Estonian Particle Verbs And Their Syntactic Analysis

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen

Institute of Computer Science
University of Tartu
Tartu, Estonia
{**kadri.muischnek, kaili.muurisep, tiina.puolakainen**}**@ut.ee**

Abstract

This article investigates the role of particle verbs in the Estonian computational syntax in the framework of Constraint Grammar, a rule-based system that performs morphological disambiguation, determines grammatical relations and analyses dependency structure of a sentence. For recognizing the particle verbs, two-fold approach is used: non-compositional particle verbs are listed in a lexicon and compositional ones are composed by the rules. The system achieves both 97.4% precision and 97.4% recall for particle verb recognition. The plans for future include building a valency lexicon of particle verbs and utilizing that in syntactic analysis.

Keywords: particle verbs, MWE, annotation, morphology, syntax, Estonian.

1. Introduction

Particle verbs (also called phrasal verbs or verb-particle constructions) are a common type of multiword expressions, that all are known to cause problems for NLP applications, e.g. (Sag et al., 2002). The problems include identifying particle verbs in texts, representing them in lexicons and handling the complex repertoire of changes they cause in the valency patterns compared to the simplex verbs.

In the present article we are going to investigate particle verbs in Estonian, a heavily inflecting language with free word order (or, rather, constituent order). We are looking at the role of particle verbs in the syntactic processing of Estonian, recognizing these multi-word expressions in texts using the rule-based Constraint Grammar approach. Utilizing linguistic knowledge about syntactic and semantic properties of particle verbs should help the syntactic analyzer not only to recognize these expressions in texts but should also contribute to the overall quality of the syntactic analysis.

The results of the analysis are promising: the system achieved both 97.4% precision and 97.4% recall for particle verb recognition.

The rest of the article is structured as follows. Section 2 gives some insights into the literature describing the representation and analysis of particle verbs in frameworks for syntactic analysis of various languages. Section 3 investigates particle verb as a linguistic phenomenon, comparing its linguistic properties in English and Estonian. Section 4 describes the formalism, namely Constraint Grammar, with emphasis on the presentation and recognition of particle verbs. Section 5 presents the results of the experiment of recognizing particle verbs in text and evaluates these results. Section 6 concludes the paper.

2. Related work

The problems related to multiword expressions have been quite popular research topic over the last decade. Several articles address the problems of extracting particle verbs from a text corpus (Baldwin and Villavicencio, 2002; Ramisch et al., 2008) or representing them in a lexicon (Villavicencio, 2003). As for parsing, James W. D. Constable and James R. Curran (2009) aim at better representation of particle verbs in the framework of Combinatory Categorial Grammar. They were able to produce more syntactically and semantically sound annotation of verb-particle constructions without losing in the overall parsing quality.

Martin Forst, Tracy Holloway King, and Tibor Laczko (2010) describe the ways in which particle verbs are implemented in English and German ParGram LFGs, and they also discuss the issues that arise with respect to particle verbs in the development of a computational LFG for Hungarian. The authors argue for uniform account for English, German and Hungarian particle verbs and propose different representation for compositional and non-compositional particle verbs so that the idiomatic particle verbs are listed in the lexicon as they may have argument structures which differ significantly from the verb's non-particle counterpart. The compositional particle verbs could be composed in the syntax and if the particle introduces an additional argument then a rule is used to create a new description for the verb which differs from the original description of the simplex verb only in the addition of the new argument.

The research on the Estonian particle verbs in the computational linguistic context has been recently conducted by Muischnek and Kaalep (2008). A prototype program based on this research performed well, achieving the recall of 90% and the precision of 92%. The program is meant for tagging multi-word verbs in morphologically annotated and disambiguated corpora.

3. Particle verbs in Estonian

A particle verb consists of a verb and a particle; the latter can often be homonymous with an adposition in Estonian. Also, in Estonian, just like in English a particle verb can be intransitive (Eng *take off*) or transitive (Eng *look it up*).

This work was supported by Estonian Ministry of Education and Research [grant SF0180078s08] and the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS).

Basically, the Estonian particle verbs also follow the three way classification adopted e.g. by Nicole Dehe in (Dehe, 2002), where a verb-particle combination can be classified as compositional, idiomatic or aspectual, depending on the way its meaning is composed from the meanings of its components. Compositional particle verbs consist of a verb and a particle, both used in their literal meaning, e.g. Eng *clean up*. The meaning of an idiomatic particle verb, on the other hand, is idiosyncratic and can't be inferred from the literal meanings of the verb and particle, e.g. Eng *give in*. Aspectual particles add change of state/telicity to verbs of process or activity, e.g. Eng *eat up*.

