An Experiment with Theme–Rheme Identification

Karel Pala and Ondřej Svoboda

Natural Language Processing Centre
Faculty of Informatics, Faculty of Arts
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pala@fi.muni.cz, xsvobo15@fi.muni.cz

Abstract. In this paper we start from the theory of Functional Sentence Perspective developed primarily by Firbas [1], Svoboda [12] and also later by Sgall et al. [9].

We make an attempt to formulate and implement a procedure for Czech allowing to automatically recognize which sentence constituents carry information that is contextually dependent and thus known to an addressee (*theme*), constituents containing new information (*rheme*), and also constituents bearing non-thematic and non-rhematic information (*transition*).

The experimental implementation of the procedure uses tools developed in NLP Centre, FI MU, particularly the morphological analyzer Majka [17], disambiguator DESAMB [16] and parser SET [5].

As a starting data resource we use a small corpus of 120 Czech sentences, which at the moment does not include a free continuous text. This is motivated by the fact that we do not use syntactically pre-tagged text but perform syntactic analysis directly using the parser SET. Thus, we offer only a very basic evaluation, which captures the main FSP phenomena and shows that the task is feasible.

The toolset developed for the experiment consists of two parts: first, a chunker, which determines word-order positions from the parse tree of a sentence, second, an FSP tagger which is the implementation of the procedure. It labels the chunks with the tags of what is further called *functional elements* (e.g. theme proper, transition, rheme proper). An experimental version is available at `http://nlp.fi.muni.cz/~xsvobo15/fsp/fsp.html`.

Keywords: rule-based parsing, chunking, functional sentence perspective

1 Introduction

The theory of Functional Sentence Perspective (FSP in the sequel) was proposed by V. Mathesius [6] and further elaborated by his pupil J. Firbas [1]. The term itself was created by J. Firbas as a more convenient English equivalent of Mathesius' Czech term *aktuální členění větné*.

The FSP theory naturally attracted other Czech researchers as well, particularly P. Sgall and E. Hajičová [9] who creatively introduced slightly different terminology: instead of FSP they started to use Topic-Focus Articulation (further TFA). The Czech equivalents of these terms are used rather interchangeably. In this paper we will prefer to use FSP as the original term as well as other terms like *theme* (topic in TFA), *rheme*

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 275–284, 2014.

(focus in TFA). In FSP also terms *transition* and *diatheme* [12] are used which do not seem to have straightforward counterparts in TFA. Then there are terms of the *context dependency* and *communicative dynamism*, introduced by J. Firbas [2]. They express the intuition that some information in a sentence is linked to the previous (verbal and also nonverbal) context and some is perceived as new.

In Firbas (and Svoboda) this is grasped in the following way: the sentence constituents bearing known or contextually dependent information are labelled as *themes*, then there are transitional elements – *transitions* and constituents carrying communicatively new (dynamic) information are called *rhemes*. Within thematic elements *themes proper* (ThPr) and *diathemes* (DTh) are further distinguished, which carry new information within the theme or refer to the new information from the previous text. *Transitions proper* (TrPr) and *rhemes proper* (RhPr) are also recognized among transitional and rhematic elements.

Some results by Karlík and Svoboda [4] offer a solution which inspired us to try a more formal formulation of the procedure able to automatically identify FSP elements in a sentence. They offer rules describing word order positions which can be occupied by individual sentence constituents and depending on their nature allowing to decide whether they can be labelled as thematic, transitional or rhematic. The first attempt to formulate the rules of Karlík and Svoboda as a formal procedure can be found in [8].

1.1 Early Experiments

There were attempts to propose an automatic procedure for TFA by Hajičová et al. [3] and Steinberger et al. [10] in the past. Steinberger's attempt was designed for German, Hajičová's proposal dealt with simple English sentences.

For both papers it is characteristic that they have an experimental nature and do not contain evaluation as we are used to it now. So it is not possible to asses at least approximately how successful the mentioned experiments were. This, however, is understandable if we take into consideration the time of their origin (almost 20 years ago).

1.2 Recent Development

Prague group members have published many papers related to the various aspects of the TFA theory recently, here we would like to mention especially the work related to the manual annotation of FSP (TFA) in Prague Dependency Treebank 3.0 (PDT), see [13].

