How to Analyze Natural Language with Transparent Intensional Logic?

Vojtěch Kovář, Aleš Horák, and Miloš Jakubíček

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

Abstract. Logical analysis of natural language text is generally an underspecified task. However, within the project aiming at automatic processing of natural language (NL) text by means of logical analysis followed with the inference process, we need to have a "de facto" standard way for analysis of each NL sentence. First steps for introducing such standard are described in the presented text.

The paper describes a semi-automatic way of building a corpus of logic formulae (constructions) in the formalism of the Transparent intensional logic (TIL) for real-world sentences in the Czech language. Output of a syntactic parser is used to determine the logical structure of the sentence and a verb valency lexicon is exploited for assigning TIL types. Using this information, an exemplary bank of TIL constructions is created automatically. This corpus of TIL constructions is then checked by human logic experts who iteratively consult the results with the respective theory of TIL transcription and the processing of input supportive lexicons. A user-friendly interface for such checking is presented at the end of the paper.

1 Introduction

Formalization of natural language utterance is one of the most important steps in automatic natural language understanding. Successful specification of this process is a necessary assumption to intelligent handling of natural language texts, ranging from text classification or intelligent information extraction to question answering systems. Since formal logical systems are one of the bases of theoretical computer science, they are well known and described and the representation of natural language sentences in a logical formalism may help automatic programs to work with "meaning" in natural language.

The respective logical system we are dealing with in this paper is called the Transparent Intensional Logic (TIL [1]), an expressive logical system introduced by Pavel Tichý [2], which works with a complex hierarchy of types, system of possible worlds and times and an inductive system for derivation of new facts from a knowledge base in development.

In the current project, we aim at development of a syntactic analyzer for Czech with the ability of generating TIL logical constructions from the parsing

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 69–76, 2010. © Tribun EU 2010

results (according to Horák's *Normal Translation Algorithm* [3]). During the process of standardization of natural language analysis in TIL, an exemplary corpus of TIL constructions, as a result of logical analysis of real-world sentences, is being created in a semi-automatic way that can be later used for human reference as well as training and evaluating future knowledge representation and reasoning tools.

2 Related Work

Published results in the area of natural language (NL) logical analysis regarding algorithms for converting natural language sentences into logical formulae cover mostly the First Order Predicate Logic [4,5]. However, it can be shown that first-order formalisms are not able to handle systematically the NL phenomena like intensionality, belief attitudes, grammatical tenses and modalities (modal verbs and modal particles in natural language). On the other hand, since TIL works with techniques designed for capturing the natural language meaning these problems either do not arise or they can be solved in an intuitive way in TIL (see [2]).

For the Czech language, no attempt of logical analysis of real-world NL texts is known at the time. There is a language abstraction known as tectogrammatical layer [6] (which is the result of work at the Institute of Formal and Applied Linguistics in Prague) and there are attempts to obtain this tectogrammatical layer automatically [7]. However, the obtained results are so far not complete enough and also the tectogrammatical layer of language description cannot be really called a logical formalism.

3 The TIL Formalism

The theory of the Transparent Intensional Logic (TIL) is formed by a higher-order logical system, which uses an extended type hierarchy. TIL was first introduced by its creator Pavel Tichý in [2] and then specified in numerous articles and books. From the last publications, we refer especially to [8] and [1].

In TIL, the meaning of a natural language expression is described as a *construction* procedure, which describes in a creative way the "definition" of the expression subject. The notation of these procedures uses a λ-calculus formalism – "*red apple*" is analysed as a class of individuals $\lambda w \lambda t \lambda x.[\mathbf{red}_{wt}\, x \wedge \mathbf{apple}_{wt}\, x]$.

The TIL type hierarchy is built over a type base of four types:

- o (omicron) – truth values *True* and *False*
- ι (iota) – class of (labels of) individuals
- τ (tau) – class of real numbers/time moments
- ω (omega) – class of possible worlds

These types can be combined to mappings in order to form all types of order 1. Together with classes of constructions organized by the order of their subconstructions, they allow us to refer to objects of higher-order (type of order n denoted as $*_n$).

Variables are the only simple constructions, complex constructions are created inductively by the following simple operations:

- *trivialization* – construction constructs the trivialized object
- *abstraction/closure* – construction of a (classic) function
- *application/composition* – construction of functional application
- *execution* and *double execution* – construction of the result of functional application(s)

Constructions that contain only objects of types of order 1 are denoted as *constructions of order 1*. Constructions of order n together with types of order n form a class of objects of order $n+1$.

