# Towards the Annotation of Penn TreeBank with Information Structure

**Bernd Bohnet**
School of Computer Science
University of Birmingham
Birmingham, UK
b.bohnet@cs.bham.ac.uk

**Alicia Burga**
DTIC
Pompeu Fabra University
Barcelona, Spain
alicia.burga@upf.edu

**Leo Wanner**
ICREA and DTIC
Pompeu Fabra University
Barcelona, Spain
leo.wanner@upf.edu

## Abstract

Information Structure (IS) determines the "communicative" segmentation of the meaning of an utterance, which makes it central to the semantics–syntax–intonation interface and therefore also to NLP. Despite this relevance, IS has not received much attention in the context of the majority of the reference treebanks for data-driven NLP that already contain a semantic and syntactic layers of annotation. We present our work in progress on the annotation of the Penn TreeBank with the thematicity dimension of the IS as defined in the Meaning-Text Theory. We experiment with tagging and transition-based parsing techniques. Especially the latter achieve acceptable accuracy with even very small training samples, which is promising for languages with scarce resources.

## 1 Introduction

The *Information Structure* (IS) (aka *Topic-Focus Articulation*, TFA (Sgall, 1967) in the Prague School and *Communicative Structure*, CommStr (Mel'čuk, 2001) in the Meaning-Text Theory) determines the "communicative" segmentation of the meaning of an utterance. This makes it central to the semantics–syntax–intonation interface (Lambrecht, 1994; Hajičová et al., 1998; Steedman, 2000; Mel'čuk, 2001; Erteschik-Shir, 2007) and therefore also to NLP. However, despite its prominence, IS has been largely ignored so far in the context of the reference treebanks for data-driven NLP: Penn Treebank (Marcus et al., 1993) and its semantic counterpart PropBank (Palmer et al., 2005) for English, Tiger (Thielen et al., 1999) for German, Ancora (Taulé et al., 2008) for Spanish, etc. To the best of our knowledge, only the Prague Dependency Treebank (PDT) (Hajič et al., 2006) is annotated with IS in terms of TFA. This is not to say that no proposals have been made for the annotation of IS in general; see, e.g., (Calhoun et al., 2005) for English, (Dipper et al., 2004) for German, (Paggio, 2006) for Danish, etc. However, in the light of the above mentioned interface, it is crucial to have the same corpus annotated with semantic, syntactic and IS structures.

In this paper, we present our work in progress on the annotation of the ConLL '09 variant of the Penn TreeBank (PTB) (Hajič et al., 2009) with the *thematicity* dimension of Mel'čuk's CommStr.[1] We have chosen thematicity because (i) it distinguishes, apart from the traditional Theme and Rheme, a Specifier element, and (ii) it is hierarchical in that a thematicity partition can be embedded into another thematicity partition. Both of these features facilitate a fine-grained communicative partition of complex utterances with subordinations and thus a more accurate and detailed projection between the different layers of the semantics–syntax–intonation interface.

Unlike most other proposals, we aim to automate the annotation procedure and experiment with tagging and transition-based parsing techniques. In this respect, our goal is similar to that of Postolache et al. (2005), who explore different classifier models for automatic labeling of tectogrammatical nodes in the PDT with Topic / Focus. But while Postolache et al. use for training 78.3% (38,737 sentences or 494,759 tectogrammatical nodes) of the TFA+tectogrammatical layer of the PDT, our training sample is infinitely smaller: we train on 360 manually annotated sentences. The purpose of the small training sample is

[1] It is important to note that thematicity does not intend to capture the IS in its entirety—as, e.g., Vallduví (1992)'s or Erteschik-Shir (2007)'s proposals do. It is just one of the eight dimensions of the CommStr, although the most central one.



*International Joint Conference on Natural Language Processing*, pages 1250–1256,
Nagoya, Japan, 14-18 October 2013.

twofold. First, to minimize the manual annotation effort, and, second, to assess whether automatic IS annotation can be bootstrapped starting from minimal resources.

In the next section, we introduce the theoretical notion of thematicity in the sense of the Meaning-Text Theory and present the criteria for the determination of its individual elements. Section 3 describes our tagging and parsing approaches to automatic annotation of thematicity. Section 4 outlines the experiments we carried out in order to assess the quality of both approaches, and Section 5, finally, discusses the outcome of these experiments and sketches some directions of future work we plan to undertake in this area.

