Semi-automatic Theme-Rheme Identification

Karel Pala and Ondřej Svoboda

NLP Centre
Faculty of Informatics, Faculty of Arts
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{pala, xsvobo15}@fi.muni.cz

Abstract. In this paper we start from the theory of the Functional Sentence Perspective developed primarily by Firbas [1], Svoboda [2] and also Sgall, Hajičová [3] and make an attempt to formulate a procedure allowing to semi-automatically recognize which sentence constituents carry information that is contextually dependent and thus known to an adressee (*theme*), constituents containing new information (*rheme*), and also constituents bearing non-thematic and non-rhematic information (*transition*). Having themes and rhemes recognized as successfully as possible we also hope to investigate thematic progression (thematic line) in texts in the future. The core of the procedure and its experimental implementation for Czech (using the bushbank corpus CBB.Blog [4] as a data source) are described in the paper. Since the task is really complicated we only offer basic evaluation, which, in our view, shows that the task is feasible.

Key words: theme-rheme, Functional Sentence Perspective, topic-focus articulation

1 Introduction

The theory of the Functional Sentence Perspective (further FSP, [1,2,3]) has its origin in Czech linguistics, particularly in Prague Linguistic Circle. It states that in natural language sentences one can distinguish known, and new information. This is in agreement with our intuition, which reflects a sequential processing of the language data we are producing and receiving in the course of an information exchange. Terminologically, this is grasped in the following way: sentence constituents bearing known or contextually dependent information are called *themes*, elements serving as a backbone of a sentence are characterized as *transitions* and constituents carrying communicatively new (dynamic) information are called *rhemes*. Within thematic elements we can further distinguish *themes proper* (ThPr) and *diathemes* (DTh) which carry new information or refer to the new information from the previous text. Being important to the word order, *transitions proper* (TrPr) and *rhemes proper* (RhPr) are also recognized among transitional and rhematic elements. With regard to the FSP theory a natural question can be asked: is it possible to implement a procedure for identification of these elements in natural language sentences

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 39–48, 2013. © Tribun EU 2013

and texts semi-automatically? Such a procedure would be obviously useful for any kind of information processing, information extraction and observing thematic progression in texts. Some results by Karlík and Svoboda [5, in Czech] offer a solution which may lead to the more formal formulation of the procedure able to 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. Here we try to reformulate their rules in a more formal way to be able to solve the task algorithmically in Czech sentences.

With regard to the work done in the FSP area we are following Firbas' and Svoboda's terminology (theme, rheme) while Prague group (Hajičová, Sgall and others) uses different terms *topic*, *focus* and call the area *topic-focus articulation* (TFA). We have to mention past attempts to propose the automatic procedure for FSP by Hajičová et al [6] and Steinberger et al [7]. Steinberger's attempt was designed for German, Hajičová's proposal dealt with simple English sentences. For both papers it is common that they do not offer any evaluation, thus it is difficult to judge at least approximately how successful the mentioned procedures were. Due to the time of their origin (approx. 20 years ago) they were not related to any corpora. Prague group members have published many papers related to the various aspects of the FSP theory, here we would like to mention especially the work related to the manual annotation of FSP (TFA) in PDT 2, see [8, in Czech]. Thus PDT contains labelling of the sentence constituents as thematic and rhematic elements and initially, we have considered the idea of the comparison of our results obtained automatically with PDT annotation obtained manually. After a closer look at the PDT annotation we, however, have come to the conclusion that this is a task for a separate paper – first, the differences in the notation have to be analyzed and only then the comparison can be tried. In any case we will pay attention to the comparison in the near future.

2 Motivation

The task described above has been considered difficult. Its successful solution will make it possible to obtain better insight into the information structure of utterances, which should allow more accurate information extraction as well as meaningful understanding of the thematic progression in natural language texts [9]. Our ambition in this paper is to show that the task is feasible, semi-automatically at least. We will concentrate on the basic aspects of the problem but we are aware of the wider context (e.g. anaphors or particles – rhematizers).

