The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 214-228


# Automatic evaluation of surface coherence in L2 texts in Czech

**Kateřina Rysová, Magdaléna Rysová, Jiří Mírovský**

Charles University in Prague, Czech Republic

Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics

{rysova|magdalena.rysova|mirovsky}@ufal.mff.cuni.cz

## Abstract

We introduce possibilities of automatic evaluation of surface text coherence (cohesion) in texts written by learners of Czech during certified exams for non-native speakers. On the basis of a corpus analysis, we focus on finding and describing relevant distinctive features for automatic detection of A1–C1 levels (established by CEFR – the Common European Framework of Reference for Languages) in terms of surface text coherence. The CEFR levels are evaluated by human assessors and we try to reach this assessment automatically by using several discourse features like frequency and diversity of discourse connectives, density of discourse relations etc. We present experiments with various features using two machine learning algorithms. Our results of automatic evaluation of CEFR coherence/cohesion marks (compared to human assessment) achieved 73.2% success rate for the detection of A1–C1 levels and 74.9% for the detection of A2–B2 levels.

## 1 Introduction

Our research is carried out on texts written during the international language examinations provided by the Test Centre of the Institute of Language and Preparatory Studies at the Charles University in Prague in line with the high ALTE (Association of Language Testers

in Europe) standards. Such type of examination is required by Czech universities (the needed CEFR level is usually B2) or often also by employers and the exam is compulsory for foreigners to be granted permanent residence in the Czech Republic (the required CEFR level is A1) or state citizenship (the required CEFR level is B1).[1] Therefore, it is of great importance to assess these examinations as objectively as possible and according to uniform criteria.

This is rather difficult because the writing samples are evaluated manually by human assessors (although according to the uniform rating grid) who naturally bring to the evaluation a subjective human factor. In the present paper, we aim at finding several objective criteria (concerning surface text coherence) for distinguishing the individual CEFR levels automatically. Specifically, we carry out a research on surface text coherence concerning various discourse phenomena (like the use and frequency of connectives etc.) and we test the possibility of their automatic monitoring and evaluating. The results of our research will become a part of a software application that will serve as a tool for objective assessment of surface text coherence, i.e. for automatic division of submitted writing samples into the suitable CEFR levels in the coherence/cohesion category.

## 2 Previous Research

There are many studies and projects dealing with automatic evaluation of various language phenomena especially for English. Many of them focus on grammatical aspects of language (e.g. on automatic evaluation of grammatical accuracy, detection of grammatical errors etc. – see [1]; [2] or [3]). On the other hand, only few of them aim at automatic evaluation of text coherence.

Text coherence may be viewed as local (in smaller text segments covering e.g. discourse relations between sentences within a paragraph) or global (coherence concerning larger text segments like correlation between a title and content etc.). Automatic evaluation of local

[1] Common European Framework of Reference for Languages (CEFR, the document of the Council of Europe) divides language learners into three broad categories (A: Basic user, B: Independent user, C: Proficient user). These categories may be further subdivided into six levels (A1, A2, B1, B2, C1 and C2).

coherence is a topic investigated e.g. by Miltsakaki and Kukich [4] analyzing student's essays or Lapata and Barzilay [5] focusing on machine-generated texts. Higgins et al. [6] examine possibilities of automatic assessment of both local and global coherence at once carried out on student's writing samples.

A specific topic of automatic evaluation of language is an analysis and assessment of L2 texts, i.e. (both written and spoken) texts by non-native speakers. Again there are many studies focusing especially on English (or languages like German or Dutch) as L2 and examining various aspects of language like automatic assessment of non-native prosody [7], automatic classification of article errors [8] or automatic detection of frequent pronunciation errors [9].

Whereas there is a number of studies focusing on automatic evaluation of texts written by non-native speakers for different languages, there is no similar research for Czech as L2/FL so far. Therefore, we open this topic for Czech by introducing automatic evaluation of surface text coherence, which has a clear potential for practical usage.