## 2 Theoretical Background

Since its introduction by the Prague School of Linguistics (Mathesius, 1929; Firbas, 1966; Daneš, 1970), a great number of models that define and determine the IS of an utterance have been proposed; for an overview, see, e.g., (Vallduví, 1992; Mel'čuk, 2001; Kruijff-Korbayová and Steedman, 2003; Zimmermann and Féry, 2009). In our work, we draw upon Mel'čuk (2001)'s model, which foresees a tripartite communicative segmentation of the meaning of an utterance: Theme–Rheme–Specifier. The Specifier (SP) sets up the context of the utterance $U$; Rheme (R) denotes the part of $U$ that the speaker presents as stated by $U$; and Theme (T) denotes that part of $U$ that the speaker presents as something about which R is stated.[2] T and R (also referred to in the literature as *topic/focus* and *topic/comment*) constitute the *Communicative Core* (CC) of a sentence.

The basic unit that we annotate with SP/T/R tags is a *proposition*. A proposition (P) is either a *full clause*, i.e., a clause that contains a finite verb, or a *reduced clause*, i.e., a clause with elision of the corresponding finite verb (as *up 150* in *Mitsubishi Estate ended the day at 2680, up 150*). The following assumptions hold: (i) each P possesses a CC; (ii) if a sentence is composed by a coordination of Ps (each with its own subject), no global CC is assigned: each of the Ps has its own CC, with the coordinative conjunction as the SP of the second P; (iii) two non-coordinated Ps (each holding an independent CC, with the subordinated conjunction as SP) form a global CC such that the one that comes first is T and the other one is R; if their linear order is altered, the T/R assigment switches although the meaning remains the same; (iv) if a sentence contains a main P and a relative subordinate P, one main CC is assigned (without taking into account the presence of the relative clause), and at the same time an embedded CC is assigned to the subordinate proposition. That is, as already mentioned in Section 1, thematicity in the Meaning-Text Theory is *per se* hierarchical in that each thematicity element (SP/T/R) can in itself be again assigned a thematicity structure.[3]

Consider (1) for illustration of the thematicity segmentation of a sample sentence (for clear and unambiguous notation, propositions are enclosed in "{...}"):

(1) {[*Years ago*]*SP,* [*he*]*T* [*collaborated with the new music gurus Peter Serkin and Fred Sherry in the very countercultural chamber group Tashi,* {[*which*]*T* [*won audiences over to dreaded contemporary scores like Messiaen's Quartet for the End of Time*]*R*}.]*R* }

For the communicative segmentation of a fragment of the PTB as gold standard, we used the following empirically determined criteria:

**Criteria for determining the Specifier:** Given that Specifiers do not express a separate message, but, rather, the context of the message to which they belong, we mark as Specifiers:

– fronted temporal, locative and manner circumstantials: {[*Apparently*]*SP* [*he*]*T* [*did so*]*R*};

– fronted AdjPs with a sentential scope: {[*Tired of the same*]*SP,* [*he*]*T* [*gave up*]*R*};

– fronted discourse markers: {[*But*]*SP* [*it*]*T* [*was neither deep*]*R*};

– circumstantials of the type *according to ...* (independently of their position): {[*About 25 % of the insiders*]*T ,* [*according to SEC figures*]*SP,* [*file their reports late*]*R*};

– phrases that introduce direct speech (independently of their position):{[*It*]*T* [*is done*]*R,* [*he said*]*SP*};

– NPs in vocative case (independently of their position): {[*Anna*]*SP,* [*he*]*T* [*did it*]*R*}.

[2] Strictly speaking, Theme and Rheme are defined over the meaning of an utterance, rather than the utterance itself. It is for brevity that we speak of an IS of an utterance.

[3] The hierarchical relations in a given thematicity segmentation are in practice controlled by indices. Thus, 'T(T1)' will stand for "theme within the theme element marked as 'T1' " and 'R(T1)' for "rheme within T1". To distinguish between elements of the same type at the same level of the hierarchy, numbers are used: 'T1' vs. 'T2' 'R1' vs. 'R2', etc.

