Discourse Theories vs. Topic-Focus Articulation
Applied to Prosodic Focus Assignment in Romanian

Neculai CURTEANU¹, Diana TRANDABĂŢ¹,², Mihai Alex MORUZ¹,²
¹Institute of Computer Science, Romanian Academy, Iaşi Branch;
²Faculty of Computer Science, University “Al. I. Cuza” Iaşi,
IAŞI, ROMANIA
{curteanu, dtrandabat, mmoruz}@iit.tuiasi.ro

Abstract. “…it is not the argument structure that triggers the intonational phrasing, but the [discourse, subclausal n.b.] relation of backgrounding.” (K. von Heusinger, 2007) Is it? If yes, how?

The present paper, maintaining our attempts for applying Prague School’s Topic-Focus Articulation (TFA) algorithm on the syntax-prosody interface of Romanian, proposes two comparative lines of investigation for the intonational focus assignment: (A) The TFA algorithm is improved at clause level with hints from Gussenhoven’s SAAR (Sentence Accent Assignment Rule), then extended to inter-clause level, i.e. complex sentences. The new shape of the TFA algorithm is applied to compute the Topic-Focus values in the discursive context, while the information-structural (IS) spans of Theme(s)-Rheme(s) are detached, at clause level, as lowest-highest degrees of Communicative Dynamism (CD) vs. Systemic Order (SO).
(B) The second approach we experiment for assigning intonational focus and phrasing is based on the combined and intensive use of discourse theories for computing the IS categories and structures: the Background-Kontrast entities (associated with the Prague School’s Topic-Focus) are obtained with Asher’s (1993) Segmented Discourse Representation Theory (SDRT) analysis, while Theme-Rheme structures within the finite clause are computed with Leong’s (2004) Inference-Boundary (IB) algorithm (of Hallidayan inspiration), applied for the first time to Romanian. Furthermore, this second direction is inspired from and joins the IS-discourse theory proposed by Heusinger (2007), which relies on SDRT inter-clausal evolution of discourse variables for computing the Background-Kontrast. While maintaining the classical SDRT (including rhetorical) discourse relations at the inter-clause level, Heusinger introduces a set of IS-semantics relations, and hands down at sub-clause level the rhetorical and focus relation particles with significant role in intonational-prosodic phrasing. Examples of these two types of research are compared to a gold, intonationally annotated set of Romanian sentences, the proposed theoretical and procedural techniques aiming to balance the pessimistic-realistic view on prosody prediction that it is the speaker-presupposition (and hearer-accommodation) which determines the IS focal scopes, rather than the bare text.

Keywords: syntax-prosody interface; prosodic focus assignment; Topic-Focus Articulation algorithm; discourse theories; Information Structure theory; Information Structure Discourse Theory; theme-rheme; background-kontrast; topic-focus; prosody prediction for Romanian.

I. INTRODUCTION

The aim of this paper is to bring a fresh balance on the edge of textual-intonational discourse theories, coming from a deeper understanding of the textual discourse theories, with the natural emphasis on the information structure (IS) discourse semantics, relating as tight as possible the textual discourse structures to the intonational-prosodic discourse phrasing.

There is a common platform, based on fundamental studies and results in the last two decades, which established a reliable association between intonational phrasing and meaning of IS notions, viz. Background-Kontrast entities (alias Topic-Focus in the Prague School’s language) and Theme-Rheme structures. Papers such as [4], [21], [28], [31], [32], [33] support and prove, including at experimental and perception level, consistent rules that assign categorical functions of tones and tunes (boundary tones, pitch accents, and contours) to the IS textual entities and spans. Of course, a human-like intonational phrasing for Romanian involves careful observing and adapting of the corresponding (sequence of) ToBI tag levels. The accuracy of the association between IS categories / structures and the corresponding (sequences of) ToBI tags for Romanian is not the purpose of our paper. Our main concern is on the textual IS-semantics side of the language interface within the (Romanian) proseody, with the aim of squeezing all the discursive text meanings which proved to be useful for predicting a human-like prosody.

The novelties we concentrate on in this paper can be summarized as follows: (i) Two lines of investigation are designed for detaching the IS categories and structures. (ii) The Topic-Focus Articulation (TFA) algorithm applied for Romanian [8], [10] receives an improved shape at clause-level, being then extended to complex, multi-clause sentences. In the Prague School’s approach, TFA assigns Focus to the intonationally stressed entities and Topic otherwise, while the Theme-Rheme clause-level spans are functions (degrees) of the communicative dynamism (CD). (iii) The second approach we support here for the intonational focus assignment is based on an intensive and combined use of discursive theories and algorithms for obtaining the textual IS-structures: Asher’s Segmented Discourse Representation Theory (SDRT) [2] is applied to control the evolution of the Background-Kontrast (Background-Focus or Topic-Focus) entities along and within (finite) clauses, while the Theme-Rheme structures are computed with a recursive version of Leong’s Inference-Boundary (IB) algorithm [24], [25], for the first-time adapted
and implemented for Romanian (to our knowledge). (iv) In [8] and [10] we emphasized the general view of the syntax-prosody interface as a complex operator between syntactic and intonational structures, subsumption-preserving. Now we refine this view on the textual-intonational discourse sub-interface for various relation-based discourse semantics, supporting the sub-clausal IS-semantics discourse relations introduced by Heusinger [21] as a good premise for an Information Structure Discourse Theory (ISDT) definition.

