Introducing EVALD – Software Applications for Automatic Evaluation of Discourse in Czech

Kateřina Rysová, Magdaléna Rysová, Jiří Mírovský, Michal Novák
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{rysova, magdalena.rysova, mirovsky, mnovak}@ufal.mff.cuni.cz

Abstract

In the paper, we introduce two software applications for automatic evaluation of coherence in Czech texts called EVALD – Evaluator of Discourse. The first one – EVALD 1.0 – evaluates texts written by native speakers of Czech on a five-step scale commonly used at Czech schools (grade 1 is the best, grade 5 is the worst). The second application is EVALD 1.0 for Foreigners assessing texts by non-native speakers of Czech using six-step scale (A1–C2) according to CEFR. Both applications are available online at https://lindat.mff.cuni.cz/services/evald-foreign/.

1 Introduction

Students of Czech often have problems with writing a comprehensive and continuous text. The reason is that creating texts is more demanding for them than creating separate sentences. The text is not simply “a cluster of sentences” but its structure has its own rules whose failure can result in the so called incoherent text, i.e. a text that is not fully functional in communication (see e.g. Halliday and Hasan, 1976). The ability of creating text should thus be encouraged already in the teaching process. At the same time, appropriate tools developed to assess such texts may reduce the amount of teachers’ manual work.

In this paper, we present the results of our investigation of automatic evaluation of texts in Czech (written by native and non-native speakers), more specifically, possibilities of automatic evaluation of text coherence. We present the linguistic features (concerning mainly discourse phenomena) that may be observed and evaluated automatically. Our experiments in this area resulted in a development of two software applications: Evaluator of Discourse 1.0 (EVALD 1.0) and Evaluator of Discourse 1.0 for Foreigners (EVALD 1.0 for Foreigners).

The EVALD applications can function as assistant tools for evaluation of essays in Czech[1] and they can be also used by students and learners who can easily verify their level of coherence in Czech.

EVALD applications are available as online services[2] and also as downloadable Docker containers.[3]

1.1 Czech as a Foreign Language – Assessing Criteria

The process of learning a foreign language is long-lasting and continuous. The learner goes through several stages from a beginner to a highly advanced language user. These stages (or phases of learning) are described in the document of the Council of Europe *Common European Framework of Reference for Languages* (CEFR). CEFR distinguishes six categories: A1 (basic language user – lower level), A2 (basic language user – higher level), B1 (independent language user – lower level), B2 (independent language user – higher level), C1 (proficient language user – lower level), C2 (proficient language user – higher level).

[1] The EVALD applications were created for assessing the coherence of authentic essays written by native and non-native speakers of Czech. In other words, they are trained to evaluate (prosaic) texts whose content and form (e.g. length) correspond to the common essays created as a comprehensive piece of writing on a given topic, e.g. during the Czech language exam. When evaluating a different type of text (e.g. too short texts, poems etc.), the software cannot work reliably because it is not trained for such text type.

[2] https://lindat.mff.cuni.cz/services/evald/index.php; https://lindat.mff.cuni.cz/services/evald-foreign/

[3] See http://ufal.mff.cuni.cz/evald/documentation for detailed installation instructions.

Proceedings of Recent Advances in Natural Language Processing, pages 634–641,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_082

Motivations behind learning foreign languages, including the languages of smaller countries, may be diverse. For example, learning Czech is useful for foreigners who want to study in the Czech Republic (most of the Czech universities require the CEFR level of B2). The knowledge of Czech is also 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). Therefore, it is of a great importance to assess these examinations as objectively as possible and according to uniform criteria.

This requirement is rather difficult to meet because the writing samples are evaluated only manually by human assessors (although according to the uniform rating grid) who naturally bring in a subjective human factor to the evaluation.

Therefore, we tried to find several objective criteria (concerning text coherence) for distinguishing the individual CEFR levels automatically. Specifically, we carried out research on text coherence concerning mainly various discourse phenomena (e.g., the use and frequency of discourse connective expressions) and we tested the possibility of their automatic monitoring and evaluation.