**Criteria for determining the Theme:** *Theme* is the part of the sentence that expresses what the Speaker is talking about. Therefore, it tends to be located in the initial part of the sentence (after the Specifiers, if there are any). In an SVO language as English, the Theme therefore coincides most of the times with the subject. Apart from this *ad hoc* criterion, a number of "hard" criteria to identify the Theme are available; among them:

– it can be identified by the question "And what about X" (then, X is Theme): {[*John*]*T* [*answered the question*]*R*}: *And what about John?* (**And what about question?*);

– it is not accessible for general negation/questioning: {[*He*]*T* [*did it*]*R*}: **Not he did it*;

– a relative clause is treated as an independent proposition, and therefore the relative pronoun is the Theme only if it is subject (otherwise, it is a focalized part of the Rheme): *the boy* {[*who*]*T* [*cooked*]*R*} vs. *the boy* {[*whom*]*R1-1,Foc* [*I*]*T* [*met*]*R1-2*};

– indefinite pronouns such as *nobody*, *somebody*, *nothing*, etc. and negative noun phrases cannot be Themes: e.g., in *None of the boys did it*, it is not *none of the boys*, which is the Theme, but rather *it*;

– sentences of the form *It + is + Adjective + Infinitive verb* reverse the typical position of the theme, with the infinitive verb being the theme: {[*It's necessary*]*R* [*to talk*]*T*};

– headings or titles are all-thematic.

Occasionally, a split of Theme can be observed: {[*Considered as a whole*]*T1-1,* [[*Mr. Lane*]*T1(SP)* [*said*]*R1(SP)*]*SP,* [*the filings required under the proposed rules*]*T1-2 "*[*will be at least as effective, if not more so , for investors following transactions*]*R .*"}

**Criteria for determining the Rheme:** The easiest way to recognize Rheme is through exclusion: if an element is not Theme nor Specifier, then it is part of the Rheme. A few explicit criteria can also be introduced:

– Rheme can be negated and/or questioned: {[*I*]*T* [*think so*]*R*}: *I don't think so* / *Do I think so?*;

– existential sentences (those that begin with *There is/are*) are all rhematic: {[*There are apples on the table*]R};

– non-fronted temporal, locative and manner circumstantials form part of the Rheme: {[*I*]*T* [*met John some months ago in the park, in a very unexpected way*]*R.*}

In addition, it is to be noted that if the Rheme contains a ditransitive verb that allows arguments to exchange syntactic positions, we assume that this exchange is motivated by different ISs; cf. {[*John*]*T* [[*gave me*]*T(R1)* [*money*]*R(R1)*]*R1*} vs. {[*John*]*T* [[*gave money*]*T(R1)* [*to me*]*R(R1)*]*R1*} (the tag 'R' is supplied with a number for unambiguous notation). Furthermore, within NPs that start with *wh*-words that do not belong to pseudo-clefting constructions, split Rhemes are observed: {[[*What*]*R1-1(T1)* [*he*]*T(T1)* [*said*]*R1-2(T1)*]*T1* [*was hilarious*]*R1*}.

## 3 Annotating PTB with Thematicity

Given that the thematicity structure as used in this paper is of a hierarchical nature, its automatic annotation can be viewed not only as a tagging but also as a (constituency) parsing task. We carried out experiments with both approaches, taking the tagger variant as baseline (the idea is to apply a simple and well researched technique in order to compare it with a more elaborated one). For both, we use the CoNLL '09 format; cf. Figure 1. A line in this format consists of an id, a word form, a lemma, a pos tag, and the dependency annotation with the head node and the edge label (all retrieved from the CoNLL Shared Task 2009 data set). The last two columns contain the gold communicative tag and the tag predicted in the course of the automatic annotation, respectively.

| id | form | lemma | pos | head | edge | com. | p.-com. |
|---|---|---|---|---|---|---|---|
| 1 | He | he | PRP | 2 | SBJ | []T | _ |
| 2 | believes | believe | VBZ | 0 | ROOT | [ | _ |
| 3 | in | in | IN | 2 | ADV | _ | |
| 4 | what | what | WP | 6 | OBJ | []T | _ |
| 5 | he | he | PRP | 6 | SBJ | [ | _ |
| 6 | plays | play | VBZ | 3 | PMOD | ]R]R | _ |
| 7 | . | . | . | 2 | P | _ | _ |

Figure 1: Example of a sentence annotated with its thematicity structure in CoNLL format.

In the case of tagging, we aim to assign to each element of a sentence a T, R or a SP tag. For this purpose, we use classifier-based sequence tagging. The tagger assigns one of the three tags to each word by going from left to right through the sentence. For the selection of the appropriate tag, the tagger considers features from a window of two words before and after the word in question.; cf. Table 1 for the features used by the tagger.

