FROM SYNTAX TREES IN ESTONIAN TO FRAME SEMANTICS[1]

Kaili Müürisep, Heili Orav, Haldur Õim, Kadri Vider, Neeme Kahusk, Piia Taremaa
University of Tartu

Abstract

This paper reports on the development of Estonian corpus with role semantic annotation. One of the main aims of the work is semantic analysis of Estonian simple sentences where the input is presented in the form of syntactic trees. The result of semantic analysis is presented in a frame semantic form. In our paper we discuss the frame-semantic annotation framework, syntactical analysis of simple sentences of Estonian and problems arising from exhaustive semantic annotation and possible applications.

Keywords: frame semantics, syntactical analysis of Estonian, motion

1. Introduction

One of the more distant goals in natural language processing has been the semantic analysis of the language, so that in addition to the recognition of structure of words and sentences, the computer could also understand the meaning of a sentence. As far as Estonian is concerned, semantic analysis is becoming a real option as the syntactic analysis of (simple) sentences on one end and the semantic wordnet on the other have reached the level that makes semantic analysis possible by combining the two.

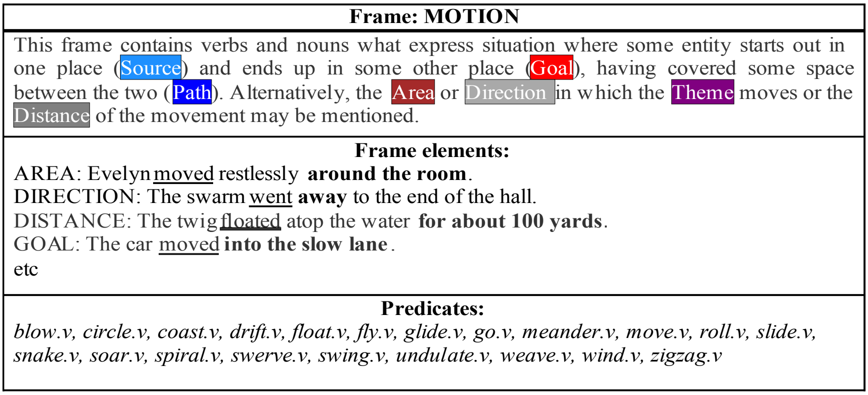
Traditionally, it is sensible to confine the semantic analysis to a narrower ontological field. We have selected the situations involving motion, as that field is one of the key fields in both theoretical semantics and its several applications (e.g. communication with robots).

Our research is based on the assumption that the predicate verb acts as the nucleus of the sentence and determines the situational type of the whole sentence. In our case, we are dealing with the frames of motion and an evoker of frame is mostly verb of motion.

We chose frame semantics as our paradigm for semantic analysis. Frame theory has proved a stimulating framework for the description of verbal meaning, both theoretically and lexicographically (see Fillmore 1985; Fillmore & Atkins 1992). Frame semantics (according to Fillmore) seeks to describe the meaning of a sentence as it is

[1] Supprted by National Programme "Language Technology Support of Estonian Language" projects No EKKTT04-5, EKKTT06-11 and EKKTT07-21; Government Target Financing project SF0182541s03 ("Computational and language resources for Estonian: theoretical and applicational aspects")

actually understood by characterising the background knowledge necessary to understand each expression. It represents this background knowledge in the form of frames, conceptual structures modelling prototypical situations. On the Figure 1 is shown an example from FrameNet about “motion” where is described the situation, frame elements and predicates connected with this situation are described.



Frame: MOTION

This frame contains verbs and nouns what express situation where some entity starts out in one place (Source) and ends up in some other place (Goal), having covered some space between the two (Path). Alternatively, the Area or Direction in which the Theme moves or the Distance of the movement may be mentioned.

Frame elements:

AREA: Evelyn moved restlessly **around the room**.
DIRECTION: The swarm went **away** to the end of the hall.
DISTANCE: The twig floated atop the water **for about 100 yards**.
GOAL: The car moved **into the slow lane**.
etc

Predicates:

blow.v, circle.v, coast.v, drift.v, float.v, fly.v, glide.v, go.v, meander.v, move.v, roll.v, slide.v, snake.v, soar.v, spiral.v, swerve.v, swing.v, undulate.v, weave.v, wind.v, zigzag.v

Figure 1. The motion frame according to FrameNet.

The annotation of predicate-argument structure in general and of FrameNet in particular, is interesting for its inter-mediate position between syntax and “deep” semantics. The semantic role labels characterise the relationship between predicate and argument as well as relationships among arguments (Burchardt et al 2006). In text, a frame is evoked by a word or expression (in our notation, FEE).

2. Description of corpus

Our preliminary base corpus consists of 370 simple sentences containing motion verbs as predicates. The sentences have been syntactically analysed and presented in the form of syntactic trees with both dependency and phrase structure annotation.

