Linguistic Logical Analysis of Direct Speech

Aleš Horák, Miloš Jakubíček, Vojtěch Kovář

Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{hales, xjakub, xkovar3}@fi.muni.cz

Abstract. Logical analysis of natural language allows to extract semantic relations that are not revealed for standard full text search methods. Intensional logic systems, such as the Transparent Intensional Logic (TIL), can rigorously describe even the higher-order relations between the speaker and the content or meaning of the discourse.
In this paper, we concentrate on the mechanism of logical analysis of direct and indirect discourse by means of TIL. We explicate the procedure within the Normal Translation Algorithm (NTA) for Transparent Intensional Logic (TIL), which covers the language analysis on the syntactic and semantic levels. Particular examples in the text are presented in syntactically complicated free-word-order language, viz the Czech language.

Key words: direct speech, indirect speech, Transparent Intensional Logic, TIL, Normal Translation Algorithm, NTA, logical analysis, syntactic analysis, parsing

1 Introduction

The analysis of natural language texts on morphological and syntactic levels already achieved application level quality, for the mainstream languages [1]. On the other hand, the analysis of various aspects on the semantic level is still on the way to quality knowledge analysis and extraction (see e.g. [2] or other SemEval 2012 task results). Standard data mining and search techniques have already reached the top of their potential and researchers and knowledge engineers employ semantics in the natural language processing [3,4,5,6]. Most current practical systems that need to utilize knowledge representation of natural language in formal logic usually do not go beyond the scope of first-order logic, even though in the language, there is a number of higher-order phenomena such as belief attitudes, grammatical tenses or intensionality, all of which cannot be addressed properly within the first-order logic.

In the following text, we are dealing with logical analysis of natural language (NL) using the formalism of the Transparent Intensional Logic (TIL, [7]), an expressive higher-order logical system introduced by Pavel Tichý [8,9], which works with a complex hierarchy of types, temporal system of possible worlds and an inference system in development.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 51–59, 2012.

The current work is a part of a long-term project aimed at providing norms for the "translation" of various NL phenomena to logical constructions, the Normal Translation Algorithm (NTA) [10,11]. The actual implementation of the system works on top of the Czech syntactic parser `synt` [12]. `Synt` is based on the robust meta-grammar formalism including context-free chart parsing enhanced with contextual actions for phrase and sentence level tests. The parser uses a meta-grammar of about 250 meta-rules for the description of the whole Czech language with automatic grammar expansion to technical parsing rules.

In the following text, we focus on the issues of analysis of complex sentences including direct discourse. We first discuss the formal definition of direct and indirect speech and their logical consequences. Then we explain in detail how the logical analysis in the `synt` parser works and how the syntactic and logical representation of direct speech is obtained. In Section 4, we describe the process of obtaining a corpus containing texts with direct speech, that was used extensively for studying various aspects of this language phenomenon and for evaluation of the parsing procedure.

2 Direct and Indirect Discourse

Direct and indirect forms of speech are related kinds of so called *reported speech*, i.e. those utterances, where the speaker refers to another utterance or utterances [13]. In the *direct speech* form, the (current) speaker uses an exact quotation of the original speaker:

> Waiter said: "Are you ready to order, sir?"
> Mr Smith replied: "Yes. I'll have the beef stew for starters and my wife would like tomato soup."

The corresponding indirect speech can look like:

> The waiter asked, whether Mr Smith was ready to order.
> He replied, that he would have the beef stew for starters and his wife would like tomato soup.

The main difference in the logical consequences lies in the change of the actual speaker positions in the reported clause. In case of the direct speech, the subject position is occupied by the original speaker and all speech aspects are related to him or her. On the other hand, the indirect form is completely related to the reporting speaker and all original speech aspects are *transformed* to this new subject. Especially, this results in higher usage of anaphoric expressions and thus higher level of ambiguity in the indirect form of reported speech.

3 Analysis of Direct Speech

In this section, we first describe the implementation of the syntactic and logical analysis in the `synt` parser and then concentrate on the additions specific to the analysis of direct speech sentences.