For training the classifier of the sequence tagger, we use the perceptron algorithm. Following Collins and Duffy (2002), averaging of the param-

| Features based on PoS tags |
|---|
| $\pi(i), \pi(i-1)\pi(i), \pi(i)\pi(i+1)$ ,$\pi(i+1)\pi(i+2)$ |
| $\pi(i-1)\pi(i-2), \pi(i-2)\pi(i-1)\pi(i)$, |
| $\pi(i-1)\pi(i)\pi(i+1), \pi(i)\pi(i+1)\pi(i+2)$ |
| **Features involving PoS tags and word forms** |
| $w(i), \pi(i-1)w(i), w(i)\pi(i+1)$ ,$w(i+1)\pi(i+2)$ |
| $\pi(i+1)w(i+2)$,$w(i-1)\pi(i-2)$,$\pi(i-1)w(i-2)$ |
| $w(i-1)w(i)$,$w(i)w(i+1)$ |
| $w(i+1)w(i+2)$,$w(i-1)w(i-2)$ |

Table 1: Features for the sequence tagger. $i$ ($i = 0, 1, \ldots$) denotes the $i$th word in the input sentence; $\pi$ and $w$ are functors to extract the PoS tags respectively word forms of the tokens.

eters obtained in the training algorithm is applied for classifying the test examples.

In the case of parsing, we aim to derive the hierarchical IS (or *communicative tree*) of a given sentence. The communicative tree $T_c$ of a sentence $x = w_1 ... w_n$ is a quintuple $T_c = (V, E, L, \delta, 0')$, such that $V = V_t \cup V_c$ is a set of nodes, with $V_t = 0, ..., n$ as a set of terminal nodes and $V_c = o', 1', ..., m'$ as a set of non-terminal communicative (label) nodes; $E \subseteq V \times V$ is a set of edges; $L$ is the set of communicative labels (in the case of thematicity: SP, T, R and P); $\delta : E \rightarrow L$ is a labeling function for nodes; $0'$ is the root node. That is, we interprete the $T_c$ as a kind of constituency tree.

For the implementation of the parser, we use the idea of transition-based parsing (Yamada and Matsumoto, 2003; Nivre et al., 2004), which uses a classifier to predict the shift/reduce actions. We draw upon the transition set of the *arc-eager parser* Nivre (2004), but with a slightly different semantics in that we define a transition system for the derivation of the $T_c$ as a quadruple $C = (S, Y, c_0, S_y)$, where $S$ is a set of parsing states; $Y$ is a set of transitions, each of which is a (partial) function t: $S \rightarrow S$; $s_0$ is an initialization function that maps a sentence $x$ to a configuration $s \in S$; and $S_y \subseteq S$ is a set of terminal states. A transition sequence for a sentence $x$ in $C$ is a sequence of pairs of states and transitions. As set $S$ of states, we use the tuple $s = (\Sigma, B, V_c, Z, E, \delta, o)$, where the stack $\Sigma$ and the input buffer $B$ are disjoint sublists of the terminal nodes $V_t$, $V_c$ is the set of communicative (label) nodes, $Z$ is the stack of communicative nodes, $E$ is the set of edges, $\delta$ is a labeling function for communicative label nodes $n \in V_c$, and $o$ is a counter for the number of pairs of delimitation brackets. The initial state for a sentence $x$ is $s_0 = ([0], [1, ..., n], \{0'\}, [0'], \{\}, \delta, 0)$. Terminal configurations have an empty buffer and only the root node $n$ is contained in the stack $\Sigma$: $s = ([0], [], V_c, Z, E, \delta, x)$. Figure 2 shows the possible transitions.

As features of the transition-based system, we use a rich feature set based on the dependency structure drawn from (Zhang and Nivre, 2011) (since we use as input a dependency structure these features are available). In addition, we use the path from the top stack element to the word of the last open bracket (as sequence of pos tags). For the training of the transition-based system, we use the perception algorithm with averaging, a beam-search with 10 elements and early update (Collins and Roark, 2004). The oracle for training of the system follows the bottom-up parsing strategy. As soon as the communicative part is completed, we remove (reduce) the nodes that belong to it from the stack. Figure 3 shows a sequence of transitions that the analyser performs to create the $T_c$ of the example sentence in Figure 1.