1.2 Czech of Native Speakers

Creating a coherent text that is fully functional in communication is not easy and obvious even for native speakers. Native speakers are also gradually learning to write a well-structured and comprehensive text.

At the same time, students' ability to create a coherent text is often examined at schools. For instance, writing an essay has been a compulsory part of the graduation examination at secondary schools in the Czech Republic for decades.

The essays by native speakers are not evaluated according to the CEFR levels, but according to a five-step scale commonly used at Czech schools (grade 1 is the best, grade 5 is the worst). EVALD 1.0 thus distinguishes 5 rating grades.

2 Previous Research

2.1 Text Coherence and Discourse

Text coherence (continuity) is a common property of each text that is fully functional in authentic communication. Linguistically, it is a complex phenomenon realized through various language aspects like semantico-pragmatic relations, coreference and anaphoric relations, lexical relations, substitution or ellipsis (see e.g. Dressler and de Beaugrande, 1972, 1981; Halliday and Hasan, 1976 or Hoey, 1979, 2001). In present-day linguistics, text coherence is thus often studied through language interactions (see e.g. Long and Chong, 2001; Camblin et al., 2007 or Hajičová, 2011).

This paper describes especially the automatically measurable aspects of semantico-pragmatic discourse relations (i.e. relations such as condition, opposition, reason, succession etc.) as one of the most important aspects of coherence. Discourse relations constitute the whole structure of a text and each text thus may be imagined as a net of semantico-pragmatic relations that are arranged hierarchically, i.e. the smallest units are linked to form higher units etc.

The right interpretation of semantico-pragmatic relations (which is a core of discourse analysis, see mainly Harris, 1952) then leads to the right interpretation of the whole text. To make this interpretation easier for the reader to comprehend, each language has its specific expressions that signal these types of relations explicitly – discourse connectives (cf. examples like *proto* "therefore", *avšak* "however", *v důsledku* "in consequence" etc.). Discourse connectives may be divided into two groups: primary and secondary (see Rysová and Rysová, 2014, 2015). In short, primary connectives are mostly grammaticalized expressions often consisting of a single word, e.g *když* "when", *protože* "because", *a* "and", *nebo* "or". On the other hand, secondary connectives are not yet fully grammaticalized, mainly multiword structures, e.g. *za podmínky, že* "on condition that", *v důsledku* "in consequence", *z tohoto důvodu* "for this reason" etc.

2.2 Studies on Automatic Evaluation of Coherence

Automatic evaluation of various language aspects (grammatical accuracy etc.) is studied in a number of projects (see e.g. Bangalore et al., 2000; Leacock and Chodorow, 2000 or Papineni et al., 2002). On the contrary, assessment of text coherence has been carried out so far rather manually by human assessors. Experiments on automatic evaluation of text coherence are thus relatively new in contemporary research. In this area, there exist only few studies (mainly for English) concerning various aspects of text coherence and cohesion.

In earlier studies, Foltz et al. (1998) and Wiemer-Hastings and Graesser (2000) have developed systems which examine text coherence in students' essays. Their systems focus on local coherence (i.e. on such aspects of coherence occurring in smaller sequence of sentence, mostly between adjacent sentences or within a single paragraph) and they measure lexical relatedness between text units by using vector-based similarity between adjacent sentences.

Their work relates in terms of similarity scoring to the TextTiling scheme (see Hearst and Plaunt, 1993 or Hearst, 1997) that may be used to recognize the subtopic of a text. Miltsakaki and Kukich (2000) also deal with text coherence in students' writing. They work with Rough Shift element of Centering Theory (Grosz et al., 1995) by examining the similarity of adjacent text units.

Possibilities of automatic evaluation of global coherence is studied by Higgins et al. (2004) focusing on four points: i) relatedness of a text to its topic, ii) relatedness to the thesis, iii) relatedness within a segment (i.e. each sentence in a text segment should be related to at least one other sentence within the segment), iv) grammar accuracy (a text is of a low coherence if it contains grammatical errors, incomplete sentences etc.).