3.1 The Synt Parser

`Synt` is a rule-based parser designed specifically for morphologically-rich free-word-order languages and currently used mainly for Czech.[1] It operates by means of a modified head-driven chart analysis with a context-free backbone interconnected with predefined in-programmed (so Turing complete) contextual actions. The contextual actions are used to capture contextual phenomena like grammatical agreement or advanced (possibly non-local) linguistic features.

The underlying meta-rules are defined in the form of a meta-grammar consisting of about 250 rules. Each rule can be attached a precedence level, a list of actions and a derivation type. The precedence level makes it possible to include mutually exclusive rules into the grammar. The backbone rules are generated from a meta-rule during the process of automatic generation of full grammar from the meta-grammar according to the derivation type (e.g. permutation of all right-hand side non-terminals, enclitics checks, etc.).

The main result of the syntactic parsing procedure is an ordered set of constituent parsing trees that is potentially very big but zipped within a *shared packed forest* structure [14] provided with a fast algorithm for extracting n best trees according to a combined ranking function.

Each of these trees can be used as an input to another set of contextual actions that transform the tree to a logical formula in the TIL formalism, using lexical type and valency information extracted from the VerbaLex verb valency lexicon [15].

3.2 Syntactic Analysis of Direct Speech

A sentence with direct speech consists of the *direct speech* segment and a *reporting clause*. We analyze the reporting clause as the head element of the whole sentence, as the direct speech part often plays the role of subject or object in the reporting clause.[2] The structure of the direct speech part can be arbitrarily complex – it can consist of one or more sentences, or of an incomplete sentence. Therefore, we analyze the content of the direct speech by the *direct_speech* non-terminal that can cover one, or more, clauses, and also expressions, where the verb is not present.

Here comes the question, what should be actually considered a sentence in the context of direct speech. One segment of direct speech with one respective reporting clause can contain an arbitrary number of sentences, or even paragraphs, so it is often not clear where the sentence boundary should be. There are two straightforward approaches:

– Consider the whole pair, i.e. complex direct speech and the respective reporting clause, as one sentence.

[1] The `synt` grammar was also adapted for the Slovak and English languages, which are subject of further development.

[2] As in e.g.: *"Go away," he said.*

– Split the direct speech to multiple sentences and consider only the sentence closest to the reporting clause as its completion.

The first solution may seem more correct, because there is an immediate relationship between the reporting clause and all parts of the direct speech; however, it would lead to sentences consisting of thousands of words or even more. Parsing such sentences would be computationally unfeasible, therefore we do not consider it a good solution. Since all the respective relations are extra-syntactic (anaphoric relations, relative tenses, ...), we have developed a combined solution – complex sentences can be contained in one direct speech segment only in case the whole is not too long, otherwise, the direct speech is split at sentence boundaries and the rest of the direct speech analysis is linked to the reporting clause via a specific link used during the logical analysis phase. Such solution is best realizable from the technical point of view and it is also closest to what the currently available sentence segmenters do.

Having the sentence unit fixed, there are three possible combinations of where the reporting clause can be placed, with regard to the direct speech segment:

– The reporting clause comes before the direct speech segment – e.g. *He asked: "Would you bring me a beer?"*
– The reporting clause comes after the direct speech segment – e.g. *"Would you bring me a beer?" he asked.*
– The direct speech segment is divided into two parts, with the reporting clause between them – e.g. *"Would you," he asked, "bring me a beer?"*

Therefore, three basic rules are needed to address these three combinations:

```
clause   →   clause   ':'   direct_speech
clause   →   direct_speech    clause
clause   →   direct_speech    clause    ','    direct_speech
```

As mentioned above, the direct speech segment then rewrites to a complex sentence in quotes. In case the content of the direct speech cannot be analyzed by the *sentence* non-terminal, we allow the direct speech to rewrite as an arbitrary sequence of characters. These two analyses are mutually exclusive, since the `non_sentence` rule of `direct_speech` is analysed on a higher (i.e. less probable) rule level and is thus pruned away in the case where both `direct_speech` rules match.

```
  direct_speech   →   '"'   sentence   '"'
9:direct_speech   →   '"'   non_sentence   '"'
  non_sentence    →   /[^"]+/
```