## 4 Experiments

Following the criteria in Section 2, four annotators in teams of two manually annotated a fragment of 435 sentences of the PTB with the thematicity structure, in a series of blocks of about 40–50 sentences. To ensure high mutual agreement, the annotation procedure went as follows. First, one of the teams provided a first round annotation of a block of sentences. This annotation was revised by the other team and the two annotations were discussed in plenum to achieve a consensus and to refine the annotation guidelines. The refined guidelines were used by the first team to annotate the next block of sentences—to be again revised by the other team and discussed in plenum. And so on.[4]

For training, we use the first 360 sentences, the next 40 sentences as development set, and the remaing 35 sentences as test set. Table 2 presents the results on the test set for the sequence tagger and the transition-based analyser. The Accuracy Score (AS) measures the correctly assigned thematicity tags on a token basis in the same way as PoS tagging is evaluated. That is, given, e.g., the sequence `[,_,[]T,[,]R]R` and the predicted sequence `_,_,[]T,[,]R`, we see 3 correct to-

[4] At this stage, we did not measure the initial agreement between the annotators since our goal was to achieve a high level of agreement in the course of the discussion. However, to follow the common practice in corpus annotation, we will provide in the near future the inter-annotator figures.

| Transition | | Condition |
|---|---|---|
| LEFT-BRACKET | $([\sigma\|i], B, V_c, [\zeta\|u'], E, \delta, o) \Rightarrow ([\sigma\|i], B, V_c, \cup\{o' \leftarrow (o+1)'\}, [\zeta\|u'\|o'], E \cup \{(i,o'),(u',o')\}, o+1)$ | $i \neq 0$ |
| RIGHT-BRACKET$_l$ | $([\sigma\|i], B, V_c, [\zeta\|u'\|o'], E, \delta, o) \Rightarrow ([\sigma\|i], B, V_c, [\zeta\|u'], E \cup \{(i,o')\}, \delta[o' \rightarrow l], o)$ | $\|\zeta\| > 1$ |
| SHIFT | $(\sigma, [i\|\beta], V_c, \zeta, E, \delta, o) \Rightarrow ([\sigma\|i], \beta, V_c, \zeta, E, \delta, o)$ | |
| REDUCE | $([\sigma\|i], B, V_c, \zeta, E, \delta, o) \Rightarrow (\sigma, B, V_c, \zeta, E, \delta, o)$ | $i \neq 0$ |

Figure 2: Possible parsing transitions.