To date, few attempts have been made to develop new methods for automatic evaluation of coherence in non-native (learners') texts, see e.g. Yannakoudakis and Briscoe (2012) investigating learners' coherence by monitoring part-of-speech distribution, number of (primary) discourse connectives (based on a fixed list) or word length.

Up to now, the research on automatic evaluation of students' essays (written by both native and non-native speakers) seems to be a relatively unexplored topic. In this paper, we thus aim to contribute to it by discussing new aspects of coherence that may be evaluated automatically (e.g. apart from primary connectives, we reflect also some types of secondary connectives, see Section 4.4).[4] At the same time, the EVALD applications are the first attempts to measure text coherence automatically on Czech data.

3 Language Material

Three corpora served as a source for basic linguistic research on coherence in Czech as well as for training of both software applications. Texts by the non-native speakers were obtained from the MERLIN corpus (Boyd et al., 2014) and CzeSL-SGT (Šebesta et al., 2014). Texts by the native speakers were taken from the corpus Skript2012 / AKCES 1 (Šebesta et al., 2016). EVALD 1.0 for Foreigners was trained on 945 texts in total and EVALD 1.0 on 1,118 texts.

4 Components of EVALD Applications

EVALD applications are based on supervised machine learning. Therefore, coherence must be first assessed on the text from the training corpora manually. Trained EVALD models then try to mimic this manual annotation using a set of linguistically motivated features. Many of the features take advantage of the linguistic information collected automatically on the texts during the pre-processing stage.

4.1 Assessment of the Texts by Linguists

Manual evaluation of texts from the corpora mentioned in Section 3 was performed by two trained evaluators.[5] The texts of non-native speakers were categorized into 6 classes differentiated by CEFR (A1, A2, B1, B2, C1 and C2) according to their coherence. The texts of native speakers were divided into 5 categories in accordance with the ratings used at Czech schools: 1 (excellent), 2 (very good), 3 (good), 4 (satisfactory), 5 (fail/unsatisfactory). The inter-annotator agreement (IAA) was measured on 100 texts of non-native speakers and on 100 texts of native speakers of Czech assessed (simultaneously) by the two evaluators. The exact IAA agreement reached 51% (on texts by non-native speakers) and 64% (on texts by native speakers). With tolerance of one level distance (e.g. evaluator 1: A1, evaluator 2: A2), the IAA agreement is 93% on texts by non-native speakers and 90% on texts by native speakers. Our IAA agreement is comparable to other similar projects, see e.g. Östling et al. (2013) reaching 45.8% of exact agreement among teachers evaluating 1,702 school essays in Swedish.

[4] The long-term investigation of discourse resulted in publication of the first discourse annotated corpus for Czech – the Prague Discourse Treebank, see the first version in Poláková et al, 2012 and the second one in Rysová et al., 2016.

[5] The evaluation of texts from the MERLIN corpus was taken directly from this corpus, as the texts from MERLIN already contained a reliable coherence evaluation according to CEFR.

4.2 Automatic Pre-processing of the Texts

A highly modular Treex processing system (Žabokrtský, 2011) was used for the automatic analysis of the text from its surface representation to deep syntactic dependency trees. The analysis pipeline for Czech comprises word tokenization, sentence splitting, part-of-speech tagging and lemmatization with MorphoDiTa (Straková et al. 2014), surface dependency parsing with MST parser adapted to Czech (Novák and Žabokrtský, 2007), and rule-based transition to deep syntactic trees. Such trees are then ready to be labeled with discourse-related annotation.

An algorithm designed by Jínová et al. (2012) was used to find intra-sentential discourse relations expressed by the primary connectives (e.g. *a* "and", *ale* "but", *protože* "because"). Inter-sentential discourse relations were addressed by our own method exploiting the list of inter-sentential primary connectives in Czech (compiled from the data of the Prague Discourse Treebank 2.0). Furthermore, an algorithm for automatic annotation of inter-sentential discourse relations expressed by secondary connectives containing pronominal anaphor (e.g. *díky tomu* "thanks to that", *kvůli tomu* "due to that", *kromě toho* "besides that") was created. It takes advantage of an adjusted version of the Treex Coreference Resolver (Bojar et al., 2012) module for demonstrative pronouns. The module does not seek an antecedent. Instead, it determines whether the pronoun refers to an entity, event or is non-anaphoric.

