Memoirs of the Scientific Sections of the Romanian Academy
Tome XXXIV, 2011

LEVELS OF FINITE CLAUSE PARSING: IMPLICATIONS AND APPLICATIONS

CRISTINA BUTNARIU, NECULAI CURTEANU and CECILIA BOLEA

Institute of Computer Science, Romanian Academy, Iaşi Branch, Romania
2, Codrescu, 700483, Iasi, Romania
E-mail: cbutnariu@iit.tuiasi.ro, cbolea@iit.tuiasi.ro
Corresponding author: curteanu@iit.tuiasi.ro

The aim of this paper is three-fold: (1) To point out the four types of inter-clause dependencies, thus levels of finite-clause parsing; (2) To discuss the essential role of the finite (and non-finite) clause structure in text processing, discourse parsing, text-to-speech modeling and prosody prediction; (3) To apply the Segmentation-Cohesion-Dependency (SCD) method based on marker classes, their semantic hierarchies, and marker sequences, combined with a machine learning technique, in order to obtain improved results on clause parsing of the Romanian Acquis corpus. A special emphasis of our analysis is devoted to the intonational role that the inter-clause lexical markers and their sequences are playing on the syntax-prosody interface (of Romanian), more precisely, for establishing the discursive-intonational contrastive foci and patterns that are necessary in prosody prediction.

Key words: machine learning clause parsing, sentence segmentation, lexical inter-clause markers, prosody prediction.

1. INTRODUCTION

The finite clause parsing problem needs a careful analysis since its implications, solutions, and applications are considerable for text and speech processing. The syntactic / semantic structure of the finite (and non-finite) clause plays the essential role of a turning point between local and global linguistic structures in all natural languages. This is because the finite clauses, organized further in sentences and paragraphs, are made up of finite predications (predicates or events) surrounded by the semantic roles, some of them recursively embedded in non-finite predications.

This paper is devoted to the task of obtaining the finite-clause structure of a text, with the emphasis on the inter-clause relationships at the specific, *lexical* level. The reason behind this approach is to use the graph-based hierarchies of finite and non-finite clauses as an essential source of semantic information for local-global *prosody prediction*. While the use of clause segmentation and (certain levels of) parsing in text processing is much better established, we want to discuss the inter-clause level of parsing that is necessary for discursive-intonational units in the prosody prediction of

the (Romanian) language. Our task is to relate the finite clauses (and recursively embedded non-finite clauses) by their inter-clausal markers and, when it is the case, to point out the rhetoric and contrastive intonational markers between and inside the finite clauses.

The clause parsing based on the inter-clause lexical markers is supported not only by the intuitive stronger and "multi-valued" semantics of these functional elements but also by the theoretical definition of the *functional generative capacity* for *phrase markers* discussed in [7]. This is a generalization of the *strong generative capacity* introduced and investigated in [18], the functional generative capacity for the inter-clause *lexical* markers opening a large window toward an improved modeling of the syntax-prosody interface, prosody prediction, thus improved taxonomy for natural language parsing, either for text or speech.

We summarize here some of the applications for the task we are dealing with, *viz.* clause-level segmentation (splitting) and parsing:

(I) Text Processing; (I.1) Parallel text alignment; **(I.1.2)** Machine translation; **(I.2)** Anaphora resolution; **(I.3)** Semantic role labeling parsing; **(I.4)** Building discourse structure; **(I.4.1)** Automatic summarization.

(II) Speech Processing; (II.1) Prosody prediction; **(II.1.1)** Rhetoric and contrastive discursive-intonational patterns; **(II.1.2)** Text-To-Speech (TTS) systems.

This hierarchical structure of applications is actually a graph with much more nodes and links.

Section 2 of the paper demonstrates the main types of inter-clause segmentation (splitting) and parsing, and the relationship to the important tasks of language processing, from classical (predicational semantics-based) text parsing, to rhetorical /contrastive discourse structure parsing, and prosody prediction for the intonational phrasing units at the finite-clause level and higher. *Section* 3 exemplifies in more detail these applications of clause segmentation and parsing, with implications for several approaches to text and speech processing, and for the syntax-prosody modeling. *Section* 4 combines the method of SCD (Segmentation-Cohesion-Dependency) configurations, which relies on the careful analysis of the sequences of intra- and inter-clause phrase markers with a machine learning algorithm, to obtain improved solutions

to the task of finite-clause parsing with lexical inter-clause markers. Applications are envisaged on parsing the Romanian Acquis Corpus, a large collection of translated and adapted laws of the European Community legislation. The clause-level parsing of the Acquis Corpus based on inter-clause lexical markers represents an exercise for the true target of the present paper approach: intonational pattern design on textual discourse structure for prosody prediction and TTS systems.