3 Word order positions

The free word-order in Czech makes it possible to combine sentence constituents rather freely. It can be observed that a finite verb takes the medium position in sentences in approx. 60%. Noun and adverbial phrases occur either before the verb or behind it but sentences with a verb at the beginning or the end appear regularly as well. The morphosyntactic cases in Czech permit to have an *object* at the beginning of the sentence and a *subject* at the end frequently. Thus we deal with sentences displaying various word-order patterns, particularly with their main types. As a resource we use Czech BushBank [10] with 31,822 syntactically tagged sentences. We consider the following basic word-order patterns:

- s(VP ADP), s(VP NP ADP), s(VP ADP NP), . . . , verb is in the initial position: 6,410 (20.1%),
- s(NP VP NP), s(ADP NP VP NP), . . . , verb is in the medial position 18,746: (58.7%),
- s(VP), s(ADP VP), s(NP ADP NP VP), s(NP NP VP), . . . , verb is in the final position: 6,666 (20.9%).

We distinguish up to five word-order positions in Czech sentences: pre-initial (usually occupied by conjunctions that are not a part of a clause), initial, post-initial (where enclitics follow Wackernagel's rule) medial and final. 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 or an adverbial phrase, or a verb. A conjunction or a particle may occur in the pre-initial position, affecting e.g. the modality of the sentence.

So we can split the problem into three tasks:

- recognizing the sentence constituents (using the IOBBER chunker[10] and SET parser [11],
- segmenting a sentence into word-order positions (pre-initial, initial, post-initial, medial, final),
- identifying what sentence constituents occupy them and deciding if they contain thematic, transitional, or rhematic elements.

3.1 Recognition of the sentence constituents

For identifying the sentence constituents and word-order positions they belong into we use partial syntactic annotation in the bushbank originally supplied by a statistical parser IOBBER [10] (used for noun phrases) and rule-based SET parser (verb phrases and clauses), disambiguated manually afterwards. At the experimental stage partial syntactic information was sufficient. Morphological information was also necessary for the task: POS of the constituents plus all respective grammatical categories contained in tags. For example, the Czech noun *žárovky* (*light bulbs*) has the tag k1gFnPc1, expressing the corresponding categories, i. e. noun, feminine, plural, nominative. For a verb *mluvíme* (*we speak*) the respective tag is k5eAaImIp1nS, which provides grammatical categories relevant from the FSP point of view:

- tense and modality (present tense – aI, indicative – mI, further Temporal and Modal Exponents – TMEs),
- person and number (1st person – p1, singular – nS, further Personal and Number Exponents – PNEs).

They are needed for recognizing some thematic and transitional elements which are a part of a verb form in Czech.

There are, however, problems with prepositional noun phrases – it is difficult to recognize which sentence constituent they belong to. So far we decided to work with the longest possible constituents but we are aware that this is just a preliminary solution, which has to be tested in detail. The accuracy of the used parsers output obviously influences the assigning of the thematic and rhematic labels to the constituents but in our view this is not so critical, though the accuracy of the parsers does not exceed 89%.

We also have to mention some particles which play a relevant role in FSP tagging and are problematic also in parsing. These particles are called rhematizers (e.g. *jen (only), právě (just), ...*) and they indicate that a sentence constituent which follows them has to be labelled as rhematic. This is captured in rules (particularly in rule 7 given below in the next Section 4. The list of rhematizers in Czech is rather small, approximately not more than 10. We do not deal with rhematizers in detail since they are handled by the parser SET as units hanging on sentence constituents with rather unspecified status. To explore rhematizers in detail is a task for future research.

An important point has to be made here – due to the complexity of the task of Th/Rh labelling we have decided to work with simple sentences at the beginning. The rule was to allow no punctuation thus, among correctly formed sentences, avoiding subordinate clauses (which would, in cases, occupy word-order positions by their own) and the need for full syntactic analysis in which case a treebank corpus or parsers would have to be used. This is motivated by the fact that we have to answer simple questions first to gain firm ground for solving more complex parts of the problem. After having managed simple sentences we can come to complex clauses taking a full approach to the problem breaking any artificial limitations we used in the experiment.

4 A procedure for assigning themes and rhemes

Having the information mentioned above we can try to formulate basic rules for determining thematic, transitional and rhematic elements in a Czech sentence:

1. The first step is to recognize clause boundaries by finding the pre-initial position occupied with a clause conjuction,

2. If an adverb of time or place appears in the initial or medial position it is labelled as DTh (diatheme),
3. Enclitics (personal and other pronouns and auxiliary forms of *být – to be*) always take the post-initial position and are labelled as ThPr (theme proper), they are also anaphoric expressions. For dealing with them we should be able to recognize anaphors and their antecendents, however, this aspect of the issue calls for more detailed analysis. There is a tool for Czech able to handle anaphors but its integration into the FSP area has to be further explored in future. This rule has strongly deterministic character,
4. Any constituent in the final position is labelled a RhPr (rheme proper) – the rule seems to work very reliably,
5. A finite verb expressing grammatical categories of the subject is labelled as ThPr (theme proper) as well as TrPr (transition proper) for bearing temporal and modal (TMEs) categories,
6. Noun phrases in the initial or medial position are usually labelled as DTh (diatheme), noun phrases in the final position are most frequently labelled as RhPr. This rule appears to be almost universal.
7. If a rhematizer occurs in a sentence it indicates that a sentence constituent which follows it has to be labelled as rhematic.