These intuitive rules of the extended type hierarchy allow us to refer to all complicated phenomena in natural language, such as belief attitudes or procedural definitions.

4 The Synt Parser

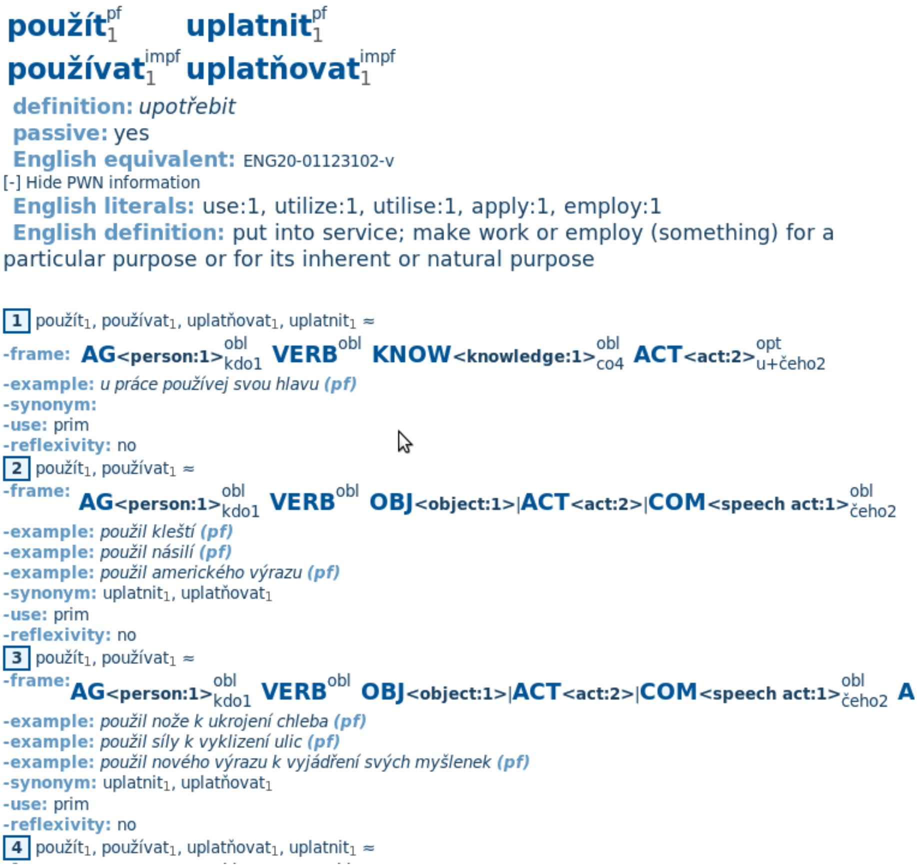
The parser used in the mentioned project is called `synt` [9]. It is based on a large CFG grammar with contextual actions (ca. 3,500 rules generated from 200 meta-rules) and an efficient variant of the head-driven chart parsing algorithm. As its input, `synt` takes a morphologically (ambiguously) annotated Czech sentence. The possibly ambiguous parsing result can be represented in various formats: output trees, a much more compact packed forest of these trees, the analysis chart, list of extracted noun and prepositional phrases, clauses, a dependency graph and some others. Moreover, the output parsing trees are ranked according to their probability and there is an algorithm for efficient obtaining one or more best unambiguous analyses.

The resulting parsing trees can be converted into the formulae in the TIL formalism, using the algorithm described in [3] including the prototype implementation. The algorithm basically „consists in assigning the appropriate (sub)constructions to analysed (sub)constituents by employing the lexicon and in the type checking which makes it possible to prune the contingencies that cannot be resolved on a lower level of the derivation tree" [3].

Therefore, for wide coverage of natural language texts we need to build a lexicon of TIL types, namely for verbs and their arguments since they contain the essential information of the sentence. This is further described in the next section.