2. LEVELS OF INTER-CLAUSE DEPENDENCY: FROM SEGMENTATION TO LEXICAL MARKER RELATIONSHIP

The next abbreviations and figures have the following meaning: **cm1, cm2, cm3, cm4** are inter-clause markers; **dm1, dm2** are rhetorical discourse markers; **cl1**, **cl2**, **cl3** are finite clauses; **seg1**, **seg2**, **seg3** are discourse segments (RST, see [16], [17]).

The captions for the Figures 1-4 above have the following explanations:

- Figure 1: The *blue-empty triangle* corresponds to the inter-clause markers that are used only for the finite-clause segmentation purpose.
- Figure 2: The *blue-filled triangle* means inter-clause markers bearing a *dependency* functional role, on two levels of specification (see level 2 and 3 of inter-clausal dependency kinds): (a) regent clause *vs.* subordinate clause, and (b) regent clause *vs.* subordinate clause of one of the three types: relative clause, completive-type clause, and modifier and circumstantial-type clause. *E.g.*, the SCD markers in class M3 in [6], [7] are inter-clause markers of kind (a) or (b), *viz.* M3 contains mainly inter-clause / discourse lexical and punctuation markers.
- Figure 3: The *green-filled triangle* represents the inter-clause markers of *lexical* level, and we use "*deşi*" (*although*), "*că*" (*that*), "*dacă-atunci-altfel*" (*co-relational*) etc. inter-clausal markers together with their inter-clausal dependency relations. These types of inter-clause lexical markers are exactly those useful in global-level prosody prediction, to which are added other intonational-focus markers (*e.g.* the classical *only*, *even*, *also*), including the rhetorical discursive ones.

- Figure 4: The *red-empty trapeze* stands for rhetorical discourse lexical markers, delimiting and hierarchically relating RST segments (usually, elementary discourse units – EDU's) [16], [17]. The papers [7] and [17] pointed out that specific rhetoric discourse markers may occur *inside* the finite clause (Figures 7 and 8 below), giving evidence that discourse structures may hand down at the intra-clause level (contradicting the believe that the finite clause is the "minimal" brick when constructing the rhetorical discursive structures). Furthermore, the prosodic discourse, *i.e.* local and global intonational structures resulted from classical syntactic / semantic structures and Information Structure (IS) functions (Background-Focus and Theme-Rheme) is shown to work equally at intra-clause [9], [14] and inter-clause levels [21], [22], [23].

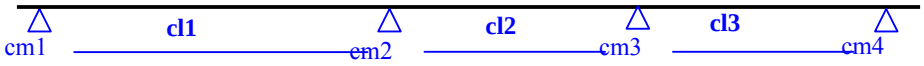

Fig. 5 - Finite-clause segmentation (no inter-clause dependencies)

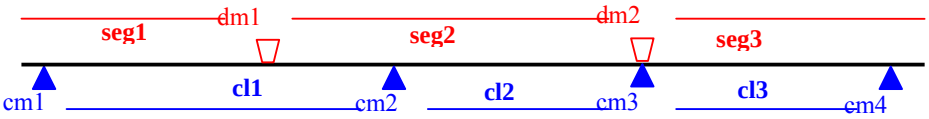

Fig. 6 - Finite-clause parsing on two dependency levels of SCD markers

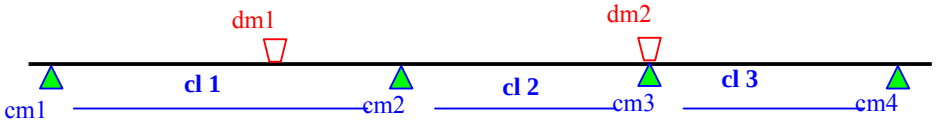

Fig. 7 - Finite-clause parsing based on lexical-marker dependencies

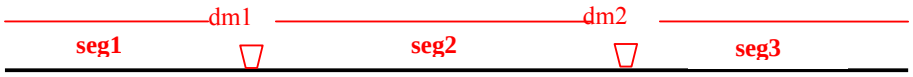

Fig. 8 - Discourse markers for rhetorical (clause-like) parsing

Numerous examples supporting the distribution of clause and discourse markers may be found in papers such as [3], [5], [7], [10], [11], 17], [20], etc.

3. CLAUSE STRUCTURE AND ITS ROLE IN TEXT AND SPEECH PROCESSING

This section proposes brief outlines on some papers dealing with clause / discourse segmentation or parsing of the types mentioned above, aiming to be significant for our future task, freely of the involved language, be it Romanian, English, German etc. We notice that there are not many papers in computational linguistics whose task is purposely devoted to obtaining the finite-clauses and their *dependency trees*, compared to the number of the papers where the clause segmentation (splitting) or clause parsing (dependency trees) are computed just as a necessary step to unrestricted text (predicational) parsing or (rhetorical) discourse structure parsing. The task of clause parsing, despite the fact that has numerous applications, has been "surprisingly neglected by researchers", as C. Orăşan remarks in [19].

