

# Data-driven Approaches for Information Structure Identification

Oana Postolache,  
Ivana Kruijff-Korbayová

University of Saarland,  
Saarbrücken, Germany

{oana,korbay}@coli.uni-saarland.de

Geert-Jan M. Kruijff

German Research Center for  
Artificial Intelligence (DFKI GmbH)  
Saarbrücken, Germany

gj@dfki.de

## Abstract

This paper investigates automatic identification of Information Structure (IS) in texts. The experiments use the Prague Dependency Treebank which is annotated with IS following the Praguian approach of Topic Focus Articulation. We automatically detect *t*(opic) and *f*(ocus), using node attributes from the treebank as basic features and derived features inspired by the annotation guidelines. We present the performance of decision trees (C4.5), maximum entropy, and rule induction (RIPPER) classifiers on all tectogrammatical nodes. We compare the results against a baseline system that always assigns *f*(ocus) and against a rule-based system. The best system achieves an accuracy of 90.69%, which is a 44.73% improvement over the baseline (62.66%).

## 1 Introduction

Information Structure (IS) is a partitioning of the content of a sentence according to its relation to the discourse context. There are numerous theoretical approaches describing IS and its semantics (Halliday, 1967; Sgall, 1967; Vallduví, 1990; Steedman, 2000) and the terminology used is diverse — see (Kruijff-Korbayová and Steedman, 2003) for an overview. However, all theories consider at least one of the following two distinctions: (i) a Topic/Focus<sup>1</sup> distinction that divides the linguistic meaning of the sentence into parts that link the sentence content

to the discourse context, and other parts that advance the discourse, i.e., add or modify information; and (ii) a background/kontrast<sup>2</sup> distinction between parts of the utterance which contribute to distinguishing its actual content from alternatives the context makes available.

Information Structure is an important factor in determining the felicity of a sentence in a given context. Applications in which IS is crucial are text-to-speech systems, where IS helps to improve the quality of the speech output (Prevost and Steedman, 1994; Kruijff-Korbayová et al., 2003; Moore et al., 2004), and machine translation, where IS improves target word order, especially that of free word order languages (Stys and Zemke, 1995).

Existing theories, however, state their principles using carefully selected illustrative examples. Because of this, they fail to adequately explain how different linguistic dimensions cooperate to realize Information Structure.

In this paper we describe data-driven, machine learning approaches for automatic identification of Information Structure; we describe what aspects of IS we deal with and report results of the performance of our systems and make an error analysis. For our experiments, we use the Prague Dependency Treebank (PDT) (Hajič, 1998). PDT follows the theory of Topic-Focus Articulation (Hajičová et al., 1998) and to date is the only corpus annotated with IS. Each node of the underlying structure of sentences in PDT is annotated with a TFA value: *t*(opic), differentiated in contrastive and non-contrastive, and *f*(ocus). Our system identifies these two TFA values automatically. We trained three different clas-

<sup>1</sup> We use the Praguian terminology for this distinction.

<sup>2</sup> The notion ‘kontrast’ with a ‘k’ has been introduced in (Vallduví and Vilks, 1998) to replace what Steedman calls ‘focus’, and to avoid confusion with other definitions of focus.

sifiers, C4.5, RIPPER and MaxEnt using basic features from the treebank and derived features inspired by the annotation guidelines. We evaluated the performance of the classifiers against a baseline system that simulates the preprocessing procedure that preceded the manual annotation of PDT, by always assigning *f(ocus)*, and against a rule-based system which we implemented following the annotation instructions. Our best system achieves a 90.69% accuracy, which is a 44.73% improvement over the baseline (62.66%).

The organization of the paper is as follows. Section 2 describes the Prague Dependency Treebank and the Praguian approach of Topic-Focus Articulation, from two perspectives: of the theoretical definition and of the annotation guidelines that have been followed to annotate the PDT. Section 3 presents our experiments, the data settings, results and error analysis. The paper closes with conclusions and issues for future research (Section 4).