5 Building the Verb Lexicon

As explained in [3], we need to have a high-coverage lexicon of TIL types of Czech verbs. For building such a lexicon, we have used the valency lexicon of the Czech verbs *VerbaLex*.



použít$_1^{\text{pf}}$ **uplatnit$_1^{\text{pf}}$**

používat$_1^{\text{impf}}$ **uplatňovat$_1^{\text{impf}}$**

definition: *upotřebit*

passive: yes

English equivalent: ENG20-01123102-v

[-] Hide PWN information

English literals: use:1, utilize:1, utilise:1, apply:1, employ:1

English definition: put into service; make work or employ (something) for a particular purpose or for its inherent or natural purpose

1 použít$_1$, používat$_1$, uplatňovat$_1$, uplatnit$_1$ ≈

-frame: **AG**<person:1>$^{\text{obl}}_{\text{kdo1}}$ **VERB**$^{\text{obl}}$ **KNOW**<knowledge:1>$^{\text{obl}}_{\text{co4}}$ **ACT**<act:2>$^{\text{opt}}_{\text{u+čeho2}}$

-example: *u práce používej svou hlavu* ***(pf)***

-synonym:

-use: prim

-reflexivity: no

2 použít$_1$, používat$_1$ ≈

-frame: **AG**<person:1>$^{\text{obl}}_{\text{kdo1}}$ **VERB**$^{\text{obl}}$ **OBJ**<object:1>|**ACT**<act:2>|**COM**<speech act:1>$^{\text{obl}}_{\text{čeho2}}$

-example: *použil kleští* ***(pf)***

-example: *použil násilí* ***(pf)***

-example: *použil amerického výrazu* ***(pf)***

-synonym: uplatnit$_1$, uplatňovat$_1$

-use: prim

-reflexivity: no

3 použít$_1$, používat$_1$ ≈

-frame: **AG**<person:1>$^{\text{obl}}_{\text{kdo1}}$ **VERB**$^{\text{obl}}$ **OBJ**<object:1>|**ACT**<act:2>|**COM**<speech act:1>$^{\text{obl}}_{\text{čeho2}}$ **A**

-example: *použil nože k ukrojení chleba* ***(pf)***

-example: *použil síly k vyklizení ulic* ***(pf)***

-example: *použil nového výrazu k vyjádření svých myšlenek* ***(pf)***

-synonym: uplatnit$_1$, uplatňovat$_1$

-use: prim

-reflexivity: no

4 použít$_1$, používat$_1$, uplatňovat$_1$, uplatnit$_1$ ≈

Fig. 1. Example of VerbaLex complex valency frames for 'používat'

5.1 VerbaLex

VerbaLex [10] is an exhaustive lexicon of Czech verb valency frames, currently containing over 10,000 Czech verbs (see Figure 1). The valency information is given in the form of syntactic (shallow) valencies, i.e. the morphological categories that a phrase needs to meet to be able to be an argument of a particular verb, as well as in the form of semantic valencies, i.e. semantic classes that a particular phrase usually belongs to in order to form a valency of that verb.

The semantic classes used in *VerbaLex* are compatible with the ones used in the Czech *WordNet* [11] so that it is relatively easy to find a semantic class of a given word and check if it fits to the valency frame of a particular verb. This is of course limited by the (Czech) *WordNet* coverage, precision of the data in both *WordNet* and *VerbaLex* and the fact that long phrases that cannot be found in *WordNet* can also stand as verb arguments in the sentence.

The information in the lexicon may also be used for pruning ambiguous syntactic analysis (by omitting analyses producing different verb phrases than the one recorded in the lexicon). This is already implemented and used in the `synt` system for the shallow valencies [12].

5.2 From VerbaLex to the Lexicon of the TIL Types

To build the lexicon of the verb types as outlined in Section 4, we have exploited the information from the *VerbaLex* lexicon. The number of arguments can be obtained from the obligatory members of the verb frame and their types can be derived from their semantic roles which is present in the valency lexicon as well.

Translating semantic roles of the possible verb arguments into the TIL types clearly needs a mapping from the semantic roles to the types. The most recent work is aimed at building such a mapping – an introductory analysis of this task was published in [13].

The resulting lexicon of the verb types is then given as the parameter to the `synt` parser that builds the corresponding TIL construction from the most probable syntactic tree for the sentence, according to the Normal Translation Algorithm. If the analysis produces a valid construction, this construction is then included into the corpus to be evaluated by human logical experts, as described in the next sections.