4.3 Feature Extraction Based on Linguistic Research

On the manually assessed texts, we carried out a comparative linguistic research, whose aim was to find language features that are distinctive for each level of text coherence. Our attention was paid to those features that are automatically detectable in the text. Based on this research, we have compiled a list of distinctive features, see Section 4.4.

For illustration, we present some partial results of this research phase on the example of connective expressions in the texts written by learners of Czech. Table 1 shows the distribution of these expressions[6] over A2, B1 and B2 categories in the MERLIN corpus, measured in terms of absolute and relative (per 100 sentences) frequencies.

The table reveals that B2 learners of Czech use markedly more connective expressions than learners at lower levels (A2 or B1). On the other hand, there is no difference between levels A2 and B1 in this respect.

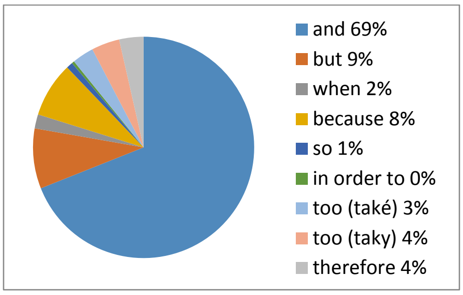
In addition, Graphs 1, 2 and 3 capture the distribution of the most frequently used connective expressions in the A2, B1 and B2 texts (the graphs are based on the values from Table 1). In all cases, the most common connective is *a* "and". However, the CEFR levels vary in the proportion of the connective *a* "and" among the other frequently used connective expressions. Basic language users (A2) use the connective *a* "and" substantially more often (compared to their use of other common connective expressions) than independent language users (B1, B2). By contrast, language users in categories B1 and B2 do not differ in this respect.

The examples above illustrate that the individual language features do not always distinguish all the coherence categories but they sometimes distinguish only some of them. It is thus interesting that some of the language phenomena are not improving evenly, when moving from A2 to B2. This observation demonstrates that the learning process of the individual language phenomena is not always linear. That is, it does not always follow the process of gradual improvement but some aspects of the second language are acquired rather "in jumps" by the learners. Similarly, we searched for further linguistic criteria according to which it would be possible to distinguish individual coherence categories of texts (written both by native and non-native speakers of Czech). Subsequently, the identified distinctive features (automatically detectable in texts) have been implemented in the EVALD software.

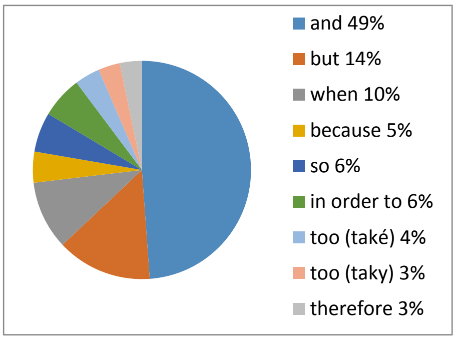
Most frequent connective means	**A2**		**B1**		**B2**	
	Tokens in 102 texts	**Tokens per 100 sentences**	**In 171 texts**	**Per 100 sent.**	**In 157 texts**	**Per 100 sent.**
a "and"	369	32	709	23	1,107	45
ale "but"	47	4	231	7	320	13
když "when"	11	1	66	2	230	9
protože "because"	43	4	136	4	104	4
tak "so"	5	0	64	2	133	5
aby "in order to"	2	0	35	1	140	6
také "too"	17	1	52	2	85	3
taky "too"	22	2	56	2	73	3
proto "therefore"	19	2	22	1	74	3
Total	**535**	**47**	**1,371**	**44**	**2,266**	**91**

Table 1: Most frequent connective expressions of A2, B1 and B2 CEFR levels in MERLIN corpus.