## 2 Prague Dependency Treebank

The Prague Dependency Treebank (PDT) consists of newspaper articles from the Czech National Corpus (Čermák, 1997) and includes three layers of annotation:

1. The morphological layer gives a full morphemic analysis in which 13 categories are marked for all sentence tokens (including punctuation marks).
2. The analytical layer, on which the “surface” syntax (Hajič, 1998) is annotated, contains analytical tree structures, in which every token from the surface shape of the sentence has a corresponding node labeled with main syntactic functions like SUBJ, PRED, OBJ, ADV.
3. The tectogrammatical layer renders the deep (underlying) structure of the sentence (Sgall et al., 1986; Hajičová et al., 1998). Tectogrammatical tree structures (TGTSSs) contain nodes corresponding only to the autosemantic words of the sentence (e.g., no preposition nodes) and to deletions on the surface level; the condition of projectivity is obeyed, i.e., no crossing edges are allowed; each node of the tree is assigned a functor such as ACTOR, PATIENT, ADDRESSEE, ORIGIN, EFFECT, the repertoire

of which is very rich; elementary coreference links are annotated for pronouns.

### 2.1 Topic-Focus Articulation (TFA)

The tectogrammatical level of the PDT was motivated by the ever increasing need for large corpora to include not only morphological and syntactic information but also semantic and discourse-related phenomena. Thus, the tectogrammatical trees have been enriched with features indicating the information structure of sentences which is a means of showing their contextual potential.

In the Praguian approach to IS, the content of the sentence is divided into two parts: the Topic is “what the sentence is about” and the Focus represents the information asserted about the Topic. A prototypical declarative sentence asserts that its Focus holds (or does not hold) about its Topic: Focus(Topic) or not-Focus(Topic).

The TFA definition uses the distinction between Context-Bound (CB) and Non-Bound (NB) parts of the sentence. To distinguish which items are CB and which are NB, the question test is applied, (i.e., the question for which a given sentence is the appropriate answer is considered). In this framework, weak and zero pronouns and those items in the answer which reproduce expressions present in the question (or associated to those present) are CB. Other items are NB.

In example (1), (b) is the sentence under investigation, in which CB and NB items are marked. Sentence (a) is the context in which the sentence (b) is uttered, and sentence (c) is the question for which the sentence (b) is an appropriate answer:

- (1) (a) Tom and Mary both came to John’s party.  
 (b) John<sub>CB</sub> invited<sub>CB</sub> only<sub>NB</sub> her<sub>NB</sub>.  
 (c) Whom did John invite?

It should be noted that the CB/NB distinction is not equivalent to the given/new distinction, as the pronoun “her” is NB although the cognitive entity, Mary, has already been mentioned in the discourse (therefore is given).

The following rules determine which lexical items (CB or NB) belong to the Topic or to the Focus of the sentence (Hajičová et al., 1998; Hajičová and Sgall, 2001):

1. The main verb and any of its direct dependents belong to the Focus if they are NB;
2. Every item that does not depend directly on the main verb and is subordinated to a Focus element belongs to the Focus (where “subordinated to” is defined as the irreflexive transitive closure of “depend on”);
3. If the main verb and all its dependents are CB, then those dependents  $d_i$  of the verb which have subordinated items  $s_m$  that are NB are called ‘proxi foci’; the items  $s_m$  together with all items subordinated to them belong to the Focus ( $i, m > 1$ );
4. Every item not belonging to the Focus according to 1 – 3 belongs to the Topic.

Applying these rules for the sentence (b) in example (1) we find the Topic and the Focus of the sentence: [John invited]<sub>Topic</sub> [only her]<sub>Focus</sub>.

It is worth mentioning that although most of the time, CB items belong to the Topic and NB items belong to the Focus (as it happens in our example too), there may be cases when the Focus contains some NB items and/or the Topic contains some CB items. Figure 1 shows such configurations: in the top-left corner the tectogrammatical representation of sentence (1) (b) is presented together with its Topic-Focus partitioning. The other three configurations are other possible tectogrammatical trees with their Topic-Focus partitionings; the top-right one corresponds to the example (2), the bottom-left to (3), and bottom-right to (4).

