Robust clause boundary identification for corpus annotation

Heiki-Jaan Kaalep, Kadri Muischnek
Dept of Computer Science
University of Tartu
Estonia
Heiki-Jaan.Kaalep@ut.ee, Kadri.Muischnek@ut.ee

Abstract

The paper describes a rule-based system for tagging clause boundaries, implemented for annotating the Estonian Reference Corpus of the University of Tartu, a collection of written texts containing ca 245 million running words and available for querying via Keeleveeb language portal. The system needs information about parts of speech and grammatical categories coded in the word-forms, i.e. it takes morphologically annotated text as input, but requires no information about the syntactic structure of the sentence. Among the strong points of our system we should mention identifying parenthesis and embedded clauses, i.e. clauses that are inserted into another clause dividing it into two separate parts in the linear text, for example a relative clause following its head noun. That enables a corpus query system to unite the otherwise divided clause, a feature that usually presupposes full parsing. The overall precision of the system is 95% and the recall is 96%. If "ordinary" clause boundary detection and parenthesis and embedded clause boundary detection are evaluated separately, then one can say that detecting an "ordinary" clause boundary (recall 98%, precision 96%) is an easier task than detecting an embedded clause (recall 79%, precision 100%).

Keywords: Estonian, corpus, clause splitting

1. Introduction

Clause splitting is often regarded a subtask of syntactic analysis, but it can also be viewed as a task in its own. For many languages, large-scale automatic syntactic analysis is still an unsolved issue, at least to some extent. Nevertheless, the information about clause boundaries is necessary for solving several tasks that otherwise don't require (full) syntactic analysis. For example, for a collocation extraction system it would be better to combine the word-forms in the whole clause instead of the usual window of 3-4 words while extracting candidate pairs of multiword verbs (e.g. particle verbs, verbal idioms, support verb constructions) from a text of a language with a free word order (e.g. German or Estonian).

Another example benefiting from annotated clause boundaries is a corpus query system. Often the users would be willing to explore co-occurrences of words, lemmas or grammatical categories and again, at least in a free word-order language, the appropriate context for retrieving many of those co-occurrences would be a clause and not the long sentence of the written language.

As a solution to that problem we have implemented a clause splitting system as a separate module. Our rule-based system needs information about parts of speech and grammatical categories coded in the word-forms, i.e. it takes morphologically annotated text as input, but requires no information about the syntactic structure of the sentence.

Among the strong points of our system we should mention identifying parenthesis and embedded clauses, i.e. clauses that are inserted into another clause dividing it into two separate parts in the linear text. That enables e.g. a corpus query system to unite the otherwise divided clause, a feature that usually presupposes full parsing.

The target language of the clause boundary identification system is Estonian, a language belonging to the Finnic group of the Finno-Ugric language family and characterized by rich morphological system and relatively free word-order.

The system described in this article has been used for annotating clause boundaries in the Estonian Reference Corpus of the University of Tartu, a collection of written texts containing ca 245 million running words and available for querying via Keeleveeb[1] language portal.

2. What is a clause or what should be treated as a clause by a clause splitting system?

According to (Ejerhed, 1996), there are many open questions, even for a single language, concerning the definition of the clause units to have as targets for clause segmentation; clause definitions and clause segmentation rules are highly language specific.

Elsevier's Encyclopaedia of Language & Linguistics (2006) gives two definitions of the concept 'clause':

- A syntactic unit consisting of subject and predicate that alone forms a simple sentence and in combination with others forms a compound sentence or complex sentence.
- In modern grammars, sometimes identified as a unit larger than a phrase but smaller than a sentence, to account for clauses that fall outside the traditional 'subject, predicate' pattern.

The clauses can be further divided into finite and infinite ones depending on the finiteness of their main verb. Some

[1] www.keeleveeb.ee

grammatical formalisms consider infinite constructions (i.e. constructions with infinite main verb) phrases rather than clauses, some other formalisms as clauses (because they can be analysed into clause elements). For the present work, we are targeting the finite clauses and a subclass of the infinite ones – namely the gerundial clauses. The first reason behind that decision is that these non-finite constructions are regarded as the most "sentence-like" by the authors of the academic grammar of Estonian (EKG II). The other reason is that these infinite clauses are always separated from the main clause with a comma and thus are more easily identifiable.

