Verb Subcategorisation Acquisition for Estonian Based on Morphological Information

Siim Orasmaa*

Institute of Computer Science, University of Tartu
J. Liivi Str 2, 50409 Tartu, Estonia
{siim.orasmaa}@ut.ee

Abstract. A method for automatic acquisition of verb subcategorisation information for Estonian is presented. The method focuses on detection of subcategorisation relations between verbs and nominal phrases. Simple comparison of verb-specific argument candidate's frequency ranking against a global frequency ranking of the candidate is used to decide whether the argument candidate is likely governed by the verb. The method also requires only limited linguistic resources from the input corpora: morphological annotations and clause boundary annotations. The results obtained are evaluated against a manually built valency lexicon.

Keywords: verb subcategorisation acquisition, morphological information, frequency ranking comparisons, Estonian.

1 Introduction

Verb subcategorisation information is important in Natural Language Processing, as it specifies morphosyntactic forms of verb arguments and therefore supports tasks related to further syntactic analysis of texts (e.g parsing, grammar building). Many existing methods for automatic verb subcategorisation acquisition require predefined subcategorisation structure (a set of subcategorisation frames) ([1],[2]) or require that the input corpus is in treebank form or fully/partially parsed [3]. However, large machine-readable lists of subcategorisation frames and robust parsers are not available for many languages. The goal of current work is to explore a possibility of subcategorisation acquisition with limited linguistic resources: a corpus having only morphological annotations (word part-of-speech and grammatical categories coded in word form) and clause boundary annotations is taken as the input for the task. The proposed method focuses on acquisition of subcategorisation relations between verbs and nominal phrases (NPs) morphologically marked with semantic cases[1].

In contrast to previous approaches, which often have employed sophisticated statistical methods for identifying verb arguments/subcategorisation frames, the current

* This work is supported by the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS) and by European Social Funds Doctoral Studies and Internationalisation Programme DoRa, which is carried out by Foundation Archimedes.

[1] What is meant by 'semantic cases' is further explained in Section 2.

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 583–590, 2013.

method uses simple comparison of verb-specific argument candidate's ranked list against a globally calculated ranked list of potential candidates to decide whether the verb tends to attract one or some of the candidates. Verb subcategorisation is represented as a ranked list of morphological cases (argument markers), where ranking of a case indicates how likely the case is governed by the verb.

The current study focuses on subcategorisation acquisition for Estonian verbs. Estonian has limited linguistic resources regarding verb valency information and the task has not been attempted for the language before.

This paper has following structure: first, relevant properties of Estonian language in the context of verb subcategorisation acquisition are described, followed by presentation of the related work. Then, the acquisition method is described and achieved results are presented and discussed. Finally, conclusions are drawn and future work is pointed out.

2 Properties of Estonian

Estonian is a language belonging to the Finnic group of the Finno-Ugric language family. As it is characterised by free word order, position of an argument in a clause cannot be taken as an important clue in subcategorisation acquisition (like it has been done in English).

In Estonian, a verb can subcategorise for following constituent types:

1. NP with a specific case-marking. For example, in sentence *Ma hoolin sinust* ‘I care about you‘, the verb *hoolima* ‘to care‘ requires that the object of caring is marked by elative case (word suffix *-st* indicates the elative in *sinust* ‘(about) you‘).
2. Adpositional phrase. Example: *Ametnikud vastutavad andmete eest* ‘Officials are responsible for the data‘, the verb *vastutama* ‘to be responsible‘ requires adpositional phrase (*andmete eest* ‘for the data‘) headed by the adposition *eest* ‘for‘.
3. Infinite verb. Example: *Anu proovis laulda* ‘Anu attempted to sing‘, the verb *proovima* ‘to attempt‘ requires that the action attempted is marked as an infinite verb (*laulda* ‘to sing‘).
4. Specific subclause type. Example: *Ta teatas, et kohtumine jääb ära* ‘He announced that the meeting was cancelled‘, the verb *teatama* ‘to announce‘ requires that-clause (started by subordinating conjunction *et*).