Sections 2 and 3 present the two lines of investigating the prosodic focus assignment for the Romanian text. These two approaches have as purpose to obtain the information-structure, Theme-Rheme, Background-Kontrast, and the corresponding intonational phrasing. The final Section 4 compares the outcomes of the two approaches with a set of manually annotated Romanian sentences. The results point out the better performance of the TFA algorithm for estimating the Background-Focus entities (in the former approach), and of the IB-algorithm for Theme-Rheme structure computing (in the latter one).

II. IMPROVED LOCAL AND EXTENDED GLOBAL TFA ALGORITHM

Prague’s School Topic-Focus Articulation (TFA) algorithm [17] attributes Topic-Focus to a sentence, considering its constituent order in communicative dynamism vs. the “standard” systemic order, in the frame of contextual boundness or non-boundness. The TFA algorithm applies 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. Examples of specificity degrees for temporal complementation are: general (“niciodată”, “mereu”), indexical (“astăzi?”, “anul acesta”), specific (“22 iunie”, “într-o frumoasă zi de mai”). The TFA algorithm uses semantic specificity degrees only for verbs and temporal / locative complements. The context of the sentence is used through some simple heuristics, focusing on whether the head of a noun group is defined or not, or whether the verb or the complementation are indexical references. The output of the TFA algorithm is determining the appurtenance of an element to a clause’s Topic or Focus. The TFA algorithm was adapted and implemented for Romanian prosodic structures in [8], [10]. The TFA algorithm was adapted and implemented for Romanian prosodic structures in [8], [10]. Examples 1 and 2 present sentences annotated with the TFA algorithm, where T stands for Topic, F for Focus, and T/F for ambiguous cases:

Example 1. [Ion] ț [a câștigat]; [competiția] TR.


Applying the TFA algorithm, a couple of issues were observed. First of all, the context considered for the boundness constraints is minimal, considering only the current sentence and the inflected form of the words, although Hajicova et al. [17] mentioned that the previous sentence is needed in order to properly analyse a verb context.

“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]).

Another problem is due to the fact that, 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 SAAR (Sentence Accent Assignment Rule) [3], [16].

In order to address these two issues, we propose a new version of the TFA algorithm, where the syntactic and semantic information required by the original algorithm (if the noun group is defined or not, if the temporal complement or the verb is indexical or specific) is completed with information structure annotation, obtained by discursive analysis. The starting point of our approach is the fact that 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.

We present below the rewritten TFA algorithm [8], [10], underlying the modifications. The notations we used are: S is the set containing the Verbal Group (VG) and all its complements from a sentence s; f denote a set containing focused constituents, t - the topic, and t/F denote ambiguous topic-focus elements: S(i) denotes the i-th complement in the surface order. verb(S) is the VG, and last(S) is the last element in the surface word order.

1. if (last(S)=verb(S))
   1.1. if (subject(S) is background)
       t=t∪{subject(S)}; f=f∪{verb(S)};
   1.2. else if (subject(S) is kontrast)
       t=t∪{verb(S)}; f=f∪{subject(S)};
   1.3. for other complementation, goto 3.2.
2. else
   2.1. if (verb(S) is general)
      /*general verbs are be, have, become, happen, etc.*/
      t=f∪{verb(S)}; else if (verb(S) is background)
      t/f=t/f∪{verb(S)}; else
      f=f∪{verb(S)};
   2.2. for (i=0 to pos(verb(S))-1)
      /*for all complements in front of the verb*/
      if (S(i) is kontrast subject or kontrast temporal complement)
      t/f=t/f∪{S(i)}; else
      t=t∪{S(i)};
   2.3. /*to the right of the verb*/
      i. if ((pos(last(S))=pos(verb(S))+1) and (last(S) is background))
      /*the verb has only one complement to the right, which is a given entity, a
      definite noun or a personal pronoun*/
      t/f=t/f∪{last(S)}; ii. if (last(S) is kontrast temp or loc)
      f=f∪{last(S)}; else if (last(S) is background temp or loc)
      t=t∪{last(S)};
   iii. let (S(i),S(j)) be the rightmost pair that
      fails to follow SO
In its present form, the extended TFA algorithm on example 4a is:

Example 4a. [Maria]_{T} [nu a fost]_{T} [în acea zi]_{T}.

Example 5a. [Ion]_{T} [a câştigat]_{T} [o competiţie]_{T}.

