

# ***Expanding Topic-Focus Articulation with Boundary and Accent Assignment Rules for Romanian Sentence***

Neculai Curteanu<sup>1</sup>, Diana Trandabăț<sup>1,2</sup>, Mihai Alex Moruz<sup>1,2</sup>

<sup>1</sup> Institute of Computer Science, Romanian Academy

<sup>2</sup> Faculty of Computer Science, University Al. I. Cuza Iași

# Motivation

- The aim of this paper is to bring a fresh balance on the edge of textual-intonational discourse theories, by using the *information structure (IS) discourse semantics*, relating as closely as possible textual discourse structures to intonational-prosodic discourse phrasing.
- Steedman (2000) supports and proves the existence of a consistent set of rules that assign categorical functions of tones and tunes to the IS textual entities and spans.
- Our main concern is the *textual IS-semantics* side of the language interface within the (Romanian) prosody, with the aim of providing human-like prosody.

# Overview

↪ There are two approaches for prosody generation from text:

↪ Background-Kontrast

- Background-Kontrast is associated to pitch accents.
- Topic-Focus in the Prague School's Topic Focus Articulation algorithm

↪ Theme-Rheme

- *Theme-Rheme* structures are contiguous, the theme appearing at the beginning of the sentence, and the *rheme* “developing” the theme within a sentence (clause)
- Theme is what the speaker choose to take as the departure point in a clause, while Rheme is what the hearer already knows or has access to.

# Overview

- *Two lines* of investigation are designed for detaching the IS categories and structures.
  - *Improving* the Topic-Focus Articulation algorithm with rules derived from predicate semantics and *creating* a set of rules for assigning Theme-Rheme on the topic/focus marked entities using the communicative dynamism.
  - The *second approach* is based on *discursive theories*: Segmented Discourse Representation Theory is applied to determine the Background-Kontrast entities within clauses; Theme-Rheme structures are computed with a recursive version of the *Inference-Boundary (IB) algorithm* [Leong; 2004];

# TFA for Romanian

- Prague's School Topic-Focus Articulation (TFA) algorithm (Hajicova et al., 1995), attributes Topic-Focus to a sentence, considering its constituent order in communicative dynamism vs. the “standard” systemic order of the clausal semantic roles.
- The TFA algorithm is applied on constituency parsed clauses, annotated with part-of-speech information.
- Several semantic features of the constituents are also considered, namely *specificity degrees*:
  - (i) *general* – low specificity, contextually non-bound;
  - (ii) *specific* – high specificity, contextually non-bound;
  - (iii) *indexical* – mid-specificity, contextually-bound.
- **Example 1:** [Privit de sus]<sub>T</sub>, [lichidul]<sub>T</sub> [părea negru]<sub>F</sub>.

# TFA extension for Romanian

➤ TFA algorithm problems:

➤ The context considered is minimal, considering only the current sentence and the inflected form of the words.

*“As for the verb, it is important to have access to the verb of the preceding utterance and to use a systematic semantic classification of the verbs. If the main verb of sentence  $n$  has the same meaning as (or a meaning included in) that of sentence  $n - 1$  (in the sense of hyponymy), then it belongs to the topic.” (Hajicova et al. [17: pp. 9]).*

➤ Following the Topic-Focus assignment, some sentences contain no focus at all, or more than one focused constituent.

➤ We need a balanced assignment of Topic-Focus entities, observing both the TFA criteria and sentence level prosody patterns such as Sentence Accent Assignment Rule.

# TFA extension for Romanian

- We propose a new version of the TFA algorithm, completed with information structure obtained from discursive analysis.
- The TFA notion of Topic has much in common with the more recently characterized concept of *Background*, while the Focus correspond to the notion of *Kontrast*.
- Considering the preceding context of a sentence and performing co-reference resolution, each constituent is annotated with *Background* or *Kontrast*.
  - The *backgrounded* entities are the ones already mentioned in the discourse
  - *Kontrasted* entities are the entities that have not been previously introduced in the discourse.

