Syntax-Prosody Interface for Romanian within Information Structure Theories

Neculai Curteanu, Diana Trandabăţ, Mihai Alex Moruz
Overview

- Outline
- Prague Schools Topic Focus Articulation
- Syntax – Prosody Interface as an IS Driven Mapping
- Syntactic Tools: Theory and Parsing
- Topic – Focus Computing for Romanian
- Conclusion
Outline

- Topic-Focus Articulation (TFA) algorithm for text-based prosody design;
- Main challenges:

Discover the relationship between these structures.
Main challenges

- What is the **organization of linguistic markers and structures** within SCD (Segmentation-Cohesion-Dependency) and FDG (Functional Dependency Grammar) parsing frameworks at:
  - **Functional level** – lexical and discourse semantics;
  - **Hierarchical level** – marker class dependencies;
  - **Sequential level** – linear precedence of marker mapping.

- Adapting and *implementing* the TFA *algorithm* for *the first time* to Romanian prosodic structures, to be continued with TFA sentence-level refinements.
Our basic ideas on developing a sound prediction for (Romanian) prosody received a strong support from the Eva Hajičová’s recent speech at her ACL Lifetime Achievement Award;

The initial intuition was to use the RST (Rhetoric Structure Theory) discourse structures for the general layers of global prosody structures of spoken discourse;

For the local, clause-level prosody prediction we grant the intonational approach derived from the intensive use of information structure (IS);

Clause-like IS is crucial for the local discourse structure and coherence, in either written or spoken language.
We propose the following image of a general parsing strategy, in parallel for both text and speech on the syntax-prosody interface:

- a lattice-type organization of textual *marker* (boundary) *classes*, delimiting syntactic structures and establishing dependencies;
- a similar set of marker classes for the *prosody side* of the interface;
- an isomorphic mapping between the two main lattice structures with the property of preserving *subsumption* between the involved (sub)structures.

Discourse prosody is constituted from *prosodic phrase groups*, whose phonologic boundaries (markers) are *prosodic fillers*, represented by prosodic word markers and prosodic phrase-discourse markers (Chiu-yu Tseng *et al.*).
Syntax-Prosody Interface as an IS-Driven Mapping

- Steedman’s CCG (Combinatory Categorial Gramar) observes similar ideas on textual and phonological boundaries (markers), which support the above proposed latticeal mapping on the syntax-prosody interface:

- *Intonation* coincides with surface structure (syntactic boundaries *subsume* the intonational ones);

- The reverse *does not* hold: not all surface syntactic boundaries are *explicitly* marked by intonational boundaries;

- The majority of IS boundaries go unmarked by explicit intonational boundaries;

- *Boundary tones* mark the right-hand side boundary of the *prosodic phrase*, differing from the *syntactic* and *discourse markers*, which bound the left-hand side of a *textual phrase*;

- IS constituents and their semantic interpretation provide the logical forms that discourse semantic functions apply to.
Syntactic Tools: Theory and Parsing

- The Segmentation-Cohesion-Dependency Linguistic Theory and Parsing Strategy;
  - Classes of boundary and dependency markers;
  - Construction of local (clause-level) and global (discourse-level) syntactic structures (predication and rhetorical semantics);
  - SCD segmentation and parsing algorithms defined on the basis of hierarchical (latticeal) dependencies between (and within) marker classes.

- The FDG parsing theory:
  - Clause acquisition and clause dependency establishing;
  - Local (subclause) phrase parsing;
Syntactic SCD-FDG trees

Syntactic Units

TFA

IS-based Syntactic Units

Intonational TFA

IS-based IU

Automatically established ToBI labels

Empirically established IU

FDG-based IU

Empirically established ToBI labels
The purpose of *Topic-Focus Articulation* (TFA) algorithm:

- to assign topic/focus (theme-rheme) features to text spans of a sentence (on the basis of IS elements);

**Input:**

- the IS-free clause syntactic units;
- SCD markers;
- certain (IS-oriented) semantic features:
  - verb specificity
  - nominal definiteness;

**Output:**

- IS-based units, which are actually syntactic constituents with their attached topic-focus values;
Input *(LISP-style functional syntax)*:
verb (topic(f), sem(interm), label (puse), rtree (loc(topic(f), sem(gen), so(10), surf(advp), label (jos)), obj(topic(f), det(1), so(3), surf(np), label (paharul), rtree(attr(golit)))))

Output *(IS-based units (TFA algorithm output), combined with automatically detected Intonational Units cf. the procedure in § 3.3)*:

**Step 1.** All words are marked as being in the focus *(puse(f), jos(f), paharul golit(f))*;

**Step 2.** The verb becomes ambiguous since intermediate *(puse(t/f), jos(f), paharul golit(f))*;

**Step 3.** ”jos” becomes topic and ”paharul golit” becomes focus *(puse(t/f), jos(t), paharul golit(f))*
O’Brien grabbed the bottle by the top and poured in the glasses a liquid of a red bright.

