# Automatic Text Analysis in Language Assessment: Developing a MultiDis Web Application

**Sara Košutar*, Dario Karl‡, Matea Kramarić*, Gordana Hržica***

* Faculty of Education and Rehabilitation Sciences, University of Zagreb
University Campus Borongaj, Borongajska cesta 83 f, 10 000 Zagreb
sara.kosutar@erf.unizg.hr
matea.kramaric@erf.unizg.hr
gordana.hrzica@erf.unizg.hr
‡Department of Data Science, InSky Solutions
Medačka ulica 18, 10 000 Zagreb
dario.karl.sl@gmail.com

**Abstract**

Language sample analysis provides rich information about the language abilities in the written or spoken text produced by a speaker in response to a language task. Language sample analysis is generally used to assess the abilities of children during language acquisition, but also the abilities of adult speakers across the lifespan. Its wide range of uses also allows for the assessment of language abilities in educational contexts such as second language acquisition or fluency, the abilities of bilingual speakers in general, and it is also used for diagnosis in speech and language pathology. Various computer programs have been developed to assist in the language sample analysis. However, these programs have been developed mainly for English and are often not fully open-access or do not provide data on population metrics, history of data uploaded by a user, and/or improvements in basic language measures. The time needed for transcription and the linguistic knowledge required for manual analysis are considered to be the main obstacles to its implementation The goal of this paper is to present a web-based application MultiDis intended for the analysis of language samples at the microstructural text level in Croatian. The application is still under development, but the current version fulfils its main purpose – it enables the (semi-) automatic calculation of measures reflecting language productivity, lexical diversity, syntactic complexity, and discourse cohesion in spoken language, and provides users with socio-demographic and linguistic metadata as well as the history of uploaded transcripts. We will present the challenges we have faced in developing the application (e.g., annotation system, text standardisation), future improvements we plan to make to the application (e.g., syntactic parsing, speech-to-text, multilingual analysis), and the possibilities of its use in the wider scientific and professional community.

## 1. Language sample analysis

Language sample analysis provides rich information about the language abilities in the written or spoken text produced by a speaker in response to a language task, e.g. storytelling, written essay, description of a picture, answering questions, etc. It is an ecologically valid means of language assessment that can be used along with standardised language tests because it provides data that tests cannot. Compared to standardised tests, language sample analysis has greater ecological validity because it reflects the natural everyday situation of language production. Consequently, it allows for a more in-depth analysis of specific morphosyntactic, semantic, and pragmatic features. Due to its lower bias, it proved to be more suitable for studying regional variations and dialects compared to standard questionnaires (e.g., Samardžić and Ljubešić, 2021). Language sample analysis is generally used to assess children's abilities during language acquisition, but also adult speakers' abilities across the lifespan (e.g., Westerveld et al., 2004). Its wide range of uses allows for the assessment of language abilities in educational contexts such as second language acquisition or fluency (e.g., Clercq and Housen, 2017), the abilities of bilingual speakers in general (e.g., Gagarina et al., 2016), and it is also used for diagnosis in speech and language pathology (e.g., Justice et al., 2006). This type of analysis is widely used in some countries, but in many countries, scientists and professionals are unaware of its benefits or find it too complex and time-consuming (see Heilmann, 2010; Klatte et al., 2022).

The process of collecting language samples involves several steps. First, a speaker is given a language task, for example telling a story based on a picture, and is recorded while performing this task. The recordings are then transcribed using special codes and are divided into smaller units of analysis, e.g., communication units (C-units; see Labov and Waletzky, 1967). Special codes mark different features of the spoken language or deviations (e.g., repetitions, omissions of vowels, use of regionally marked words, morphosyntactic errors, etc.). When written language samples are collected, the speaker responds to the task in writing, but all further steps are the same. Once the transcripts are produced, they can be analysed in various computer programs that enable (semi-)automatic calculation of different language measures.