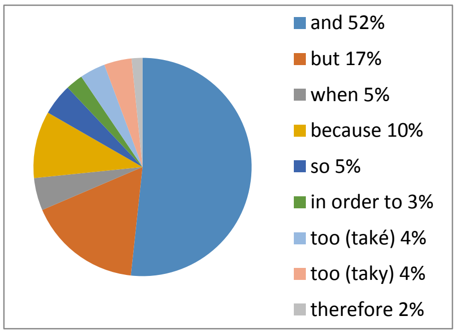
[6] The table captures the occurrence of the most frequent connective expressions, i.e. those that have more than 100 tokens in the MERLIN corpus texts.



Graph 1: Distribution of most frequent connective expressions in A2 of MERLIN texts.



Graph 2: Distribution of most frequent connective expressions in B1 of MERLIN texts.



Graph 3: Distribution of most frequent connective expressions in B2 of MERLIN texts.

4.4 Final List of Linguistic Features Evaluated by Both Software Applications

EVALD 1.0 and EVALD 1.0 for Foreigners work with the following language features.

Surface features consist of those using only tokenization and sentence segmentation, i.e. not any advanced part of the text analysis such as syntactic parsing and discourse parsing:

• number of all connective words per 100 sentences; • number of coordinating connective words per 100 sentences (e.g. *a* "and", *ale* "but", *nebo* "or"); • number of subordinating connective words per 100 sentences (e.g. *aby* "in order to", *když* "when", *protože* "because"); • number of tokens (words) per sentence (i.e. sentence length); • richness of the vocabulary (i.e. variety of lemmas).

Advanced features extract information from the automatically parsed tree structures and from automatically annotated discourse relations:

• 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 connectives *a* "and", *ale* "but", *protože* "because", *také* "too", *potom* "then", *pak* "then", *když* "when", *nebo* "or", *proto* "therefore", *tak* "so", *aby* "in order to", *totiž* "that is"; • ratio of inter-sentential discourse relations expressed by secondary connectives containing pronominal anaphor; • number of predicate-less sentences per 100 sentences (i.e. constructions without finite verbs like *Lovely!*); • ratio of discourse relations from class of Temporal, Contingency, Contrast and Expansion relations; • ratio of occurrence of the most common connective within all connectives in a text; • ratio of occurrence of the first and second most common connective within all connectives in a text.

4.5 Machine Learning Modelling

For machine learning experiments, the Random Forest algorithm implemented in WEKA toolkit (Hall et al., 2009) was chosen, as it provided the best results in the initial stages of the experiments among many other algorithms available in WEKA. Details on machine learning experiments and discussion on them can be found in Rysová et al. (2016).

5 Experiments and Evaluation

EVALD applications were trained and evaluated using 10-fold cross validation. In the first experiment, the whole datasets described in Section 3 were used (L1: 1,118 texts, L2: 945 texts) and for the evaluation, the F-score was measured, see Table 4 (the column F-Score).

Since the available data do not reflect the real distribution of the individual assessed classes in population, the data sets were uniformed for the second experiment – the instances in all classes were reduced to achieve a uniform distribution of them across the individual categories (L1: 475 texts, L2: 600 texts; more details are given in Novák et al., 2017). Confusion matrices are presented in Tables 2 and 3. The information for overall accuracy on these uniformed data sets is captured in the second column of Table 4.

EVALD 1.0		Assessed automatically				
		1	2	3	4	5
Assessed manually	1	84	1	1	9	0
	2	82	5	2	6	0
	3	53	9	6	27	0
	4	20	0	3	66	6
	5	2	0	0	35	58

Table 2: EVALD 1.0 – Confusion Matrix for Random Forest.

EVALD 1.0 for Foreigners		Assessed automatically					
		A1	A2	B1	B2	C1	C2
Assessed manually	A1	64	27	5	2	2	0
	A2	30	35	17	14	2	2
	B1	5	14	62	14	4	1
	B2	1	3	0	81	5	10
	C1	11	22	8	37	6	16
	C2	0	0	3	10	2	85

Table 3: EVALD 1.0 for Foreigners – Confusion Matrix for Random Forest.