The idiomatic particle verbs form a closed set; the compositional and aspectual particle verbs can, at least theoretically, be produced by combining all semantically relevant verbs and particles, for example all motion verbs with all directional particles or some aspectual particles with otherwise non-aspectual verbs. Which particles and verbs really do combine and form a compositional or aspectual particle verb is a very interesting research question that can be answered using big text corpora, but this remains outside the scope of this article.

As for the particle placement and word order, the verb and the particle do not need to be adjacent to each other in Estonian and the order of the components may vary depending on the clause type, resembling much the behaviour of particle verbs in German (examples 1 vs 2). However, differently from German, Estonian particle and verb combinations are spelled as two words even if the particle immediately precedes the verb.

(1) Ma **vaatan** need paberid homseks **üle**.
I look these papers tomorrow-TR over
'I shall look over those papers by tomorrow'

(2) Kui sa need paberid **üle vaatad**, siis on
If you these papers over look then is
kõik valmis.
everything ready
'Once you have looked over those papers, we will be done.'

Most of the verb particles are homonymous with pre- or postpositions (Estonian has both of them), creating a disambiguation problem, similar to the one concerning the English word *over* in examples 3 vs 4.

(3) He **looked over** the papers in less than 10 minutes.

(4) He **looked** over the fence and saw his neighbour.

Just like in English examples the word-forms *look* and *over* form a particle verb *look over* in example (3), but don't belong together in the same way in example (4), the Estonian verb *vaatama* 'to look' and adverb *üle* 'over' form a particle verb in the examples (1) and (2), but not in example (5), where *üle* is a preposition.

(5) Ta **vaatas** üle aia ja nägi oma
s/he looked over fence-GEN and saw own
naabrit.
neighbour-PART
'S/he looked over the fence and saw his/her neighbour.'

As a pre- or postposition has to be adjacent to the noun phrase that is the constituent of the adpositional phrase, they are usually easier to detect. In (6), however, the invariable word *üle*, which can function both as a particle and a preposition, is positioned before the noun *jõu* 'force' in genitive case form, as if *üle* were a preposition in the prepositional phrase *üle jõu* 'exceeding capabilities'. Actually, *üle* functions as a particle in this clause, forming the particle verb *läks üle* 'went over, switched'.

(6) Meelitustelt **läks** ta **üle** jõu
flattery-PL-ABL went s/he over force-GEN
kasutamisele.
utilization-ALL
'S/he switched from flattery to violence'

Many of these invariable words that can function both as particles and as pre- or postpositions are quite frequent in texts. The most frequent simplex verbs are also the most frequent verbal components forming various verbal multiword expressions, among them particle verbs. Sentences of the written language tend to make an extensive use of inserted infinitival clauses and nominalizations. All this results in sentences where there are possible components of several particle verbs scattered across the clause.

Like in English, also in Estonian the valency (i.e. the possible arguments and their morphological coding) of a particle verb can differ from that of the simplex verb. E.g. Eng *make* is a transitive verb, whereas *make off* is an intransitive particle verb. In Estonian, transforming a simplex verb into a particle verb can alter the valency pattern in several ways: a transitive verb may transform to an intransitive particle verb and an intransitive verb may transform to a transitive particle verb, various adverbial arguments could be inserted or deleted etc.

4. Recognition of particle verbs by the syntactic analyzer

The syntactic analyzer of Estonian (Müürisep et al., 2003) is based on the Constraint Grammar (CG) formalism (Karlsson et al., 1995) and its last version uses VISL CG-3 format and software[1]. The analyzer consists of separate sets of grammar rules for determination of clause boundaries, morphological disambiguation, syntactic mapping and syntactic function assignment. A set of rules for recognition of dependency relations is currently under development. Recent syntactic tagging of the Estonian corpus of written text showed that the recall of shallow syntactic analysis is 92.6% and precision 72%.