Language sample analysis provides information about language abilities at two levels of text structure (Gagarina et al., 2012). First is the microstructural level, which refers to the internal linguistic organisation and includes text length, vocabulary use, morphosyntax, cohesive devices, etc. At the microstructural level, one can observe, for example, which language structures have emerged during language acquisition or how complex they are in terms of their internal features. The macrostructural analysis allows for assessing the ability of the hierarchical organisation of the text (e.g., in storytelling, whether the speaker has expressed a goal, an attempt, an outcome, etc.). At the macrostructural level, one can examine how successfully a speaker connects sentences according to a language task. By examining these elements, one gains insight into the quality of an individual's language when performing a particular language task, but also indirectly information on her or his language skills in general.

### 1.1. Language measures

Different aspects of microstructure correspond to several dimensions, such as productivity, lexical diversity, and syntactic complexity. A set of (semi-)automatic measures has been proposed to assess language abilities at the microstructural level. Productivity refers to the amount of language (words or utterances) produced (Leadholm and Miller, 1992). Measures of productivity include the total number of C-units or the total number of words (TNW). C-units are often used instead of utterances in spoken language analysis (see MacWhinney, 2000). The basic criteria for dividing a sequence of spoken words into utterances are intonation and pauses. However, transcribers may rate the utterances differently against these criteria, which results in lower inter-rater reliability (Stockman, 2010). C-units consist of one or more clauses. A clause is any syntactic unit consisting of at least one predicate. A complex sentence with one or more dependent clauses constitutes one C-unit, while a compound sentence is divided into two or more C-units, depending on the number of independent clauses. Studies have shown that measures of productivity can distinguish children with typical language status from children with developmental language disorders (DLD; Wetherell et al., 2007), bilingual from monolingual children (Hržica and Roch, 2021), and adult speakers according to their language skills (Nippold et al., 2017).

Measures of lexical diversity are used to assess vocabulary abilities. The more diverse the vocabulary produced, the greater the lexical diversity. Measuring lexical diversity is more complex and therefore methodologically challenging. Traditional measures include the number of different words (NDW; Miller, 1981) and the type-token ratio (TTR; Templin, 1957). Types and tokens can be easily calculated automatically, whereas lemmas are more difficult to calculate automatically, and require specialized natural language processing tasks. In particular, this requires morphological analyses such as lemmatisation, part-of-speech (POS) tagging, or morphological segmentation. In languages with rich morphology, the lemma-token ratio would be more appropriate, but due to the time-consuming nature of the task, this has rarely been done (see Balčiūnienė and Kornev, 2019). Another problem with measures of lexical diversity and measures of productivity is that they are affected by the length of a language sample (Malvern et al., 2004; McCarthy, 2005).

To overcome these limitations, alternative measures have been developed, such as D (Malvern and Richards, 1997) and moving average type-token ratio (MATTR; Covington and McFall, 2010). The measure D is based on modelling the decrease in TTR with the increasing size of the language sample using mathematical algorithms. MATTR calculates TTR for text windows of a fixed size, e.g., 500 words. The window moves through the text and calculates TTR for words 1-501, 2-502, etc. At the end of the text, all TTRs are averaged to determine the final score. However, it is not yet clear which of these measures provides more reliable results, as the results of validation studies vary (see deBoer, 2014; Fergadiotis et al., 2015). Regardless of methodological limitations, these measures can distinguish the abilities of children and adults with typical language status from children or adults with DLD (e.g., Hržica et al., 2019; Kapantzoglou et al., 2019). Measures of lexical diversity have also been found to correlate with standardised vocabulary tests in bilingual children (e.g., Hržica and Roch, 2021).

Syntactic complexity refers to the range of syntactic structures and the degree of sophistication of these structures in language production (Ortega, 2003). It is usually measured by calculating the average length of the C-unit. The length of the C-unit increases when there is a dependent clause or when the syntax within the clause is more complex, for example when the clause is extended by adding attributes, appositions, or adjectives. Measures of syntactic complexity have been shown to distinguish between different groups of speakers, including children with DLD and adults of different ages (e.g., Rice et al., 2010; Nippold et al., 2017). In addition to the average length of syntactic units, other measures of syntactic complexity include clausal density (i.e., the total number of main and subordinate clauses divided by the total number of C-units) and mean length of clause (main or subordinate), and they are also commonly used (e.g., Scott and Stokes, 1995; Norris and Ortega, 2009). Because of the variety of measures and the different methods of calculation, little is known about which measures are appropriate concerning typological differences between languages, and some of these measures are not always automatic.