- (2) Q: Which teacher did Tom meet?  
A: Tom<sub>CB</sub> met<sub>CB</sub> the teacher<sub>CB</sub> of chemistry<sub>NB</sub>.
- (3) Q: What did he think about the teachers?  
A: He<sub>CB</sub> liked<sub>NB</sub> the teacher<sub>CB</sub> of chemistry<sub>NB</sub>.
- (4) Q: What did the teachers do?  
A: The teacher<sub>CB</sub> of chemistry<sub>NB</sub> met<sub>NB</sub> his<sub>CB</sub> pupils<sub>NB</sub>.

## 2.2 TFA annotation

Within PDT, the TFA attribute has been annotated for all nodes (including the restored ones) from the tectogrammatical level. Instructions for the assignment of the TFA attribute have been specified in

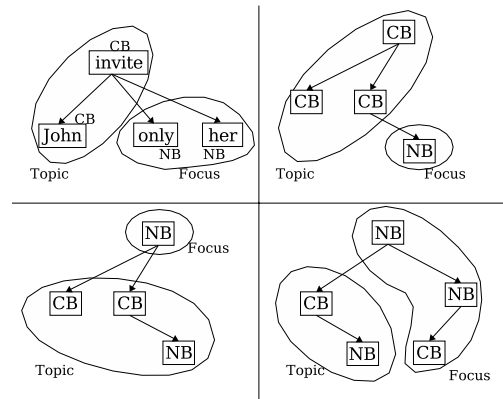


Figure 1: Topic-Focus partitionings of tectogrammatical trees.

(Buráňová et al., 2000) and are summarized in Table 1. These instructions are based on the surface word order, the position of the sentence stress (intonation center – IC)<sup>3</sup> and the canonical order of the dependents.

The TFA attribute has three values:

1. t — for non-contrastive CB items;
2. f — for NB items;
3. c — for contrastive CB items.

In this paper, we do not distinguish between contrastive and non-contrastive items, considering both of them as being just t. In the PDT annotation, the notation t (from topic) and f (from focus) was chosen to be used because, as we mentioned earlier, in the most common cases and in prototypical sentences, t-items belong to the Topic and f-items to the Focus.

Prior the manual annotation, the PDT corpus was preprocessed to mark all nodes with the TFA attribute of f, as it is the most common value. Then the annotators corrected the value according to the guidelines in Table 1.

Figure 2 illustrates the tectogrammatical tree structure of the following sentence:

- (5) Sebevědomím vtroků to ale neotřáslo.  
self-confidence bastards it but not shake  
'But it did not shake the self-confidence of those bastards'.

<sup>3</sup> In the PDT the intonation center is not annotated. However, the annotators were instructed to use their judgement where the IC would be if they uttered the sentence.

|    |   |   |
|----|---|---|
| 1. | The bearer of the IC (typically, the rightmost child of the verb)   | f |
| 2. | If IC is not on the rightmost child, everything after IC  | t |
| 3. | A left-side child of the verb (unless it carries IC)  | t |
| 4. | The verb and the right children of the verb before the f-node (cf. 1) that are canonically ordered  | f |
| 5. | Embedded attributes (unless repeated or restored)   | f |
| 6. | Restored nodes  | t |
| 7. | Indexical expressions ( <i>já</i> I, <i>ty</i> you, <i>těd</i> now, <i>tady</i> here), weak pronouns, pronominal expressions with a general meaning ( <i>někdo</i> somebody, <i>jednou</i> once) (unless they carry IC) | t |
| 8. | Strong forms of pronouns not preceded by a preposition (unless they carry IC)   | t |

Table 1: Annotation guidelines; IC = Intonation Center.

Each node is labeled with the corresponding word’s lemma, the TFA attribute, and the functor attribute. For example, *votroků* has lemma *votrok*, the TFA attribute *f*, and the functor *APP* (appurtenance).