In example 3a, the sentence contains an t/f, which will simply be turned into an f, resulting:

Example 3b. [Vinul]_{T} [nu avea]_{T} [culoarea obişnuită]_{T}.

Example 4a has no t/f tag to rely on. In this case, step 4.2. demands the appliance of SAAR rules. In deciding the focus based on the SAAR-derived rules, we mark first all the arguments and the adjunct in the clause, and then place the focus on the (rightmost, if more) adjunct in the clause. If there are no adjuncts, the (rightmost, if more) argument receives the focus. If the sentence contains neither adjuncts nor arguments, the verb will be focused. The result of applying step 4.2 of the extended TFA algorithm on example 4a is:

Example 4b. [Maria]_{T, ARG} [nu a fost]_{T} [în acea zi]_{T, ADV}.

Example 5a activates step 5 from the TFA algorithm. Since both the verb “a câştigat” and the argument “o competiţie” are marked as focus, applying SAAR rules decide that the argument will be focused, rather then the verb, the latter receiving a T/F value.

Example 5b. [Ion]_{T} [a câştigat]_{T,F} [o competiţie]_{T}.

Hajicova et al. [16] states that “In its present form, however, the algorithm has several limitations. It can process only simple sentences.”. Thus, a further step in the development of the TFA algorithm for Romanian was the extension of the TFA algorithm to complex clauses. A simple and effective method is splitting the complex clause into simple clauses, and then applying the algorithm recursively on each clause. In correlated adjacent clauses, the algorithm is simply applied consecutively for each clause. More attention needs to be assessed in applying the algorithm on subordinate clauses, since the subordinate clause is to be treated as the corresponding complement of the verb, completing thus the regent sentence. For example, the complex clause in example 6a contains two simple clauses, marked by 1/ and 2/:

Example 6a. [Ion]_{T} [ştia]_{1} / că [[întârziase]_{1} [mult]_{2}].

This example can be transformed into a simple clause, by reducing the completive clause to a direct argument:

Example 6b. [Ion]_{T} [ştia]_{1} [asta]_{1} / [ceea ce a descoperit]_{2} / [în faţa întregii audienţei]_{1, 2}.

This compression of the complex clauses into mere complements allow for the appliance of the TFA algorithm on complex sentences. The benefits of this approach become evident if more than one subordinate clause is considered, as in examples 7.

Example 7a. [Ion]_{T} [a anunţat]_{1} / [ceea ce a descoperit]_{2} / [în faţa întregii audienţei]_{1, 2}.

The complex clause in example 7a is formed by two simple clause, but one of them is intercalated into the second (the clause “ceea ce a descoperit” is embedded into the second clause “Ion a anunţat * în faţa întregii audienţei”). If we consider the embedded completive clause a simple complement, the TFA algorithm will yield the result presented in 7b.
**Exemple 7b.** [Ion]_T [a anunțat]_F [ceva]_TF [în fața întregii audienței].

Analyzing the sentence in example 7b, we notice that we have two constituents marked with F, so step 5 in the TFA algorithm will activate, transforming the verb’s focus into T/F.

**Exemple 7c.** [Ion]_T [a anunțat]_TF^1 [cea ce a descoperit]_TF^2 [în fața întregii audienței].

The examples above show that the improved and extended version of the TFA algorithm works on both simple and complex clauses / sentences.

**Theme-Rheme assignment on TFA structures**

A different layer of annotation needed in order to decide the prosodic structure of a clause is the Theme-Rheme disjunction [31]. Usually, the Theme-Rheme structures are contiguous, the theme appearing at the beginning of the sentence, as that part of the sentence structure which is being presented by the speaker as readily available in the hearer’s memory, while the rheme “develops” the theme within a sentence (clause), as what is being asserted about the theme. However, sometimes the rheme’s contiguous structure is broken by a disorder in the surface representation of the systemic order (SO) [18], i.e. a different communicative dynamism (CD) than the standard systemic order, when the speaker intend to place the intonational center differently than on the last element of the sentence (in declarative clauses).

We developed a set of simple Theme-Rheme assignment rules that apply on the structures marked with the TFA algorithm. The rules, based on the communicative dynamism, are:

(i) The constituents before the verb form the theme.

(ii) The constituents placed after the verb form the rheme.

(iii) The verb is included in the theme if it is marked with topic (t), in the rheme otherwise.

(iv) 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 5c presents the application of the Theme-Rheme assignment algorithm, where the constituent placed before the verb represents the theme, and the rheme contains the verb (marked with T/F) and the following argument.

**Exemple 5c.** [[Ion]_T]_theme [a câștigat]_TF [o competiție]_TF rheme-

A more complex situation is presented in example 8a, where the sentence is marked with the output of steps 1-3 of the TFA algorithm.