4.1 An Example

The example sentence is taken from the corpus CBB.Blog [10] (shown in the original vertical format):

Motorka si razí cestu temným údolím a na plastovou pokrývku doráží neúprosný déšť. (*A motorbike was making its way through a dark valley and rain drops were beating the plastic cover mercilessly.*)

The individual clauses are found by means of finding sentence-level coordinate conjunctions which belong to pre-initial positions (*a* by rule 1). In these simple clauses, noun (*Motorka*) and prepositional (*na plastovou pokrývku*) phrases are recognized as sentence constituents on their own, occupying initial positions and by rule 6 they are found diathematic (DTh). A reflexive pronoun *si* in the first clause falls by Wackernagel's rule in the post-initial position and is considered a theme proper (ThPr, rule 3). By the same rule 6, the noun phrases *cestu* and *temným údolím* at the end of a clause (final position) are labelled rhemes proper (RhPr). In medial positions, finite verbs *razí* and *doráží* bearing TMEs (`mI` present tense, indicative mode, `aI` imperfect aspect) are assigned transitions proper (TrPr) and themes proper (ThPr) for having PNEs (`p3` 3rd person, `nS` singular).

The segmentation into word-order positions is a task for the first tool, saving also a summary of the contents, whether e.g. a sentence-level conjunction, a noun phrase, an adverb or a verbal exponent (PNE or TME) is present. The second tool then uses the combination of the supplied information and the position in a sentence to assign one or more FSP elements if a rule is defined. Development was carried in Python for the speed of development and the

```
<s>
Motorka    Motorka    k1gFnSc1P?       clause:ff.syntax.1268  k1c1gFnS:ff.syntax.1270
si         se         k3xPyFc3P-       clause:ff.syntax.1268  yFxPk3c3:ff.syntax.1271
razí       razit      k5eAaImIp3nSP?   clause:ff.syntax.1268  vp:ff.syntax.1269
cestu      cesta      k1gFnSc4P-       clause:ff.syntax.1268  k1c4gFnS:ff.syntax.1272
temným     temný      k2eAgNnSc7d1P-   clause:ff.syntax.1268  k1c7gNnS:ff.syntax.1273
údolím     údolí      k1gNnSc7P-       clause:ff.syntax.1268  k1c7gNnS:ff.syntax.1273
a          a          k8xCP-
na         na         k7c4P-           clause:ff.syntax.1274  k7c4:ff.syntax.1276
plastovou  plastový   k2eAgFnSc4d1P-   clause:ff.syntax.1274  k7c4:ff.syntax.1276
pokrývku   pokrývka   k1gFnSc4P-       clause:ff.syntax.1274  k7c4:ff.syntax.1276
doráží     dorážet    k5eAaImIp3nSP?   clause:ff.syntax.1274  vp:ff.syntax.1275
neúprosný  úprosný    k2eNgInSc1d1P-   clause:ff.syntax.1274  k1c1gInS:ff.syntax.1277
déšť       déšť       k1gInSc1P-       clause:ff.syntax.1274  k1c1gInS:ff.syntax.1277
.          .          kIx.P-
</s>
```

Fig. 1: Output of the parsers IOBBER and SET as stored in the CBB Corpus

temporary nature of both tools as we plan to offer basic FSP annotation directly in a parser's output.

5 Results and Evaluation

Presently, we have performed a basic experiment, in which FSP labels have been assigned to sentence constituents with some limitations: in the experiment we decided to work only with simple sentences containing coordinate clauses and with omitted punctuation, for which our two tools word-order segmenter and FSP tagger (both written in Python) provide: identification of the word-order positions in sentences taken from the CBB corpus (9,513 sentences) (by segmenter), assigning thematic and rhematic labels to the sentence constituents occuring in these sentences (by FSP tagger, see Figure 2 and Figure 3).