6 Creating the Corpus

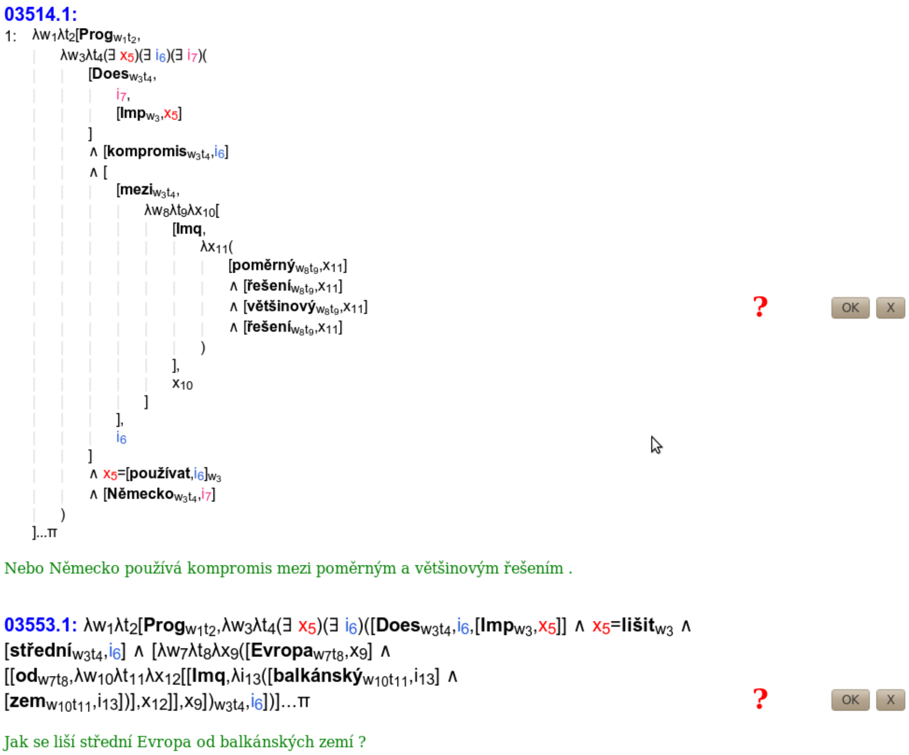
The corpus of TIL constructions is built on top of the morphologically annotated corpus DESAM [14]. The unambiguous morphological annotation (manually checked by linguistic experts) will minimize the errors on the morphological analysis level and therefore the overall result quality will be better (compared to using e.g. ambiguous or automatically disambiguated morphological information).

Furthermore, in the initial stage of the project we confine ourselves to analyze simple sentences in present, past and future tense, which means that the whole sentence contains exactly one clause with exactly one verb. Such sentences are supposed to be handled well by the parser and the selection will therefore help to further reduce number of errors in the constructions generated automatically.

In the future, the logical analysis in `synt` will cover more complicated phenomena, such as relative subordinate sentences or complex sentences with temporal events including direct speech.

7 Evaluating the Automatic Constructions

As mentioned above, all constructions on the output of the parser are included into the corpus of constructions. However, such a resource contains a lot of errors that may come from various levels of the analysis. Therefore, logic specialists are asked to provide feedback to the parser developers so that errors can be fixed in the right places.



03514.1:

```
1:  $\lambda w_1 \lambda t_2$[Prog$_{w_1 t_2}$,
        $\lambda w_3 \lambda t_4 (\exists\, x_5)(\exists\, i_6)(\exists\, i_7)($
            [Does$_{w_3 t_4}$,
                $i_7$,
                [Imp$_{w_3}$,$x_5$]
            ]
            $\wedge$ [kompromis$_{w_3 t_4}$,$i_6$]
            $\wedge$ [
                [mezi$_{w_3 t_4}$,
                    $\lambda w_8 \lambda t_9 \lambda x_{10}$[
                        [Imq,
                            $\lambda x_{11}$(
                                [poměrný$_{w_8 t_9}$,$x_{11}$]
                                $\wedge$ [řešení$_{w_8 t_9}$,$x_{11}$]
                                $\wedge$ [většinový$_{w_8 t_9}$,$x_{11}$]
                                $\wedge$ [řešení$_{w_8 t_9}$,$x_{11}$]
                            )
                        ],
                        $x_{10}$
                    ]
                ],
                $i_6$
            ]
            $\wedge$ $x_5$=[používat,$i_6$]$_{w_3}$
            $\wedge$ [Německo$_{w_3 t_4}$,$i_7$]
        )
    ]...π
```

Nebo Německo používá kompromis mezi poměrným a většinovým řešením .