In the last decades of the 20th century, various computer programs have been developed to support language sample analysis (overview: Pezold et al., 2020), but they are often not user-friendly. More recently, web-based programs have been introduced that allow for the analysis of language use at different linguistic levels (e.g., Coh-Metrix; McNamara et al., 2014). The measures are based on basic calculations (e.g., TTR, MLU), but there are also advanced measures based on language technologies such as the annotation of morphological, syntactic, and semantic features. Such applications are mainly developed for English or other widely spoken languages and are often not fully open-access. There is an increasing awareness of the importance of language sample analysis as a complementary method in language assessment. The time needed for transcription and the linguistic knowledge required for manual analysis are considered to be the main obstacles to its implementation (Pezold et al., 2020). Therefore, the development of a tool for the automatic calculation of language measures could make naturalistic language assessment more feasible.

## 2. Goal of the paper

The goal of this paper is to present a web-based application MultiDis, intended for the analysis of language samples at the microstructural level in Croatian, which enables the (semi-)automatic calculation of measures reflecting language productivity, lexical diversity, syntactic complexity, and discourse cohesion in spoken and written language. We will present the challenges we have faced in developing the application, future improvements we plan to make to the application, and the possibilities of its use in the wider scientific and professional community.

## 3. Development of the MultiDis web application

Existing computer-based resources used to analyse children's or adults' language abilities are either developed for English only or do not provide data on population

metrics, history of data uploaded by a user, and/or improvements in basic language measures such as NDW or TTR. The Computerized Language Analysis (CLAN; MacWhinney, 2000), for example, is a freely available desktop application whose users are expected to have a high level of language and transcription expertise. Text Inspector (2018), on the other hand, is a web-based application, but it is only designed for the text analysis of the English language and the target users are mainly first or second language acquisition teachers. We aim to develop a web-based application that fosters the analysis of language samples in Croatian. Our target users work at least partly with spoken language (e.g., language diagnostics performed by speech and language pathologists), so the application should support both written and spoken language analysis. The application is currently being developed, and we will present the coding system, language resources, data collection and language measures that have been implemented so far.

## 3.1. Annotation codes

Considering that our target users mostly work with spoken language, there are several codes which can be used to annotate the data. Computer programs for language analysis such as CLAN (MacWhinney, 2000) have an entire system of very specific annotation codes. In the MultiDis web application, a new and simpler system of annotation codes was developed to provide a faster and more organised annotation process. The system of the codes was designed to include several categories with individual codes and subsets of codes. The main idea is to have a system of annotation codes that can be changed over time according to the following criteria:

hierarchical (with categories and subcategories of codes)
extensible (adding new categories and codes)
easily customizable system (each category has a recognizable first character).

To date, the following categories have been established: *phonotactic codes* include conversation markers and elements of communication; *citation codes* indicate references to another utterance within the language sample; *phonetic codes* indicate pronunciation and other elements specific to spoken language; *sociolinguistic codes* indicate dialectisms, neologisms, foreign words, etc.; *correction codes* indicate errors made at a particular level of linguistic structure – phonological, morphosyntactic and/or lexical. There is also an additional code for corrections – a marker that can be used to exclude a particular segment from the transcript and provide a correct or standardised form that the application will use to standardise any text before moving on to a later stage of language analysis. A full description of codes is available on the web page of the application: http://www.multidis.com.hr/statistics/.

An example of multiple annotation codes would be a sentence in (1), that would look like (2) in the following uploaded transcript. Angle brackets point to a segment that needs to be excluded and round brackets point to a 'standardised' form of that segment. In addition, the @d code preceding the token *ćuko* 'dog' refers to a dialectism. The application will convert the sentence in (2) into the standardised form or the sentence as in (3), mapping the dialectism and providing this information in the final analysis report.