PDT 3.0 contains annotation of the sentence constituents on three levels: morphological, analytical (syntactic) and tectogrammatical. We will touch here the tectogrammatical level, on which TFA elements (topic, focus and contrast) and communicative dynamism are manually annotated. Procedures for automatic topic/focus bipartition of sentences have been proposed and tested [15,14].

Initially, we considered to compare the PDT annotation obtained manually with our results. After a closer look at the PDT annotation we, however, came to the conclusion that this would be a completely separate task:

- first, apart from the terminological differences there are also differences in the notation that would require more detailed analysis,

- second, we, in fact, looked at some example sentences in the PDT Annotators Manual for T-level and found relevant terminological differences preventing us from trying to use PDT data for comparison in this paper, [7],
- third, TFA annotation in PDT is closely linked to the tectogrammatical level, which we do not work with,
- fourth, TFA annotation in PDT works with the terms *context (non)boundness* which we do not use in the same sense, and semantical relation of aboutness, which is difficult to grasp formally.

The mentioned points show that the more detailed comparison would be very stimulating but, as we hinted, it is a time consuming task for future. It would also require to create some reasonable test data, on which broader agreement could be hopefully reached. We also observe that within FSP theory it is not necessary to work with the tectogrammatical level.

2 Motivation

The task described above has been considered difficult but also challenging. Its successful solution will make it possible to obtain better insight into the information structure of utterances, which should allow for more accurate information extraction as well as meaningful understanding of the thematic progression in natural language texts [11]. Our ambition in this paper is to show that the automatic identification of themes and rhemes is feasible on the basic level at least. We concentrate on the basic aspects of the problem but are well aware of the wider context (e.g. anaphors or particles functioning as rhematizers). So far we work with some methodological constraints, see below.

Our approach is motivated by the fact that we try to answer simple questions first to gain firm ground for solving more complex parts of the problem in the next step. After having managed simple sentences we can come to complex clauses though the basic types of the Czech clauses are handled already.

3 Resources

Though the idea of using the PDT data as a resource for our experiments came to us as quite tempting we had to abandon it as we hinted above. One methodological decision was adopted: due to the experimental setup we decided not to work with free text yet, thus we have prepared a small corpus containing a collection of 120 sample Czech sentences representing various syntactic structures which allow us to test the FSP tagger and improve it step by step to be able to process free text, ultimately. The sentences in our experimental corpus are partly sample sentences displaying relevant syntactic structures and partly sentences taken randomly from online newspapers (such as iDNES or Lidové noviny).

4 Word Order Positions

The free word order in Czech makes it possible to combine sentence constituents quite freely. However, the internal word order within noun, adjective and adverbial phrases is practically fixed in Czech.

It can be observed that a finite verb takes the medium position in Czech sentences in approx. 60%. The morphosyntactic cases in Czech permit to have a direct object in accusative case or indirect object in dative case at the beginning of the sentence and subject in nominative case at the end frequently. The same can be said about adverbial constituents expressed either by adverbs or prepositional groups in various cases, most frequently in locative.

Following [4] we distinguish up to five word-order positions in Czech sentences: pre-initial (usually occupied by conjunctions), initial, post-initial (where enclitics follow Wackernagel's rule), medial and final. In this point we differ from TFA as it is annotated in PDT. The order of enclitic elements in Czech is strictly given: auxiliary forms of verb *být – to be*) are followed by reflexives (pronouns or particles), then by personal, adverbial and demonstrative pronouns.

Unlike the medial position, the initial and the final positions must always be present (even in the form of a merged initial-final position) and can contain only one sentence constituent. The initial, medial and final positions may be occupied by a noun phrase, an adverbial phrase or a verb. A conjunction or a particle may occur in the pre-initial position.

5 Levels of Analysis

To finally obtain labelled sentence constituents in their word order positions, several discrete steps, using automated tools, must be successfully performed. We will use the following sentence as an example:

Přijdu do školy, až napíšu ten text. which translates as *I will come to school when I finish writing the text.*

- First, tokenization (by `unitok.py`) of the input text takes place, yielding a basic vertical text.
- The vertical is extended with complex POS tags and lemmas by the morphological analyser Majka [17] and morphosyntactic disambiguator Desamb [16].
- Using an experimental grammar, partly written by one of the authors, the morphologically annotated input is unambiguously parsed by the SET parser to produce a dependency tree, see Figure 1.
- The chunker processes the obtained trees and segments the sentence(s) into word-order positions along with indication of the constituent type, e.g. a conjunction, a noun phrase or a relative clause.
- Finally, the FSP tagger labels the word order positions with functional elements, taking the information from the chunker and the position in a clause into account.