One problem arises in the case where the direct speech is interrupted by the reporting clause, but it forms one logical unit, e.g. in the sentence shown above: *"Would you," he asked, "bring me a beer?"*. For example, the manual for annotators of the Prague Dependency Treebank [16] deals with this direct speech type by

using non-projective constituents.[3] In the `synt` parser, the intra-clause position of the reporting clause is analysed in a way similar to a *parenthesis*, i.e. a part of the original clause, which can be inserted between any two sentence constituents.

3.3 Logical Analysis of Direct Speech

The analysis of direct speech is not so loaded with the anaphora resolution problem as the indirect speech form, however, we can encounter situations, where the actual content of the direct speech clause is logically less related or even completely irrelevant. Let us have a look at the following examples

Peter said: “Hand me the book.” (1)

Peter asked: “Hand me the ...” (2)

Peter thought: “The unicorn!” (3)

Peter screamed: “Aaaargh!” (4)

The example sentence (1) forms the standard reporting utterance with the two parts of reporting clause and direct speech reported clause. However, all the remaining examples fail on the syntactic level to be analysed as a (complete) clause. The sentence (2) contains an incomplete (probably interrupted) reported clause, sentence (3) shows, that Peter’s thought is related with and individual object, and last, the sentence (4) represents an example of an object, which cannot be analysed even on the morphological level and stays here for a non-verbal sound.

The logical analysis of direct speech sentences in `synt` is related to the procedure of analysis of complex sentences, see [17]. The construction generated by this procedure for the sentence (1) can look like:[4]

$$
\begin{aligned}
&\lambda w_1 \lambda t_2 \Big[\mathbf{P}_{t_2}, \Big[\mathbf{Onc}_{w_1}, \lambda w_3 \lambda t_4 (\exists x_5)(\exists c_6)(\exists i_7) \Big(\\
&\quad \big[\mathbf{Does}_{w_3 t_4}, i_7, [\mathbf{Perf}_{w_3}, x_5] \big] \ \wedge \\
&\quad \wedge [\mathbf{Peter}_{w_3 t_4}, i_7] \ \wedge \ x_5 = [\mathbf{say}, c_6]_{w_3} \ \wedge \\
&\quad \wedge c_6 = {}^0 \Big[\lambda w_8 \lambda t_9 (\exists x_{10})(\exists i_{11}) \Big(\big[\mathbf{Does}_{w_8 t_9}, Ty, [\mathbf{Perf}_{w_8}, x_{10}] \big] \ \wedge \\
&\qquad \wedge x_{10} = [\mathbf{hand\_sb\_st}, Já, i_{11}]_{w_8} \ \wedge \ [\mathbf{book}_{w_8 t_9}, i_{11}] \Big) \Big] \\
&\quad \Big) \Big], \mathbf{Anytime} \Big] \ldots \pi
\end{aligned}
\tag{5}
$$

$\mathsf{Peter}/(o\iota)_{\tau\omega}$; $\mathsf{say}/((o(o\pi)(o\pi))_{\omega} * _n)$; $\mathsf{hand\_sb\_st}/((o(o\pi)(o\pi))_{\omega}\iota\iota)$; $\mathsf{book}/(o\iota)_{\tau\omega}$;

[3] See [16, Section 3.6.1]

[4] The `synt` outputs are translated from the Czech language for the demonstration purpose here in the paper.

Type	*Id*	*Word/Phrase*	*Reference*
sentence	sent1	Peter said: "Hand me the book."	
clause	m1	Peter said	
np	m2	Peter	
clause	m3	_ Hand me the book	
pron_pers_zero	m_zerosubj1	_	
pron_pers_strong	m4	me	m2
np	m5	book	

Fig. 1. The Saara system – anaphora resolution of the example sentence[6] ("mN" in the table refers to the term "markableN.")