**Exemple 8a.** [Ion]_T [a câștigat]_TF [la competiție]_TF [o cupă].

Since steps 1-3 of the algorithm result in two focuses, a further refinement is needed, so step 5 will be applied, yielding the final TFA distribution:

**Exemple 8b.** [Ion]_T [a câștigat]_TF [la competiție]_TF [o cupă].

Following the Theme-Rheme assignment rules for the example 8b, the structures will be:

**Exemple 8c.** [[Ion]_T]_theme [a câștigat]_TF rheme [la competiție]_TF rheme [o cupă] rheme.

III. **DISCOURSE THEORIES COMPETING TO SQUEEZE HUMAN-LIKE TUNES AND TONES FROM TEXT STYLING**

The key problem for efficiently modeling the syntax-prosody interface is to come more insightful on the sub-interface: textual discourse – intonational discourse. On each side of this interface map there are important “white spots” which need to be explored. Such a not enough known realm, of special interest for our problem, should be an Information Structure Discourse Theory (ISDT), compatible and exploiting all the types of textual and spoken discourse semantics that could be productive for an efficient IS semantics.

Continuing to be inspired by the Prague School’s TFA [17], [18], in this paper we bet on two ideas: (a) Asher’s SDRT [2], enhanced from Kamp’s DRT [22], is endowed with both the discursive evolution of reference entities and relations among discourse structures. This makes SDRT a “healthy” environment for an ISDT development, thus for computing the Background-Kontrast entities and Theme-Rheme structures. (b) Among some important discursive IS semantics (Prague School’s TFA [17], [18], M. Steedman’s CCG [31], [32], [33], RST-oriented approaches [27], [35], Van Valin’s RRG [36], as well as [3], [12], [16] etc.), we chose the sub-clausal IS discourse relations proposed by Heusinger in [21] as being equally suitable to SDRT and IS semantically promising. Subsection 3.4 contains a brief description of the sub-clausal IS relations supported in [21].

A. **Classical and Novel Discourse Theories**

Various discourse relations (for those discourse theories oriented on relational-based discourse-semantics) are trying to model distinct discursive meanings of textual spans. Of special interest for our discussion is whether the involved discourse relations are acting not only on inter-clause level, but also at intra-clause textual structures, since the intonational and prosodic functions are working particularly on sub-clausal structures. We remind here just some of the well-known theories, with their types of discourse semantics, ontological trace(s), and possible subsumption-based relationships between them.

(D1) The (classical) predicational semantics may be seen as the discursive level of the classical syntax, closely related to the logical form and semantic roles (abbreviated PRED_Dsem). The second-order predicational relations, either of syntactic flavor (Subjective clause, Predicative clause, Direct Completive clause, Time Completive clause, Location Completive clause etc.), or of semantic flavor (Agentive clause, Direct Object clause, Temporal clause etc.), come from the clause-level (first-order and finite) predicational semantics, which are also projected at the sub-clausal level on non-finite
clauses whose semantic phrase head is a Noun, (non-finite) Verb, or Adjective bearing the predicational feature [17], [9].

(D2) The well-known Rhetorical Structure Theory (RST) [26], [27] points out those textual spans rhetorically related, called RS segments and built on finite-clause type units. What we proved since 2005 [7] is that RST acts equally on sub-clause structure level, a fact that is important for the compatibility of RST and IS-discourse relations, on one hand, and for exploiting the RS markers as intonational focus particles, on the other. The rhetorical discourse semantics is abbreviated as RST<sub>DSem</sub>. It is clear that PRED<sub>DSem</sub> subsumes (<sub>sub</sub>) RST<sub>DSem</sub>. The subsumption relations still hold for sub-clausal structures [7].

(D3) Asher’s SDRT, while inheriting the control of inter-clause referential variables (and lexical chains), which we use to compute Background-Kontrast entities as in subsection 3.2, borrows discourse relations from RST and classical predicational semantics, inclusive at sub-clause level. SDRT is the natural environment for the development of IS-discourse relations in [21].

(D4) In the SDRT setting, Heusinger [21] introduces five IS-discourse relations to be applied to intonational-prosodic phrasing and hands them down at sub-clause level. They are briefly described in subsection 3.4.

The subsumption relations between different discourse semantics could be conjectured as follows:

\[
PRED_{DSem} <_{sub} RST_{DSem} \text{ (high probability)}; \quad SDRT_{DSem} <_{sub} \text{IS}_{DSem} \text{ (probable)}; \quad PRED_{DSem} <_{sub} \text{IS}_{DSem} \text{ (highly probable)}; \quad RST_{DSem} <_{sub} \text{IS}_{DSem} \text{ (very probable)}; \quad RST_{DSem} <_{sub} \text{IS}_{DSem} \text{ (probable, on certain classes of common structures)}.\]

Actually, for all these subsumptions, the structure classes on which they are evaluated should be specified, e.g. clausal, and/or subclausal, discursive semantics type, predicational semantics, intonational-prosodic structure level etc.