| | | |
|---|---|---|
| 1 | SHIFT | $([0], [\text{He,believes, ...}], \{0'\}, [0'], \{\}, \delta, 0) \Rightarrow ([0\|\text{He}], [\text{believes, ...}], \{0'\}, [0'], \{\}, \delta, 0)$ |
| 2 | LEFT-BRACKET | $\Rightarrow ([0\|\text{He}], [\text{believes, ...}], \{0', 1'\}, [0'\|1'], \{(\text{He}, 1'), (0', 1')\}, \delta, 1)$ |
| 3 | RIGHT-BRACKET$_T$ | $\Rightarrow ([0\|\text{He}], [\text{believes, ...}], \{0', 1'\}, [0'], \{(\text{He}, 1')\}, \{(1' \rightarrow T)\}, 1)$ |
| 4 | REDUCE | $\Rightarrow ([0], [\text{believes,in, ...}], \{0', 1'\}, [0'], \{(\text{He}, 1'), (0', 1')\}, \{(1' \rightarrow T)\}, 1)$ |
| 5 | SHIFT | $\Rightarrow ([0,\text{believes}], [\text{in, ...}], \{0', 1'\}, [0'], \{(\text{He}, 1'), (0', 1')\}, \{(1' \rightarrow T)\}, 1)$ |
| 6 | LEFT-BRACKET | $\Rightarrow ([0,\text{believes}], [\text{in, ...}], \{0', 1', 2'\}, [0'\|2'], \{(\text{He}, 1'), (0', 1'), (\text{believes}, 2'), (0', 2')\}, \{(1' \rightarrow T)\}, 2)$ |
| 7 | SHIFT | $\Rightarrow ([...,\text{in}], [\text{what, ...}], \{0', 1', 2'\}, [0'\|2'], \{(\text{He}, 1'), (0', 1'), (\text{believes}, 2'), (0', 2')\}, \{(1' \rightarrow T)\}, 2)$ |
| 8 | SHIFT | $\Rightarrow ([...,\text{in,what}], [\text{he, ...}], \{0', 1', 2'\}, [0'\|2'], \{(\text{He}, 1'), (0', 1'), (\text{believes}, 2'), (0', 2')\}, \{(1' \rightarrow T)\}, 2)$ |
| 9 | LEFT-BRACKET | $\Rightarrow ([...,\text{in,what}], [\text{he, ...}], \{0', 1', 2', 3'\}, [0'\|2'\|3'], \{..., (\text{believes}, 2'), (\text{what}, 3'), (2', 3')\}, \{(1' \rightarrow T)\}, 3)$ |
| 10 | RIGHT-BRACKET$_T$ | $\Rightarrow ([...,\text{in,what}], [\text{he, ...}], \{0', 1', 2', 3'\}, [0'\|2'], \{..., (\text{what}, 3'), (2', 3')\}, \{(1' \rightarrow T), (3' \rightarrow T)\}, 3)$ |
| 11 | REDUCE | $\Rightarrow ([...,\text{in}], [\text{he, ...}], \{0', 1', 2', 3'\}, [0'\|2'], \{(..., (\text{what}, 3'), (2', 3')\}, \{(1' \rightarrow T), (3' \rightarrow T)\}, 3)$ |
| 12 | SHIFT | $\Rightarrow ([...,\text{in,he}], [\text{plays}], \{0', 1', 2', 3'\}, [0'\|2'], \{..., (\text{what}, 3'), (2', 3')\}, \{(1' \rightarrow T), (3' \rightarrow T)\}, 3)$ |
| 13 | LEFT-BRACKET | $\Rightarrow ([...,\text{in,he}], [\text{plays}], \{0', 1', 2', 3', 4'\}, [0'\|2'\|4'], \{..., , (2', 3'), (\text{he}, 4')\}, \{(1' \rightarrow T), (3' \rightarrow T)\}, 4)$ |
| 14 | SHIFT | $\Rightarrow ([...,\text{in,he,plays}], [], \{0', 1', 2', 3', 4'\}, [0'\|2'\|4'], \{..., (2', 3'), (\text{he}, 4')\}, \{(1' \rightarrow T), (3' \rightarrow T)\}, 4)$ |
| 15 | RIGHT-BRACKET$_R$ | $\Rightarrow ([0,\text{believes,in,he,plays}], [], \{0', 1', 2', 3', 4'\}, [0'\|2'], \{..., (2', 4')\}, \{..., (4' \rightarrow R), \}, 4)$ |
| 16 | RIGHT-BRACKET$_R$ | $\Rightarrow ([0,\text{believes,in,he,plays}], [], \{0', 1', 2', 3', 4'\}, [0'], \{..., (2', 4')\}, \{..., (4' \rightarrow R), (2' \rightarrow R)\}, 4)$ |
| 19 | REDUCE,...,REDUCE | $\Rightarrow ([0], [], \{0', 1', 2', 3', 4'\}, [0'], \{..., (\text{he}, 4'), (2', 4'), (0', 2')\}, \{..., (4' \rightarrow R), (2' \rightarrow R)\}, 4)$ |

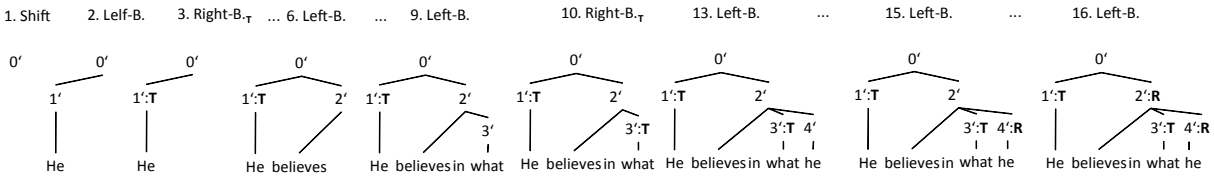


Figure 3: Transition sequence for the sentence: *He believes in what he plays*.

kens out of 5, i.e., an accuracy score of 60%. Note that the assignment `]]R` instead of `]R` is considered as wrong. In contrast to this simple score, the labeled bracket score (LBS) and the unlabeled bracket score (UBS) consider the bracketing; the scores are calculated with the `evalb`-script as used for the evaluation of phrase structure parsers.

| System | AS | LBS | UBS |
|---|---|---|---|
| sequence tagger | 71.74 | 51.78 | 53.29 |
| transition-based | 88.67 | 68.95 | 74.33 |

Table 2: Accuracy scores for the assignment of the communicative labels.

