# Semi-automatic Theme-Rheme Identification

K. Pala, O. Svoboda

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

# Outline

- Theory of FSP – Firbas, Svoboda (1964, 1982)
- In Prague – Sgall, Hajičová, TFA (1980, 1993), only topic and Focus
- FSP elements (Firbas), Th, ThP, Dth, Tr, Rh, RhP
- Partial syntactic analysis, IOBBER, SET
- Simple sentences
- Word-order positions: (pre-)initial, post-initial, medial, final
- Rules formulated by Svoboda and Karlík
- Tools – segmenter and FSP tagger
- Results, evaluation

# Introduction

- Known, transitional and new information
- Our intuition, which reflects sequential processing of the language data
- Known information is typically contextually dependent
- Transitional elements are mostly verbs
- Elements carrying new information typically appear in the final part of the sentence
- Can thematic, transitional and rhematic elements be identified semi-automatically
- Two experiments – Hajičová et al 1993 for English, Steinberger 1994 for German
- New attempts appear with regard to PDT which contains manually annotated TFA – help for annotators

# Data

- We have decided to start with simple sentences, complex clauses will come later
- We have used output from the parser IOBBER (Grác, Radzisewski, 2013) and SET (Kovář, Jakubíček)
- Czech Bushbank – 31 822 sentences
- In 20,1 % sentences verb occurs in the initial position
- In 58,7 % sentences verb appears in the medial position
- In 20,9 % sentences verb occurs in the final position
- Enclitic elements behave very regularly, they follow Wackernagel's rule
- They standardly appear in the post-initial position

# Typical Word-order Positions

- Pre-initial – particles, conjunctions
- Initial – noun constituents
- Post-initial - enclitics
- Medial – verbs, adverbs
- Final – noun constituents
- Rhematizers – typically appear in front of the respective constituent (*jen, jenom, právě, zrovna*)
- Relations between sentence constituents and word-order positions – what can appear on what position?

# Rules for FSP labeling

a) First: clause boundaries are recognized by finding the pre-initial position occupied with a clause conjuction or particle,

b) If an adverb of time or place appears in the initial or medial position it is labelled as DTh (diatheme),

c) 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). This rule has strongly deterministic character,

d) Any constituent in the final position is labelled a RhPr (rheme proper) (the rule works very reliably),

e) 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) exponents, and also personal and numeral ones (PNEs)

f) 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.

g) If a rhematizer occurs in a sentence it indicates that a sentence constituent which follows it has to be labelled as rhematic.

# The segmenter

- The segmenter splits sentences found in CBB.Blog into chunks (doing
- some post-processing):
- – read a sentence (delimited by a full stop)
- – skip it if it contains punctuation (relative clauses, etc.)
- – if the first token is a coordinate conjunction,
- a pre-initial position is found
- – next, a constituent in an initial position is recognized:
- – a finite verb
- – a noun (prepositional) phrase (minimal/maximal by a switch)
- – an adverb, a conjunction

# Segmenter 2

- other unit (frequently (a part of) NP) not marked as such,
- infinitives with no complements
- in a loop across all tokens:
- enclitics are found both using a hard-wired list and a rule
- (those following a preposition)
- verbs, noun/prepositional phrases, conjunctions, adverbs,
- particles are marked to belong to a single position
- other elements are reported but marked anyway
- reflexive pronouns are split from the end of noun phrases

# FSP tagger

- The FSP tagger names word-order positions found by the segmenter and assigns sentence constituents:
- processes sentences possibly found in a coordination separately (using conjunctions already marked by the segmenter or punctuation present in the manually tagged reference corpus)
- an initial position is processed in the following way:
- if it contains a finite verb it is assigned a theme proper and a transition proper
- a noun or prepositional phrase is labeled as diathematic here
- – an infinitive is also labeled as diathematic

# FSP parser 2

Conjunctions and particles are marked as a part of transition proper

- a final position is looked for from the end of a clause – if enclitics are there, the final position is renamed to an initial-final
- all enclitics are labelled as themes proper
- if there are elements between the initial/post-initial and final position they are treated as belonging to a medial position
- they are labeled in a similar fashion as in the initial position except for an infinitive marked as a transition (non-proper)
- the final position is marked as a rheme proper; if a finite verb is there, theme proper and transition proper label is added as well

# Evaluation

- Sample of 300 sentences – processed manually
- Various types of the errors have been discovered:
- Errors in corpus – e.g. tokenization, punctuation
- Errors coming from the parsers – noun and prepositional constituents analyzed separately
- Failure of the FSP rules – full and acceptable
- Results depend on the types of errors
- Their acceptability varies – it has to be analyzed further

# Table 2

300 100.0%

| | | |
|---|---|---|
| A: Correctly labelled sentences | 203 | 67.6% |
| Ap Correctly labelled sentences, partly | 60 | 20.2% |
| N Sentences with Rh not recognized | 18 | 6.0% |
| Np Sentences with n.f. errors in labeling | 19 | 6.2% |
| | 300 | 100 % |

# Conclusions

The task consisted in the experiment with semi-automatic identification:

- of word-order positions

- theme-rheme tagging in Czech

- starting point – rules by Svoboda and Karlík (1982)

- segmenter and FSP tagger

- in our view, the results show that the attempt makes sense

- new problems have appeared not touched in the FSP theory yet

  (infinitives, rhematizers, pp-attachment, MWEs and idioms, types of errors and their relation to the evaluation)

- relation to the output of parsers