Both lexicon-based and generating approaches are used for recognizing the particle verbs in text. The initial list of particle verbs derives from a monolingual dictionary (Langemets, 2009), the current version of the lexicon contains 2175 entries. By set-combining rules that combine

[1] VISL project homepage at the Institute of Language and Communication, University of Southern Denmark: http://beta.visl.sdu.dk/

145 verbs of movement with 28 directional particles and another list of otherwise non-aspectual verbs with aspectual particles 4060 particle verbs can be additionally generated and recognized. Altogether those lists contain 6235 particle verbs.

The particle verbs in the lexicon contain 150 different particles, each of them composing 1-120 different particle verbs. For example particle *välja* 'out' is a component of 120 particle verbs, while particle *üleval* 'above' occurs only as a component of one idiomatic particle verb, *üleval pidama* 'stay in line'.

The parser also uses a valency lexicon. The valency lexicon of simplex verbs contains 4252 entries, for particle verbs it holds valency patterns for 41 particle verbs so far.

As described in the Section 3, many verb particles can also have a pre- or postpositional readings or even readings of a case form of a noun in some cases. During morphological disambiguation the context of verb particles is often very ambiguous, for example, the cases of adjacent nouns are often unclear and other essential ambiguities as verb-noun ambiguity can occur in some sentences. For these reasons, some of verb particles get incorrect readings and conversely, some adpositions or nouns receive particle reading by mistake. For minimizing the number of such errors, a new group of correcting rules was introduced that corrects erroneous particle or adposition readings based on more clear context knowledge.

Also, the same particles can combine with different verbs to form different particle verbs. Particles can also combine with verb nominalizations forming compounds. So additional rules were devised for distinguishing particle readings from pre- or postposition or noun readings in specific contexts and recognizing also nominalizations of particle verbs.

Special rules were introduced for two large groups of relatively regularly combining particle verbs: 1) motion verbs plus directional particles and 2) all verbs plus some perfective (aspectual) particles, for example the "universal perfectivizer" *ära*. Not all of those verbs and particles can really combine with each other, but if they are both present in a clause, they quite certainly form a regular particle verb; excluding contexts where potential particles are in fact pre- or postpositions.

Several stages of analysis are needed for the complete identification of particle verbs in the text. First, morphological disambiguation rules resolve ambiguities between particle and pre- or postposition or noun readings. After morphological disambiguation a new stage is introduced that specializes on determining which particles and verbs present in the clause actually combine to form a particle verb. In example (7) there are two verbs, *minema* 'to go' and *tulema* 'to come' and also an aspectual particle *ära*. Two particle verbs, *ära tulema* 'come off, come away' and *ära minema* 'go away' are potentially possible, but in this particular context, only *ära tulema* is really present in the clause.

(7) Koju minnes **tuli** saapal tald **ära**.
home going came boot-ALL sole off
'While s/he was going home a sole came off his/her boot'

Finally dependency-relation rules link components of particle verbs assigning a verb as a head and particle as dependent. Altogether, the grammar for the identification of the Estonian particle verbs consists of approximately 700 rules.

The example in Figure 1 illustrates the parsing process. The sentence *Kutsun Eesti elanikke üles oma arvamust avaldama* 'I call the residents of Estonia to voice their opinions' includes a particle verb *üles kutsuma*. The wordforms are in separate lines followed by their morphological description, valency information, the syntactic label and link to the upper node in the dependency tree. In the first stage of analysis the morphological disambiguation rules select the correct morphological readings and remove the ones which do not fit the context. Special rules add valency information (see tags <PhVerb>, <üles> <NGP–P>). After that additional disambiguation and tag correction takes place. During the next step, shallow syntactic rules add possible syntactic tags (see @FMV, @NN>, @OBJ in the example) to every reading and then start to remove the ones which can not appear in the current context. We call these rules constraints. In the last step the dependency analysis takes place. Each word in the sentence gets a link to its head (see tags #No->No). The first word *kutsun* 'call' is a finite main verb which is a root of the whole sentence. The particle component belonging to it *üles* 'up' is in the position 4. It has a special syntactic tag @Vpart and a link #4->1 pointing to its head. *Elanikke* 'residents' is an object which belongs to the verb *kutsun* 'call', while *arvamust* 'opinion' is an object which belongs to infinite verb *avaldama* 'voice, present'.