# TFA extension for Romanian

↪ We added supplementary steps to the TFA algorithm, in order to assign *one* and *only one* focus to a clause using a set of rules derived from SAAR (Sentence Accent Assignment Rule).

↪ **SAAR-derived rules:**

- If the verb has adjuncts, then the adjuncts will receive focus.
- If the verb has arguments, then the arguments will receive focus.
- If the verb has more than one adjunct or more than one argument, then the rightmost one receives the focus.

# TFA extension for Romanian

- ~ **Example 2a.** [Vinul]<sub>T</sub> [nu avea]<sub>T</sub> [culoarea obișnuită]<sub>T/F</sub>.
- ~ **Example 2b.** [Vinul]<sub>T</sub> [nu avea]<sub>T</sub> [culoarea obișnuită]<sub>F</sub>.
  
- ~ **Example 3a.** [Maria]<sub>T</sub> [nu a fost]<sub>T</sub> [în acea zi]<sub>T</sub>.
- ~ **Example 3b.** [Maria]<sub>T-ARG</sub> [nu a fost]<sub>T</sub> [în acea zi]<sub>F-ADJ</sub>.
  
- ~ **Example 4a.** [Ion]<sub>T</sub> [a câștigat]<sub>F</sub> [o competiție]<sub>F</sub>.
- ~ **Example 4b.** [Ion]<sub>T</sub> [a câștigat]<sub>T/F</sub> [o competiție]<sub>F</sub>.
  
- ~ **Example 5a.** [Ion]<sub>T</sub> [a câștigat]<sub>F</sub> [la competiție]<sub>T</sub> [o cupă]<sub>F</sub>.
- ~ **Example 5b.** [Ion]<sub>T</sub> [a câștigat]<sub>T/F</sub> [la competiție]<sub>T</sub> [o cupă]<sub>F</sub>.

# Theme-Rheme assignment on TFA structures

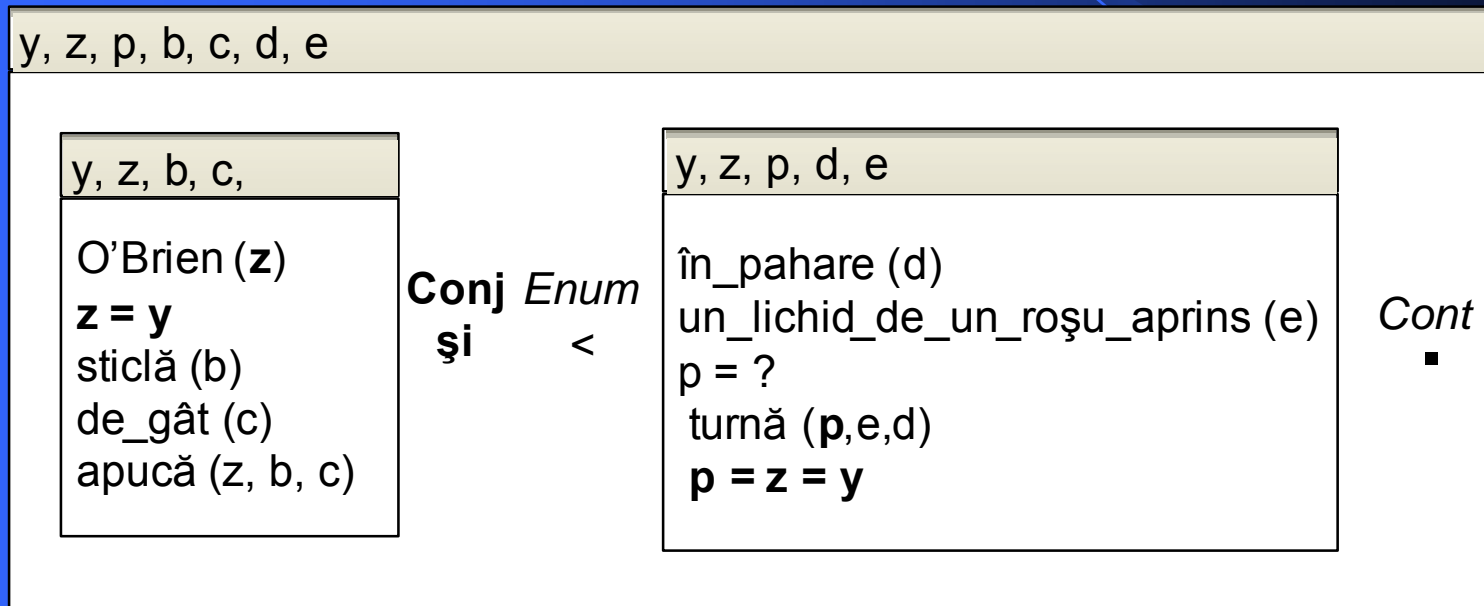
- We developed a set of simple *Theme-Rheme* assignment rules that apply on the structures marked with the TFA algorithm.
  - The constituents before the verb form the *theme*.
  - The constituents placed after the verb form the *rheme*.
  - The verb is included in the theme if it is marked with topic (t), in the *rheme* otherwise.
  - Within the discovered *rheme*, if a constituent is marked as topic, and it is not the last constituent of the sentence, it belongs to the *theme*, since it marks a violation of the systemic order.
- **Example 6.**  $[[\text{Ion}]_T]_{\text{theme}} [[\text{a câștigat}]_{T/F} [\text{o competiție}]_F]_{\text{rheme}}$ .
- **Example 7.**  $[[\text{Ion}]_T]_{\text{theme}} [[\text{a câștigat}]_{T/F}]_{\text{rheme}} [[\text{la competiție}]_T]_{\text{theme}} [[\text{o cupă}]_F]_{\text{rheme}}$ .