Current work focuses on relations where verb subcategorises for NP with a case marking. In Estonian, nouns and adjectives decline in 14 morphological cases. Traditionally, 3 cases are considered grammatical cases (nominative, genitive and partitive) and 11 cases semantic ones. Because grammatical cases have multiple syntactic functions (marking subject, object and genitive attribute), current work leaves these out and concentrates on syntactically less ambiguous semantic cases.

3 Related Work

For languages with available syntactic resources such as computational valency lexicons, grammars or annotated treebanks, subcategorisation acquisition task is often viewed as task of finding verb subcategorisation frames (SCFs). Methods proposed by Manning [1], and Briscoe and Carroll [2] for verb SCF acquisition in English assumed that SCFs were known in advance (e.g from valence dictionaries). They collected evidence of verb co-occurrence with SCFs from corpora and used statistical hypothesis testing to decide whether a particular verb subcategorises for a certain SCF. Both approaches relied on automatic corpus preprocessing: Manning used partial parsing (part-of-speech tagging and limited chunking), and Briscoe and Carrol applied full parsing.

More recently, Lippincott et al. [4] showed for English that state-of-the-art verb subcategorisation acquisition can be done without the parsed input, just by learning grammatical relations from POS tags within a close proximity of a verb. They used an unsupervised probabilistic model for the task and their model did not need a predefined subcategorisation frame inventory.

Several authors also consider languages with limited syntactic resources. Aldezabal et al. [3] addressed the task on Basque. They noted that adjuncts are a substantial source of noise in SCF acquisition, especially in the context of limited resources, and therefore they focused on the argument/adjunct distinction task. They used partial parsing of the input corpora to obtain instances of verbs together with their dependents and applied statistical filtering methods to distinguish arguments from adjuncts. Kermanidis et al. [5] experimented on subcategorisation acquisition on Modern Greek and English, using only limited preprocessing of input corpora (morphological and part-of-speech tagging, phrase chunking). While these preprocessing settings are similar to the settings in current work, notable difference is that Kermanidis et al. [5] still used phrase chunking (which is not readily available for Estonian), while they did not use information about clause boundaries to limit the set of possible argument candidates.

4 The Method

The method used in current work is based on the empirical notion that the total frequency of a morphological case in a corpus is the sum of verb-specific case frequencies (case occurrences in verb contexts). Because occurrences of a case do not distribute evenly across all verbs, verb-specific case frequency can be used as an indicator of subcategorisation relation with the verb. To confirm this indication, only cases that are more frequent in context of a verb than would be expected by their total frequency are brought out as cases possibly governed by the verb.

The method requires that the input corpus has been annotated for basic linguistic information: sentence boundaries, morphological information (word lemmas, part of speech tags, morphological case and conjugation information) and clause boundaries inside sentences.

In principle, if one tries to find words having possible subcategorisation relation with the main verb in sentence, one could include all the words co-occurring in same sentence with the verb, or one could use a fixed-size window (e.g take N words from the left and the right context of the verb). However, such approaches will be problematic in case of complex sentences consisting of more than one clause, as words from other clause could not have subcategorisation relation with the verb and will add noise to the co-occurrence counts. Therefore the method proposed in this work is based upon linguistically motivated clause boundary annotation, introduced by Kaalep and Muischnek [6]. In addition to separating different clauses, the annotation also marks embedded clauses and thus allows uniting clause parts that have been cut by an embedded clause. For example, from sentence *The house, in which we lived with Piret, belonged to a childless old couple*, two separate clauses must be extracted: *The house belonged to a childless old couple* and *in which we lived with Piret*, so the verbs *lived* and *belonged* can be associated only with the words belonging to their clause-context.