The grammar is required to have the following properties:

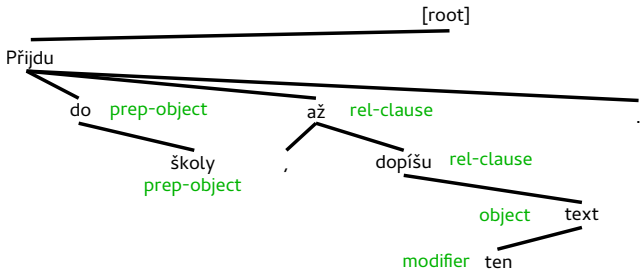

Fig. 1. Dependency tree created by SET with our grammar (note that the labels of edges are technical and unused currently)

- ability to recognize a coordination of top-level clauses within compound sentences,
- to link main types of subordinate clauses to their head words through subordinating expressions.

If these conditions are met, a single unit (e.g. a NP/PP, an adverb, a clause) can be extracted by the chunker from the parse tree to form, as a whole, a word-order position. Components of a verb phrase (VP) are, in contrast, separated into positions of their own.

To delimit individual sentences, the chunker first finds the root node. In accordance with the experimental grammar design, the root is either a finite verb (VF), or a coordinate conjunction with VFs as immediate children. The subtrees of VFs represent individual sentences.

Among a head VF's children, other parts of a verb phrase (e.g. auxiliars and modals) are found. Other subtrees of a VP (e.g. NPs represented by their head, a noun) then represent the rest of chunks, or sentence constituents. The chunks occupy word-order positions on their own. Even particles, not usually considered sentence constituents, are included.

The tagger's output format is the vertical text with tokens belonging to the word-order positions enclosed in XML notation. The XML opening tag (named after the position) contains morphosyntactic information added by the tagger and the FSP label explaining the choice, e.g.

```
<initial NP="k1gFnSc1" diatheme="NP in initial">
```

For testing, a web interface has been developed. Apart from plain text, vertical text may be passed to the tools in the pipeline.

The resulting XML is presented graphically after being XSL-transformed to HTML. The positions are shown as boxes, with the name above the content, and the labels underneath. Additional information is displayed in tooltips. Debugging output is also available in separate windows.

It has to be noted that in the course of the syntactic analysis we face usual problems with the ambiguity of prepositional phrase attachment. So far we have decided to work with the longest possible constituents but we are aware that this is just one of the heuristic solutions which has to be further tested.

We also have to mention some particles which play a relevant role in FSP tagging but are not easy to handle in parsing. In this context we speak about *rhematizers*, e.g. *jen* (only), *právě* (just), *i* (even), which indicate that a sentence constituent which follows them has to be labelled as rhematic. This is captured by rules of the FSP tagger (particularly in rule 4 given below in the next Section 6. The number of rhematizers in Czech is rather small, approximately 20.

```
<s>
    <initial PNE="p1nS" TME="eAaPmI" theme-proper="PNE"
             transition-proper="TME" verb="přijít">
Přijdu  přijít  k5eAaPmIp1nS  0  -1,
    </initial>
    <medial NP="k7c2" diatheme="NP">
do      do      k7c2          1  0,prep-object
školy   škola   k1gFnSc2      2  1,prep-object
    </medial>
    <final clause="až" rheme-proper="final position">
,       ,       kIx,          3  4,
až      až      k8xS          4  0,rel-clause
dopíšu  dopsat  k5eAaPmIp1nS  5  4,rel-clause
ten     ten     k3xDgInSc4    6  7,modifier
text    text    k1gInSc4      7  5,object
    </final>
.       .       kIx.          8  0,
</s>
```

Fig. 2. Vertical representation of word-order positions carrying FSP labels.