The list of verbs for the corpus was automatically extracted from the Estonian WordNet (http://www.cl.ut.ee/ressursid/teksaurus/) and it includes all verb senses belonging to the MOTION hierarchy. The top verbs of the hierarchy which include almost all the senses of motion verbs are the following:

- liigutama – *’make move, displace, move – cause to move*’ with 123 synsets in subtree;
- liikuma – *’move, change position’* with 223 synsets in subtrees.

Using the list of verbs we extracted a subcorpus of simple sentences with motion verbs from the corpus of sample sentences of „Types of Simple Sentences in Estonian” by Huno Rätsep (printed in 1978). These are constructed context-free sentences and contain a lot of content words. For example:

Tankid lähenesid tee poolt kõrgendikule.
Tanks approached from the road to the hill.

Uurija lähenes küsimusele oskuslikult.
Researcher approached to the problem skilfully.

Generated simple sentences match very well to our goal, because they modify often the linguistic expression of the same situation and therefore help us to discover morphosyntactic features which specify or define specific roles or even frames.

3. Towards deeper syntactic analysis

The corpus has been disambiguated morpho-syntactically semi-automatically using the annotation scheme of Constraint Grammar of Estonian (Müürisep et al 2003), which gives shallow dependency-oriented description to the sentences. As it is essential to have the phrase structure description for the annotation of frames, we used the same method as building a treebank Arborest (Bick et al 2004): the Constraint Grammar annotation scheme was transformed to the combination of VISL-style phrase structure and dependency structure annotation using VISL phrase structure parser. Figure 2 illustrates the textual format of the output of the PS parser. The example (1)

(1) **Peeter hiilis linnaääri mööda koosolekult koju püssi järele.**
Peter sneaked by the suburbs from the meeting to home for the gun.

is analysed as follows: it is finite clause consisting of subject (S) which is a proper noun (prop); predicate (P) which is a finite verb (v-fin); adverbial (A) which is a postpositional phrase (pp) with two daughters: head as postposition and dependent as noun; two adverbials which are nouns and another adverbial which is postpositional phrase.

```
STA:fcl                                               #STAtement:finite clause
S:prop('Peeter+0',prop,sg,nom,.cap)     Peeter        #Subject:poper noun
P:v-fin('hiili+s',main,indic,impf,ps3,sg,af,.inV)  hiilis        #Predicate:finite verb
A:pp                                                  #Adverbial:postposition phrase
=D:n('linna-ää;r+i',com,pl,part)       linnaääri      #Dependent:noun
=H:pst('mööda+0',post,.part)    mööda                 #Head:postposition
A:n('koos-olek+lt',com,sg,abl)  koosolekult           #Adverbial:noun
A:n('kodu+0',com,sg,adit)       koju                         #Adverbial:noun
A:pp                                                  #Adverbial:postposition phrase
=D:n('püss+0',com,sg,gen)       püssi                 #Dependent:noun
=H:pst('järele+0',post,.gen)    järele                       #Head:postposition
FST:punc('.',Fst)       .                             #FullSTop:punctuation mark
```

Figure 2. Textual output of Constraint Grammar to VISL-tree converter.

In the course of the task, there was a phrase structure grammar created, containing ca 40 rules. The formalism allows mother-from-daughters rewriting rules, addressing function and form tags, as well as word forms and base forms. As the phrase structure of the sentences in the corpus was in most cases very simple, almost half the rules describe word order configurations.

The output of the parser was checked manually, eliminating few errors and ambiguities and then the VISL-annotation was transformed to TIGER XML format (Mengel, Lezius 2000). Figure 3 demonstrates the graphical format of the output.

