Towards the Automatic Extraction of Term-defining Contexts in Lithuanian

Agnė BIELINSKIENĖ, Loïc BOIZOU, Jolanta KOVALEVSKAITĖ and Andrius UTKA

Centre of Computational Linguistics, Vytautas Magnus University, Kaunas, Lithuania

Abstract. The paper presents the ongoing research to apply a pattern-based approach for Lithuanian which can help to automatically extract term-defining contexts from a specialized corpus of education and science. The stages of research include analysis of constituting elements in definitional patterns; formalization of definitional patterns; automatic extraction of term-defining contexts. The first evaluation shows that despite the relatively low frequency of term-defining contexts, their quality can be high enough to serve as a starting point for definitions.

Keywords. knowledge-rich context, term-defining context, specialized corpus, semantic relations, pattern-based approach

Introduction

The paper is dealing with the problem of extraction of terminologically relevant contexts for Lithuanian terms in the specific domain of education and science. The list of terms used in this study was extracted automatically [1] from the 4 million word specialized corpus of Education and Science. The ongoing research is an attempt to apply a corpus-based descriptive approach, in order to obtain contextual information that would help to define the extracted terms for the terminology dictionary of Education and Science.

The definitional knowledge about terms can be extracted from corpora, as it is the context where the term meaning is created [2]. In the knowledge-rich contexts [3] important conceptual characteristics are expressed by semantic relations (hyponymy, meronymy etc.) which can be identified by patterns. Accordingly, it is possible to obtain the relevant terminography-oriented knowledge about the term concept from a corpus and to use this information to provide a starting point for dictionary definitions.

Much of corpus based-descriptive research has been conducted with English data [2] [3], however the methodology has hardly been applied to Slavic [4] or Baltic languages. The paper presents the first attempts to apply a pattern-based approach for...
Lithuanian which can help to automatically extract term-defining contexts (TDCs) from a specialized corpus. The paper presents the following topics: 1) description of term-defining criteria; 2) description of term-defining patterns; 3) automatic context extraction tool; 4) evaluation of results.

1. Description of Patterns for TDCs

The methodology, which has recourse to a rule-based approach, is based on the identification of a set of patterns in a corpus. These patterns are used for the extraction of TDCs, where possible dictionary definitions can be found.

In knowledge-rich contexts [3], conceptual characteristics of a term are typically presented in two ways: through semantic relations (e.g. hypernymy and hyponymy) and through attributes. We also distinguish the contexts between term-defining and term-explanatory. In this paper, however, we focus and analyse them, as these contexts present the most useful information for dictionary definitions, i.e. hypernymic and hyponymic relations. Hypernymic and hyponymic relations can be identified and extracted automatically from a corpus using different methods of relationship extraction [see, for example, [5], p. 77]. We use pattern-based approach since linguistic patterning works well for hypernymy-hyponym relationship.

We have analysed concordances of a restricted set of terms in the Lithuanian Corpus of Education and Science and identified the four textual elements which constitute definitional patterns in Lithuanian terms:

- verb lexical items, such as bate (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 (e.g. Noun ... yra 'is' ... Term, Term ... yra 'is' ... Noun).
- noun lexical items, such as terminas (term), apibrėžimas (definition), which are often included in definitions (e.g. Term apibrėžimas with Term ).
- 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 (e.g. Term:nominative – Noun:nominative).

We have identified 18 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 – Na (Tn – Na)

T – single or multiword term, N – noun, n – nominative, a – accusative, i – instrumental, g – genitive, … – gaps with inserted words.

English translations are given in the present tense of the third person singular, but plural and other tenses are possible.

For both models with the long hyphen (–), the equivalent model with a short hyphen (–) is included. The latter is not standard, but it is a pragmatic decision to deal with texts with improper typography.
8. \( Nn \ldots \quad Tn \ (Nn \ldots \quad Tn) \)
9. \( Tn: \quad Nn \ (Tn: \quad Nn) \)
10. \( Nn: \quad Tn \ (Nn: \quad Tn) \)
11. \( Tn \ldots \ būtī \quad N \ (Tn \ldots \ is \quad N) \)
12. \( N \ldots \ būtī \quad Tn \ (N \ldots \ is \quad Tn) \)
13. \( \text{terminas} \ "Tn" \ (the \ term \ "Tn") \)
14. \( \text{terminas} \ Tn \ (the \ term \ Tn) \)
15. \( \text{Tg \ terminas} \ (the \ Tg \ term) \)
16. \( \text{apibrėžimas} \ "Tn" \ (the \ definition \ "Tn") \)
17. \( \text{apibrėžimas} \ Tn \ (the \ definition \ Tn) \)
18. \( T \ \text{apibrėžimas} \ (the \ T \ definition) \)

Thus, most of the patterns for the extraction of TDCs are not strictly grammatical, lexical or typographic, they include multiple elements, and can therefore be called paralinguistic [3]. In most cases, the case is relevant. Furthermore, some pattern elements may not be contiguous: considering the phrase length in Lithuanian, we decided that it is reasonable to allow 1 to 5 word gaps.