6 FSP tagger

The procedure first determines the nature of the chunks, matching them with word-order positions. The implemented procedure consists of the following main points:

1. If an input sentence is compound it is split into separate clauses by means of coordinate conjunctions or punctuation.
2. Some coordinate conjunctions, such as *a* (and) or *ale* (but), which stand in the front of a clause, create a pre-initial position (they can occupy no other position).
3. If a (group of) enclitics is found the chunks then form the post-initial position. This position is also optional.
4. The other chunks represent the initial, medial (optional) and final positions with one exception: if the post-initial position is the last in the clause the initial and final positions merge into one initial-final position.

Following the raising communicative dynamism scale, basic rules are applied to label the positions as thematic, transitional or rhematic:

1. All elements in the post-initial position and also finite verbs (for expressing the gender and number of a subject) in other positions are labelled as themes proper (ThPr).
2. Noun phrases (NP/PP), adverbs, infinitives, relative clauses and some particles are labelled as diathemes (DTh). Depending on the word-order position, some of them will be relabelled as transitional or rhematic in the next steps.
3. Finite verbs (bearing temporal and modal grammatical categories) and some particles are labelled as transitions proper (TrPr).
4. A position containing a rhematizer before a NP/PP is labelled as rheme proper (RhPr). The final position is labelled as transitional if it contains a finite verb. If a rhematizer was not found the position is labelled as RhPr.

Fig. 3. Graphical output of the chunker and FSP tagger

6.1 Example

The procedure applied to the example sentence finds a single clause in the input. The full stop is left aside. No element is found to mark a pre-initial or post-initial position. The first, last and middle chunks fall into the initial, final and medial positions, respectively.

The labelling of themes proper is performed, marking the the verb *přijdu* (I will come), whose ending *-u* expresses the subject (1st person and singular number), as ThPr. A NP is labelled as a diatheme. The verb is labelled additionally as TrPr. Finally, the clause standing in the final position receives the RhPr label.

One can argue that the label of *do školy* (to school) should be a *rheme* (Rh) rather than a diatheme because it expresses a rather important argument of the verb. On the other hand, if the school had already been mentioned the labelling would have been correct. Working with context and exploiting verbal valencies is, however, a subject to further experimenting.

7 Results and Evaluation

Presently, we have performed a basic experiment, in which sentence constituents have been labelled automatically with the FSP tags. As we have said above we have decided to work with some methodological limitations, particularly:

- In the experiment we work with simple sentences which contain the basic types of the dependent clauses (relative, content and adverbial ones).
- We do not work yet with a continuous text but only with a collection of the sample sentences, each considered out of context, to create a baseline we can build upon.

We have developed two tools, a chunker which processes the output from the parser SET and provides identification of the word-order positions in sentences taken from the the sample corpus (120 sentences so far). The output from the chunker is then handled by the FSP tagger assigning FSP labels to the sentence constituents occuring in the corpus sentences, see Figure 2 and Figure 3).

The results in Table 1 are very basic by their nature, they indicate that the success in FSP labelling is 88%. The number of serious errors (incorrect assignment of the Rh label) can be considered acceptable. They are basically caused by the quality of the used grammar.

Sentences in the testing corpus were evaluated by authors and care was taken to treat them in the same way as the procedure to account for the lack of context-sensitivity of the original formal description of the relation of word-order positions and the distribution of FSP elements in a Czech sentence.

Table 1. The results of experimental FSP labelling

	Sentences total	120	100.0%
A	Correctly analyzed (incl. marginal errors)	106	88.3%
N	Fatal errors	14	11.7%
A–	Marginal errors	13	10.8%
	All errors	27	22.5%

The line A comprises sentences, in which the FSP labels have been assigned correctly. As to errors we can clearly distinguish two sorts of errors:

- N, the result is completely negative, i.e. the FSP tagger does not assign the labels Th/Tr/Rh at all or assigns them incorrectly,
- *A*–, partial errors, here the FSP tagger assigns the label Rh correctly but single errors may occurr with other labels (Th, Tr, Dth).

In our view, the distinction between errors of the type N and A- has to be made, we are convinced that sentences with partial errors can be still considered acceptable. Thus we can conclude that the situation with evaluation is not black and white and will require further analysis.

It has to be remarked that there are several language phenomenona that lower the success rate of chunking and tagging:

- semi-sentential infinitive constructions and, similarly, nominal valencies are not well recognized currently,

– PP attachment causing e.g. adverbial NPs to be connected to other constituents than they belong to,
– coordinated relative clauses.