As we may see, the verbal object x_5 in this construction is connected with an argument of the higher-order type $*_n$ representing a (trivialized) construction of order n. In this case the construction c_6 generates a *proposition* (of type π, or $o_{\tau\omega}$), which keeps all the properties related to the meaning of Peter's speech.

The corresponding anaphoric expressions that connect the reporting and reported speech can be identified using the automatic anaphora resolution tool Saara [18] that works in relation to the `synt` syntactic parser. An example of the anaphoric links from Saara can be seen in Figure 1. This allows us to link the variable *Já* from the construction (5) with the subject variable i_7 there.

The appropriate type of the verb say is obtained during the logical analysis of lexical items in `synt` by means of consulting the VerbaLex verb valency lexicon [19]. The entry related to the verb *říct* (*say*) is presented in Figure 2. Each verb frame participant is labelled with a two-level semantic role, which can be used for specific information regarding the TIL type of each lexical item. Currently, the verb arguments denoted by the 1st-level role COM[7] are analysed as the TIL higher-order type $*_n$.

The analysis of the other three example sentences (2), (3) and (4) does not contain two clauses, as the reported part fails to form a (whole) clause. In the case of the sentence (3), the reported part could be analysed as a noun phrase denoting and individual concept, but the sentences (2) and (4) even do not provide any such characteristics. In such cases, the analysis does not analyse the (incomplete) content of the direct speech part and the resulting construction related the reporting verb only to the (individual) expression in the form of a

[6] Again, the sentence words are translated from Czech in which the tools operate.
[7] "something that is communicated by or to or between people or groups"

Fig. 2. VerbaLex entry related to the verb *říct* (*say*).

string of characters. For example, the sentence (4) thus receives the analysis:

$$\lambda w_1 \lambda t_2 \Big[\mathbf{P}_{t_2}, \Big[\mathbf{Onc}_{w_1}, \lambda w_3 \lambda t_4 (\exists x_5)(\exists c_6)(\exists i_7) \Big(\\ \big[\mathbf{Does}_{w_3 t_4}, i_7, [\mathbf{Perf}_{w_3}, x_5] \big] \ \wedge \\ \wedge \, [\mathbf{Peter}_{w_3 t_4}, i_7] \ \wedge \ x_5 = [\mathbf{scream}, c_6]_{w_3} \ \wedge \\ \wedge \, c_6 = {}^{00}\text{“Aaaargh”} \\ \Big) \Big], \mathbf{Anytime} \Big] \ldots \pi \\ \text{Peter} / (o\iota)_{\tau\omega}; \text{scream} / ((o(o\pi)(o\pi))_\omega * _n); \text{“Aaaargh”} / \iota \tag{6}$$

Due to the “polymorphic” nature of the higher-order type $*_n$ the type of the verbal object can accept the argument in any of the cases of the direct form.

4 Direct Speech Corpus

In order to study the issues of syntactic and logical representation a corpus of direct speech of about 20,000 sentences has been created. It is obvious that the definition of direct speech is quite broad and allows speculative interpretations

as to what should be considered as direct speech. To be able to build the corpus automatically we therefore restrained ourselves only to direct speech which is introduced and finished by quotes. We used the czTenTen corpus from which we selected candidate sentences using the Corpus Query Language (CQL [20]).

Obviously, the formulation of the CQL query was subject to a trade off between precision and recall. After numerous trials following query was concluded:

```
<s/> containing
(<s>
[word!="\""]* [k!="k1"] "\"" [word!="\""]+
[k="k5"] [word!="\""]+ "\"" [word!="\""]*
</s>)
```

The resulting corpus was then used as a testbed for the study of syntactic and logical properties of the direct speech form in common texts.

5 Conclusions

We have described an efficient conversion of Czech sentences with direct speech into logical formulae in the formalism of Transparent Intensional Logic (TIL), as a part of the Normal Translation Algorithm project. We have described the parser used, the process of syntactic analysis and creation of the logical formulae from the constituent syntactic trees. We have also described a corpus of Czech direct speech which has been newly created for purposes of studying the phenomenon and for evaluation.