## 5 Discussion and future work

The results of our experiments show that the interpretation of the annotation of PTB with IS as a transition-based parsing task is promising. Acceptable accuracy scores are achieved already with a very small training set. A direct comparison with other works on automatic annotation with IS, as e.g., (Postolache et al., 2005) with TFA, is not possible since the data sets and the annotation schemata are different; see, e.g., (Hajičová, 2007) for a precise outline of the criteria for the annotation of TFA in the Prague school and a juxtaposition of TFA and the CommStr. However, it is instructive to observe that the AS we achieve with the transition parser is about the same as Postolache et al.'s accuracy with RIPPER and MAXENT models and only slightly below their performance with C4.5. This means that parsing is a valid alternative to token-oriented classification. However, we must be aware that parsing can only be applied if we assume the IS structure to be hierarchical (more precisely, a tree). The parser performance in terms of LBS and UBS, which capture the "bracketing" of the transitivity elements within the structure, are somewhat lower and can still be improved with a larger training sample (see below).

To have a clearer idea what the most recurrent mistakes of our IS parser are and whether they can be avoided (e.g., by a larger training sample) we carried out an error analysis of the resulting automatic annotation. This analysis has shown that sentences with clear (lexical, syntactic,

and/or punctuation) thematicity markers are analyzed correctly. Cf., e.g., (2) and (3):

(2) [*Indeed*]SP, [*the government*]T [*is taking a calculated risk*]R

(3) [*At the same time*]SP, [*the government*]T [*didn't want to appear to favor GM by allowing a minority stake* [*that*]T [*might preclude a full bid by Ford*]R]R

In more complex sentences, the algorithm does not accurately detect the propositions involved, triggering the following errors: (a) consecutive themes (even if the second one begins with a verb) (4); (b) consecutive rhemes (even if there is no theme and there is no verb in the first rheme) (5); (c) reduced clauses are not labelled as embedded rhemes (6).

(4) [*In a prepared statement*]SP, [*GM*]T [*suggested its plans for Jaguar*]T [*would be more valuable in the long run than the initial windfalls investors*]T [*migh reap from a hostile Ford bid*]R.

(5) [*Erwin Tomash, the 67-year-old founder of this maker of data communications products and a former chairman and chief executive*]R, [*resigned as a director*]R.

(6) [*In national over-the-counter trading*]R, [*SFE technologies shares*]T [*closed yesterday at 31.25 cents a share,* **up 6.25 cents**]R.

Another detected error is that just the verb is labelled as rheme, which also brings embededdness problems (7).

(7) " [*Our intensive discussions with Jaguar , at their invitation*]R , " [*GM said*]R , " [*have as their objectives to create a cooperative business relationship with Jaguar* [*that*]T [*would*]R *provide for the continued independence of this great British car company...*]R.

Apart from these major errors, a number of minor errors can be detected—e.g., subordinate conjunctions are assumed as part of the theme if they are initial (8); initial locative or temporal specifiers are always labelled as part of the rheme (9); initial specifiers are confused with themes (10); etc.

(8) [*Although GM*]T [*has U.S. approval to buy up to 15% of Jaguar's stock*]SP, [*it*]T [[*hasn't yet disclosed how many shares it now owns*]R.

(9) [*After a stronger - than - expected pace early this year*]R , [*analysts*]T [*say the market , after a series of sharp swings in recent months, now shows signs of retreating*].

(10) [*Under the circumstances*]T , *Dataproducts said* , [[*Mr. Tomash*]T [*said*]R]SP [*he*]T [*was unable to devote the time required because of other commitments*]R.

Finally, we found some few cases of non-annotated parts (*Dataproducts said* in (10)) or over-generated levels of embededness.

All these errors are likely to be straighten out with a larger training sample. The dependency curve between the accuracy ($y$-axis) and the size of the training set ($x$-axis) in Figure 4 shows that an increase of the size of the training set (e.g., to 1000 sentences) will further improve the scores. We are about to do this and apply the retrained transition-based analyser to the entire PTB. The information about how the resulting IS-bank can be accessed will be posted at http://www.taln.upf.edu/resources.

Our future work involves the extension of the annotation by other dimensions of the CommStr and a study of the correlation between the various dimensions of the IS and prosody. We assume that in particular the hierarchical structure of thematicity will correlate well with the prosodic structure of both simple and parenthetical or subordinate constructions, and thus contribute to a better quality in speech synthesis. Our positive experience with it in natural language text generation, where it guides syntactic realization, confirms this assumption.