The early paper [4] is dealing with the *first parsing algorithms* for the Romanian clause and sentence, based on two interleaving Top-Down algorithms. The following excerpt is illustrative:

"**Exemplul 7.**

*/ Pesemne [1]/ (**că**) de atâţia ani [2]/ (**de când**) trăia bine (,) [3]/ Aristide ajunsese [2]/ (**să**) creadă [4]/ (**că**) şi alţii o duceau tot aşa [5]/ (**şi**) (**că**) nu le pria două mese una după alta (,) [6]/ (**cum**) nu i-ar fi priit lui (.) [7]/ "*

Arbore corect:

```
(FRAZA
    (RESTPROP NIL) (COORD NIL) (DEL NIL) (PROP P1)
        (  (RESTPROP T) (COORD NIL) (DEL CĂ) (PROP REST)
            (  (RESTPROP NIL) (COORD NIL) (DEL DE CÂND) (PROP P3)  )
              (RESTPROP NIL) (COORD NIL) (DEL CĂ) (PROP P2)
                  (  (RESTPROP NIL) (COORD NIL) (DEL SĂ) (PROP P4)
                      (  (RESTPROP NIL) (COORD NIL) (DEL CĂ) (PROP P5)
                        (RESTPROP NIL) (COORD ŞI) (DEL CĂ) (PROP P6)
                            (  (RESTPROP NIL) (COORD NIL) (DEL CUM)
(PROP P7)  )  )  )  )  )
```

Arbore alternativ :

```
(FRAZA
    (RESTPROP NIL) (COORD NIL) (DEL NIL) (PROP P1)
        (  (RESTPROP T) (COORD NIL) (DEL CĂ) (PROP REST)
```

```
( (RESTPROP NIL) (COORD NIL) (DEL DE CÂND) (PROP P3) )
  (RESTPROP NIL) (COORD NIL) (DEL CĂ) (PROP P2)
      ( (RESTPROP NIL) (COORD NIL) (DEL SĂ) (PROP P4)
          ( (RESTPROP NIL) (COORD NIL) (DEL CĂ) (PROP P5)
            (RESTPROP NIL) (COORD ŞI) (DEL CĂ) (PROP P6) )
      ( (RESTPROP NIL) (COORD NIL) (DEL CUM) (PROP P7) )
) ) )
```

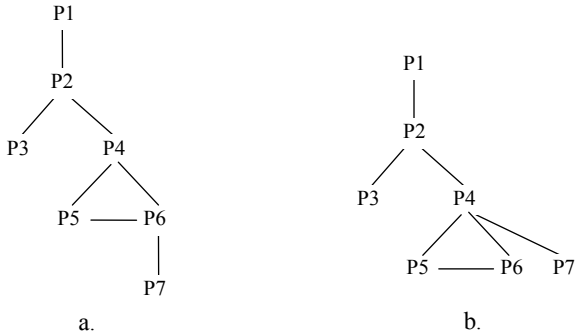

Fig. 9 - The two dependency trees of the outcome LISP structures

Without any translation into English, this fragment from [4] shows two alternative trees as the result of the Romanian finite-clause parsing. The linguistic solution is based on inter-clause and intra-clause markers (cue phrases, including the predicational feature occurrence, see [7], [8]), while the processing tool was an ATN (Augmented Transition Network) finite automaton-style interpreter.

Several authors dealt with finite-clause (not discourse structure) segmentation and parsing purposely. [5] is probably the first paper containing depth-first and breadth-first algorithms for Romanian finite clause segmentation and parsing based on hierarchically organized classes of lexical and grammatical markers.

The papers [19] and [20] devised two novelties to the solution of finite clause segmentation: applying a hybrid method relying on both linguistic rules and machine learning, and porting results from the English language to the Romanian. [19] obtains a good outcome for clause splitting in the (unrestricted text) English sentence (around 88%) and accentuates this necessary step for a large registry of natural language applications: machine translation, parallel text alignment [25], TTS systems, discourse structure processing. [20] continues Orasan's hybrid method based on machine learning and transfers the technique to the Romanian sentence with good results, but avoids to

establish the *types of* inter-clause *dependencies*, the relationship(s) between (or among, for the co-relation case) finite clauses of the Romanian sentence.

[6] continues [5], with a Java implementation of the SCD method in [7] for the clause segmentation, based on the SCD classes of intra- and inter-clause lexical markers. Evaluation of the clause segmentation algorithm applied on 15,000 clauses from the "1984" (G. Orwell) corpus provided very good results (96.6% precision and 95% recall). An interesting comparison between the SCD-based clause segmentation algorithm [7] and Marcu's RST clause-like segmentation at elementary discourse units [17] is enclosed in [6].