03553.1: $\lambda w_1 \lambda t_2$[**Prog**$_{w_1 t_2}$,$\lambda w_3 \lambda t_4 (\exists\, x_5)(\exists\, i_6)($[**Does**$_{w_3 t_4}$,$i_6$,[**Imp**$_{w_3}$,$x_5$]] $\wedge$ x_5=**lišit**$_{w_3}$ $\wedge$ [**střední**$_{w_3 t_4}$,i_6] $\wedge$ [$\lambda w_7 \lambda t_8 \lambda x_9$([**Evropa**$_{w_7 t_8}$,$x_9$] $\wedge$ [[**od**$_{w_7 t_8}$,$\lambda w_{10} \lambda t_{11} \lambda x_{12}$[[**Imq**,$\lambda i_{13}$([**balkánský**$_{w_{10} t_{11}}$,$i_{13}$] $\wedge$ [**zem**$_{w_{10} t_{11}}$,i_{13}])],x_{12}]],x_9])$_{w_3 t_4}$,i_6])]...π

Jak se liší střední Evropa od balkánských zemí ?

Fig. 2. Corpus of TIL constructions – detailed view of a selected sentence

7.1 The TIL Corpus Web Interface

For this purpose, a web interface for the corpus was developed[1] to make the work of the logic specialists easier and to provide a visualisation of correct, incorrect and not yet checked constructions.

As can be seen in Figure 2, each sentence is displayed as the TIL construction (automatically created by the parser) as well as the plain text that enables easy reading. On the right side of each sentence, there are two buttons for marking the constructions accepted or wrong. There is also a status indicator showing if the sentence has already been checked and what was the decision.

If the decision was negative, the user (logic specialist) is asked to provide a brief description of the error. The decisions are stored in the database as well as the error descriptions and information about the user; all of this is immediately available to the parser developers. Also, the history of changes is kept by the means of the `git` versioning system.

[1] `http://corpora.fi.muni.cz/til`

8 Conclusions

In the paper, we have described the first steps to create an exemplary corpus of natural language data annotated for their logical structure. As a process parallel to the corpus creation, the parser producing the logical formulae is being improved using the feedback from corpus human experts. At the end of this process, we hope to have a high-quality corpus of TIL constructions as well as a wide-coverage parser producing TIL constructions with a reasonable precision.

The work has however just been started. The future will hopefully bring intensive development of the parser including adapting it to more complicated sentences as well as incremental increase of the quality of the corpus.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project P401/10/0792.

References

1. Duzı, 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).
2. Tichý, P.: The Foundations of Frege's Logic. de Gruyter, Berlin, New York (1988).
3. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2002).
4. Pease, A., Li, J.: Controlled english to logic translation. Theory and Applications of Ontology: Computer Applications (2010) 245–258.
5. Yusuke Miyao, Alastair Butler, K.Y., Tsujii, J.: A Modular Architecture for the Wide-Coverage Translation of Natural Language Texts into Predicate Logic Formulas. In: Proceedings of Pacific Asia Conference on Language, Information and Computation, PACLIC 2010, Institute for Digital Enhancement of Cognitive Development, Waseda University (2010) 481–488.
6. Mikulová, M. e.a.: Annotation on the Tectogrammatical Level in the Prague Dependency Treebank: Annotation Manual. Universitas Carolina Pragensis (2008).
7. Klimeš, V.: Transformation-based tectogrammatical analysis of Czech. In: Text, Speech and Dialogue, Springer (2006) 135–142.
8. Tichý, P.: Collected Papers in Logic and Philosophy. Prague: Filosofia, Czech Academy of Sciences, and Dunedin: University of Otago Press (2004).
9. 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.
10. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Proceedings of the Computer Treatment of Slavic and East European Languages 2005, Bratislava, Slovakia (2005) 107–115.
11. Pala, K., Smrž, P.: Building the Czech WordNet. Romanian Journal of Information Science and Technology **7**(2–3) (2004) 79–88.

12. Jakubíček, M.: Enhancing Czech Parsing with Complex Valency Frames. Master's thesis, Masaryk University, Brno (2010).
13. Horák, A., Pala, K., Duží, M., Materna, P.: Verb Valency Semantic Representation for Deep Linguistic Processing. In: Proceedings of the Workshop on Deep Linguistic Processing, ACL 2007, Prague, Czech Republic, the Association for Computational Linguistics (2007) 97–104.
14. Pala, K., Rychlý, P., Smrž, P.: DESAM — annotated corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530 Lecture Notes in Computer Science 1338.