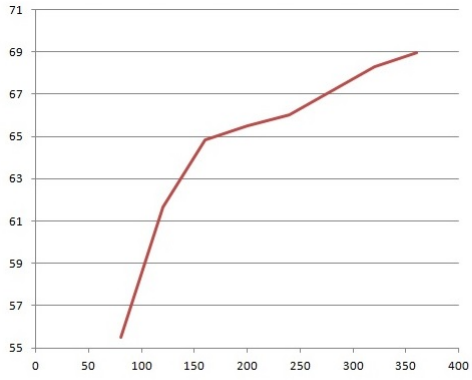


Figure 4: Dependency between the size of the training set and accuracy.

## References

S. Calhoun, M. Nissim, M. Steedman and J. Brenier 2005. A Framework for Annotating Information Structure in Discourse. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 45–52. Ann Arbor, MA: Association for Corpus Linguistics.

M. Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP 2002.*, pages 1–8.

M. Collins and B. Roark. 2004. Incremental parsing with the perceptron algorithm. In *Proceedings of ACL*, pages 112–119.

F. Daneš. 1970. Zur linguistischen Analyse der Textstruktur. *Folia Linguistica*, 4:72–78.

S. Dipper, M. Götze, M. Stede, and T. Wegst. 2004. ANNIS: A Linguistic Database for Exploring Information Structure. In S. Ishihara, M. Schmitz and A. Schwarz (eds.). *Working Papers of the SFB 632, Interdisciplinary Studies on Information Structure 1*, pages 245–279. Potsdam: University of Potsdam.

N. Erteschik-Shir. 2007. *Information Structure: The Syntax-Discourse Interface*. Oxford University Press, Oxford.

J. Firbas. 1966. Non-thematic Subjects in Contemporary English. *Travaux Linguistiques de Prague*, 2:229–236.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CONLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 2009 CoNLL Shared Task*, pages 1–18.

J. Hajič, J.Panevová, E. Hajicová, P. Sgall, P. Pajas, J. Štepánek, J. Havelka, M. Mikulová, Z. Žabokrtský. 2006. *Prague Dependency Treebank 2.0*. Charles University Prague.

E. Hajičová. 2007. The Position of TFA (Information Structure) in a Dependency Based Description of Language . In K. Gerdes, T. Reuther, and L. Wanner (eds.). *MTT 2007. Proceedings of the 3rd International Conference on Meaning-Text Theory*. Wiener Slawistischer Almanach, Sonderband 69, Munich & Vienna.

E. Hajičová, B. Partee, P. Sgall. 1998. Topic-Focus Articulation, Tripartite Structures, and Semantic Content. Kluwer Academic Publishers, Dordrecht.

I. Kruijff-Korbayová and M. Steedman. 2003. Discourse and Information Structure. *Journal of Logic, Language and Information*, 12(3):249–259.

K. Lambrecht. 1994. *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge University Press, Cambridge.

M.P. Marcus, B. Santorini and M.A. Marcinkiewicz. . 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

V. Mathesius. 1929. Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen*, 155:202–210.

I.A. Mel'čuk. 2001. *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Benjamins Academic Publishers, Amsterdam.

J. Nivre, J. Hall, and J. Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of CoNLL*, pages 49–56.

J. Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together (ACL)*, pages 50–57.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics. *Computational Linguistics*, 31(1):71–105.

P. Paggio 2006. Annotating Information Structure in a Corpus of Spoken Danish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1606–1609.

O. Postolache, I. Kruijff-Korbayová and G.-J. M. Kruijff 2005. Data-driven approaches for information structure identification. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 9–16.

P. Sgall. Functional Sentence Perspective in a Generative Description of Language. 1967. *Prague Studies in Mathematical Linguistics*, 2:203-225.

M. Steedman. Information Structure and the Syntax-Phonology Interface. 2000. *Linguistic Inquiry*, 31(4):649-685.

M. Taulé, M.A. Martí and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. Marrakesh, Morocco.

C. Thielen, A. Schiller, S. Teufel and C. Stöckert. Guidelines für das Tagging deutscher Textkorpora mit STTS. 1999. Institute for Natural Language Processing, University of Stuttgart. http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/.

E. Vallduví 1992. *Information Component*. Garland, New York and London.

H. Yamada and Y. Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of IWPT*, pages 195–206.

M. Zimmermann and C. Féry 2009. *Information Structure: Theoretical, Typological, and Experimental Perspectives*. Oxford University Press, Oxford.

Y. Zhang and J. Nivre. 2011. Transition-based parsing with rich non-local features. In *Proceedings of ACL*.