For the present work we generally do not identify the clause type, but we do differentiate between parenthesis and embedded clause vs. "ordinary" clauses. By parenthesis and embedded clause we mean a clause that is inserted into another clause dividing the latter into two halves. For example, a relative clause following its head noun is often an embedded clause. (According to the rules of Estonian orthography, a relative clause is always separated from the main clause by comma(s).) Another typical example is parenthesis that is always separated from the "outside" clause by dashes, brackets or commas. Identifying these clauses enables us to re-unite the divided clause for the further applications.

For example, in sentence (1), there is a verbal idiom *jalga laskma* 'to run away, lit. to shoot the foot' and the components of that idiom are separated from each other by a relative clause *kes mu autot rammis* 'who rammed my car' modifying the subject of the main clause, *taksojuht* 'taxi-driver'. Identifying simply two clause boundaries (shown as vertical bars in the example) in that sentence would make identifying the multi-word item *jalga laskma* impossible; only recognizing the relative clause as a parenthetical clause enables to treat the sequence *Siis lasi taksojuht jalga* 'Then the taxi-driver ran away' as one clause.

(1) Siis lasi taksojuht, | kes mu autot rammis, | jalga.
Then shoot taxi-driver who my car-PART rammed foot-PART
'Then the taxi-driver, who had rammed my car, ran away.

3. The algorithm

3.1 Basic assumptions

When creating the algorithm, we assumed that when a writer is creating a sentence, he has a certain repertoire of devices – conjunctive words, punctuation marks, word order etc – for signalling the beginning and/or end of the coherent sub-sentential units we are interested in. A complex sentence has to contain some elements from this repertoire, and if one meets them, he can be sure that he is facing a certain type of sentence structure, even if some otherwise compulsory element (e.g. a finite verb) is missing. E.g. the Estonian sentence (2) contains no finite verb, and only two infinite ones, but still has clearly three separate units, signalled by commas and conjunctive words *kui* 'if, when', *siis* 'then' and *et* 'that, for':

(2) Kui osta külmkapp, siis ikka selleks, et toitu säilitada.
If buy-INF refrigerator then still that-TRANSL for food-PART store-INF
The reason for buying a refrigerator is to store food, what else?

When confronted with a complex sentence, the program typically faces more than one conjunctive word and/or punctuation mark, and it is not clear from the beginning which ones are used for separating clauses, and which ones for subclausal coordination. The general idea of the algorithm is to proceed step by step, identifying more clear-cut cases of clause boundaries and nested constructions, before moving on to less obvious ones. A sentence may be traversed several times.

As an input, the program receives a morphologically analysed and disambiguated text, i.e. every wordform has its lemma and grammatical categories determined. (We rely on the tools *etmrf* and *t3mesta* by Filosoft Ltd. to perform these tasks.)

In short, the algorithm proceeds as follows. First, some fail-safe parentheses (cf part 2) are tagged. Second, the verb forms that might be suitable for acting as main verbs in clauses (i.e. verbal centres) are tagged. Third, the punctuation marks and conjunctive words are tagged as potential clause boundary indicators. Fourth, these potential clause boundaries are classified step-by-step into true boundaries and non-boundaries. Finally, relative clauses are marked as embedded clauses, so that subsequent tools, e.g. a corpus query processor or a multi-word verb phrase tagger are able to look at a coherent span of words, omitting the inserted construction.

3.2 Step by step details

3.2.1 Step 1

Text in brackets, even if it is only one word, is marked as a parenthetical unit. Brackets are a fail-safe indicator for separating their contents from the rest of the sentence.

3.2.2 Step 2

The following types of verb forms are tagged as possible verbal centres of clauses:

1) A finite verb form

2) A past participle, either a personal (*nud*- participle) or an impersonal one (*tud*-participle), if it acts as a part of a compound verb form. The latter consists of a negation word *ei* and/or a finite form of the verb *olema* '*to be*', followed by a participle, e.g. *ei kolinud* '*did not move*', *ei kolitud* '*were not moved*', *olen kolinud* '*I have moved*', *olin kolitud* '*I had been moved*', *ei olnud kolinud* '*had not moved*' etc. The negation word and the auxiliary verb *olema* '*to be*' have to precede the participle; otherwise the participle would be interpreted as an attribute, e.g. *kolinud*

pere '*a family that has moved*', *kolitud mööbel* '*moved furniture*'.
However, it is normal that a complex sentence in present perfect, past perfect or having negation consists of several clauses and only the first one has the compulsory auxiliary verb or the negation word, while the rest of the clauses contain only the participle, as in sentences (3), (4).