Table 1: The results of first the experimental Th/Rh labelling (recall)

All sentences in CBB.Blog	32,287	
Simple sentences	9,513	100.0%
Labelled (fully or partially)	8,481	89.2%
Not labelled	1,032	10.8%

The results in Table 1 are very basic by their nature, they say that recall is 89.2% which is the result that has exceeded our expectations. It is partially limited by the quality of the parser's output.

```
<s>
    <initial NP="k1c1gFnS" diatheme="NP">
Motorka    Motorka   k1gFnSc1P?      clause:ff.syntax.1268  k1c1gFnS:ff.syntax.1270
    </initial>
    <post-initial theme-proper="post-initial position">
si         se        k3xPyFc3P-      clause:ff.syntax.1268  yFxPk3c3:ff.syntax.1271
    </post-initial>
    <medial PNE="k5eAaImIp3nSP?" TME="k5eAaImIp3nSP?" theme-proper="PNE"
            transition-proper="TME" verb="razit">
razí       razit     k5eAaImIp3nSP?  clause:ff.syntax.1268  vp:ff.syntax.1269
    </medial>
    <medial NP="k1c4gFnS" diatheme="NP">
cestu      cesta     k1gFnSc4P-      clause:ff.syntax.1268  k1c4gFnS:ff.syntax.1272
    </medial>
    <final NP="k1c7gNnS" rheme-proper="final position">
temným     temný     k2eAgNnSc7d1P-  clause:ff.syntax.1268  k1c7gNnS:ff.syntax.1273
údolím     údolí     k1gNnSc7P-      clause:ff.syntax.1268  k1c7gNnS:ff.syntax.1273
    </final>
    <pre-initial conjunction="a">
a          a         k8xCP-
    </pre-initial>
    <initial NP="k7c4" diatheme="NP">
na         na        k7c4P-          clause:ff.syntax.1274  k7c4:ff.syntax.1276
plastovou  plastový  k2eAgFnSc4d1P-  clause:ff.syntax.1274  k7c4:ff.syntax.1276
pokrývku   pokrývka  k1gFnSc4P-      clause:ff.syntax.1274  k7c4:ff.syntax.1276
    </initial>
    <medial PNE="k5eAaImIp3nSP?" TME="k5eAaImIp3nSP?" theme-proper="PNE"
            transition-proper="TME" verb="dorážet">
doráží     dorážet   k5eAaImIp3nSP?  clause:ff.syntax.1274  vp:ff.syntax.1275
    </medial>
    <final NP="k1c1gInS" rheme-proper="final position">
neúprosný  úprosný   k2eNgInSc1d1P-  clause:ff.syntax.1274  k1c1gInS:ff.syntax.1277
déšť       déšť      k1gInSc1P-      clause:ff.syntax.1274  k1c1gInS:ff.syntax.1277
    </final>
.          .         kIx.P-
</s>
```

Fig. 2: Word-order positions identified and FSP elements recognized

As to the assessment of the accuracy of the Th/Rh labelling, i.e. to evaluation how successful the labelling was, we have selected a sample containing 300 sentences each with assigned Th/Rh and evaluated it manually. The results can be seen in Table 2 and they show that our experiment is making sense.

Table 2: The accuracy of the first experimental Th/Rh labelling

	Sample sentences	300	100.0%
A	Correctly labelled sentences	203	67.6%
Ap	Correctly labelled sentences, partly	61	20.2%
N	Sentences with Rh not recognized	18	6.0%
Np	Sentences with errors in labelling	19	6.2%

The presented results can be characterized as more than promising – first two lines include sentences in which the Th/Rh labels have been assigned sucessfully, the line A includes sentences in which Th/Rh labels have been assigned completely, line Ap contains sentences in which label Rh is assigned correctly but some sentence constituents are not labelled, however, the result can be still considered acceptable. Similarly, line N comprises sentences in which no Th/Rh labels are assigned to sentence constituents, i.e. this result is completely negative. In Np there are sentences where some sentence constituents are labelled correctly but Rh is not identified. On the whole, we can take A and Ap together obtaining 87.7% sentences processed successfully. This result can be considered suitable for possible future applications.

It has to be remarked that there is a phenomenon that lowers accuracy of tagging – it is related to the infinitives in Czech. They can be parsed in different ways that call for a deeper analysis – this causes errors in Th/Rh labelling.