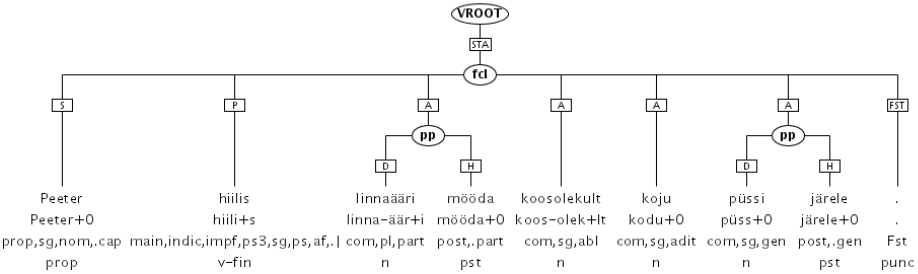


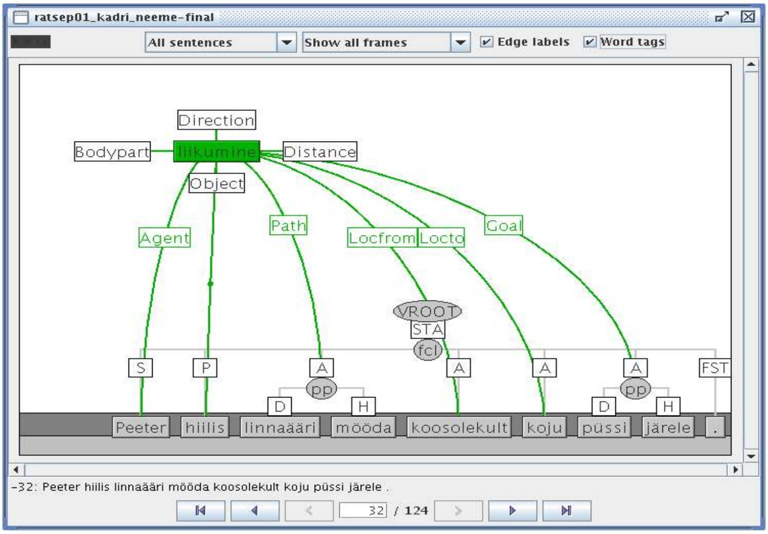
Figure 3. The sample sentence in TIGER XML format.

4. SALTO tool and the work flow

The generated trees are subjected to frame semantic analysis, using the SALTO graphic interface (Burchardt et al 2006) that enables the graphic association of frames and semantic roles with syntactic tree structures. It was developed at Saarland University for the annotation of semantic roles in the frame semantics paradigm in a simple drag-and-drop fashion. SALTO can also handle unparsed sentences, consisting only of a sentence node and the terminals, so that annotation of data without syntactic analysis is possible as well. Still, we decided to use syntactically analysed sentences, because it allows us to assign frame nodes to phrases, not only words, and one of our main goals is to deal with syntax-semantics interface.

Annotation proceeds according to the principle “one predicate at a time” and the uses of each predicate are annotated by two independent annotators. Annotator has to find frame-evoking element (FEE) in the sentence. By right-clicking on the FEE the terminal user chooses appropriate frame. After that, user selects each frame element and drags the elements to appropriate nodes. Figure 4 you see a simple annotation instance for the verb “hiilima” (to sneak).

In sentence (1), the word hiilima (*to sneak*) is associated with frame LIIKUMINE (*moving*). After evoking the frame MOVING, annotator can associate the semantic roles (frame elements) with appropriate nodes. In this case, AGENT with Peeter, PATH with postpositional phrase 'linnaääri mööda' (*by the suburbs*), LOCFROM with 'koosolekult' (*from the meeting*), LOCTO with 'koju' (*to home*) and GOAL with 'püssi järele' (*for the gun*).



-32: Peeter hiilis linnaääri mööda koosolekult koju püssi järele .

Figure 4. Annotation instance using SALTO-tool.

As suggested by the authors of SALTO, our work flow consist of two independently annotated files, the process of merging them, and conflict resolution. All of our annotators participated in the conflict resolution, and the result was freezed as golden standard. Each sentence need not to have only one frame, so we can annotate many frames in one sentence.

In principle, the SALTO tool allows the frames to span over the sentence boundaries, so we can have the agent, for example, in the next sentence, but due to the characteristics of our corpus we have not used this feature.

5. Problems

5.1. Frame type depends on morphosyntax

(2) **Lennuk**	**heitis**	**pomme.**
Bomber-nom	throw-pst3sg	bomb-pl.**part**
SUBJ	V	OBJ-partial

The bomber was throwing bombs.

(3) **Lennuk**	**heitis**	**pommid**	**kõrgusest**	**metsatukale.**
Bomber-nom	throw-pst3sg	bomb-pl.**nom**	height-ela	coppice-all
SUBJ	V	OBJ	ADVL-Locfrom	ADVL-Locto

The bomber threw the bombs from the height to the coppice.

The partitive case of the word ’bomb’ makes it partial object in morphosyntax (sentence 2) and adds a frequentative meaning to the verb. This means that situation (as well as frame type) becomes Action, not Motion as in (3) where throwing happens once.

5.2. Frame elements

Originally, we relied on FrameNet for determining semantic roles, but by now, we have modified those to meet the goals of our project. For example, we use the following terms for describing components of motion events: AGENT (agent of motion, who moves himself), CAUSER (initiator of motion, who is not agentive), and OBJECT (which is moving entity) and so on.

It’s not always easy to decide whether the motion is agentive or nonagentive, for example

(4) **Lennuk**	**lendab**	**Londonist**	**Pariisi.**
plane	fly-3sg	London-ela	Paris-ill
AGENT ??	FEE	Locfrom	Locto

The plane flies from London to Paris.