# 3 Text Coherence

There are many approaches to text coherence as well as capturing and monitoring coherence relations in large corpora, such as Rhetorical Structure Theory (RST, [10]), Segmented Discourse Representation Theory (SDRT, [11]) and the project Penn Discourse Treebank (PDTB, [12]). The PDTB approach inspired also the annotation of discourse in the Prague Dependency Treebank for Czech (PDT, [13]) – the only corpus of Czech marking relations of text coherence relations.

In this paper, we use the PDT way of capturing coherence relations. We focus on the aspects of surface coherence (cohesion), i.e. on the surface realization of coherence relations that may be processed automatically (like signalization of discourse relations by discourse connectives, distribution of inter- and intra-sentential discourse relations, distribution of semantico-pragmatic relations like contingency, expansion etc.).

# 4 Language Material: Corpus MERLIN

For our analysis, we use the language data of the corpus MERLIN [14][2] containing altogether 2,286 writing samples by non-native speakers (learners) of Czech, German and Italian.

German and Italian texts of the corpus were collected by TELC (The European Language Certificates) and Czech texts were provided by the Test Centre of the Institute of Language and Preparatory Studies at the Charles University in Prague. Both institutions (as full members of The Association of Language Testers in Europe (ALTE)) offer internationally recognized language exams in accordance with the high ALTE standards.

All texts forming the corpus MERLIN were created as out-puts of standardized tasks aligned to the Common European Framework of Reference for Languages (CEFR) – it means that all writing samples are evaluated across the CEFR levels, in the MERLIN case as A1–C1.[3]

The evaluation reflects both an overall level (general linguistic range) and the individual rating criteria including vocabulary range, vocabulary control, grammatical accuracy, surface coherence (cohesion), sociolinguistic appropriateness and orthography.

MERLIN uses two rating instruments: an assessor-oriented version of the holistic scale (see Alderson [15]) for the general linguistic range and an analytical rating grid closely related to CEFR rating table[4] used in the process of scaling the CEFR descriptors, see [16] and [17].

## 4.1 Sample of Learners' Texts

Example 1 demonstrates a Czech writing sample from the corpus MERLIN (the overall CEFR rating of this text is A2, i.e. basic user – elementary level):

(1) *Čau Martine,*

*Chci Tě zaprvé poděkovat že si mě pozval. Já ještě potřebuju ale vědet kdy to začíná? Abychom jsem mohl vědět kdy musím z domova odejít. Kdo ještě přijde, budou tam Tomáš a Lukáš, jestli ano, tak fajn. Budou tam tvoje rodiče, Radek chtěl vědět.*

[2] http://merlin-platform.eu/index.php

[3] Corpus MERLIN does not contain C2 texts at the moment.

[4] Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001)

*Uvidím tě později*
*David*

Literal translation into English:[5]

*Hello Martin,*
*First, I want to thank you that you have invited me. But I need to know when it begins? In order to know when I must leave my home. Who will come – Tomáš and Lukáš as well? If yes, it is fine. Your parents will be there? Radek wanted to know.*
*See you later*
*David*

The writing sample in Example 1 is provided with the MERLIN evaluation criteria presented in Table 1, i.e. with the assessments by the trained human evaluators.

Table 1: Evaluating table for the MERLIN writing sample in Example 1

| | |
|---|---|
| Overall CEFR rating | A2 |
| Grammatical accuracy | A2 |
| Orthography | B1 |
| Vocabulary range | A2 |
| Vocabulary control | A2 |
| Coherence/Cohesion | A2 |
| Sociolinguistic appropriateness | A1 |

## 4.2 Levels of Coherence in MERLIN

The writing sample in Example 1 was assigned A2 level for Coherence/Cohesion. Corpus MERLIN contains altogether 441 writing samples in Czech across the A1–C1 levels. Their distribution concerning Coherence/Cohesion is captured in Table 2.

[5] The original Czech text contains some errors in morphology and spelling that are not represented in the English translation.

Table 2: Distribution of Czech writing samples across CEFR levels of coherence in corpus MERLIN

| Coherence level | Number of texts |
|---|---|
| A1 | 1 |
| A2 | 102 |
| B1 | 172 |
| B2 | 157 |
| C1 | 9 |
| Total | 441 |

# 5 The Experiment

Our goal was to experimentally verify whether and to what extent the human annotation of the Coherence/Cohesion CEFR mark can be simulated by automatic methods. We tried to find possible distinctive criteria/features for automatic detection of the individual CEFR levels in this category.