Classical and consistent papers for (English) clause parsing may be considered [10] and [11]. The purpose of finite clause parsing for unrestricted text in [10] was exactly as the task of the present paper, *i.e.* to improve the detection of larger prosodic units for an experimental TTS system. Even at that time, the author acknowledges correctly that the task of clause parsing for unrestricted text aiming to find the relevant sentence internal syntactic boundaries turned out to be a very difficult problem. [10] considers that the correct evaluation of the syntactic boundaries, in particular clause boundaries, is a prerequisite for the automatic insertion of boundary tones, pauses, and sentence boundaries, making the prosodic prediction to be more natural for the synthetic speech systems. The in-depth relationship between clausal and intonational structures is pointed out, statistics of clause boundaries co-occurring with intonational units and boundaries, for various types of finite clauses and contextual situations, is provided.

[11] brings also meaningful remarks, apparently well-known but, in fact, often ignored, with strange and overcharged computational consequences for language technology.

Ejerhed's papers [10], [11] provide fundamental results for our problem and emphasizes right contexts of its applications: speech synthesis (improved prosody in TTS systems), speech recognition (automatic segmentation of input to speech recognizers), semantic role labeling [13], machine translation (using clauses as translation units) etc.

It is worth to mention the viewpoint of [1] for our problem: Clause-level parsing can often be exploited when available, and such a parser would be undoubtedly useful.

The question is: many researchers are not really convinced that the benefits of a finite-clause parser outweigh its cost in terms of computational expense, depending on the application envisaged. For instance, phrase boundaries are not better prosodic predictors for sentence-level intonational boundaries, but they are "cheaper" than the clausal ones, thus easier to be used instead for prosody prediction algorithms.

One of the most frequent application of clause-level parsing is related to (rhetorical) discourse parsing. There are two main trends for discourse unit recognition and (rhetorical relation) dependency establishing, thus discourse parsing of (narrative) text spans. The *first* direction is represented by those papers (and the discursive theories they support) in which for computing the *elementary discourse units* (EDUs) one has to compute the finite clause at the beginning, *e.g.* [5], [6], [7]. This approach to discourse theory (in particular, RST [16], [17]) relies on the idea that an EDU has to contain at least a finite predication, actually, a finite clause (more generally, an EDU may contain just a sub-clausal phrase, as shown in [7], [14], [17]).

The *second* trend in discourse theory is to compute the (rhetorical) discourse segments *without* the finite clause parsing *ab initio*. This is the case of [3], [17], [21] etc. An interesting and comprehensive discussion on the role of classical, finite clause, and what D. Marcu calls "*clause-like*" spans for rhetorical discourse structures and parsing can be found in [17 :125-126]. Subsection 4.2 in [7] (*A Special Case: Sub-Clausal Discourse Segments*) discusses the situation of handing down of the rhetoric discourse markers *inside* the finite clause, and provides several analyses and examples (from [17]) supporting this particular case, very important for the discourse-prosody interface economy [9], [14].

Some significant studies that model the speech discourse based on prosodic-phonological entities and markers, relevant for textual discursive units, are coming from Chinese, around the group of Chiu-yu Tseng and collaborators [22], [23].

Potsdam Project [26] has as major purpose to develop a linguistic model for the articulation of *information structure* (IS) on syntax-prosody interface. [26] provides many insightful results on key subjects, balancing the gravity center of the research between German and Japanese. Of special interest is the prosody of *that*-complements and the relative clauses, *e.g.* [15]. There is a contrast between the RST of the textual

discourse [16], [17], for which the relative clause is embedded into its nuclear segment (and semantically pruned from the discourse tree at text summarization), and the special role in sentence-level intonational phrasing played by *that*-complements and relative clauses, in all languages. A working hypothesis [12] for *that*-complements is that they are expected to be *new* (thus Focus, Kontrast), whereas the *relative clause* (actually, its semantic nominal head) may be either *new* or *given*, in the IS sense.

Another interesting subject in [26] is whether the Chomsky's phase structure [2] can be (or not) the natural attachment site of IS-related elements.

Maybe the most exiting conjecture studied within [26] is that IS and syntax may work in parallel. Consistent arguments supporting that IS concepts do not play indeed an immediate role in syntax are discussed.

4. CLAUSE PARSING WITH SCD MARKERS FOR THE ACQUIS CORPUS

We propose a surface approach to clause parsing that operates with *cue phrases* that are formal indicators of the clause boundaries, which we further denote '*clause markers*'. The basic idea is that clause markers have a double role: they represent clause boundaries and they have an inter-clausal semantic role, in the sense that they indicate dependency types between clause units.