In the first processing step, the method extracts from corpus clauses that contain a finite verb (belonging to the grammatical category of indicative mood active voice). Clauses containing more than one verb will be discarded as they would require additional analysis to determine, which verb governs which nouns. Clauses that contain potential phrasal verbs are also left out. It is done because phrasal verbs (constructions verb + adverb, such as *üle ajama* ‘spill over‘) change subcategorisation structure of a clause, so the resulting structure is different than the structure for a single verb. Clauses with phrasal verbs are filtered out using a list of phrasal verb expressions compiled from the Explanatory Dictionary of Estonian (EKSS) [7] and from the Estonian-Russian dictionary (EVS) [8].

Next, the following information is extracted from each clause: the lemma of the finite verb and the morphological cases of declinable words (nouns, adjectives, numerals and pronouns). While counting morphological cases co-occurring with a verb in a clause, only word types are counted. E.g if there are 5 clauses where verb *haarama* ‘to grab‘ co-occurs with declinable word *käega* ‘by hand‘, then the word *käega* increases counts of its morphological case, comitative, only by 1. This way of counting aims at reducing the bias introduced by lexical items that co-occur frequently with the verb and form idiomatic expressions.

After the counting phase, each verb is associated with a list of case frequencies. Relatively high frequency of a case in context of a verb indicates that it can be in a subcategorisation relation with the verb. However, there can be other reasons for relatively high frequency:

A) The case could indicate a subcategorisation relation with some non-predicate clause member (a noun, an adjective or an adverb) co-occurring with the predicate (finite verb). For example, in clause *see võimaldab disketile salvestamist* ‘this enables saving to disk‘, the verb *võimaldama* ‘to enable‘ only governs nominalisation *salvestamist* ‘(of) saving‘ and the noun *disketile* ‘(to) disk‘ is governed by the nominalisation.

B) The case could be subcategorised not by the verb alone, but by a multiword verb construction. For example, verb *hakkama* ‘to begin‘ with noun *silma* ‘(into) eye‘ forms an idiomatic expression (*silma hakkama* ‘meet the eye‘) which has different subcategorisation structure than the verb alone. Thus, a case frequently co-occurring with *silma hakkama* does not reflect subcategorisation structure of the verb *hakkama*.

C) The case could have overall high frequency in the corpus, so the high frequency in the context of the verb does not necessarily indicate a subcategorisation relation.

For example, the inessive case is the most frequent semantic case in the corpus[2]; however, it often indicates location of action, and because many verbs can optionally specify location of action, the inessive can indicate an adjunct rather than an argument of a verb.

Current work does not address the situations of type A, as these situations would require syntactic analysis. Filtering out phrasal verbs and counting only unique declinable words in the verb's context should reduce the number of situations of type B. In order to address the effects of overall high frequency (C), the list of frequency-sorted cases associated with a verb is compared to the list of frequency-sorted cases from the whole corpus, and only important ranking changes are brought out.

To get the list of total case frequencies, cases are counted in all obtained clauses. To be in accordance with verb specific case counting, here also only word form types contribute to the count of their respective case.

In the following example, (1) is a list of frequency sorted cases from all clauses and (2) is a list of frequency sorted cases associated with verb *tutvustama* ‘to introduce‘. Both lists are sorted in case frequency descending order. Asterisk and number following a case denote the increase in rank, when compared to list (1).

(1) *total*: nom; gen; part; in; el; ad; all; com; ill; tr; abl; es; ter; ab;

(2) tutvustama: nom; part*1; gen; all*3; es*7; in; com*1; ad; ill; el; abl

In the final step, a list of frequency sorted cases associated with a verb is further filtered: only semantic cases that had their rank increased are kept in the list. After the final step, the list of cases associated with the verb *tutvustama* ‘to introduce‘ is:

(3) tutvustama: all; es; com

This result shows 3 subcategorisation possibilities of the verb *tutvustama*. The allative case (word suffix *-le*) marks a person to whom someone/something is introduced, e.g *Ta tutvustas sind meile* ‘She introduced you to us‘. The essive case (word suffix *-na*) marks the role in which someone/something is introduced, e.g *Ta tutvustas end arstina* ‘She introduced herself as a doctor‘. The comitative case (word suffix *-ga*) has two roles: it can mark a person/group to whom someone/something is introduced (*Ta tutvustas sind kõigiga* ‘She introduced you to everyone‘), and if that role is already occupied by allative case, it marks a manner of introducing (*Ta tutvustas sind meile uhkusega* ‘She introduced you to us proudly‘).