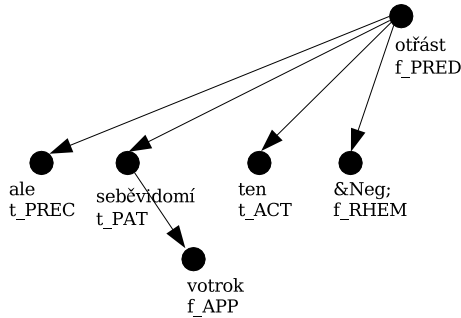


Figure 2: Tectogrammatical tree annotated with t/f.

In order to measure the consistency of the annotation, Interannotator Agreement has been measured (Veselá et al., 2004).<sup>4</sup> During the annotation process, there were four phases in which parallel annotations have been performed; a sample of data was chosen and annotated in parallel by three annotators.

| AGREEMENT | 1     | 2     | 3     | 4     | AVG   |
|-----------|-------|-------|-------|-------|-------|
| t/c/f     | 81.32 | 81.89 | 76.21 | 89.57 | 82.24 |
| t/f       | 85.42 | 83.94 | 84.18 | 92.15 | 86.42 |

Table 2: Interannotator Agreement for TFA assignment in PDT 2.0.

The agreement for each of the four phases, as well as an average agreement, is shown in Table 2. The second row of the table displays the percentage of nodes for which all three annotators assigned the

<sup>4</sup> In their paper the authors don’t give Kappa values, nor the complete information needed to compute a Kappa statistics ourselves.

same TFA value (be it t, c or f). Because in our experiments we do not differentiate between t and c, considering both as t, we computed, in the last row of the table, the agreement between the three annotators after replacing the TFA value c with t.<sup>5</sup>

### 3 Identification of topic and focus

In this section we present data-driven, machine learning approaches for automatic identification of Information Structure. For each tectogrammatical node we detect the TFA value t(topic) or f(ocus) (that is CB or NB). With these values one can apply the rules presented in Subsection 2.1 in order to find the Topic-Focus partitioning of each sentence.

#### 3.1 Experimental settings

Our experiments use the tectogrammatical trees from The Prague Dependency Treebank 2.0.<sup>6</sup> Statistics of the experimental data are shown in Table 3.

Our goal is to automatically label the tectogrammatical nodes with topic or focus. We built machine learning models based on three different well known techniques, decision trees (C4.5), rule induction (RIPPER) and maximum entropy (MaxEnt), in order to find out which approach is the most suitable for our task. For C4.5 and RIPPER we use the Weka implementations (Witten and Frank, 2000) and for MaxEnt we use the openNLP package.<sup>7</sup>

<sup>5</sup> In (Veselá et al., 2004), the number of cases when the annotators disagreed when labeling t or c is reported; this allowed us to compute the t/f agreement, by disregarding this number.

<sup>6</sup> We are grateful to the researchers at the Charles University in Prague for providing us the data before the PDT 2.0 official release.

<sup>7</sup> <http://maxent.sourceforge.net/>

| PDT DATA     | TRAIN            | DEV             | EVAL            | TOTAL           |
|--------------|------------------|-----------------|-----------------|-----------------|
| #files       | 2,536<br>80%     | 316<br>10%      | 316<br>10%      | 3,168<br>100%   |
| #sentences   | 38,737<br>78.3%  | 5,228<br>10.6%  | 5,477<br>11.1%  | 49,442<br>100%  |
| #tokens      | 652,700<br>78.3% | 87,988<br>10.6% | 92,669<br>11.1% | 833,356<br>100% |
| #tecto-nodes | 494,759<br>78.3% | 66,711<br>10.5% | 70,323<br>11.2% | 631,793<br>100% |

Table 3: PDT data: Statistics for the training, development and evaluation sets.