# Combined Discourse Theories for determining IS-structures

➤ In this section we approach two ideas:

- Segmented Discourse Representation Theory (SDRT - Asher, 1993) allows for discursive evolution of reference entities and relations among discourse structures.
  - SDRT, while inheriting the control of inter-clause referential variables, which we use to compute Background-Kontrast entities, borrows discourse relations from RST and classical predicational semantics.
  - The natural environment for the development of IS-discourse relations.
- The sub-clausal IS discourse relations proposed by Heusinger (2007).
  - In the SDRT setting, Heusinger introduces *five* IS-discourse relations to be applied to intonational-prosodic phrasing: Non-restrictive Modification, Enumeration, Frame-Setting, Backgrounding, Topicalization.

# Background-Kontrast Entities Acquired using SDRT



↪ [O'Brien]<sub>B</sub> [apucă]<sub>K</sub> [sticla]<sub>K</sub> [de\_gât]<sub>K</sub> [și]<sub>B</sub> [Ø = O'Brien]<sub>B</sub> [turnă]<sub>K</sub> [în pahare]<sub>K</sub> [un lichid de un roșu aprins.]<sub>K</sub>

# Theme-Rheme Computing with Inference-Boundary Algorithm

- Leong (2004) proposes the *Inference-Boundary* algorithm for delimiting the Theme and Rheme spans of the clause.
- We propose an extended, recursive form of the IB-algorithm for computing the Theme-Rheme spans within Romanian sentences.
- The input of the algorithm:
  - Segmentation of the finite clauses in the sentence; the non-finite clauses (if any) are recognized within the finite clauses;
  - A constituent parsing of the elements in the clauses;
  - The *syntactic* roles of the syntactic constituents;
  - The systemic order (SO) for each predication.

# Theme-Rheme Computing with Inference-Boundary Algorithm

## ~ Textual theme:

- **A** = {Continuatives (exclamations): *hai, hei, uau, ...*}
- **B** = {Conjunctives; e.g. *și, sau, dar*, punctuation marks, etc.}
- **C** = {Conjunctive-adjuncts; e.g. *deși, deci, din cauză că*, etc.}

## ~ Interpersonal Theme:

- **E** = {Vocatives; e.g. *Ion!, Maria!*}
- **F** = {Modal adjuncts, mood or comment adjuncts; e.g. *probabil, desigur, posibil*, etc.}
- **G** = {Finite operators; auxiliaries}