The speed and the precision of the whole process is sufficient and promises its future usage in automatic reasoning and intelligent question answering. In the future, we will mainly concentrate on exploiting these results in real-word applications, which mainly means integrating the information gained from logical analysis into the complex pipeline of linguistic processing, including anaphora resolution or inter-sentence relationship analysis.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013, by EC FP7 project ICT-248307 and by the Czech Science Foundation under the project P401/10/0792.

References

1. Matsuzaki, T., Tsujii, J.: Comparative parser performance analysis across grammar frameworks through automatic tree conversion using synchronous grammars. In: Proceedings of the 22nd International Conference on Computational Linguistics. (2008)

2. Specia, L., Jauhar, S., Mihalcea, R.: Semeval-2012 task 1: English lexical simplification. In: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), Montreal, Canada. (2012)
3. d'Amato, C., Fanizzi, N., Fazzinga, B., Gottlob, G., Lukasiewicz, T.: Ontology-based semantic search on the web and its combination with the power of inductive reasoning. Annals of Mathematics and Artificial Intelligence (2011) 1–39
4. Hoxha, J., Junghans, M., Agarwal, S.: Enabling semantic analysis of user browsing patterns in the web of data. arXiv preprint arXiv:1204.2713 (2012)
5. Christensen, J., Soderland, S., Etzioni, O., et al.: An analysis of open information extraction based on semantic role labeling. In: Proceedings of the sixth international conference on Knowledge capture, ACM (2011) 113–120
6. Efrati, A.: With semantic search, google eyes competitors. The Wall Street Journal (March 15, 2012)
7. Duží, M., Jespersen, B., Materna, P.: Procedural Semantics for Hyperintensional Logic. Foundations and Applications of Transparent Intensional Logic. Volume 17 of Logic, Epistemology and the Unity of Science. Springer, Berlin (2010)
8. Tichý, P.: The Foundations of Frege's Logic. de Gruyter, Berlin, New York (1988)
9. Tichý, P.: Collected Papers in Logic and Philosophy. Prague: Filosofia, Czech Academy of Sciences, and Dunedin: University of Otago Press (2004)
10. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. PhD thesis, Masaryk University, Brno (2002)
11. Horák, A.: Computer Processing of Czech Syntax and Semantics. Librix.eu, Brno, Czech Republic (2008)
12. Horák, A., Kadlec, V.: New Meta-grammar Constructs in Czech Language Parser synt. In: Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2005, Karlovy Vary, Czech Republic, Springer-Verlag (2005) 85–92
13. Coulmas, F.: Direct and indirect speech. Volume 31. De Gruyter Mouton (1986)
14. Kadlec, V.: Syntactic analysis of natural languages based on context-free grammar backbone. PhD thesis, Masaryk University (2008)
15. Horák, A., Pala, K.: Building a large lexicon of complex valency frames. In: Proceedings of the FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages, Lund University, Sweden, Tartu, Estonia (2007) 31–38
16. Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Štěpánek, J., Pajas, P., Kárník, J.: Anotace na analytické rovině – Návod pro anotátory (2005) `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer`.
17. Horák, A., Jakubíček, M., Kovář, V.: Analyzing time-related clauses in transparent intensional logic. In: Proceedings of Recent Advances in Slavonic Natural Language Processing 2011, Brno, Czech Republlic, Masaryk University (2011) 3–9
18. Němčík, V.: The Saara Framework: An Anaphora Resolution System for Czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing 2009, Brno, Czech Republlic, Masaryk University (2009) 49–54
19. Hlaváčková, D., Horák, A., Kadlec, V.: Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In: Proceedings of Text, Speech and Dialogue 2006, Brno, Czech Republic, Springer-Verlag (2006) 79–85
20. Jakubíček, M., Rychlý, P., Kilgarriff, A., McCarthy, D.: Fast Syntactic Searching in Very Large Corpora for Many Languages. In: PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Tokyo (2010) 741–747