B. Background-Kontrast Entities Acquired by SDRT Means

The examples discussed in this sub-section are application of the SDRT analysis. We consider the following necessary sets for variable evolution control:

- \{x, y, z, p, q, r, s, t, u, v, w, k, l, m, n, o, ...\} - the set of free variables.

- \{a, b, c, d, e, f, g, ...\} is the set of constants.

### Example 1

\[
\begin{align*}
x, y & \quad \text{Winston (k)} \quad \text{il (y)} \quad \text{urămi (x,y)} \quad \text{FrS} \quad \text{cu_coada_ochiului (a)} \\
\end{align*}
\]

Example 1. [Winston]<sub>k</sub>, [il urămi]<sub>k</sub>, [cu coada ochiului]<sub>k</sub>

In the given examples, the equalities indicate the coreference chains, thus the backgrounded elements. The other variables or constants correspond to the Kontrast (Focus).

Section 4 will discuss the assignment of Background-Kontrast using SDRT compared with the gold annotated sentences.

### Example 2

\[
\begin{align*}
\end{align*}
\]

### Example 3

\[
\begin{align*}
\end{align*}
\]

On the SDRT representations in examples 1 – 3, we also marked Heusinger’s IS-discourse relations [21], outlined in § 3.4. In the analyzed examples, we could identify the IS-relations of non-restrictive Modification (Mod), Enumeration (Enum), Frame-Setting (FrS) and Backgrounding (Bkg), but not the Thematisation (Thm) relation. The explanation we claim is that the first four IS-relations are deductible from the
predicational-discursive structures (semantics) of the text, among intra- or inter-clausal structures or entities. Moreover, the relations Mod, Enum, and FvS are of predicational-semantics nature, while the Bkg relation is of discursive nature, pointing out the Given-New (contextually bound or non-bound) entities.

The special Thm relation, with the special IS role of acquiring the Theme-Rheme structures and their intonational-prosodic boundaries and contours, does not have an explicit correspondent into the predicational-discursive semantics, but only within the IS-semantics that is responsible for the intonation-prosody. Computing the Thm relation, viz. Theme-Rheme structures, with the IB-algorithm should be the solution to complete the IS-specific and effective difference between the syntactic-discursive structures that strictly subsumes the intonational-prosodic ones (see [8], [10]).

C. Theme-Rheme Computing with Inference-Boundary (IB) Algorithm

P. A. Leong [24], [25] provides a comprehensive analysis of the informational-structure (IS) concepts of Theme-Rheme, relying mainly on the Halliday’s theory [19] that the Theme is a clause-initial element (and the constraining one) for the development of the clauseal message. The proposed Inference-Boundary (IB) model and algorithm [24], [25] shows further that Theme is capable of generating a boundary of acceptability for delimiting the Theme and Rheme spans (portions) of the clause. Behind the Leong’s IB-model stands the principle of Acceptable Message Development (AMD), which states that the Theme (or thematic head) of a clause must be acceptably developed by the Rheme, in the context of how the clause message can progress after the Theme.

Our interest concentrates on the relationship between Background-Kontrast, assimilated to Given-New in the discourse-oriented IS-theories [5], [16], [21], [32] and as Topic-Focus in the Prague School’s TFA model and algorithm [17], [18], on one hand, and Theme-Rheme theories, on the other hand. This is because Background-Kontrast (or Topic-Focus) are associated to pitch accents (and sequences of ToBI tags) within Theme-Rheme boundary tones, simple arguments receive certain boundary tones (and sequences of ToBI tags), while to Theme-Rheme contours are assigned other specific sequences of ToBI tags [31], [33], [28], [4], [5]. Halliday [19] defines Theme as what the speaker choose to take as the departure point in a clause (actually, in a predication [9]), while Rheme is what the hearer already knows or has access to. What IB-model and algorithm [24], [25] bring further is to consider the clause thematic element as having the constraining force to determine the boundary within which the Rheme can occur and the clauseal message is developed.

We propose here an extended, recursive form of the IB-algorithm for computing the Theme-Rheme portions within the finite and non-finite clause, adapted to Romanian. To our knowledge, it is for the first time a version of the IB-algorithm [24], [25] is applied to Romanian, for computing the Theme-Rheme clause-level spans (portions).

Recursive Inference-Boundary (IB) Algorithm for Theme-Rheme identification in finite clause

The algorithm will receive as input the following analyses of the sentence: (a) Segmentation of the finite clauses in the sentence; the non-finite clauses (if any) are recognized within the finite clauses; (b) A constituent parsing of the elements in the clauses; (c) The syntactic roles of the syntactic constituents; (d) The systemic order (SO) for each predication (e.g. [17], [8]).