5 Evaluation

5.1 Corpus

For subcategorisation acquisition, a 5.8 million word fiction subcorpus of the Reference Corpus of Estonian [9] was chosen, because lexicographers often take examples of verb

[2] Assuming the corpus introduced in the next section. Other examples in the current section are also based on this corpus.

usage (including subcategorisation examples) from fiction texts, and so the results of the system can be more easily compared to examples listed in dictionaries.

After clauses containing multiple verbs and potential phrasal verbs had been filtered out, total 486,192 clauses were obtained, associated with 4677 different verbs. 4542 (97 %) of these verbs co-occurred with at least one declinable word, and 3534 verbs (76 %) co-occurred with at least one declinable word in a semantic case.

5.2 Automatic Evaluation

In order to evaluate performance of the method, the obtained verb subcategorisation information is compared to subcategorisation information in manually built valency lexicon of syntactic analyser for Estonian [10]. The valency lexicon specifies in detail the morphological case alternation related to object of a clause, and also brings out the semantic cases that are subcategorised by the main verb. However, cases listed in the lexicon do not form a complete subcategorisation frame of a verb and it is not specified, whether a case marks an obligatory or an optional argument of the verb.

Because the cases having higher ranking in the results list are interpreted as being more likely governed by the verb, the evaluation method must take this into account. So, each case that occurs in the lexicon, but has a low ranking in the results list or does not appear in the results at all, must be penalised. Similar situation appears in the evaluation of information retrieval systems, where one typically obtains a list of documents as a result of a query and wants to ensure that all the documents relevant to the query appear at the top of the document list. One of the frequently used measures in such setting is mean average precision (MAP), which aggregates results across multiple recall levels and queries to provide a single-figure precision measure [11]. This measure is also used here.

Calculating MAP requires first finding an average precision (AP) for each verb and then taking the mean of all found APs. For a single verb, a precision is calculated at each position in the results list where some case from the lexicon appears, and these precisions are then averaged over all the cases in the lexicon to get the AP. If some case from the lexicon is missing in the results list, the precision is taken 0 at that point. For example, the semantic cases associated with the verb *helistama* ‘to phone‘ in the lexicon are $L = \{all, ill\}$ and the list acquired from the corpus is $C = \{all, ad, ill, abl\}$. The precision for $all \in L$ is 1.0 (as all has top ranking in C) and the precision for $ill \in L$ is $\frac{2}{3}$ (because ill is the 3rd case in C and one redundant case appears before ill). The average precision on detection of semantic cases governed by the verb *helistama* is $\frac{1+2/3}{2} = 0.83$. In order to find the MAP score, such calculations are done for each verb and then the mean of all verb specific APs is taken.

Only verbs governing at least one semantic case were taken from the lexicon for evaluation. Also, lexicon verbs that did not appear in the clauses extracted from the corpus were discarded from the evaluation. This gave a total of 413 verbs for evaluation. These verbs were split into 3 similar size groups by their occurrence frequency (high, medium and low), and MAP scores were calculated for each group separately and for all verbs together.

Results in Table 1 show that the method is sensitive to verb frequency: for verbs occurring less than 17 times, governed cases are detected only with mean average

Table 1. Mean Average Precision on detection of semantic cases governed by verbs. Verbs are divided into groups by their frequency in the corpus.