(1) *Dečko i ćuko su ulovili žabicu*. 'The boy and the dog caught the frog'
(2) *Dečko i <ćuko> (@d pas) su ulovili žabicu*. 'The boy and the dog caught the frog'
(3) *Dečko i pas su ulovili žabicu*. 'The boy and the dog caught the frog'

The annotation system and parsing rules for the transcripts were implemented using common Regular expressions (regex) in Python (Van Rossum, 2020). Regular expressions allow the system to recognise specific codes, save the data and convert the language into a standard form, so that existing language resources, such as tokenizers and lemmatizers, achieve a higher hit rate and precision. After annotation and parsing, the application will provide a standardised language text on which further language sample analysis is performed.

## 3.2. Language resources

The next step in the development of the application was the integration of an open-source Python library. We started with Stanza (Qi et al., 2020) to solve the following tasks common in natural language processing:

lemmatisation
POS tagging
syntactic parsing (sentence and clause segmentation).

In the early stages of developing the MultiDis web application, one of the main linguistic resources used was Stanza, a Python natural language processing toolkit for human language developed at Stanford University (Qi et al., 2020). Stanza enables quick out-of-the-box processing of multilingual texts. Since we plan to test our use case – based on the analysis of children's spoken language – on multiple languages, Stanza has an advantage over several other natural language processing models, frameworks and neural pipelines, such as Podium (Tutek et al., 2021), CLASSLA (Ljubešić and Dobrovoljc, 2019) or BERTić (Ljubešić and Lauc, 2021). Lemmatisation and POS tagging are fairly accurate (> 85 % of the cases), as they do not interfere with the computation of currently implemented language measures, though the process of delimiting the boundaries of C-units has been an obstacle that is currently being resolved. We are also exploring other options and planning further analysis and accuracy testing for this task. Since the language samples that the application will analyse are non-literary texts, we also plan to explicitly compare the aforementioned tools in the tasks of lemmatisation, POS tagging and morphosyntactic description (MSD) using our datasets to improve the application's baseline accuracy in these tasks. The standard for POS tagging is MulTextEast language resources (Erjavec, 2010), version 4 for the Croatian language. In this way, a token *ćuko* 'dog' is annotated as a dialectism using the annotation codes for the transcript parsing, and the standardised form *pas* 'dog' receives a morphosyntactic tag *Ncmsn* (nominative case, common noun, masculine, singular).

## 3.3. Data collection – manual annotation of transcripts with the new coding system

In the next step of developing the MultiDis web application, it was important to test the annotation system and the parsing of the language samples, as the aim was to obtain a standardised text with the data on the participants'

socio-demographic and language characteristics, parsed with the appropriate annotation codes and available to the user along with the morphosyntactic data. Before running the analysis, the texts were manually transcribed by students and volunteers within the courses *Computer Analysis of Child Language* and *Volunteering* at the Department of Speech and Language Pathology at the Faculty of Education and Rehabilitation Sciences, University of Zagreb. The test transcripts are the result of a storytelling task, mostly *Frog where are you?* (Mayer, 1969) and *Multilingual Assessment Instrument for Narratives* (MAIN; Gagarina et al., 2012; Gagarina et al., 2019; Hržica and Kuvač Kraljević, 2012, 2020). After the implementation of annotation codes, these transcripts have been successfully standardised and prepared for the final analysis. Any other transcript can be uploaded to the application and the user can only receive data about their uploaded transcripts and not about the transcripts of other users.

### 3.4. Automation of language measures

Using the standardised text and the provided language data from the previous step in the analysis, the next task of the MultiDis web application is to provide users with a detailed analysis of language measures. It is important to note that the measures are currently calculated intertextually, but we plan to compare the individual results with the population results, as well as with the baseline data. The application incorporates diverse measures that can be used in the language assessment such as productivity, lexical diversity, syntactic complexity and discourse cohesion. The list of language measures included in the MultiDis web application is available in Table 1.