All our models use the same set of 35 features (presented in detail in Appendix A), divided in two types:

1. Basic features, consisting of attributes of the tectogrammatical nodes whose values were taken directly from the treebank annotation. We used a total of 25 basic features, that may have between 2 and 61 values.
2. Derived features, inspired by the annotation guidelines. The derived features are computed using the dependency information from the tectogrammatical level of the treebank and the surface order of the words corresponding to the nodes.<sup>8</sup> We also used lists of forms of Czech pronouns that are used as weak pronouns, indexical expressions, pronouns with general meaning, or strong pronouns. All the derived features have boolean values.

### 3.2 Results

The classifiers were trained on 494,759 instances (78.3%) (cf. Table 3) (tectogrammatical nodes) from the training set. The performance of the classifiers was evaluated on 70,323 instances (11.2%) from the evaluation set. We compared our models against a baseline system that assigns focus to all nodes (as it is the most common value) and against a deterministic, rule-based system, that implements the instructions from the annotation guidelines.

Table 4 shows the percentages of correctly classified instances for our models. We also performed a

<sup>8</sup> In the tectogrammatical level in the PDT, the order of the nodes has been changed during the annotation process of the TFA attribute, so that all t items precede all f items. Our features use the surface order of the words corresponding to the nodes.

10-fold cross validation, which for C4.5 gives accuracy of 90.62%.

| BASILINE | RULE-BASED | C4.5  | RIPPER | MAXENT |
|----------|------------|-------|--------|--------|
| 62.66    | 58.92      | 90.69 | 88.46* | 88.97  |

Table 4: Correctly classified instances (the numbers are given as percentages). \*The RIPPER classifier was trained with only 40% of the training data.

The baseline value is considerably high due to the topic/focus distribution in the test set (a similar distribution characterizes the training set as well). The rule-based system performs very poorly, although it follows the guidelines according to which the data was annotated. This anomaly is due to the fact that the intonation center of the sentence, which plays a very important role in the annotation, is not marked in the corpus, thus the rule-based system doesn't have access to this information.

The results show that all three models perform much better than the baseline and the rule-based system. We used the  $\chi^2$  test to examine if the difference between the three classifiers is statistically significant. The C4.5 model significantly outperforms the MaxEnt model ( $\chi^2 = 113.9$ ,  $p < 0.001$ ) and the MaxEnt model significantly outperforms the RIPPER model although with a lower level of confidence ( $\chi^2 = 9.1$ ,  $p < 0.01$ ).

The top of the decision tree generated by C4.5 in the training phase looks like this:

```

coref = true
|   is_member = true
|   |   POS = ...
|   is_member = false
|   |   is_rightmost = ...
coref = false
|   is_generated = true
|   |   nodetype = ...
|   is_generated = false
|   |   iterativeness = ...

```

It is worth mentioning that the RIPPER classifier was built with only 40% of the training set (with more data, the system crashes due to insufficient memory). Interestingly and quite surprisingly, the values of all three classifiers are actually greater than the interannotator agreement which has an average of 86.42%.

What is the cause of the classifiers' success? How come that they perform better than the annotators themselves? Is it because they take advantage of a

large amount of training data? To answer this question we have computed the learning curves. They are shown in the figure 3, which shows that, actually, after using only 1% of the training data (4,947 instances), the classifiers already perform very well, and adding more training data improves the results only slightly. On the other hand, for RIPPER, adding more data causes a decrease in performance, and as we mentioned earlier, even an impossibility of building a classifier.

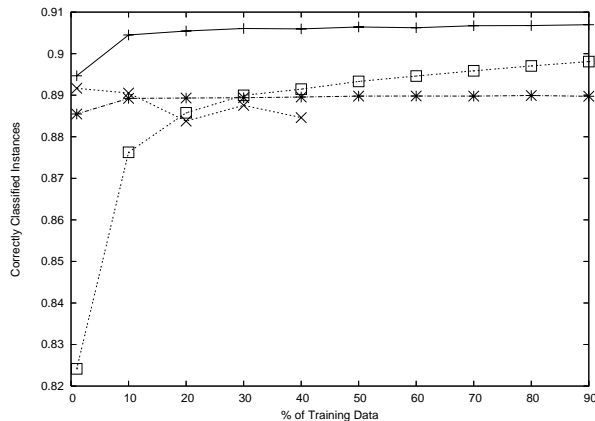


Figure 3: Learning curves for C4.5 (+), RIPPER( $\times$ ), MaxEnt( $*$ ) and a naïve predictor ( $\square$ ) (introduced in Section 3.3).