(3) Ta oli avanud akna ja lõhkunud ukse.
He had opened the window and broken the door.

(4) Ta oli kiiresti avanud akna ja lõhkunud ukse.
He had quickly opened the window and broken the door.

How can we determine that the latter participles present a compound verb form in this case, not an attribute? A simple rule would be: if there is a compound verb form in a sentence, and you later meet the same type of the participle (*nud*- or *tud*-participle respectively), positioned immediately after either a comma or a conjunction word, then this participle acts as a verb. This rule in turn assumes that we have already reliably identified an instance of the compound verb form, a task that is not trivial. The present algorithm requires the negation word or the auxiliary verb to be adjacent to the participle; otherwise, the participle is labelled as an attribute. The requirement of adjacency is in fact too restrictive (see (4)), and as a consequence, not every instance of the compound verb is recognised. This in turn will result in labelling later participles incorrectly as attributes, and possibly failing to tag the clause boundaries. How to overcome this limitation in a principled way is an issue for future work.
3) The gerund (*des*-form) (5) and the negative form of a gerund (*mata*-form), represented morphologically by the abessive case form of the supine (6) may also act as a verbal centre of a non-finite clause. Similarly, the *maks*-form, morphologically the translative case of the supine, having the meaning "*in order to VERB*", may act as a verbal centre of a non-finite clause (7). However, in order to qualify, they have to follow a comma, or, in case of the *des*- and *mata*-form, be the first word in the sentence.

(5) Ta vahetas töökohta, kolides teise linna.
He changed his job, moving to another town.

(6) Ta vahetas töökohta, kolimata teise linna.
He changed his job, without moving to another town.

(7) Ta lahkus täna varakult, kolimaks oma uude koju.
He left today early in order to move to his new home.

3.2.3 Step 3
Conjunctions *ja* '*and*', *ning* '*and*', *ega* '*neither*', *või* '*or*' and punctuation marks *,;;.?!* are marked as potential clause boundaries.

3.2.4 Step 4
The start and ending of direct speech are tagged as sure clause boundaries. The start is signalled by a colon, followed by quotation marks. Likewise, the end is signalled by a comma, period, question mark or an exclamation mark, followed by quotation marks.

3.2.5 Step 5
If it is possible to establish how quotation marks go in pairs, and both inside the quotation marks and outside are possible verbal centres of clauses, then the enclosed unit may be tagged as a parenthetical clause. Otherwise, quotation marks do not signal a clause boundary, e.g. "War and Peace".

3.2.6 Step 6
Colon and semicolon are tagged as sure clause boundaries.

3.2.7 Step 7
Some (combinations of) conjunction words following a comma or a dash are considered to be so strong signals of clause boundaries that the existence of verbal clause centres need not be checked. The assumption here is that such strong signals mean that the following words of the sentence are clearly not connected with the previous ones as strongly as those are connected among themselves. Those conjunction words are: *ja* '*and*', *ning* '*and*', *ega* '*neither*', *või* '*or*', *et* '*that*, *for*', *kui* '*if*, *when*', *millal* '*when*', *kus* '*where*', *kuhu* '*where to*', *kust* '*from where*', *sest ~ kuna* '*because, as*', *kuid ~ ehkki* '*although, albeit*', *siis* '*then*', *kuni* '*as long as*', *nagu ~ otsekui ~ justkui* '*as if*', kuidas '*how*', *kas* '*whether*, *if*'. Interrogative-relative pronouns *mis* '*what*', *kes* '*who*', *missugune ~ milline* '*what kind*, *what type*' in any case form are also considered to signal sure clause boundaries. If there is an intervening word between a comma and the conjunction *et*, then this does not undermine the existence of a sure clause boundary, e.g. *ilma et* '*without*', *nii et* '*so that*', *ainult et* '*only that*' etc.
Conjunctions *aga ~ kuigi* '*although, albeit*' after a comma are considered sure clause boundaries only if a verbal centre can be established after them before the next potential clause boundary. This way, constructions (8), (9) are not tagged as clauses. The fact that the four words, all having a similar meaning '*although*' (*kuid*, *ehkki*, *aga*, *kuigi*) are considered of unequal predictive value for clause boundaries, is currently based on rough corpus statistics; a closer look at their behaviour is definitely needed.