## 5.1 Processing the Data

The first step was to parse the data (441 texts) from the raw text up to the deep syntactico-semantic (tectogrammatical) layer in the annotation framework of the Prague Dependency Treebank (PDT)[6] following the theoretical framework of the Functional Generative Description, see Sgall [18, 19]. To parse the data, we used the current version of Treex, a modular system for natural language processing [20], with a pre-defined scenario for Czech text analysis, which includes tokenization, sentence segmentation, morphological tagging,[7] surface

[6] The Prague Dependency Treebank [13] is a corpus of Czech newspaper texts (containing almost 50 thousand sentences) with a multi-layer annotation: morphological, surface syntactic and deep semantico-syntactic. On top of the dependency trees of the tectogrammatical layer, the PDT contains also manual annotation of discourse relations including annotation of discourse connectives.

[7] with recognition of unknown words (by heuristic guessing), which is very helpful for L2 texts with high number of typos

syntactic parsing and deep syntactic parsing.

On top of the automatically parsed dependency trees of the tectogrammatical layer, we automatically annotated explicit discourse relations (i.e. relations expressed by discourse connectives). As a theoretical background for capturing discourse relations in text, we employed the approach described in Poláková et al. [21] and used first in the annotation of the Prague Discourse Treebank 1.0 (PDiT; [22]) and later in the Prague Dependency Treebank 3.0 [13]. It is an approach similar to (and based on) the approach used for the annotation of the Penn Discourse Treebank 2.0 (PDTB; [12]). Both these approaches are lexically based and aim at capturing local discourse relations (between clauses, sentences, or short spans of texts), which is in accordance with our project and aims.[8]

For automatic annotation of intra-sentential discourse relations, we used a slightly modified algorithm originally designed by Jínová et al. [23] for a pre-annotation of intra-sentential discourse relations in the Prague Dependency Treebank. For automatic annotation of inter-sentential discourse relations, we devised and implemented an algorithm based on combining features from the automatically parsed deep-syntax dependency trees and lists of common Czech inter-sentential connectives and their most frequent discourse types (senses) extracted from the PDT using the query engine PML-Tree Query [24].

## 5.2 Features and Methods

To select features for automatic assessment of Coherence/Cohesion text levels, we first carried out a linguistic analysis of a couple of sample texts. Then we extracted (values of) these features from the automatically parsed texts. We established a relatively simple baseline and experimented with several other sets of features, as described below and summarized in Table 3.

The Baseline consists of a single feature that uses a list of 45 most frequent discourse connectives first extracted from the discourse annotation in the PDT 3.0 and complemented by a few informal variants that are likely to appear in texts written by non-native speakers

[8] If we aimed at evaluating the global coherence of texts, other theories would be more appropriate, such as the Rhetorical Structure Theory (RST; [10]), which tries to represent a document as a single tree expressing the hierarchy of discourse relations both between small and larger text segments.

(e.g. *teda* as an informal variant of *tedy* [*so, therefore*]). The feature counts number of occurrences of these connective words in the tested text, without trying to distinguish their connective and non-connective usages, and normalizes the count to 100 sentences. The Baseline is thus as follows:

- number of all connective words per 100 sentences

Another set of features – called Surface features – consists of features that only use tokenization and sentence segmentation. They do not use any advanced part of the text analysis such as syntactic parsing and discourse parsing. These features include also the baseline feature and all together are:

- number of all connective words per 100 sentences
- number of coordinating connective words per 100 sentences
- number of subordinating connective words per 100 sentences
- number of tokens per sentence

Other features extract information from the automatically parsed tree structures and from automatically annotated discourse relations. Together with the surface features they form a feature set called All features. Here is a list of the additional features:

- number of intra-sentential discourse relations per 100 sentences
- number of inter-sentential discourse relations per 100 sentences
- number of all discourse relations per 100 sentences
- number of different connectives in all discourse relations
- ratio of discourse relations with connective *a* [*and*]
- number of predicate-less sentences per 100 sentences
- ratio of discourse relations from class Temporal
- ratio of discourse relations from class Contingency
- ratio of discourse relations from class Contrast
- ratio of discourse relations from class Expansion