### 3.3 Error Analysis

If errors don't come from the lack of training data, then where do they come from? To answer this question we performed an error analysis. For each instance (tectogrammatical node), we considered its *context* as being the set of values for the features presented in Appendix A. Table 5 displays in the second column the number of all contexts. The last three columns divide the contexts in three groups:

1. Only t — all instances having these contexts are assigned t;
2. Only f — all instances having these contexts are assigned f;
3. Ambiguous — some instances that have these contexts are assigned t and some other are assigned f.

The last row of the table shows the number of instances for each type of context, in the training data.

|            | All             | Only t           | Only f          | Ambiguous         |
|------------|-----------------|------------------|-----------------|-------------------|
| #contexts  | 27,901          | 9,901            | 13,009          | 4,991             |
| #instances | 494,759<br>100% | 94,056<br>19.01% | 42,048<br>8.49% | 358,655<br>72.49% |

Table 5: Contexts & Instances in the training set.

Table 5 shows that the source of ambiguity (and therefore of errors) stays in 4,991 contexts that correspond to nodes that have been assigned both t and f. Moreover these contexts yield the largest amount of instances (72.49%). We investigated further these ambiguous contexts and we counted how many of them correspond to a set of nodes that are mostly assigned t ( $\#t > \#f$ ), respectively f ( $\#t < \#f$ ), and how many are highly ambiguous (half of the corresponding instances are assigned t and the other half f ( $\#t = \#f$ )). The numbers, shown in Table 6, suggest that in the training data there are 41,851 instances (8.45%) (the sum of highlighted numbers in the third row of the Table 6) that are exceptions, meaning they have contexts that usually correspond to instances that are assigned the other TFA value. There are two explanations for these exceptions: either they are part of the annotators disagreement, or they have some characteristics that our set of features fail to capture.

|                     | $\#t > \#f$   | $\#t = \#f$  | $\#t < \#f$  |
|---------------------|---|--|--|
| #ambiguous contexts | 998   | 833  | 3,155  |
| #instances          | t=50,722<br>f= <b>4,854</b><br>all=55,576<br>11.23% | t= <b>602</b><br>f= <b>602</b><br>all=1,204<br>0.24% | t= <b>35,793</b><br>f=266,082<br>all=301,875<br>61.01% |

Table 6: Ambiguous contexts in the training data.

The error analysis led us to the idea of implementing a naïve predictor. This predictor trains on the training set, and divides the contexts into five groups. Table 7 describes these five types of contexts and displays the TFA value assigned by the naïve predictor for each type.

If an instance has a context of type  $\#t = \#f$ , we decide to assign f because this is the most common value. Also, for the same reason, new contexts in the test set that don't appear in the training set are assigned f.

The performance of the naïve predictor on the evaluation set is 89.88% (correctly classified instances), a value which is significantly higher than

| Context Type | In the training set, instances with a context of this type are: | Predicted TFA value |
|--------------|---|---------------------|
| Only t       | all t   | t                   |
| Only f       | all f   | f                   |
| #t > #f      | more t than f   | t                   |
| #t = #f      | half t, half f  | f                   |
| #t < #f      | more f than t   | f                   |
| unseen       | not seen  | f                   |

Table 7: Naïve Predictor: its TFA prediction for each type of context.

the one obtained by the MaxEnt and RIPPER classifiers ( $\chi^2 = 30.7$ ,  $p < 0.001$  and respectively  $\chi^2 = 73.3$ ,  $p < 0.001$ ), and comparable with the C4.5 value, although the C4.5 classifier still performs significantly better ( $\chi^2 = 26.3$ ,  $p < 0.001$ ).