(8) aitab lastel, aga ka täiskasvanutel, head tervist säilitada
helps the children, but also grown-ups , good health to keep

(9) sisaldab vett, kuigi väikestes kogustes
contains water, albeit in small amounts

3.2.8 Step 8

Look at the remaining possible clause boundaries. If it has a possible verbal centre on both sides, tag this boundary as a sure one.

What should one do if the sentence contains a phrase without a verbal centre, enclosed by possible clause boundaries, and there is a possible verbal centre on both sides, cf. *endine üliõpilane* '*a former student*' in (10)?

(10) Mari on tegelikult Maiu, endine üliõpilane, | elab Tammsaare teel ja | armastab laulda rahvalaule.
Mari is actually Maiu, a former student, who lives in Tammsaare road and loves to sing folk songs.

Which of the possible boundaries, the one before the phrase on the one after it, should be chosen as the sure clause boundary? The rule here is that if a possible verbal centre is preceded by a comma, a conjunction *ja* '*and*' or *ning* '*and*', then this can be tagged as a sure clause boundary. So in example (10) the sure boundary is in front of *elab* '*lives*' (marked with a vertical bar).

3.2.9 Step 9

Delete possible clause boundaries, if they appear to separate coordinated list elements, i.e. if the words on both sides are in the same case, e.g. *rohelistes, punastes ja sinistes pükstes* 'in green, red and blue trousers'. Sometimes after this deletion it becomes clear how to classify the remaining possible clause boundaries.

Consider sentence (11) as an example. It is divided into 4 potential clauses by three conjunction words *ja*, *ja* and *ning*. Only the first and fourth part contains a potential verbal centre, meaning that the sentence actually consists of two clauses. Where should one put the clause boundaries? *Pikkade* and *pingeliste* are in the same case form (plural genitive), so the *ja* between them will be disqualified as a possible clause boundary. The same will happen to *tujukus ning isepäisus*. As a result, we are left with exactly one potential clause boundary, with a potential verbal centre on both sides, and we can tag it as a sure boundary (shown as a vertical bar in (11))

(11) Ma ei nurisenud pikkade ja pingeliste tööpäevade üle ja | Presidendi tujukus ning isepäisus ei häirinud mind
I did not grumble about the long and strenuous working days and the President's moodiness and capriciousness did not bother me.

3.2.10 Step 10

In addition to those which have been marked as parenthesis in step 1, some clauses can be marked as embedded ones. In order to qualify they must be enveloped by a single clause, meaning that there should not be a verbal centre on both sides of the embedded one. A suitable candidate for an embedded clause starts with an interrogative-relative pronoun *mis* '*what*', *kes* '*who*', *missugune ~ milline* '*what kind*, *what type*' in any case form, or with a conjunction word *kus* '*where*', *kuhu* '*where to*', *kust* '*from where*', *et* '*that*, *for*', or *millal* '*when*', preceded by a comma; in other words, it should be a relative clause, cf. (12).

(12) Mees , <embedded> kes tuli vastu </embedded> , kandis musta kaabut
The man who approached wore a black hat.

4. Evaluation

A 16,000-word test corpus, consisting in equal proportions of fiction, newspaper and popular science texts, was used for the evaluation. The precision and recall achieved by our system are presented in Table 1.