5. Results and Evaluation

The evaluation of particle verb recognition was conducted on the 94,279-word manually annotated (for syntactic functions and dependency relations) corpus. The text contained 1379 particle verbs, of which 1344 were correctly recognized at the level of syntactic function assignment. Altogether the analyzer recognized 1379 particle verbs, so 35 of them were recognized as particle verbs by mistake. In terms of recall (ratio of correctly recognized particle verbs and all correct particle verbs in the text) and precision (ratio of correctly recognized particle verbs and all particle verbs recognized by the analyzer in the text) we have observed the recall of 97.4% and the precision of 97.4% for recognizing particles. That means that 2.6% of particles were not recognized as components of particle verbs and another 2.6% were incorrectly labelled as components of particle verbs. 1349 particle components of particle verbs got correct dependency relation with correct verb giving 97.8% recall of dependency relations, whereas out of 2.2% errors 1.6% (21 particles) were linked to the wrong verb and 0.7% (9 particles) were not recognized as verb particles at all. 1.9% (26 particles) were not recognized as verb particles, but were linked to the correct verb as adverbials. 95.9% (1323) of particle verbs received both the correct syntactical reading and dependency relation.

Errors occur mainly due to insufficient size of the lexicon that does not contain infrequently used particle verbs.

```
"<Kutsun>"                        'call-1.sg'
    "kutsu" Ln V main indic pres ps1 sg ps af <NGP-P> <PhVerb> <üles> <0> @FMV #1->0
"<Eesti>"                         'Estonia-sg.gen'
    "Eesti" L0 S prop sg gen cap @NN> #2->3
"<elanikke>"                      'resident-pl.part'
    "elanik" Le S com pl part @OBJ #3->1
"<üles>"                          'up'
    "üles" L0 D @Vpart #4->1
"<oma>"                           'own-sg.gen'
    "oma" L0 P pos det refl sg gen @NN> #5->6
"<arvamust>"                      'opinion-sg.part'
    "arvamus" Lt S com sg part @OBJ #6->7
"<avaldama>"                      'voice-inf'
    "avalda" Lma V main sup ps ill <NGP-P> <All> @ADVL #7->1
"<.>"
    "." Z Fst #8->8
```

Figure 1: Parse tree of the sentence *Kutsun Eesti elanikke üles oma arvamust avaldama.* 'I call the residents of Estonia to voice their opinions.'

	Number of particle verbs in the text	Recognition of syntactic function	Recognition of dependency relations	Recognition of both syntactic functions and dependency relations
number	1379	1344	1349	1323
recall %	(100)	97.4	97.8	95.9

Table 1: Recognition of syntactic functions and dependency relations of particle verbs.

The lexicon of particle verbs has to be increased, that is especially true for non-compositional idiomatic particle verbs.

Typical source of errors are sentences with inserted infinitival clauses or nominalizations where the particle-verb pair intersects with other verbs or participles or other nominalizations that could possibly also combine with same particle.

For example, all particle verbs from example sentences in section 3 get correct analyses, but in example (8) the word-form *riiuli* is erroneously analysed as the object of the particle verb *peale panema* 'put on'; actually in this sentence peale is a postposition. In both cases the word-form *riiuli* is a noun in genitive case, that can function both as a genitive attribute and a noun in a postposition phrase. If *riiuli* were the grammatical object, the meaning of the sentence would be 'Let's put the shelf on top of the big room'.

(8) **Paneme** suurde tuppa riiuli peale ...
put-2.PL big-ILL room-ILL shelf-GEN onto
'Lets put (it) on shelf in the big room'

Like in the example (8), also in the examples (9 vs 10) semantics plays main role in deciding which reading is correct in particular context. These examples illustrate that there are cases while it is impossible to choose between readings of particle and preposition without semantic knowledge. Both examples contain verb *elama* 'to live' and an uninflecting word *üle* 'over' that can function both as a preposition and as a particle. In the sentence (10), *üle* and *elama* form a particle verb meaning 'to experience', but in the sentence (9) *üle* is part of a prepositional phrase *üle tänava* 'across the street'.

(9) **Elasin** üle tänava
live-1.SG over street-GEN
'I lived across the street' (üle is functioning as a preposition)

(10) **Elasin** **üle** vapustuse
live-1.SG over shock-GEN
'I experienced a shock' (üle is functioning as a particle)

The overall recall and precision in shallow syntax have improved by 1.4% (in recognition of syntactic functions in general) that is quite good result taking into account the relatively small proportion of particle verbs in the text. In some particular sentences a progress could be seen, in terms of overall better annotation of the syntactic structure, but the proportion of such cases was not high.