As we can see in Tables 2 and 3, many of the errors are a result of misclassifying to a neighbouring class. Therefore, we measured not only the exact accuracy of the classifier, but also its accuracy with tolerance of "one-level" error (e.g. a human annotator classifies the text as B1 and EVALD as B2), see the third column of Table 4.

	Accuracy (Random Forest algorithm with 10-fold cross-validation method)		
	F-score on the whole data set	Exact accuracy on balanced data set	Accuracy with tolerance of "one-level" error on balanced data set
EVALD 1.0	**44.9**	**46.1**	**80.8**
EVALD 1.0 for Foreigners	**51.3**	**55.5**	**82.5**

Table 4: Accuracy of EVALD applications.

6 Discussion

The accuracy presented in Table 4 is an encouraging result because it is natural even among human evaluators (teachers) to hold the assessment within one level distance, see Section 4.1.

The EVALD accuracy is also comparable to other automatic systems developed for other languages. However, the comparison is rather difficult because the existing systems focus rather on evaluation of grammatical, lexical and semantic aspects than on text coherence (which is a complex language phenomenon more difficult to monitor than e.g. spelling, grammar errors etc.). At the same time, the other systems often do not evaluate all the existing levels (e.g. A1–C2).

For example, Vajjala and Lõo (2013) reach 79% accuracy in automatic evaluation of A2–C1 CEFR levels in Estonian using especially morphosyntactic and lexical features. Volodina et al. (2016) present 67% accuracy in evaluation of A1–C1 of CEFR levels in Swedish using count-based, lexical, syntactic, morphological, and semantic features. Hancke and Meurers (2013) demonstrate 62.7% accuracy of A1–C1 CEFR levels in German with syntactic, lexical and morphological features. Concerning automatic evaluation of native speakers' essays, Östling et al. (2013) reach 62.2% accuracy in automatic evaluation of high school essays written in Swedish, using especially grammatical and lexical features.

The research on EVALD applications is going to be further deepened by implementing features concerning other language phenomena. Currently, the applications are being enriched by features reflecting coreference and anaphora (see Novák et al., 2017). In the next step, the applications will also be enriched by features concerning sentence information structure.

7 Conclusion

In our paper, we have introduced two software applications that automatically estimate a coherence level of the text created by native or non-native speakers of Czech. Their accuracy achieves around 80% with one-level error tolerance. As far as we know, a similar tool for Czech has not yet been developed. The EVALD applications are unique especially in the way that they evaluate text coherence, which is not yet fully explored topic not only for Czech but also in international context.

Acknowledgement

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

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the INLG*, pages 1–8.

Robert-Alain de Beaugrande and Wolfgang Dressler. 1981. *Introduction to Text Linguistics.* London: Longman.

Ondřej Bojar, Zdeněk Žabokrtský et al. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).* European Language Resources Association, İstanbul, Turkey, pages 3921–3928.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association, Reykjavík, Iceland, pages 1281–1288.

Christine C. Camblin, Peter C. Gordon, and Tamara Y Swaab. 2007. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, 56(1):103–128.

Common European Framework of Reference for Languages. Language Policy Unit, Strasbourg. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf. [cite 2017-04-30]

Wolfgang Dressler. 1972. *Einführung in die Textlinguistik.* Tübingen: Niemeyer.

Peter Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3):285–307.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21(2): 203–225.

Eva Hajičová. 2011. On interplay of information structure, anaphoric links and discourse relations. In *Societas linguistica europaea SLE 2011, 44th Annual Meeting, Book of Abstracts*. Universidad de la Rioja, Center for Research in the Applications of Language, Logrono, Spain, pages 139–140.

Mark Hall. Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer. Peter Reutemann, Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1), pages 10–18.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research,* 54–56.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of ACM SIGIR*, pages 59–68.

Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Derrick Higgins et al. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the NAACL*, pages 185–192.