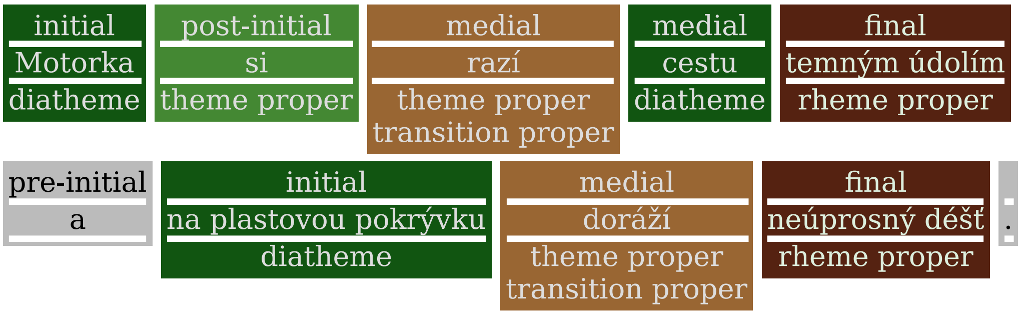

Fig. 3: Graphical output of the FSP parser

It can be seen how the Th/Rh labels are assigned to the individual sentence constituents and how the word-order positions are recognized. In the application it is possible to click on words in the sentence – then tags become visible. In this way one can observe the necessary details of the analysis.

6 Conclusions

In the paper we have been dealing with the task consisting of the identification of word-order positions and semi-automatic theme-rheme tagging in Czech. Starting from the work of Karlík and Svoboda [5] we formulate rules capturing behaviour of the constituents in Czech sentences with regard to the word-order positions they occupy. The rules form a procedure for labelling thematic, transitive and rhematic elements in Czech sentences. The experimental version of the procedure has been implemented as a tool having two modules: the segmenter processing simple Czech sentences with the standard word-order and FSP tagger tagging sentence constituents as thematic and rhematic. We are well aware of the experimental character of the presented results but, in our view, they show that it makes sense to go in the indicated direction.

Acknowledgements This work has been partially supported by the Ministry of Education of CR under the LINDAT-Clarin project LM2010013.

References

1. Firbas, J.: Functional sentence perspective in written and spoken communication. Cambridge University Press (1992, reprinted 1995)
2. Svoboda, A.: Kapitoly z funkční syntaxe (Chapters from the Functional Syntax). In: Spisy pedagogické fakulty v Ostravě (Writings of the Pedagogical Faculty in Ostrava), svazek (vol.) 66. (1989)
3. Hajičová, E., Buráňová, E., Sgall, P.: Aktuální členění věty v češtině (Functional Sentence Perspective in Czech). Academia, Prague (1980)
4. Grác, M.: Rapid Development of Language Resources. PhD thesis, Masaryk University, Brno (2013) Available on-line: `http://is.muni.cz/th/50728/fi_d/?lang=en`.
5. Karlík, P., Svoboda, A.: Skladba češtiny pro cizince (Czech Syntax for Foreigners), Brno (1982)
6. Hajičová, E., Sgall, P., Skoumalová, H.: An automatic procedure for topic-focus identification. In: Journal of Computational Linguistics, Vol. 21, Issue 1, March 1995, MIT Press Cambridge (1993) 81–94
7. Steinberger, R., Bennett, P.: Automatic Recognition of theme, focus and Contrastive stress. In: Proceedings of the Conference Focus and NLP. (1994)
8. Veselá, K., Havelka, J.: Anotování aktuálního členění věty v Pražském závislostním korpusu (2003) ÚFAL/CKL TR-2003-20, available on-line: `http://ufal.mff.cuni.cz/pdt2.0/publications/VeselaHavelkaTR2003.pdf`.
9. Svoboda, A.: Diatheme: a study in thematic elements, their contextual ties, thematic progressions and scene progressions based on a text from Ælfric. Univerzita J.E. Purkyně, Brno (1981)

10. Radziszewski, A., Grác, M.: Using Low-Cost Annotation to Train a Reliable Shallow Czech Parser. In: Proceedings of the TSD Conference, Heidelberg, Springer (2013) 575–584 Available on-line: `http://www.phil.muni.cz/plonedata/wkaa/BSE/BSE_2003-29_Scan/BSE_29_09.pdf`.
11. 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. (2011) 161–171
12. Golková, E.: Bibliography of the publications of Professor Jan Firbas. In: Brno Studies in English: 29. (2003) 99–108 Available on-line: `http://www.phil.muni.cz/plonedata/wkaa/BSE/BSE_2003-29_Scan/BSE_29_09.pdf`.
13. Šmerk, P.: Unsupervised Learning of Rules for Morphological Disambiguation. In: Lecture Notes in Computer Science, Springer Verlag (2004) 211–216