We have to conclude the results section conceding that one of our hypotheses, namely the assumption that recognizing particle verbs and using the knowledge about their valency patterns considerably improves the quality of syntactic analysis, did not fulfil completely. We are on the opinion, that the reasons for that are the following.

First, the valency lexicon of particle verbs is too small, containing the valency information for only 41 particle verbs.

Second, in Estonian a complement, also a grammatical object, can quite easily be omitted, and the sentence is still grammatical.

Third, it is difficult to differentiate between adjuncts and complements basing on their morphological form only. For example, an allative is a frequent case for complements of particle verbs; but its prototypical function is to code locative meaning and so it is also a frequent case for locative adverbials.

Fourth, original syntactic constraints need an additional fine tuning, in order to take into account the new lexical information. So one could say, that compiling a bigger valency lexicon of particle verbs and using this information in the course of syntactical analysis remains to be done in the future.

Still, the recall of 97.4% of correctly recognized particle verbs is a good result and makes it possible to use this annotation for practical linguistic needs.

6. Conclusion

This article focused on the possibilities of increasing the quality of the syntactic analysis by recognizing particle verbs and so being able to use the information about their valency patterns for the shallow and dependency parsing. The language we are working with is Estonian, a language characterized by rich inflectional system and free word order.

A two-fold approach is used for recognizing the particle verbs: the regularly combining units are produced by the rules and the idiosyncratic ones are listed in the lexicon.

As our results indicate, our lexicon and rule based approach can be regarded as successful. 97.4% of particle verbs receive correct analysis at shallow syntactical level, 97.8% of particle verbs get correct dependency relations (i.e. the particles get combined with correct verbs) and 95.9% receive both correct syntactic function and dependency relation tags, what makes it possible to use annotated data for practical linguistic purposes.

The work in near future will concentrate on modeling the changes the particles cause in valency patterns of verbs and utilizing the knowledge about the valency patterns of the particle verbs in syntactic analysis.

7. References

Baldwin, Timothy and Aline Villavicencio, 2002. Extracting the Unextractable: A Case Study on Verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*.

Constable, James and James Curran, 2009. Integrating Verb-Particle Constructions into CCG Parsing. In *Proc. of the Australasian Language Technology Association Workshop 2009*. Sydney, Australia.

Dehe, Nicole, 2002. *Particle Verbs in English. Syntax, information structure and intonation*. John Benjamins.

Forst, Martin, Tracy Holloway King, and Tibor Laczkó, 2010. Particle verbs in computational LFGs: Issues from English, German, and Hungarian. In Miriam Butt and Tracy Holloway King (eds.), *Proc. of the LFG '10 Conference*. Ottawa, Canada: CSLI Publications.

Karlsson, F., A. Anttila, J. Heikkilä, and A. Voutilainen, 1995. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Langemets, Margit (ed.), 2009. *Eesti keele seletav sõnaraamat (Explanatory Dictionary of Estonian)*. Tallinn, Estonia: Eesti Keele Sihtasutus.

Muischnek, Kadri and Heiki-Jaan Kaalep, 2008. Multiword verbs of Estonian: a database and a corpus. In *Proc. of the LREC Workshop "Towards a Shared Task for Multiword Expressions"*. Marrakech; Morocco.

Müürisep, Kaili, Tiina Puolakainen, Kadri Muischnek, Mare Koit, Tiit Roosmaa, and Heli. Uibo, 2003. A new language for Constraint Grammar: Estonian. In *Proc. of International Conference Recent Advances in Natural Language Processing RANLP 2003*. Borovets, Bulgaria.

Ramisch, Carlos, Aline Villavicencio, Leonardo Moura, and Marco Idiart, 2008. Picking them up and Figuring them out: Verb-Particle Constructions, Noise and Idiomaticity. In Kristina Clark, Alex; Toutanova (ed.), *Proc. of the Twelfth Conference on Natural Language Learning (CoNLL 2008)*. Manchester, UK: Association for Computational Linguistics.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger, 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*.

Villavicencio, Aline, 2003. Verb-Particle Constructions and Lexical Resources. In *Proc. of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

List of abbreviations

ALL allative case; GEN genitive case; ELA elative case; ILL illative case; INE inessive case; KOM komitative case; NEG negative; PART partitive case; PL plural; 1.SG 1st person singular; 2.PL 2nd person plural