These three sets of features (Baseline, Surface, All) were predefined before the experiments with the machine learning methods. We also experimented with other sets of features (Set 1

Table 3: Various sets of features used in the experiments.

| Feature | Baseline | Surface | Set 1 | Set 2 | All |
|---|---|---|---|---|---|
| number of all connective words per 100 sentences | + | + | + | - | + |
| number of coordinating connective words per 100 s. | - | + | + | - | + |
| number of subordinating connective words per 100 s. | - | + | + | - | + |
| number of tokens per sentence | - | + | + | + | + |
| number of intra-sentential discourse relations per 100 s. | - | - | + | + | + |
| number of inter-sentential discourse relations per 100 s. | - | - | + | + | + |
| number of all discourse relations per 100 sentences | - | - | + | + | + |
| number of different connectives in all discourse relations | - | - | + | + | + |
| ratio of discourse relations with connective *a* [*and*] | - | - | + | + | + |
| number of predicate-less sentences per 100 sentences | - | - | - | - | + |
| ratio of discourse relations from class Temporal | - | - | - | - | + |
| ratio of discourse relations from class Contingency | - | - | - | - | + |
| ratio of discourse relations from class Contrast | - | - | - | - | + |
| ratio of discourse relations from class Expansion | - | - | - | - | + |

and Set 2 in Table 3), trying to find the best sets of features for the learning algorithms. As for selection of these features as well as for testing the algorithms with these features we used the 10-fold cross validation on all the data, results on these two sets may be slightly biased.

We used two machine learning algorithms – Random Forest and Multilayer Perceptron,[9] namely their implementation in the Waikato Environment for Knowledge Analysis – Weka toolkit [25].[10]

We trained and tested the algorithms with 10-fold cross validation on all the available data from the MERLIN corpus (441 instances), using the sets of features defined in Table 3.

[9] These two algorithms provided the best results among several other algorithms that we tried in the preliminary stage of the research; therefore, in the subsequent experiments, we used these two algorithms.

[10] Weka toolkit ver. 3.8.0, downloaded from http://www.cs.waikato.ac.nz/ml/weka/.

Table 4: Results of the experiments – accuracy, number of correctly classified instances and number of incorrectly classified instances. Statistically significant improvements over the respective baselines (tested with paired t-test) are marked with * for significance level 0.1 and ** for significance level 0.05.

| Experiment | Accuracy (%) | Correct | Incorrect |
|---|---|---|---|
| Random Forest, Baseline | 57.1 | 252 | 189 |
| Random Forest, Surface features | 62.6 | 276 | 165 |
| Random Forest, Set 1 | ** 73.0 | 322 | 119 |
| Random Forest, Set 2 | * 67.1 | 296 | 145 |
| Random Forest, All features | ** 70.3 | 310 | 131 |
| Multilayer Perceptron, Baseline | 60.8 | 268 | 173 |
| Multilayer Perceptron, Surface features | * 66.2 | 292 | 149 |
| Multilayer Perceptron, Set 1 | ** 71.9 | 317 | 124 |
| Multilayer Perceptron, Set 2 | ** 73.2 | 323 | 118 |
| Multilayer Perceptron, All features | * 68.0 | 300 | 141 |

As the data are relatively small, we chose the 10-fold cross validation instead of setting aside an evaluation test data, which in this case would be too small.

## 5.3 Results and Evaluation

Table 4 gives an overview of the performance of the two algorithms run with different feature sets.[11] The table gives the accuracy, i.e. the percentage of correctly classified instances, and also the absolute numbers of correctly and incorrectly classified instances. Statistically significant improvements over the baselines are marked with * for significance level 0.1 and ** for significance level 0.05.

[11] Please note again that feature sets Baseline, Surface and All were set beforehand, thus the results of the algorithms using these feature sets may be considered more reliable than for feature sets Set 1 and Set 2, which were defined by subsequent experimenting with the two algorithms in an attempt to find the best set of features for each of them (again using the 10-fold cross validation on all the data).