The running of the algorithm is dependent on a number of sets, given below:

A = {Continuatives (such as exclamations, interjections, etc.); e.g. hai, hei, vai, salut}  
B = {Conjunctives; e.g. și, sau, dar, punctuation marks, etc.}  
C = {Conjunctive-adjuncts; e.g. deși, cu toate că, deci, din cauză că, Pred Occur, i.e. predication occurrence marker [7], [9], etc.}  
D = {Wh-relatives; e.g. care, pe al cărei, dintre cei cărora, etc.}  
E = {Vocatives; e.g. Ion!, Maria!}  
F = {Modal adjuncts, mood or comment adjuncts; e.g. probabil, desigur, posibil, etc.}  
G = {Finite operators; viz. modal auxiliaries (e.g. a trebui, a putea), “a fi” auxiliary, “a avea” auxiliary, semi-auxiliaries (e.g. a vrea, a voii)}  
H = {Wh-question words; e.g. cine, ce, cui, unde, cum, de ce, etc.}  
I_{part} = {NGs; viz. participant Top_Th}  
J_{circ} = {PPs, NGs; viz. circumstance Top_Th}  
K_{proc} = {VGs; viz. process Top_Th, if VG = predication}

For all FinCl (Finite Clauses) in the sentence  
If Mood(FinCl) = Declarative then for all constituents w_i(1...s) in FinCl  
//Textual Theme  
If w_i in A or B or C then w_i = Tex_Th  
//Topical Theme  
If w_i in D then w_i = Text_Th or w_i = Top_Th  
//Interpersonal Theme  
If w_i in E or F or G then w_i = Int_Th  
//Non-Finite/Interrogative then  
If w_i in H w_i = Int_Th or w_i = Top_Th  
If w_i in I_{part} or J_{circ} or K_{proc} then w_i = Top_Th, endfor

If a Non-Finite clause is embedded in the current FinCl, it is parsed the same way as a FinCl;  
If several Tex_Th or Int_Th are in sequence they are enclosed into the same Theme set;  
If Mood(FinCl) = Interrogative then  
If FinCl = Polar_Interr then Top_Th = Subject(FinCl)  
If FinCl = Content Interr then Int_Th = Wh_Word(FinCl), Top_Th = Subject(FinCl)  
If Mood(FinCl) = Imperative then  
If FinCl = Inclusive then Top_Th = Subject(FinCl); swap the predicate with the subject for computing Theme and Rheme;  
If FinCl = Exclusive then Top_Th = Subject(FinCl)

If w_i = Top_Th then TH = Tex_Th  
If w_i = Tex_Th then [w_i ... w_s] = Tex_Th  
If w_i = Tex_Th, i = 1,p then [w_i ... w_s] = Tex_Th
The algorithm first separates the topical Theme from the rest of the clause. This separation is based on the fact that the Theme is always at the beginning of the discourse commenced by the clause; the rest of the clause is composed of Rheme or Pseudo-Rheme segments. Each argument and adjunct is a separate segment (with the exception of the argument or adjunct which is directly adjacent to a transitive verb), and they are separated according to their corresponding systemic order (SO).

Below are a number of examples of Theme-Rheme separation for a series of Romanian sentences. The finite clause is marked by //.

**Remarks concerning the IB-algorithm (for Romanian)**

1. The **golden rule** for detecting the topical themes within a sentence is to segment the sentence into finite clauses, and to start searching at the beginning of each clause until one finds the topical theme (Top-TH) of that clause. The thematic portion of the clause has to enclose, obligatorily, a topical theme and, optionally, the textual (Tex-TH) and/or interpersonal (Int-TH) theme(s). The right-hand part of the topical portion constitutes the rhematic portion (segment) of the clause, until the clause is up.

2. Whether no overt (lexically present) Top-TH is obtained, then a covert one (such as the empty grammatical subject or zero anaphora) has to play the Top-TH role.

3. Of special importance are the clause-markers in the working sets above (including the covert predicational feature [9] in set C, occurring within the single-predicate clauses below), D (Wh-relatives) and H (Wh-Questions). The two clauses in the example below bear covert topical themes.


4. It is possible that a **thematic portion** of a clause to contain *more than one* topical theme, provided that all the thematical themes in the thematic portion are marked, thus our the Systemic Order (SO) [18], [8] of the involved predication. In the Example 4 below, the complete thematic portion is composed of *three* individual Top-TH(s) related among them by the IS-discourse relation called enumeration in [21].

**Example 4a.** [Azi,]Top-TH [pe stradă,]Top-TH2 [într-un copac.]Top-TH3 mușnea jumătate un pisoil. Furthermore:

**Example 4b.** [Marie,]Top-TH1 [mașina,]Top-TH2 [Ion]Top-TH3 i-a dat-o.

5. Naturally, a thematic portion can stretch over all the three types of themes:

**Example 5.** [În fine,]Tex-TH [Ion,]Int-TH [de ce]Int-TH [tu = Ø]Top-TH [n-ai venit ieri la școală?]RH

6. A finite clause may have several topical themes (and accordingly, rhemes), each one corresponding to the enclosed complements that are non-finite clauses. Furthermore, some non-finite clauses may bear not a topical theme, viz.