Michael Hoey. 1979. *Signalling in discourse*. Birmingham UK: English Language Research Unit, University of East Angllia.

Michael Hoey. 2001. *Textual Interaction: An Introduction to Written Discourse Analysis*. London: Routledge.

Pavlína Jínová, Jiří Mírovský, and Lucie Poláková. 2012. 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.* Organizing Committee, Mumbai, India, pages 43–58.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Claudia Leacock and Martin Chodorow. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 140–147.

Debra L. Long, D. L and Jennifer L. Chong. 2001. Comprehension skill and global coherence: A paradoxical picture of poor comprehenders abilities. *Journal of Experimental Psychology Learning, Memory and Cognition, 27*(14):24–1429.

MERLIN for CEFR-related language learning, teaching, and testing. 2014. WWW: <http://merlin-platform.eu/index.php>

Eleni Miltsakaki and Karen Kukich. 2000. Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000*, Athens, Greece.

Michal Novák, Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. 2017. Incorporating Coreference to Automatic Evaluation of Surface Coherence in Essays. In *Proceedings of the Statistical Language and Speech Processing (SLSP 2017).*

Václav Novák and Zdeněk Žabokrtský. 2007. Feature Engineering in Maximum Spanning Tree Dependency Parser. *Lecture Notes in Computer Science*, 4629(17):92–98. Springer, Berlin.

Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia.

Kishore Papineni et al. 2002. BLUE: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318.

Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Eva Hajičová, Jiří Mírovský, Anna Nedoluzhko, Magdaléna Rysová, Veronika Pavlíková, Jana Zdeňková, Jiří Pergler, and Radek Ocelák. 2012. *Prague Discourse Treebank 1.0.* Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdit/.

Kateřina Rysová, Jiří Mírovský, Michal Novák, and Magdaléna Rysová. 2016. *EVALD 1.0.* Data/software, ÚFAL MFF UK, Prague, Czechia, http://hdl.handle.net/11234/1-1820. [https://ufal.mff.cuni.cz/evald].

Kateřina Rysová, Jiří Mírovský, Michal Novák, and Magdaléna Rysová. 2016. *EVALD 1.0 for Foreigners.* Data/software, ÚFAL MFF UK, Prague, Czechia, http://hdl.handle.net/11234/1-1821. [https://ufal.mff.cuni.cz/evald/evald-10-foreigners].

Kateřina Rysová, Magdaléna Rysová, Jiří Mírovský. Automatic evaluation of surface coherence in L2 texts in Czech. 2016. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing ROCLING XXVIII*. 2016. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), pages 214–228.

Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. 2016. *Prague Discourse Treebank 2.0.* Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-1905.

Magdaléna Rysová and Kateřina Rysová. 2015. Secondary Connectives in the Prague Dependency Treebank. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala University, Uppsala, Sweden, pages 291–299.

Magdaléna Rysová and Kateřina Rysová. 2014. The Centre and Periphery of Discourse Connectives. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC 28).* Bangkok, Thailand: Department of Linguistics, Faculty of Arts, Chulalongkorn University, pages 452–459.

Jana Straková, Milan Straka and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, Baltimore, Maryland, pages 13–18.

Karel Šebesta, Zuzanna Bedřichová, Kateřina Šormová et al. 2014. *AKCES 5 (CzeSL-SGT).* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0023-95B1-E.

Karel Šebesta, Hana Goláňová, Jana Letafková et al., 2016. *AKCES 1.* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11234/1-1741.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. Proceedings of the third workshop on NLP for computer-assisted language learning. NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107, pages 113–127.

Elena Volodina, Ildikó Pilán, and David Alfter. 2016. Classification of Swedish learner essays by CEFR levels. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), CALL communities and culture – short papers from EUROCALL 2016, pages 456–461.

Peter Wiemer-Hastings and Arthur Graesser. 2000. Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8(2):149–169.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, pages 33–43.

Zdeněk Žabokrtský. 2011. Treex – an open-source framework for natural language processing. In *Information Technologies – Applications and Theory*. Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, pages 7–14.