4.1. THE SCD CLAUSE MARKERS AND MARKER SEQUENCES

In previous research on clause parsing [6], [7], it was proposed the use of a hierarchy of markers with three levels, according to the linguistic role each marker is playing: *M1 class* contains intra-clause phrase-level markers, *M2 class* comprises predicate and semantic-role intra-clause markers, while *M3 class* covers specific inter-clause cue phrase markers. In this paper, we are not going to use marker classes, but instead to use individual lexical markers at inter-clause level. The clause lexical markers that are relevant for clause segmentation and dependency are mainly punctuation marks, coordinating/subordinating conjunctions, adverbs, pronouns and verbs, as illustrated in Table 1.

Table 1.

Clause markers

Marker type	**Lexical Marker**
Finite verbal group	am ajuns, să meargă, mănânc
punctuation marker	, : ! . ?
coordinating conjunction	şi, sau, ori
subordinating conjunction	ca să, că, dacă, de, să
Pronoun	care, ce, cel ce, cât, orice
Adverb	când, unde, cum, cât

In one of the tasks in CoNLL-2010 workshop, clause segmentation (for English) is treated as a classification problem [24]. The idea is to build a classifier to decide whether a certain word in a sentence represents a clause boundary. In this paper we take a somehow similar approach. We start off with a list of possible inter-clause markers for Romanian, which was constructed in [6]. For each marker candidate in the sentence, we learn from the training corpus the contexts in which that marker represents a clause boundary. The *context* we choose is the *sequence of markers between two finite verb groups*. Because each clause contains exactly one finite verbal group [7], [19], this means that the clause boundaries should be searched between two verbal groups. A similar strategy is employed by the clause identification algorithm for English and Romanian proposed in [5], [6], and [20]. The author of [20] uses a predefined list of clause markers, but she distinguishes between unambiguous clause markers (subordinating conjunctions) and ambiguous markers (punctuation and coordination conjunctions). Unambiguous clause markers are treated as valid clause boundaries, while for ambiguous markers she trains a classifier to decide whether or not it represents a valid clause boundary in a given context.

4.2. THE GOLD CORPUS

In order to perform training, we created a gold corpus extracted from *Acquis Communautaire* (AC) corpus for the Romanian language. This corpus represents a collection of legislative texts for the European Union. The corpus contains thousands of documents sorted by year, available in multiple languages. Each document has a particular structure: it contains seven types of (juridical-semantics oriented) segments, and each segment with none or several sub-segments. For this paper we selected a

number of 720 sentences from the most recent collection of documents in the Romanian version of Acquis corpus (years 2004 and 2005) that were among the largest text files.

To prepare the corpus for annotation, a number of preprocessing steps are performed. First, we automatically identify the juridical segments and sub-segments and retain those that contain full sentences. On each sentence we perform POS tagging using a POS tagger for Romanian language, which is available as a web application [27]. Next, the **finite verbal group (FVG)** is recognized based on linguistic knowledge, which is described in [8], and the markers are recognized using a pre-defined list from [6]. After the pre-processing is performed, one of the authors hand-annotated the clause markers that represent valid boundaries, identified full clauses, and performed attachment disambiguation.

4.3. THE CLAUSE-IDENTIFICATION ALGORITHM

We use a memory-based learning approach: in the learning phase we store all the cases in a database. In the test phase, we look for the most frequent case from its similar ones. The algorithm works with marker sequences. Each sentence must be represented as a sequence of symbols that contains: lexical clause markers, finite verbal groups, and '*' symbol for any other words in the sentence. Bellow, we have an example of mapping a sentence to a sequence of lexical markers and symbols.

*[**Dacă** urmează]^{C1}[**să** fie întreprinse acţiuni în urma nerespectării prezentului regulament conform alineatului (1)]^{C2}[**şi dacă** este necesar]^{C3}[**ca** animalele să fie transportate cu încălcarea unor dispoziţii din prezentul regulament]^{C4}[**,** autoritatea competentă eliberează o autorizaţie pentru transportul de animale.]^{C5}*

[Dacă FVG] [să FVG *] [şi dacă FVG][ca * FVG *] [, * FVG * .]

In the learning phase, from the prepared gold corpus, we automatically induce a set of marker sequences that capture the clause boundaries between two finite verbal groups. These sequences may contain both clause-start and clause-end boundaries ("[", "]") and a score associated with each sequence, which represents its frequency of occurrence in the corpus. Examples of marker sequences are provided bellow.