Table 5: Confusion matrix for Random Forest with Set 1 (classes in the rows classified as classes in the columns).

| | A1 | A2 | B1 | B2 | C1 |
|---|---|---|---|---|---|
| A1 | **0** | 1 | 0 | 0 | 0 |
| A2 | 0 | **56** | 36 | 10 | 0 |
| B1 | 0 | 25 | **123** | 24 | 0 |
| B2 | 0 | 2 | 12 | **143** | 0 |
| C1 | 0 | 0 | 0 | 9 | **0** |

Table 6: Confusion matrix for Multilayer Perceptron with Set 2 (classes in the rows classified as classes in the columns).

| | A1 | A2 | B1 | B2 | C1 |
|---|---|---|---|---|---|
| A1 | **0** | 1 | 0 | 0 | 0 |
| A2 | 0 | **67** | 31 | 4 | 0 |
| B1 | 0 | 29 | **120** | 23 | 0 |
| B2 | 0 | 3 | 18 | **136** | 0 |
| C1 | 0 | 0 | 0 | 9 | **0** |

The confusion matrix for the Random Forest algorithm run with features from Set 1 is given in Table 5. The confusion matrix for the Multilayer Perceptron algorithm run with features from Set 2 is given in Table 6. We can count from the tables that if we allow for "one level" error in the classification (i.e. for example if we consider classification A2 instead of B1 still correct), the accuracy of the algorithms is 97.3% and 98.4%.

The tables also demonstrate that the algorithms have never classified levels A1 and C1 correctly. The reason is that these levels are represented by very small numbers of texts in the corpus (1 writing sample of A1 and 9 of C1) and therefore they do not provide a sufficient language material for training. If the texts of A1 and C1 levels are excluded from the experiments, the succession rates for detection of A2/B1/B2 levels achieve slightly higher

results: Random Forest reaches 74.7% over Set 1 and Multilayer Perceptron 74.9% over Set 2. In this case, if we allow for "one level" error, the results are 97.2% for Random Forest and 98.4% for Multilayer Perceptron.

Linguistically, the experiments demonstrate that the most relevant features of surface coherence the (human or automatic) assessors should take into account are especially the following: frequency of connective words (expressing inter- or intra-sentential discourse relations such as *and*, *but*, *because*, *although* etc.); richness or variety of connective words (there is a difference between texts using almost exclusively the conjunction *and* and texts with a bigger diversity of connective words) and lexical richness of text spans (measured as word count per sentence).

# 6 Conclusion

In the paper, we have presented experiments on automatic evaluation of surface text coherence in writing samples by non-native speakers of Czech, more specifically on automatic detection of the individual CEFR levels. The main aim of our research was to examine to what extent the human assessment of surface text coherence can be simulated by automatic methods.

We have used several distinctive features concerning discourse and observed which combination of them reaches the best results for the two selected algorithms.

The algorithm Random Forest achieved the highest succession rate (73%) with Set 1 and the algorithm Multilayer Perceptron with Set 2 (73.2%). With "one level" error in the classification, the accuracy of the algorithms is 97.3% and 98.4%.

The experiments were carried out on the language data of the corpus MERLIN containing altogether 441 writing samples across A1–C1 levels of coherence. However, levels A1 and C1 are rather rare (1 text of A1 and 9 of C1). If we exclude these two levels from the experiments and focus only on detection of A2/B1/B2 levels, Random Forest reaches 74.7% of success rate over Set 1 and Multilayer Perceptron 74.9% over Set 2.

## Acknowledgment

The authors acknowledge support from the Ministry of Culture of the Czech Republic (project No. DG16P02B016 *Automatic evaluation of text coherence in Czech*).

This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

## References

[1] S. Bangalore, O. Rambow, and S. Whittaker, "Evaluation metrics for generation," in *Proceedings of the First International Conference on Natural Language Generation – Volume 14*. Morristown, AJ, USA: Association for Computational Linguistics, 2000, pp. 1–8.