**Example 6b.** Interpretarea [lui]Top-TH impresionând tot publicul.

7. English and Romanian may differ substantially on the non-marked (SO-ordered) or marked (non-SO, thus CD-ordered) Top-TH(s) of different phrases (groups) in the clause.

**Example 7a.Eng.** [Did]Int-TH [John]Top-TH [recognize his guilt?]RH

**Example 7a.Rom.** [Și-a recunoscut]Top-TH [Ion vina?]RH

**Example 7b.Rom.** [Ion]Top-TH [și-a recunoscut vina?]RH

**D. Sub-clausal IS-semantics Discourse Relations**

Heusinger's proposal [21] for an IS-semantics Discourse Theory (ISDT) relies on SDRT inter-clausal evolution of discourse variables for computing the Background-Kontrast. While maintaining the classical SDRT (including rhetorical) discourse relations at the inter-clause level, [21] introduces a set of IS-semantics *intra-clausal* relations, also handing down at sub-clausal level the rhetorical relations [26]. RST, SDRT relations at sub-clause and higher-clause level [7], [26], along with focus sensitive particles and expressions, e.g. only, just, merely, also, too, even, specifically etc., negation, comparative constructions, counterfactual conditions, generalized quantifiers, special adverbs etc. [3], [12], have a significant role in intonational-prosodic phrasing. All these textual and prosodic markers need to be integrated into a coherent ISDT of the textual and spoken discourse, which the core proposal in [21] should be naturally integrated in.

Since intonational phrasing is often a prosodic function on sub-clause units, Heusinger [21] introduces five discourse relations aiming to complete the relational semantics of IS-
discourse. These relations are rather exemplified than formally defined.

(IS-R1) Non-restrictive Modification is a typical modifier (or relative) phrase relation.

Modif_Ex. ...[domenial] Modif \([\text{căruia i-am dedicat tot efortul}]\)

The relative pronoun \(\text{căruia}\) introduces (in cross-agreement for Romanian) the relative clause that follows, which is a modifier (as a whole) of the noun domainul.

(IS-R2) Enumeration relates two or several conjuncts as sub-clause phrases. Logical conjunction raising at predicative (or predicational) level shows that enumeration may link, order-free or order-dependent, two or several phrases on all the levels of syntactic organization.

Enum_Ex1. \([\text{Copiii}}\text{Enum l., femeile}}\text{Enum şi bătrânilii}] [\text{s-au refugiat din calea apelor.}]\)

Enum_Ex2. \([\text{Cei doi}}\text{Enum [au dansat]}\text{Enum l., s-au împrietenit]}\text{Enum şi s-au căsătorit]} [\text{după doi ani.}]\)

(IS-R3) Frame-Setting. This is a (newly introduced) relation whose purpose is to emphasize the role of circumstantial complements (and completive clauses) in the sentence prosody.

FrSet_Ex. \([\text{De două săptămâni}}\text{Veda [Ion n-a mai trecut pe la facultate.]}\]

The temporal complement of \(\text{de două săptămâni}\) is not the (proper) topical theme, which is Ion. According to the SAAR rules in Section 2, this temporal complement is intonationally focused. The following two IS-relations are the most specific to IS-semantics.

(IS-R4) Backgrounding. This relation is a generalization of the Background concept in SDRT, viz. a relationship established between defined and referred categories and phrases. Co-occurrence, instantiation, or anaphora are particular cases of Backgrounding.

BackGr_Ex1. \([\text{Profesori şi elevi ai liceului, cum ar fi G. Ihrăleanu, E. Lovinescu, R. Beligan etc.,] Background]\ [\text{au devenit mari personalităţi ale culturii române.}]\)

BackGr_Ex2. \([\text{Dinspre Ion}}\text{Background l., un copil mărunţi şi emotiv]} [\text{, veneau sunete minunate.}]\)

In the examples above, Backgrounding relates the first two spans in square brackets.

(IS-R5) Topicalization (or, better, Thematization). This relation should be associated with the IS concepts of Theme-Rheme, while Backgrounding would involve the IS-notions of Background-Kontrast.

ThRel_Ex. \([\text{Mănâile lor}}\text{ThRel [erau pline de bătături, înălţaţii mirosind a pământ proaspăt.]}\]

Thematization relates the square bracketed spans 1-2 and 1-3. For the Theme-Rheme computing and intonational behavior, see subsection 3.3. It should be natural that these IS-discourse relations to be not the only ones useful to describe the IS-semantics. There are many other focus sensitive entities and structures whose IS roles and intonational-prosody functions are waiting to be revealed within a comprehensive ISDT.