FVG] [să FVG	40
FVG *] [şi dacă FVG	12
FVG][ca * FVG	20

FVG], * FVG	10
FVG][, * FVG	18

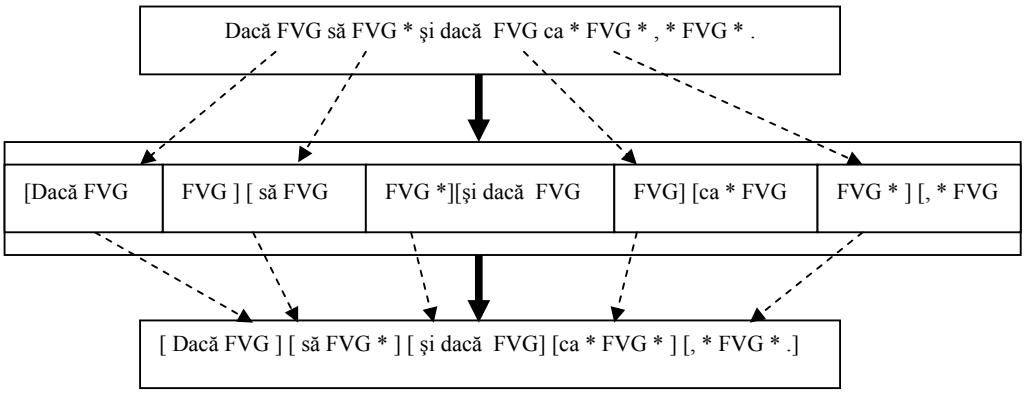

Fig. 10 - The segmentation algorithm for the above test example sentence

Given a test sentence, the clause boundary detection algorithm consists of the following steps:

1. Transform the sentence into a sequence of markers and split the sentence into sub-sequences between two consecutive finite verbal groups;

2. Find the clause boundary candidates that correspond to each sub-sequence(s) that match the sub-sequences extracted in the learning phase and their score;

3. Generate all the possible combinations of clause boundaries associated with the sentence and retain those combinations of clause boundaries that represent valid clause units;

4. Select the candidate with the highest score.

4.3. PROBLEMATIC MARKER SEQUENCES

The frequency function of marker occurrences in our corpus follows a Zipfian or power-low distribution. A large proportion of marker sequences occur only once in our training corpus, as can be seen in the diagram bellow. In our learning model, this generates a coverage problem: sequences in the test data will not be found in the training

data.

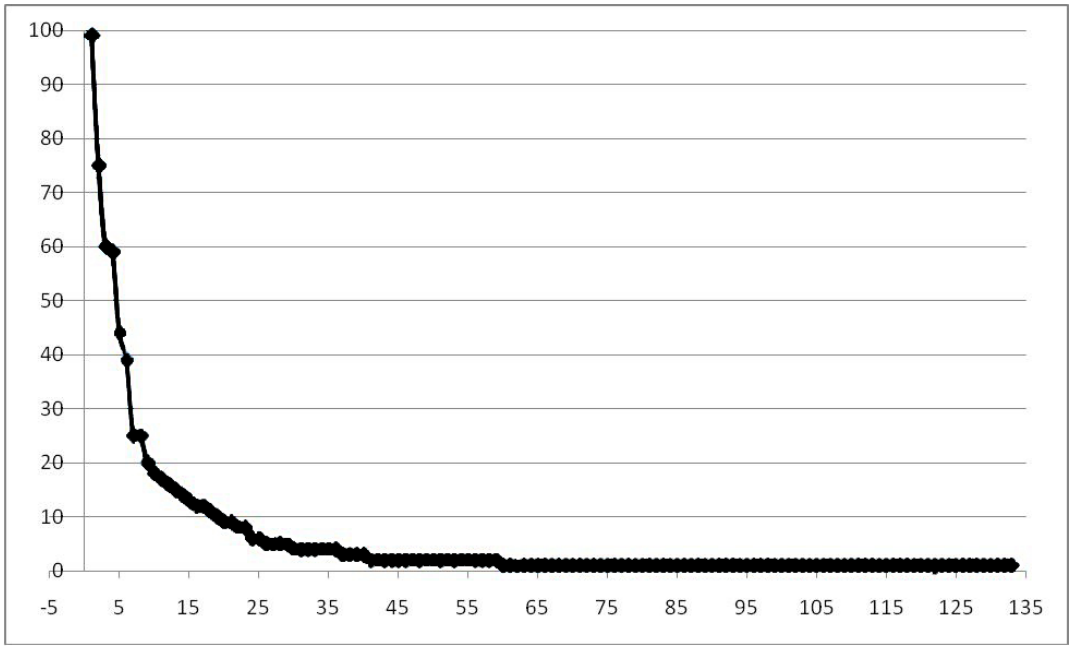

Fig. 11 - The distribution of sequences of markers. The X-axes indicates the rank of a marker in the frequency table, while the Y-axes represents the total number of the marker occurrences.

In this paper we have used a back-up rule to deal with unseen marker sequences: we ignore each unknown marker in the sequence and replace it with a '*' character. Thus, every unseen sequence of markers is reduced to a default sequence of markers that consist only of '*'characters, which can be frequently found in the training set.

Each marker carries a certain degree of ambiguity. Lexical markers such as *which* (Romanian 'care'), *if* (Romanian 'dacă') represent strong indicators of clause boundaries, while others such as *and* (Romanian 'şi') and ',' (comma) are inherently ambiguous. In case of ambiguity, the most frequent sequence in the corpus that represents a valid clause segmentation is chosen.