The question arises whether the plane is agentive or nonagentive mover – is it AGENT, OBJECT or INSTRUMENT? Apparently is here transference of meaning that the properties of agent – the pilot – have been transferred to actual instrument of moving (plane). The other possible explanation is that the plane is perceived as an object with ability to move. So it’s ontological knowledge, which isn’t expressed in this sentence very clearly.

(5) **Tuul**	**puhus**	**vaasi**	**laualt**
wind	blow-pst3sg	vase-gen	table-abl
CAUSER ??	FEE	OBJECT	Locfrom

The wind blew the vase off the table.

Here the question is the same – is the wind agentive initiator of motion? We’ve decided to take it as a nonagentive motion where the wind is CAUSER.

6. Future plans

In the next phase, we plan to include complex sentences in the corpus, create a phrase structure grammar for their analysis, and start analysing the influences resulting from the co-existence of different frames.

According to our plans, the research should be carried on at least in two directions.

First, different problems connected with conceptual-formal representation of frames should be dealt with. One of the most important problems is connected with semantic inferences. In different types of motion events different inferences about the participants are possible. We give just an example. When we take events where Agent, Instrument and Object are involved, then it appears that these entities can participate in actual motion in different ways. On the one hand, for instance, in the case of *throw* the Agent nor the Instrument does not have to change places at all, only the Object. On the other hand, in the case of *bring/take (something somewhere)* all three entities have to change places. This kind of information which concerns the dynamics of motion events and their participants should be attainable from frames, when we want our system to be able to “compute” where a concrete entity involved in motion event was/is before and after the event.

Second, the corpus as the basis of analysis should be enlarged and not simply quantitatively, but in a systematic way. Thus far we have dealt with simple affirmative sentences. We have to add sentences containing different kinds of negation, sentences in others than affirmative form (questions, orders, conditionals etc.), complex sentences with clauses connected by different semantic relations (e.g. *If A, then B; A, although B*).

Taken together, these goals point at the direction we are willing to take in moving from simple frames of single sentences towards complex frame structures of connected discourse.

References:

E. Bick, H. Uibo, K. Müürisep, 2004. Arborest – a VISL-Style Treebank Derived from Estonian Constraint Grammar Corpus. Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004). Tübingen, Germany.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, M. Pinkal, 2006. SALTO – A Versatile Multi-Level Annotation Tool. Proceedings of LREC 2006, Genoa Italy, pp 517–520.

Ch. Fillmore1985. Frames and the Semantics of Understanding. - Quaderni di Semantica. Vol. VI, pp 222–254.

Ch. Fillmore, S.B.T. Atkins, 1992. Towards a Frame-based Lexicon: the Semantics of RISK and its Neighbours. In: Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization. A. Lehrer & E.F. Kittay (eds.). Lawrence Erlbaum Associates: Hillsdale, New Jersey. pp 75–102.

A. Mengel, W. Lezius, 2000. An XML-based representation format for syntactically annotated corpora. In Proceedings of the International Conference on Language Resources and Evaluation, Athens, Greece, pp 121–126.

K. Müürisep, T. Puolakainen, K. Muischnek, M. Koit, T.t Roosmaa, H. Uibo, 2003. A New Language for Constraint Grammar: Estonian. International Conference Recent Advances in Natural Language Processing. Proceedings. Borovets, Bulgaria, 10-12 September 2003, pp 304–310.

H. Rätsep, 1978. Eesti keele lihtlausete tüübid. Eesti NSV Teaduste Akadeemia Emakeele Seltsi Toimetised nr. 12. ”Valgus”, Tallinn.

About the authors

Kaili Müürisep is a senior researcher (PhD) of language technology, University of Tartu, Estonia. Main research interests: parsing of written and spoken Estonian. E-mail: kaili.muurisep@ut.ee.

Heili Orav is a researcher (Ph.D.) of Computational Linguistics, University of Tartu, Estonia. Main research interests: (computational, cognitive) semantics, computational lexicology. E-mail: heili.orav@ut.ee.

Haldur Õim is Professor Emeritus of General Linguistics, University of Tartu, Estonia. Main research interests: theoretical semantics, semantics-syntax and semantics-pragmatics interfaces, computational semantics. E-mail: haldur.oim@ut.ee.

Kadri Vider is a researcher (M.A.) of Computational Linguistics, University of Tartu, Estonia. Main research interests: lexical semantics, wordnets, word sense disambiguation. E-mail: kadri.vider@ut.ee.

Neeme Kahusk is a researcher (M.B.) of Computational Linguistics, University of Tartu, Estonia. Main research interests: semantics, computer lexicons, word sense disambiguation, psycholinguistics. E-mail: neeme.kahusk@ut.ee.

Piia Taremaa is a student for Master degree at Department of General Linguistics, University of Tartu, Estonia. Main research interest is cognitive linguistics. E-mail: piia.taremaa@ut.ee.