## ~ Topical Theme:

- **D** = {Wh-relatives; e.g. *care, pe al căreia, dintre cei cărora*, etc.}
- **H** = {Wh-question words; e.g. *cine, ce, cui, unde, cum, de ce*, etc.}
- **Ipart** = {NGs; viz. participant Top\_Th}
- **Jcirc** = {PPs, NGs; viz. circumstance Top\_Th}
- **Kproc** = {VGs; viz. process Top\_Th, if VG = predication}

# Theme-Rheme Computing with Inference-Boundary Algorithm

- The three presented themes are contiguous. The final theme of the finite clause is the succession of the three themes.
- **Rheme:**
  - all arguments and adjuncts that are included into a Theme using the sets above.
  - If within the Rheme the systemic order is broken, then the respective constituents constitute a Pseudo-Rheme
- The basic rule for detecting the themes within a sentence is to start searching at the beginning of each clause until one finds the topical theme.
- The thematic portion of the clause has to enclose a topical theme and, optionally, the textual and/or interpersonal theme.
- If no topical theme is present, a covert one (such as the empty grammatical subject or zero anaphora) has to be the topical theme of the sentence.

# Remarks concerning the IB-algorithm (for Romanian)

~ [O'Brien]<sub>TH</sub> [apucă]<sub>RHSegm</sub> [sticla de gât]<sub>RHSegm</sub><sup>1/</sup> [și ∅]<sub>TH</sub>  
[turnă în pahare]<sub>PseudoRHSegm</sub> [un lichid de un roșu  
aprins.]<sub>RHSegm</sub><sup>2/</sup>

~ [Lui Winston]<sub>TH</sub> [i se trezără niște vagi amintiri,]<sub>RHSegm</sub><sup>1/</sup>  
[ceva]<sub>TH</sub> [ce văzuse demult]<sub>PseudoRHSegm</sub> [— o sticlă  
mare]<sub>RHSegm</sub> // [făcută]<sub>RHSegm</sub> [din lumini electrice]<sub>RHSegm</sub><sup>2/</sup>  
[care]<sub>TH</sub> [parcă se mișca]<sub>RHSegm</sub> [în sus și în jos]<sub>RHSegm</sub> //  
[turnându-și conținutul]<sub>RHSegm</sub> [într-un pahar.]<sub>RHSegm</sub><sup>3/</sup>

# Results and discussions

- In order to assess the results of the two presented algorithms, we compared the output with the manual annotation of the same sentences (a set of spoken Romanian sentences extracted from George Orwell's novel *1984*).
- The sentences were marked with Theme-Rheme boundaries and with ToBI pitch accents by the Group of Speech Processing within the Institute of Computer Science.

# Comparing the Background-Kontrast assignment

- ~ The manual annotation allows for multiple high pitch accents, as the SDRT-based approach does, while the TFA algorithm restricts to only one focus in each clause.
- ~ The SDRT-based approach, however, distinguishes many empty elements as Background, while marking most of the lexicalized constituents as Kontrast (Focus).
- ~ As for verbs, the TFA mostly marks verbs as *f* or *t/f*; the SDRT-based method mostly marks verbs as *Kontrast*; in the gold annotation, the verb forms a unit with the next constituent, marked usually with HH or HL tones.
- ~ The last constituent of the sentence is also differently treated: in the TFA approach it is mostly focused; the SDRT-based analysis has no restriction for the rightmost constituent, being considered exactly as the other constituents; in the gold annotation, the last constituent marks very clearly the descending tendency of declarative sentences in Romanian.

# Comparing the Theme-Rheme assignment

- The annotator has preferred to place as many Themes as possible, and only the last intonational group was considered Rheme; this tendency is reversed in the automatic annotation.
- There are many more similarities between the three annotations, as almost all the sentences start with Theme and end with Rheme.
- Differences appear in the marking of the middle constituents, and in establishing the boundaries for the Theme-Rheme spans.
- The analysis performed on the two proposed methods, revealed that TFA performs better in the topic-focus assignment task, while the IB approach is more reliable for the Theme-Rheme assignment.
- Prediction of prosody from text can be substantially improved by a deeper understanding of the textual discourse theories.

Thank you!