[2] M. Chodorow and C. Leacock, "An unsupervised method for detecting grammatical errors," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL 2000)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 140–147.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Philadelphia, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[4] E. Miltsakaki and K. Kukich, "Evaluation of text coherence for electronic essay scoring systems," *Natural Language Engineering*, vol. 10, no. 1, pp. 25–55, 2004.

[5] M. Lapata and R. Barzilay, "Automatic Evaluation of Text Coherence: Models and Representations," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 5, Edinburgh, 2005, pp. 1085–1090.

[6] D. Higgins, J. Burstein, D. Marcu, and C. Gentile, "Evaluating multiple aspects of coherence in student essays." in *Proceedings of HLT–NAACL*, Boston, 2004, pp. 185–192.

[7] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic Assessment of Non-Native Prosody for English as L2," in *Proceedings of Speech Prosody*, vol. 100973, no. 1, Chicago, 2010, pp. 1–4.

[8] A. M. Pradhan, A. S. Varde, J. Peng, and E. Fitzpatrick, "Automatic Classification of Article Errors in L2 Written English," in *Twenty-Third International FLAIRS Conference*, Florida, USA, 2010.

[9] K. P. Truong, A. Neri, F. De Wet, C. Cucchiarini, and H. Strik, "Automatic detection of frequent pronunciation errors made by L2-learners," in *Proceedings of InterSpeech*, Lisbon, Portugal, 2005, pp. 1345–1348.

[10] W. C. Mann and S. A. Thompson, "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[11] N. Asher, *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer Academic Publishers, 1993.

[12] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse Treebank 2.0," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, Eds. Marrakech: European Language Resources Association, 2008, pp. 2961–2968.

[13] E. Bejček, E. Hajičová, J. Hajič, P. Jínová, V. Kettnerová, V. Kolářová, M. Mikulová, J. Mírovský, A. Nedoluzhko, J. Panevová, L. Poláková, M. Ševčíková, J. Štěpánek, and Š. Zikánová, "Prague Dependency Treebank 3.0," Data/software, Prague, 2013.

[14] A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Stindlová, and C. Vettori, "The MERLIN corpus: Learner language and the CEFR." in *Proceedings of LREC 2014*, 2014, pp. 1281–1288.

[15] J. C. Alderson, "Bands and scores," *Language testing in the 1990s*, pp. 71–86, 1991.

[16] B. North, "The CEFR levels and descriptor scales," in *Unpublished manuscript, from a paper presented at the 2nd International Conference of ALTE, Berlin, Germany*, 2005.

[17] ——, *The development of a common framework scale of language proficiency.* Peter Lang New York, USA, 2000.

[18] P. Sgall, "Generativní systémy v lingvistice [Generative systems in linguistics]," *Slovo a slovesnost*, vol. 25(4), no. 274–282, 1964.

[19] ——, *Generativní popis jazyka a česká deklinace [Generative Description of Language and Czech Declension].* Prague: Academia, 1967.

[20] Z. Žabokrtský, "Treex – an open-source framework for natural language processing," in *Information Technologies – Applications and Theory*, M. Lopatková, Ed., vol. 788. Košice, Slovakia: Univerzita Pavla Jozefa Šafárika v Košiciach, 2011, pp. 7–14.

[21] L. Poláková, J. Mírovský, A. Nedoluzhko, P. Jínová, Š. Zikánová, and E. Hajičová, "Introducing the Prague Discourse Treebank 1.0," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing.* Nagoya: Asian Federation of Natural Language Processing, 2013, pp. 91–99.

[22] L. Poláková, P. Jínová, Š. Zikánová, E. Hajičová, J. Mírovský, A. Nedoluzhko, M. Rysová, V. Pavlíková, J. Zdeňková, J. Pergler, and R. Ocelák, "Prague Discourse Treebank 1.0," Data/software, Prague, 2012.

[23] P. Jínová, J. Mírovský, and L. Poláková, "Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT," in *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, E. Hajičová, L. Poláková, and J. Mírovský, Eds. Bombay: Coling 2012 Organizing Committee, 2012, pp. 43–58.

[24] J. Štěpánek and P. Pajas, "Querying Diverse Treebanks in a Uniform Way," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010).* Valletta, Malta: European Language Resources Association, 2010, pp. 1828–1835.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.