8 Conclusions

We have been dealing with the task consisting of the identification of word-order positions and automatic theme-rheme tagging in Czech. Starting from the work of Karlík and Svoboda [4] we attempted to formulate formal rules capturing behaviour of the constituents in Czech sentences with regard to the word-order positions they occupy. On this ground the rules form an algorithm for labelling thematic, transitional and rhematic elements in Czech sentences. The first experimental version of the procedure has been implemented, consisting of the two modules:

– the chunker processing simple Czech sentences with canonical (standard) word order,
– the FSP tagger tagging sentence constituents as thematic, transitional and rhematic.

We are well aware of the experimental and modest character of the presented results but, in our view, they show that it makes sense to go in the indicated direction. In the further research we will pay attention to the phenomena that so far prevent the FSP tagger from handling the continuous text with reasonable success.

Acknowledgements

This work has been partially supported by the Ministry of Education of Czech Republic under the project Lindat-Clarin.

References

1. Firbas, J.: On the problem of non-thematic subjects in contemporary English (English summary of "k otázce nezákladových podmětů v současné angličtině", ib. pp. 22–42 and 165–73). Časopis pro moderní filologii 39, 171–3 (1957)
2. Firbas, J.: Functional sentence perspective in written and spoken communication. Cambridge University Press (1992, reprinted 1995)
3. Hajičová, E., Sgall, P., Skoumalová, H.: An automatic procedure for topic-focus identification. In: Journal of Computational Linguistics, Vol. 21, Issue 1, March 1995. pp. 81–94. MIT Press Cambridge (1995)
4. Karlík, P., Svoboda, A.: Skladba češtiny pro cizince [Czech Syntax for Foreigners]. Univerzita J.E. Purkyně, Faculty of Arts, Brno (1982)
5. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for Czech. In: Human Language Technology: Challenges for Computer Science and Linguistics. p. 161–171 (2011)
6. Mathesius, V.: O tak zvaném aktuálním členění větném [on the so-called functional sentence perspective]. Slovo a slovesnost 5, 171–4 (1939)

7. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová-řezníčková, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z.: Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Tech. rep., ÚFAL MFF UK, Prague, Czech Republic (2005), http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html
8. Pala, K., Svoboda, O.: Semi-automatic theme–rheme identification. In: Proceedings of the Raslan Workshop. pp. 39–48. Karlova Studánka (2013)
9. Sgall, P.: Towards a definition of focus and topic. Prague Bulletin of Mathematical Linguistics 31, 32, 3–25, 24–32 (1979, 1980)
10. Steinberger, R., Bennett, P.: Automatic recognition of theme, focus and contrastive stress. In: Proceedings of the Conference Focus and NLP (1994)
11. Svoboda, A.: České slovosledné pozice z pohledu aktuálního členění. Slovo a slovesnost 45, 22–34, 88–103 (1984), http://kramerius.lib.cas.cz/search/i.jsp?pid=uuid:c9de3a32-530d-11e1-1418-001143e3f55c
12. Svoboda, A.: Kapitoly z funkční syntaxe. In: Spisy pedagogické fakulty v Ostravě. vol. 66 (1989)
13. Veselá, K., Havelka, J.: Anotování aktuálního členění věty v pražském závislostním korpusu (2003), http://ufal.mff.cuni.cz/pdt2.0/publications/VeselaHavelkaTR2003.pdf, ÚFAL/CKL TR-2003-20
14. Zikánová, Š., Týnovský, M.: Identification of topic and focus in czech: Comparative evaluation on prague dependency treebank. In: Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure (Formal Description of Slavic Languages 7.). pp. 343–353. Peter Lang, Frankfurt am Main (2009)
15. Zikánová, Š., Týnovský, M., Havelka, J.: Identification of topic and focus in czech: Evaluation of manual parallel annotations. The Prague Bulletin of Mathematical Linguistics No. 87 pp. 61–70 (2007)
16. Šmerk, P.: Unsupervised learning of rules for morphological disambiguation. In: Lecture Notes in Computer Science. vol. 3206, pp. 211–216. Springer Verlag (2004)
17. Šmerk, P.: Majka – fast morphological analyzer. In: Proceedings of the Raslan Workshop. pp. 13-16. Masarykova univerzita, Brno (2009)