2. Extraction of TDCs

In order to automate the extraction of TDCs, the patterns have been formalized for a tool called *Qafe* (previously *PatternConcordancer*). This tool, which has been specifically designed for the task, searches for given patterns in a morphologically annotated corpus and returns a list of sentences containing the required patterns as concordances.

*Qafe* combines a pattern matcher with a device which allows defining underspecified lexical items. It gives a rich expressive power to patterns and enables testing some alternative formalisations of patterns in order to find the optimal one.

2.1. Pattern Structure

The basic expressions, which constitute patterns, are lexical items with a free degree of determination: grammatical word, word form, lexeme, abstract word defined by a combination of grammatical features. This flexibility arises from a specific Haskell module developed in the Centre of Computational Linguistics of the Vytautas Magnus University (Lithuania). This module, called *Tefirt*, defines a ternary lexical data type made of three parts: lemma, contextual form, and grammatical information. None of the three components has to be explicitly defined. Besides, even grammatical features can be partly specified.

Basic expressions are combined by direct concatenation, but it is possible to allow gaps between basic expressions. The maximum gap size is bounded by the size of the sentence, unless the limit is explicitly specified in the pattern as a maximum number of inserted word.

2.2. Input

The following set of data is required by *Qafe* as an input: 1) a file containing the list of patterns to be searched for, 2) a folder containing a morphologically annotated corpus, 3) a file with the list of terms. Indeed, *Qafe* allows separating the list of terms and the
patterns, with a view to reduce manual work. In this case, the patterns include a variable item to be instantiated during the program execution, so that the same pattern has not to be explicitly repeated for each term.

The search is performed with a view to improve efficiency. Taking into account the significant number of multiword terms derived from a given terminological headword, the program starts to filter concordances with the headword and then attempts to find multiword terms in this far smaller subset.

2.3. Output

Results are given either as plain text, or as an *Microsoft Excel* xml file. In Excel files, results for each term are given in a separate sheet, with each concordance in a different line. The first cell indicates the extracted pattern and the extracted term appears in bold in the nearby concordance. It makes easier to sort, check and evaluate the results.

Some bugs and shortcomings have to be addressed: 1) the sentences with an asterisk are not handled properly; 2) 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); 3) the way the extracted pattern is indicated for each concordance has to be improved to be more user-friendly.

3. Evaluation of Automatically Extracted TDCs

57 two-word terms have been chosen for evaluation. Firstly, we have established the number of patterns used and their frequency (section 3.1). Secondly, we have analysed how many of extracted TDCs are relevant (section 3.2). Finally, we tried to establish limitation criteria for the patterns, in order to improve the relevance among automatically extracted patterns.

3.1. Accuracy and Productivity of Patterns

For the 57 two-word terms the experimental tool *Qafe* has used 13 out of 18 patterns. In Table 1 we present the patterns, their frequencies of occurrence, and the number of relevant contexts. Additionally, Table 1 shows the productivity of each pattern, i.e. the number of different terms which appear with each pattern.

The prototypical definitional patterns are the most frequent: these are the patterns with verbs *būti* (is) and the patterns with hyphens (which replace *būti*). Among the most frequent patterns are the ones where terms are used after the predicate: [Nn ... – ... Tn] and [N ... būtī ... Tn]. Patterns with more definite lexical expressions (e.g. *T apibrėžimas, T terminas*) are very rare.

The most frequent patterns also produce the largest numbers of relevant text-defining contexts. However, the detailed analysis of each pattern has shown that more than half of TDCs cannot be considered as relevant, i.e. they cannot be defined as term-defining (see detailed discussion in section 3.2).

The selected patterns occurred with most of the terms (Table 1). However, 13 terms out of 57 do not appear with the given pattern and, therefore, no defining contexts could be found for them. Most of these terms are rare in the corpus, but the
term frequency is not always the crucial factor regarding TDCs, as it is well known that
the definitions can be of uneven quality, and “high-quality definitions are the exception
rather than the rule” [3].