To find out whether the naïve predictor would improve if we added more data, we computed the learning curve, shown in Figure 3. Although the curve is slightly more abrupt than the ones of the other classifiers, we do not have enough evidence to believe that more data in the training set would bring a significant improvement. We calculated the number of new contexts in the development set, and although the number is high (2,043 contexts), they correspond to only 2,125 instances. This suggests that the new contexts that may appear are very rare, therefore they cannot yield a big improvement.

## 4 Conclusions

In this paper we investigated the problem of learning Information Structure from annotated data. The contribution of this research is to show for the first time that IS can be successfully recovered using mostly syntactic features. We used the Prague Dependency Treebank which is annotated with Information Structure following the Praguian theory of Topic Focus Articulation. The results show that we can reliably identify t(opic) and f(ocus) with over 90% accuracy while the baseline is at 62%.

Issues for further research include, on the one hand, a deeper investigation of the Topic-Focus Articulation in the Prague Dependency Treebank of Czech, by improving the feature set, considering also the distinction between contrastive and non-contrastive t items and, most importantly, by investigating how we can use the t/f annotation in PDT (and respectively our results) in order to detect the

Topic/Focus partitioning of the whole sentence.

We also want to benefit from our experience with the Czech data in order to create an English corpus annotated with Information Structure. We have already started to exploit a parallel English-Czech corpus, in order to transfer to the English version the topic/focus labels identified by our systems.

## References

- Eva Buráňová, Eva Hajičová, and Petr Sgall. 2000. Tagging of very large corpora: Topic-Focus Articulation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 139–144.
- Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of valency and Meaning. Studies in Honor of Jarmila Panevová*. Karolinum, Prague.
- Eva Hajičová and Petr Sgall. 2001. Topic-focus and saliency. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 268–273, Toulouse, France.
- Eva Hajičová, Barbara Partee, and Petr Sgall. 1998. Topic-focus articulation, tripartite structures, and semantic content. In *Studies in Linguistics and Philosophy*, number 71. Dordrecht: Kluwer.
- M. Halliday. 1967. Notes on transitivity and theme in english, part ii. *Journal of Linguistic*, (3):199–244.
- Ivana Kruijff-Korbayová and Mark Steedman. 2003. Discourse and Information Structure. *Journal of Logic, Language and Information*, (12):249–259.
- Ivana Kruijff-Korbayová, Stina Ericson, Kepa J. Rodrigues, and Elena Karagjosova. 2003. Producing Contextually Appropriate Intonation in an Information-State Based Dialog System. In *Proceeding of European Chapter of the Association for Computational Linguistics*, Budapest, Hungary.
- Johanna Moore, Mary Ellen Foster, Oliver Lemon, and Michael White. 2004. Generating Tailored, Comparative Description in Spoken Dialogue. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*.
- Scott Prevost and Mark Steedman. 1994. Information Based Intonation Synthesis. In *Proceedings of the ARPA Workshop on Human Language Technology*, Princeton, USA.
- Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.
- Petr Sgall. 1967. Functional sentence perspective in a generative description. *Prague Studies in Mathematical Linguistics*, (2):203–225.
- Mark Steedman. 2000. Information Structure and the syntax-phonology interface. *Linguistic Inquiry*, (34):649–689.
- Malgorzata Stys and Stefan Zemke. 1995. Incorporating Discourse Aspects in English-Polish MT: Towards Robust Implementation. In *Recent Advances in NLP*, Velingrad, Bulgaria.
- Enrich Vallduví and Maria Vilku. 1998. On rheme and kontrast. In P. Culicover and L. McNally, editors, *Syntax and Semantics Vol 29: The Limits of Syntax*. Academic Press, San Diego.
- Enrich Vallduví. 1990. *The information component*. Ph.D. thesis, University of Pennsylvania.
- František Čermák. 1997. Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics*, (2):181–197.
- Kateřina Veselá, Jiří Havelka, and Eva Hajičová. 2004. Annotators’ Agreement: The Case of Topic-Focus Articulation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2004)*.
- Ian H. Witten and Eibe Frank. 2000. *Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