Group	Verbs in group	Verb frequency range	MAP
Low frequency verbs	140	1–16	56.0%
Medium frequency verbs	136	17–95	79.6%
High frequency verbs	137	98–5415	82.9%
All verbs	413	1–5415	72.7%

precision 56.0%. However, considering medium and high frequency verbs, the method has rather promising mean average precisions (79.6% and 82.9% respectively).

These results support previous research, which has found that simple co-occurrence frequency can be an effective indicator for subcategorisation relations. Kermanidis et al. [5] compare different statistical filtering methods (log likelihood ratio, T-score, binomial hypothesis testing, and filtering by relative frequency threshold) and report that filtering by relative frequency threshold, despite its simplicity, nearly outperforms the other statistical methods.

As lexicons listing the complete subcategorisation frames are not available for Estonian, it is not possible to estimate which is the percentage of subcategorization frames covered by the proposed method.

However, the case lists obtained in this work can be used to aid valency lexicon building: from cases acquired with the method, lexicographer can choose cases for further studying and for including into the lexicon. Case ranking (which can be made more informative by bringing out exact occurrence counts) supports this, as one can have higher confidence about high ranked cases being governed by the verb.

6 Conclusions

In this paper, a method for automatic acquisition of verb subcategorisation information for Estonian has been presented. The focus of the method is on detection of subcategorisation relations between verbs and NPs. The method requires only minimal linguistic annotation (morphological and clause boundary annotations) of the input corpus, and uses simple comparison of verb-specific argument candidate's frequency ranking against total frequency ranking of the candidate to decide whether the candidate is possibly governed by the verb. Verb subcategorisation is represented as a ranked list of morphological cases (argument markers), where ranking of a case indicates the likelihood of the case being governed by the verb. Ranking performance of the method was evaluated against a manually built valency lexicon and mean average precision 72.7% was measured. In future work, the plan is to extend the set of argument types used in the method and also to experiment with other statistical filtering methods used in literature to see whether these methods will produce comparable rankings.

Appendix: List of Case Abbreviations

ab	abessive	*ill*	illative
abl	ablative	*in*	inessive
ad	adessive	*com*	comitative
all	allative	*nom*	nominative
el	elative	*part*	partitive
es	essive	*ter*	terminative
gen	genitive	*tr*	translative

References

1. Manning, C.: Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In: Proceedings of 31st Meeting of the Association of Computational Linguistics, Columbus, Ohio, pp. 235–242 (1993)
2. Briscoe, T., Carroll, J.: Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC, pp. 356–363 (1997)
3. Aldezabal, I., Aranzabe, M., Gojenola, K., Sarasola, K., Atutxa, A.: Learning Argument/Adjunct Distinction for Basque. In: Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition, ULA 2002, Philadelphia, Pennsylvania, vol. 9, pp. 42–50 (2002)
4. Lippincott, T., ÓSéaghdha, D., Korhonen, A.: Learning Syntactic Verb Frames Using Graphical Models. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Korea (2012)
5. Kermanidis, K., Fakotakis, N., Kokkinakis, G.: Automatic acquisition of verb subcategorization information by exploiting minimal linguistic resources. Corpus Linguistics 9(1), 1–28 (2004)
6. Kaalep, H.-J., Muischnek, K.: Robust clause boundary identification for corpus annotation. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey (2012)
7. EKSS: Eesti kirjakeele seletussõnaraamat. ETA KKI, Tallinn (1988–2000)
8. EVS: Eesti-venesõnaraamat I. Eesti Keele Instituut, Tallinn (1997)
9. Kaalep, H.-J., Muischnek, K., Uiboaed, K., Veskis, K.: The Estonian Reference Corpus: Its Composition and Morphology-aware User Interface. In: Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT, pp. 143–146 (2010)
10. Müürisep, K.: Parsing Estonian with Constraint Grammar. In: Online proceedings of NODALIDA 2001, Uppsala (2001), `http://stp.ling.uu.se/nodalida01/pdf/myyrisep.pdf`
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)