If the relation “\(\text{k} \rightarrow \text{m}\)" should be interpreted as “is entailed from” or “has the origin in", we estimate the following connections for IS\_Dsem relations in [21] to hold: (a) Unrestricted-Modification \(\rightarrow\) PRED\_Dsem; (b) Enumeration \(\rightarrow\) PRED\_Dsem; (c) Frame-Setting \(\rightarrow\) SAAR \(\leftrightarrow\) PRED\_Dsem; (d) Backgrounding \(\leftrightarrow\) IS\_Dsem \(\leftrightarrow\) SDRT\_Dsem; (e) Thematization \(\leftrightarrow\) IS\_Dsem.

IV. COMPARISON AMONG RESULTS AND A GOLD SET

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 [1]. The first set of examples compares the Background-Kontrast (Topic-Focus) assignment using the two different methods: TFA and SDRT.

TFA Topic – Focus (Background – Kontrast)

Winston\([\text{il urmări}}\text{[cu coada ochiului,}\text{TF}]

O’Brien\([\text{apucă}}\text{[sticla}}\text{[de găt}]\text{TF} [\text{şi turnă}]\text{TF} [\text{în pahare}]\text{TF [un lichid de un roşu aprins.]}\text{TF}]

[I] \text{Lui Winston\([\text{[i se treziră}}\text{[nişte vagi amintiri,}]\text{TF} / [\text{ceva}}\text{[ce văzuse}}\text{TF [demult]}\text{TF [– o sticlă mare făcută din luminii electrice]TF [care]}\text{TF [parcă se mişca]TF [în sus şi în jos]TF [turnânduşi-şi conţinutul într-un pahar.]}\text{TF}]

SDRT Background – Kontrast

Winston\([\text{ki urmări}}\text{[cu coada ochiului,}]\text{K}

O’Brien\([\text{apucă}}\text{[sticla}}\text{[de găt] K [şi]}\text{K [turnă] K [în pahare]}\text{[un lichid de un roşu aprins.]}\text{K}

[I] \text{Lui Winston\[\text{[i se treziră}}\text{[nişte vagi amintiri,}]\text{K} / [\text{ceva}}\text{[ce văzuse}}\text{K [demult]}\text{K [– o sticlă mare făcută din luminii electrice]K [care]}\text{K [parcă se mişca]K [în sus şi în jos]K [turnânduşi-şi conţinutul într-un pahar.]}\text{K]

Gold Background – Kontrast

Winston\([\text{[il urmări}}\text{[cu coada ochiului,} Hi-L,\text{TF}]

O’Brien\([\text{apucă}}\text{Hi [sticla}}\text{Hi [de găt] Hi [şi]}\text{Hi [turnă] Hi [în pahare] Hi [un lichid Hi [de un roşu Hi aprins,]}\text{Hi-L,}\text{TF}]

[I] \text{Lui Winston\text{[i se treziră}}\text{[nişte vagi amintiri,}]\text{[ceva}}\text{[ce văzuse}}\text{[demult]}\text{[– o sticlă mare] [făcută] [din luminii electrice]K [care]}\text{K [parcă se mişca]K [în sus şi în jos]K [turnânduşi-şi conţinutul într-un pahar]}\text{K,TF]

Analyzing the examples, we observed that the manual annotation allows for multiple high pitch accents (H tones) into a clause, as the SDRT-based approach does, while the TFA algorithm restricts to only one focus in each clause. However, the SDRT-based approach distinguishes many empty elements as Background, while marking most of the lexicalized constituents as Kontrast (Focus).

Another difference between the three methods resides in the way they treat the verbs. Thus, in the TFA, the verb is mostly marked as \(f \) or \( ttf \); in the SDRT-based method it is mostly Kontrast, since not referred in the previous sentence; in the
gold annotation, the verb forms an unity 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 (being the rightmost constituent is one of the major reason for assigning it the focus value); 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.

The second set of examples compares the Theme-Rheme annotation:

TFA Theme - Rheme

[Winston]theme [îl urmări cu coada ochiului.]rheme

IB Theme - Rheme

See examples IB1, IB4 and IB5 in section 3.3.

Gold Theme - Rheme

[Winston îi urmări cu coada ochiului.]theme
[O’Brien apucă sticlă de gâr]theme și turnă în pahare]theme [un lichid de un roșu aprins.]theme 1/

In analyzing the gold Theme-Rheme, we notice that 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, where the Theme marking is minimized. In the Theme-Rheme case, there are much more similarities between the three annotations than in the Background-Kontrast case. Almost all the sentences start, in all three annotations, with Theme and end with Rheme. The differences consist in the marking of the middle constituents, and in establishing the boundaries for the Theme-Rheme portions.

The analysis performed on the two proposed methods, compared to the gold annotation, revealed that the TFA performs better in the topic-focus assignment task, while the IB approach is more reliable for the Theme-Rheme assignment. Further combinations of the methods need to be investigated.

We believe that the prediction of prosody from text can be substantially improved by a deeper understanding of the textual discourse theories, with the natural emphasis on the information structure (IS) discourse semantics, using also the communicative dynamism of the surface representations of a sentence.

REFERENCES