## Appendix A

In this appendix we provide a full list of the feature names and the values they take (a feature for MaxEnt being a combination of the name, value and the prediction).

| BASIC FEATURE                   | POSSIBLE VALUES   |
|---------------------------------|---|
| nodetype                        | complex, atom, dphr, list, qcomplex   |
| is_generated                    | true, false   |
| functor                         | ACT, LOC, DENOM, APP, PAT, DIR1, MAT, RSTR, THL, TWHEN, REG, CPHR, COMPL, MEANS, ADDR, CRIT, TFHL, BEN, ORIG, DIR3, TTILL, TSIN, MANN, EFF, ID, CAUS, CPR, DPHR, AIM, EXT, ACOMP, THO, DIR2, RESTR, TPAR, PAR, COND, CNCS, DIFF, SUBS, AUTH, INTT, VOCAT, TOWH, ATT, RHEM, TFRWH, INTF, RESL, PREC, PRED, PARTL, HER, MOD, CONTRD |
| coref                           | true, false   |
| afun                            | Pred, Pnom, AuxV, Sb, Obj, Atr, Adv, AtrAdv, AdvAtr, Coord, AtrObj, ObjAtr, AtrAtr, AuxT, AuxR, AuxP, Apos, ExD, AuxC, Atv, AtvV, AuxO, AuxZ, AuxY, AuxG, AuxK, NA  |
| POS                             | N, A, R, V, D, C, P, J, T, Z, I, NA   |
| SUBPOS                          | NN, AA, NA, RR, VB, Db, Vp, C=, Dg, PD, Vf, J, J, P7, P4, PS, CI, TT, RV, PP, P8, Vs, Cr, AG, Cn, PL, PZ, Vc, AU, PH, Z:, PW, AC, NX, Ca, PQ, P5, PJ, Cv, PK, PE, P1, Vi, P9, A2, CC, P6, Cy, C?, RF, Co, Ve, II, Cd, Ch, J*, AM, Cw, AO, Vt, Vm  |
| is_member                       | true, false   |
| is_parenthesis                  | true, false   |
| sempos                          | n.denot, n.denot.neg, n.pron.def.demon, n.pron.def.pers, n.pron.indef, n.quant.def, adj.denot, adj.pron.def.demon, adj.pron.indef, adj.quant.def, adj.quant.indef, adj.quant.grad, adv.denot.grad.nneg, adv.denot.ngrad.nneg, adv.denot.grad.neg, adv.denot.ngrad.neg, adv.pron.def, adv.pron.indef, v, NA                        |
| number                          | sg, pl, inher, nr, NA   |
| gender                          | anim, inan, fem, neut, inher, nr, NA  |
| person                          | 1, 2, 3, inher, NA  |
| degcmp                          | pos, comp, acomp, sup, nr, NA   |
| verbmod                         | ind, imp, cdm, nr, NA   |
| aspect                          | proc, cpl, nr, NA   |
| tense                           | sim, ant, post, nil, NA   |
| numertype                       | basic, set, kind, ord, frac, NA   |
| indeftype                       | relat, indef1, indef2, indef3, indef4, indef5, indef6, inter, negat, total1, total2, NA   |
| negation                        | neg0, neg1, NA  |
| politeness                      | polite, basic, inher, NA  |
| deontmod                        | deb, hrt, vol, poss, perm, fac, decl, NA  |
| dispmod                         | disp1, disp0, nil, NA   |
| resultative                     | res1, res0, NA  |
| iterativeness                   | it1, it0, NA  |
| DERIVED FEATURE                 | POSSIBLE VALUES   |
| is_rightmost                    | true, false   |
| is_rightside_from_verb          | true, false   |
| is_leftside_dependent           | true, false   |
| is_embedded_attribute           | true, false   |
| has_repeated_lemma              | true, false   |
| is_in_canonical_order           | true, false   |
| is_weak_pronoun                 | true, false   |
| is_indexical_expression         | true, false   |
| is_pronoun_with_general_meaning | true, false   |
| is_strong_pronoun_with_no_prep  | true, false   |