4.4. THE DEPENDENCY ALGORITHM

The clause dependency algorithm determines the structure of the sentence. The dependency between clause units is seen as an attachment disambiguation problem. Each dependency relation is labeled with a clause marker. Once we have the clause boundaries detected, we learn from the corpus the types of regent-subordinate

dependencies governed by each lexical marker. An example of a parsed sentence is given bellow.

[[***Dacă*** *urmează [****să*** *fie întreprinse acţiuni în urma nerespectării prezentului regulament conform alineatului (1)]^{C2}]^{C1} [* ***şi dacă*** *este necesar [* ***ca*** *animalele să fie transportate cu încălcarea unor dispoziţii din prezentul regulament]^{C4}]^{C3}****,*** *autoritatea competentă eliberează o autorizaţie pentru transportul de animale.]^{C5}*

[[Dacă FVG[să FVG *]] [şi dacă FVG[ca * FVG *]] , * FVG * .]

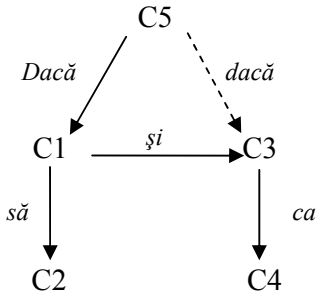

Fig. 12 - The dependency algorithm for the test example in the sentence above

In the example given there are four regent-subordinate dependencies: *Dacă*(C5, C1), *şi_dacă*(C5, C3), *să*(C1, C2) and *ca*(C3, C4). What we learn in this process is the probability of a marker (or a marker sequence) to indicate a certain dependency relation between two (or more) clauses. For instance, in the example above, we learn from the corpus that in 70% of the cases the sequence [* FVG]$_{C1}$ [*să* FVG *]$_{C2}$ is equivalent to *să*(C1, C2) and indicates that clause C2 is subordinated to C1, as shown in Figure 12.

4.5. EVALUATION

In order to estimate the accuracy of our model, the *leave-one-out* cross validation technique is used: each file in the gold corpus is taken as a test file, while the other files are used for training. On average, we obtained an accuracy of 83% for clause segmentation and 60% for the attachment disambiguation algorithm (clause dependencies). Most of the errors in clause boundary detection appear for embedded clauses and in cases where boundaries are represented by ambiguous clause markers, such as punctuation marks. For the dependency algorithm, most of the errors occur when the structure of the sentence is complex and there are at least three dependency

levels in the hierarchy (e.g. when the regent and the subordinate clauses are separated by other clauses).

5. CONCLUSIONS

We described a surface method based on lexical-level markers for the Romanian clause parsing. The present paper continues [5], [6], [7] with classes of inter-clause lexical markers for the clause segmentation task, but also carry on [19] and [20], by introducing machine learning techniques for the first time in the SCD context. Another novelty is that clause parsing is achieved by machine learning not only for the lexical markers making up clausal boundaries, but also on the clause dependency relations represented by these inter-clause markers. We emphasize that the parsing process evaluation has still important resources to be improved (larger lexical marker databases, more efficient learning techniques). The evaluation for the Romanian clause segmentation is comparable with similar approaches for the same problem, *e.g.* [20]. The dependency algorithm produces also encouraging results, the present method being the first clause parsing of Romanian which uses lexical-level markers, aiming to be utilized for prosody prediction of the Romanian clause and sentence speech patterns.

Clearly, more linguistic information is required to reach the state-of-the art results for clause parsing obtained in the CoNLL task for the English language [24]. Previous work on clause parsing shows that information about the boundaries of noun phrase structures increases the accuracy of the clause parsing algorithm. For future work we plan to develop a hybrid model that integrates complex linguistic information.

Authors' contributions: N. Curteanu developed the concepts behind SCD segmentation and parsing using lexical markers; he also made the theoretical discussion on markers and their role at inter-clause and discourse levels. C. Butnariu designed and implemented the learning model for clause segmentation and parsing and did the review on clause segmentation. C. Bolea created the resources for the training model and analysed the results.

REFERENCES

1. ALLEN J., *Linguistic Aspects of Speech Synthesis*, Proceedings of the National Academy of Sciences of the USA, 1995, **92**, 9946-9952.

2. CHOMSKY N., *Derivation by phase* (M. Kenstowicz , ed.), 2001, 1–52.

3. CRISTEA, D., POSTOLACHE O., PISTOL I., *Summarisation through Discourse Structure*, Proceedings of CiCling 2005, Springer LNSC, 2005, vol. **3406**.

4. CURTEANU N., *Algorithms for the syntactic analysis of the Romanian senstence and clause*, Proceedings INFO-IAŞI'83, Iaşi, 1983, 533-548 (in Romanian).