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Translation</th>
<th>Occurrences</th>
<th>Relevant context</th>
<th>Number of different occurring terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>T = būti N</td>
<td>(T = is N)</td>
<td>88</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>N = – T</td>
<td>(T = – N)</td>
<td>44</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>N = būti T</td>
<td>(N = is T)</td>
<td>32</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>N = T</td>
<td>(N = T)</td>
<td>19</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>T = N</td>
<td>(T = N)</td>
<td>10</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>T = N</td>
<td>(T includes N)</td>
<td>6</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>N = T</td>
<td>(N = T)</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>T = N</td>
<td>(T = N)</td>
<td>8</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>T = N</td>
<td>(T constitutes N)</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>T = N</td>
<td>(T is considered as N)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T = term T</td>
<td>the term T</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T = definition T</td>
<td>the definition T</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

3.2. Analysis of Extracted TDCs

In order to evaluate efficiency of the tool Qafe and the selected patterns, we have
carried out the analysis of automatically extracted TDCs. By applying above mentioned
selection criteria for TDCs, we have manually identified relevant and irrelevant
elements of the contexts.

Relevance of the contexts is presented in Table 1: 45 per cent (102 cases) of all
identified contexts (227 cases) are relevant. The performed analysis has helped us to
improve patterns and create some filtering techniques in order to limit “noise” and to
increase the proportion of relevant contexts.

3.2.1. Relevant TDCs

The analysis of relevant contexts has shown that they may coincide with the classic
Aristotelian form of the definition: \( T \) \( (\text{term}) = X \) \( (\text{hypernym}) + \) differentiating
characteristics. The two next examples present respectively a complete and a partial
correspondence with this traditional definition structure. The partial one lacks the
second component, i. e. the specific difference:

\[
\text{Mokslo institutas yra akademijos padalinys, kuriame atliekami ilgalaikiai atskir}
\]

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

\[
\text{Mokslų akademija yra juridinis asmuo.}
\]

The academy of science is a legal entity.

Contexts that contain a list of hypernyms are also considered to be relevant
contexts that partially coincide with the classic form of a definition:
The Academy of Science is the founder of The Library of the Academy of Science, the Archive of the Academy of Science, The Chamber of the Academy of Science and other institutions.

The definition in its classical form is most often produced by the pattern \([Tn ... \ būti ... N]\). The fifth most frequent pattern is a variant of the first one, where the verb \(būti\) is replaced by a hyphen \(-[Tn -... Nn]\).

A given term may occur in a position that is not typical to the classic form of a definition. Therefore, we had to include the patterns, where hyponym is not on the left, but on the right of hypernym. In this case, the term occurs at the end of the sentence:

\([Nn: ... - ... Tn]\) Tyrimui buvo pasirinktos septynios dalykinės sritys: verslas, edukologija, gamtos mokslai, istorija, matematika, fizika.

\([Nn: ... Tn]\) For the research seven fields have been chosen: business, education, natural sciences, history, mathematics, physics.

It has turned out that classic definitions have not only lexical and grammatical, but also typographical elements. This is well represented in contexts of the pattern \([Nn: ... - ... Tn]\), where the construction \([Z ( kitaip – T) – Y]\) can be frequently found:

\([Nn: ... - ... Tn]\) Moksliniai tyrimai ( kitaip – mokslinis darbas) – tyrimai, skirtomi į fundamentinius ir taikomuosius mokslinius tyrimus.

\([Nn: ... - ... Tn]\) Scientific research (or scientific work) – research that are divided into fundamental and applied scientific research.

Whereas patterns express semantic relations, we tried to establish the means of their expression. We will analyse several such cases. For example, although the patterns \([Tn ... \ apimti ... Na]\) and \([Tn ... \ sudaryti ... Na]\) allow identifying meronymy, some of these contexts also contain hypernyms, cf.:

\([N ... \ būti ... Tn]\) Mokslo institutas sudaro mokslo laboratorijos, skyriai ir kitų padaliniai bei mokslininkų grupės.

\([N ... \ to be ... Tn]\) In faculties there are departments, scientific centres, laboratories and other subdivisions, as well as groups of scientists.

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 kitų padaliniai.