- IP (*Intonational Phrase*) acquisition;
- ip (*intermediate phrase*) acquisition;
- RU (*Rhythmical Unit*) acquisition.

Intonational TFA

- Assigning topic/focus values to the intonational units computed from the FDG tree;

- Cases:
  - The intonational unit is equal to the topic/focus unit;
  - The intonational unit contains several topic/focus units;
  - The intonational unit is contained in the topic/focus unit.
The Analysis Corpus

150 sentences from George Orwell’s “1984”:

- At syntactic level, we added an FDG tree for each sentence, in addition to POS, NG/VG (Noun/Verb Group), and SCD markings;

- At phonological level, for each sentence, ToBI tone labels and phonological entity boundaries were assigned for the spoken text.

Examples of (manual) intonational unit annotation:


Conclusion

- TFA theory and algorithm is a procedure intervening exactly on the core of syntax-prosody interface, for assigning proper intonational stress (ToBI labels) to text spans, adapted and implemented for the first time to Romanian;

- Finding the set(s) of rules by which one could evaluate the focused elements of a spoken utterance is one of the keys of the prosody prediction;

- The contrastive analysis realized on the 150 Romanian sentence corpus for IS-parsing and intonational units’ detection helped us to adapt and use (I)TFA algorithms as the major breakthrough for Romanian prosody prediction;

- As future work, our aim is to extend the (I)TFA algorithm for Romanian to the global (sentence) level, to be continued with TFA clause-level refinements, its rhetorical-level extension, and embedding it into local-global linking algorithms.
Thank you!
*Winston him followed with the back of his eye.

Input:
verb(topic(t/f), sem(interm), label (urmări), rtree (sbj(topic(t), sem(gen), so(1), surf(np), label (Winston)), manner(topic(t(f)), det(1), so(6), surf(nc), label (coada), ltree(prep(cu)) rtree(attr(ochiului))))

Output:
[[[Winston(t), îl urmări(t/f)]_RU, [cu coada ochiului (t(f))]_RU]_ip]_IP)
* The Brotherhood it is called.

**Input:**
verb (topic(f), sem(interm), label (i se spune), ltree (mod(topic(f), det(1), so(6), surf(np), label (Frăția))))

**Output:**
([[frăția(f)]_RU, [i se spune(t)]_RU_ip]_IP)
*It had a smell sweet and sour.

**Input:**
verb (topic(f), sem(gen), label(avea), rtree(obj(topic(f), det(0), so(3), surf(np), label(miros)), ltree(det(un)), rtree(attr(dulce-acrisor)))))

**Output:**
([[[avea(t), un miros]_{RU} [dulce-acrisor(f)]_{RU}_{ip}]_{IP}])
Watched from above, the liquid seemed black.

Input:
verb (topic(f), sem(gen), label(părea negru), ltree(mod(topic(f), so(6), surf(ap), label (privit), rtree(mod(topic(f), so(6), surf(advp), label (sus), ltree(prep(de)))))), act(topic(f), det(1), so(1), surf(np), label (lichidul))))

Output:
([[[privit de sus(t)]_RU, [lichidul(t), părea negru(f)]_RU]_IP]_IP)
Winston lifted the glass to his lips with a certain anxiousness.

**Input:** verb (topic(f), sem(interm), label (îşi duse), ltree (act (topic(f), det(1), so(1), surf(np), label (Winston))), rtree (obj (topic(f), det(1), so(3), surf(np), label (paharul)), loc (topic(f), sem(gen), det(0), so(10), surf(np), label (buze), ltree(prep(la))), mod (topic(f), det(0), so(6), surf(np), label (rabădare), ltree(prep(cu), det(o), det(oarecare)))))

**Output:** ([Winston(t), îşi duse(t/f)]RU, [paharul(t)]RU, [la buze(t)]RU_ip, [[cu o oarecare nerăbdare]RU_ip]IP)