5. CURTEANU N., *From Morphology to Discourse through Marker Structures in the SCD Parsing Strategy*, Language and Cybernetics, Akademia Libroservo, Prague, 1994, 61-74.

6. CURTEANU N., ZLAVOG E., BOLEA C., *Sentence-Based and Discourse Segmentation / Parsing with SCD Linguistic Strategy*, in 'Intelligent Systems' Conference Volume (H.-N. Teodorescu et al., eds), Performantica Press, Iaşi (Romania), 2005, 153-168.

7. CURTEANU N., *Local and Global Parsing with Functional (F)X-bar Theory and SCD Linguistic Strategy,* I-II, Computer Science Journal of Moldova, Academy of Science of Moldova, 2006, **14**, *1* (40): 74-102; *2* (41): 155-182.

8. CURTEANU N., MORUZ A., TRANDABĂŢ D., BOLEA C., DORNESCU I., *The Structure and Parsing of Romanian Verbal Group and Predicate*, Advances in Intelligent Systems and Technologies ECIT2006 – 4th European Conference on Intelligent Systems and Technologies, Iasi, Romania, 2006, 93-105.

9. CURTEANU N., TRANDABĂŢ D., MORUZ A., *Expanding Topic-Focus Articulation with Boundary and Accent Assignment Rules for Romanian Sentence*, Text, Speech and Dialogue, Proceedings of the 12th International Conference TSD 2009 (V. Matousek, P. Mautner et al., eds.), Plzen, Czech Republic, Lecture Notes in Computer Science, no. 5729, 2009, 226-233.

10. EJERHED E., *Finding clauses in unrestricted text by finitary and stochastic methods*, Proceedings of the 2nd Conference on Applied Natural Language Processing, Austin, Texas, 1988, 219 –227.

11. EJERHED, E., *Finite State Segmentation of Discourse into Clauses*, Workshop on Extended Finite State Models of Language, Proceedings of the ECAI 96 (A. Kornai, ed.), 1996, Budapest, Hungary.

12. FÉRY C., SHINICHIRO I., *How Information Structure Shape Prosody*, Information Structure from Different Perspectives (M. Zimmermann & C. Féry, eds.), Oxford University Press, 2007.

13. GILDEA D., JURAFSKY D., *Automatic labeling of semantic roles*, Computational Linguistics, 2002, **23**, *3*, 245-288.

14. VON HEUSINGER K., *Discourse Structure and Intonational*, Topic and Focus: Intonation and Meaning. Theoretical and Crosslinguistic Perspectives (D. Büring & M. Gordon & Ch. Lee, eds.), Dordrecht, Springer, 2007, 265-290.

15. KALAND C., VAN HEUVEN V. J., *The structure-prosody interface of restrictive and appositive relative clauses in Deutch and German*, Speech Prosody, Chicago, 2010.

16. MANN W., THOMPSON S., *Rhetorical Structure Theory*: *A Theory of Text Organization*, Research Report RS-87-190, Information Sciences Institute, University of Southern California, Marina del Rey, California, 1988, 80 pp.

17. MARCU D., *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph. D. Thesis, Univ. of Toronto, Canada, 1997, 341.

18. MILLER P., *Strong Generative Capacity*. *The Semantics of Linguistic Formalism*. CSLI Publications, Stanford, California, 1999.

19. ORASAN C., *A hybrid method for clause splitting in unrestricted English texts.* Available at: http://www.wlv.ac.uk/sles/compling/papers/orasan-00.pdf , 2000.

20. PUSCASU G., *A Multilingual Method for Clause Splitting*, Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, Birmingham, UK, 2004.

21. SORICUT S., MARCU D., *Sentence Level Discourse Parsing using Syntactic and Lexical Information*, Research Report, Information Sciences Institute, University of Southern California, Marina del Rey, California, 2003.

22. TSENG C.-Y., SAN Z.-Y., CHANG C.-H., TAI C.-H., *Prosodic Fillers and Discourse Markers-Discourse Prosody and Text Prediction*, The Second International Symposium on Tonal Aspects of Languages (TAL 2006), France, 2006, 109-114.

23. TSENG C.-Y., SU Z.-Y., LEE L.-S., *Prosodic Patterns of Information Structure in Spoken Discourse. A Preliminary Study of Mandarin Spontaneous Lecture vs. Read Speech*, Speech Prosody, Chicago, 2010.
24. TJONG E. F., SANG K., DÉJEAN H., *Introduction to the CoNLL-2001 shared task: Clause identification*, Proceedings of CoNLL 2001, Toulouse, France, 53–57.
25. TUFIS D., *From Word Alignment to Word Senses, via Multilingual Wordnet*, Computer Science Journal of Moldova, 2006, **14** (1):3-33.
26. ⁂ Potsdam Project, http://www.sfb632.uni-potsdam.de/projects_a1eng.html, 2007
27. ⁂ http://www.racai.ro/webservices/

Received November 22, 2010