Table 1 presents the joint results for "ordinary" clause boundary detection and for parenthesis and embedded clause boundary detection. If these two types of clause boundaries are evaluated separately, then one can say that detecting an "ordinary" clause boundary (recall 98%, precision 96%) is an easier task than detecting parenthesis and embedded clause (recall 79%, precision 100%). Mistaking a start- or endpoint of a parenthesis or embedded clause for an "ordinary" clause boundary is the most frequent mistake made by the system. Among other frequent mistakes are the ones caused by erroneous

text class	tokens	sentences	clause boundaries found by the system	correct clause boundaries found by the system	clause boundaries not detected by the system	recall	precision
newspapers	5205	328	308	294	15	95%	95%
popular science	5944	439	333	318	18	95%	95%
fiction	4926	286	440	427	21	95%	97%
ALL	16075	1053	1081	1039	54	95%	96%

Table 1. Precision and recall of clause boundary identification

morphological analysis, especially a word-form erroneously tagged as a finite form of a verb can result in a false clause boundary tag. Mistakes in punctuation, notably spurious and missing commas, also cause mistakes in clause boundary identification.

5. Related work

Since Eva Ejerhed (1988) brought up the clause boundary detection as a separate research subject, both rule-based and machine learning methods, later also hybrid methods have been used for solving the clause identification or clause splitting task.

CoNLL has organized a shared task on clause identification (Tjong Kim Sang and Déjean 2001) aimed at discovering clause boundaries with machine learning methods. The Penn Treebank clause segmentation (Bies et al 1995) was used as gold standard.

The research resembling our's has been carried out by Vladislav Kubon et al. (2007) and Georgiana Puscasu (2004). Kubon et al. segmented Czech sentences by applying rules to morphologically analysed text. The rules made extensive use of the Czech strict rules for punctuation and information about the finiteness of verb-forms in the text. Georgiana Puscasu (2004) has developed a hybrid clause splitting system for Romanian texts and ported it to English texts. The system also takes morphologically annotated text as input and uses mostly information about punctuation, conjunctions and verb finiteness for identification of clause boundaries.

As for Estonian, two Constraint Grammar based systems – a shallow parser (Müürisep 2000) and a morphological disambiguator (Puolakainen 2001) – contain a special module of constraints (i.e. rules) for clause boundary identification, but the authors don't evaluate the rules for clause boundary identification apart from the rest of the system. The aforementioned Constraint Grammar parser for Estonian does not distinguish parenthesis and embedded clauses from "ordinary" ones and as a consequence, the former always divides the main clause into two disconnected parts.

6. Conclusion

This paper presented a clause boundary identification system developed for Estonian texts, deployed in annotating the Estonian Reference Corpus. The system is able to identify parenthesis and embedded clauses and thus enables re-uniting the clauses, divided by the inserted constructions. The system achieves 95% recall and 96% precision.

7. Acknowledgements

This work has been supported by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS and by the Estonian Ministry of Education and Science (grant SF0180078s08).

8. References

Bies, A., Fergusson, M., Katz, K. and MacIntyre, R. (1995). Bracketing Guidelines for Treebank II Style Penn Treebank Project. Technical Report, University of Pennsylvania. Internet document at ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz (06.10. 2011)

Ejerhed, E. (1988). Finding clauses in unrestricted text by finitary and stochastic methods. In: *Proceedings of ANLP '88*, pp. 219–227.

Ejerhed, E. (1996). Finite state segmentation of discourse into clauses. In: *Proceedings of the ECAI'96 Workshop on Extended Finite state models of language*. ECAI'96, Budapest, Hungary

Encyclopedia of Language & Linguistics (2nd ed.) Editor-in-chief Keith Brown. Elsevier.

EKG II = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. (1993). *Eesti keele grammatika II Süntaks*. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

Kubon, V., Lopatková, M., Plátek, M., Pognan, P. (2007). A Linguistically-Based Segmentation of Complex Sentences. In: *Proceedings of FLAIRS Conference*, pp 368–373.

Müürisep, K. (2000). Eesti keele arvutigrammatika: süntaks. Dissertationes mathematicae Universitatis Tartuensis 22. Tartu: TÜ kirjastus.

Puolakainen, T. (2001). Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes mathematicae Universitatis Tartuensis 27. Tartu: TÜ kirjastus.

Puscasu, G. (2004). A Multilingual Method for Clause Splitting. In: *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, Birmingham, UK. http://clg.wlv.ac.uk/papers/puscasu-04a.pdf (04. 09. 2011)

Tjong Kim Sang, E., Déjean, H. (2001). Introduction to the CoNll Shared Task: Clause identification. In: *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, pp 53–57, Toulouse, France, July 2001.