

A number of irrelevant contexts comes from a totally different issue, e.g. usage of hyphens, unnecessary spaces before and after typographical elements. Other sources of inaccuracies are related to

1) statistical and table data in texts;
2) non-standard punctuation, that accidentally coincides with a pattern.

Conclusion

- A relatively high proportion of irrelevant contexts shows that some additional limitations (linguistic and typographical) need to be introduced, in order to reduce noise;

- despite the relatively low frequency of term-defining contexts, their quality is high enough to be useful for lexicographers.