\([N ... \ to be ... Tn]\) In faculties there are departments, scientific centres, laboratories and other subdivisions.
Typographical elements such as punctuation help to identify a term which is included in a row of hyponyms, e.g.:

*Mokslo ir studijų institucijos – universitetai, mokslo institutai, kolegijos – kiekvienais metais Statistikos departamentui pateikia ataskaitas apie MTEP išlaidas.*

*Scientific and study institutions – universities, scientific institutes, colleges – report each year about R&D expenses.*

The initial analysis of relevant terms has been aimed at investigating whether extracted contexts match term-defining criteria: hypernym plus differentiating characteristics. Some of extracted contexts pass these criteria only partially, however these contexts have also turned out to be very important, as they have at least one element of definition (the hypernym). Often such constructions are enumerations that allow getting some additional information that is related with the analysed term. Besides, we have established that extracted contexts express semantic relations that are characteristic for TDCs.

### 3.2.2. Irrelevant Contexts

Irrelevant contexts are those that do not conform to the classic definition formula. In this case, it has been very important to evaluate semantic relations (especially hyponymy).

It has been determined that 55 per cent of all cases have been irrelevant (127 contexts from 227). The irrelevant cases have occurred either due to grammar, or due to the absence of necessary semantic relations (hyponymy) or due to other inaccuracies, that are related to different ways of expression or errors.

The pattern `[Tn ... būti ... N]` has produced the biggest number of irrelevant contexts. The irrelevant contexts have been produced due to the fact in the Lithuanian grammar that the verb *būti* (to be) is often used as an auxiliary, which constitutes a sentence predicate in combination with a participle. These constructions account for 68 per cent of irrelevant contexts for this pattern. In such a case, the structure of the sentence does not match the definition formula since a hypernym of the term is missing:

*Aukštas mokslas yra dalinais mokamai – studentai moka mokestį už mokslą.*

*Higher education is partially paid, as students pay a fee for their education.*

Thus, contexts of the pattern that contain participles of certain forms may be filtered out. However, some specific lexical items owe special attention, since certain participles (*būti + žinomas ‘is known’, įvardijamas ‘is named’, laikomas ‘is considered’*) may form TDCs.

On the other hand, it should be admitted that in spite of mismatch with the classical definition formula, the contexts with participles contain some useful information about the concept of a term and could be considered as knowledge-rich contexts. Therefore, we are considering of labelling them as explanatory contexts.

The other important factor, that makes contexts irrelevant, is the absence of semantic relations between hypernym and hyponym. This is the essential condition of formulation of a classical definition. The situation occurs when a context matches the predefined pattern, but a hypernym is missing:
In the case of the pattern [Nn ... – ... Tn], the hyphen performs very often the sole combining function without a hypernym, thus these contexts do not match the definition formula and they are not term-defining, e.g.:

Šiai pareigybei turėti būt keliamas tik vienas kvalifikacinis reikalavimas – turėti mokslo laipsnį
This specialization requires only one requirement of qualification – to have a scientific degree.

It must be noted that in the two last examples that lack a hypernym, the typography linker (:) or (–) combines predicative structures instead of noun syntagms (which are relevant for term extraction).

A number of irrelevant contexts comes from a totally different issue related to 1) statistical and table data in texts; 2) non-standard punctuation, that accidentally coincides with a pattern; 3) technical errors in the corpus.

Obviously it is not possible to easily compile filtering rules for all context variations. The free word order in the Lithuanian language and naturally occurring contexts are difficult to formalize, in addition to errors or random matches. Nevertheless, we estimate that one third of irrelevant contexts may be filtered out.

Conclusions

1. Although the precise quantitative evaluation is on-going, the first results are promising. Extracted concordances are already successfully used as terminographical data despite their uneven quality and quantity. Usually the TDCs are infrequent, but their quality can be high enough to serve as a starting point for definitions. Thus, even a small number examples can provide essential information for term definitions.

2. The most productive definitional patterns allow recognising semantic relations in naturally occurring contexts. However, a relatively high proportion of irrelevant contexts shows that some additional limitations (linguistic and typographical) need to be introduced, in order to reduce “noise”. After initial analysis, we may claim that some pattern limitations can be applied, e.g. to be + Participle.

3. In this paper, only patterns for TDCs were discussed. But additionally (especially when the corpus does not provide contexts that match defining patterns), the identification of other relevant contexts (explanatory contexts) may give terminography-oriented knowledge. The description and formalization of the patterns for these other kinds of contexts is a part of our future research.

Acknowledgments

The research described in this paper is funded by a grant (No. LIT-2-44) from the Research Council of Lithuania in the framework of the project “Automatic Identification of Education and Science Terms” (ŠIMTAI 2).
References