| Category | Measure | Description |
|---|---|---|
| Language productivity | Number of communication units (NCU) | The total number of communication units |
| | Total number of words (TNW) | The total number of tokens (repeated tokens are excluded) |
| | Number of different words (NDW) | The total number of word forms – types |
| Lexical diversity | Type-token ratio (TTR) | The total number of tokens divided by the total number of types |
| | Index of lexical diversity D* | Based on the VOCD algorithm calculates the probability of the next token in a sequence based on an arbitrarily chosen *n*-token sample from the text |
| | Moving average type-token ratio (MATTR) | Based on a window length pre-defined by a user, the text is divided into segments and for each window length, the TTR is calculated – the average TTR ratio of each segment is the measure of MATTR |
| Syntactic complexity | Mean length of the communication unit | The total number of words is divided by the total number of communication units |
| | Clausal density | The total number of main and subordinate clauses is divided by the total number of communication units |
| | Mean length of clause | The total number of tokens is divided by the total number of clauses |
| Discourse cohesion | Ratio of connectives | The total number of connectives is divided by the total number of C-units. |

| | | |
|---|---|---|
| | Ratio of different connectives** | The total number of one type of connective is divided by the total number of all other types of connectives in the text |

Table 1: List of language measures implemented in the MultiDis web application (*being tested; **in the process of implementation).

The process of automatic analysis of language measures is based on precise segmentation of C-units and clauses, as well as on the results of tokenisation and lemmatisation. Each simple sentence (e.g., *The dog is playing with the frogs*), each complex sentence containing a subordinate clause or a parenthetical phrase (e.g., *When the dog chased the cat away, the birds were happy*), and each clause of a compound sentence was considered as one C-unit (e.g., *One goat is in the water and the other is grazing grass*). Given the fact that we need 100% accuracy on this task, at this stage, we are still in the process of developing an automatic way of detecting connectives in the text as well as clause delimiters. Thus, a user still has to manually divide the text into C-units following the above-mentioned criteria before uploading a language sample to the application. This also means that the user can change any automatically parsed C-unit. Collecting a larger amount of data will make it possible to train and apply an appropriate machine learning model to enable automatic segmentation of C-units and clauses.

At the current stage of developing the application, a user can obtain the results of all available language measures based on C-unit segmentation, as well as the morphosyntactic data and the data provided by the annotation codes. It is important to note that the MATTR measure does not have a fixed window length; instead, there is a default window size that contains 10% of the total number of tokens, and the user can manually adjust the window size. In this way, we have avoided the possibility for the results on MATTR to be the same as the results on TTR for language samples with less than 500 tokens, and we have allowed the user to define the best window size for this measure. Measure D and the number of different connectives are currently being implemented and tested before these results are made available to users. The remaining measures listed in Table 1 have been successfully implemented.

## 4. Technical specifications of the MultiDis web application

The MultiDis web application is deployed on the Croatian Academic and Research Network (CARNET) server as a monolithic Docker service. All requests are first forwarded to a Nginx service for the static files and only then to the application itself via a Gunicorn service (Python Web Server Interface Gateway HTTP Server). The application and the entire backend logic are written in the Python programming language (Van Rossum, 2020) within the Django web framework. All data is stored in a MySQL database instance on the server. As mentioned earlier, a Stanza PyTorch model (Qi et al., 2020) is run with the application to infer the language data and provide morphosyntactic information. Other open-source libraries and packages used are python-docx, NumPy and Pandas.

The application is designed so that each segment can be improved, without compromising our main goals or the user's experience. In this sense, we can also include written language samples and provide new annotation codes and categories for written language or implement measures that are only used in the analysis of adult language. Lemmatisation and POS tagging can be improved by replacing the existing model with a new, customized and open-source model that can be extended to languages other than Croatian.

## 5. Future extensions

The MultiDis web application is still under development, but the current version fulfils its main purpose – it allows for (semi-)automatic analysis of spoken language, and provides users with socio-demographic and linguistic metadata as well as the history of uploaded transcripts. In addition to the implementation of a service for the automatic determination of C-units and clause boundaries, additional data will be made available to users, such as the analysis of Croatian dialects and reference data for language measures, at least for some populations and some text types. Several other options are also being considered, such as fully automatic parsing of the original language sample without the manual annotation codes and an experimental speech-to-text service. As the tools and resources to develop this application are also available for other languages, the application could be scaled for multilingual analysis, preferably in collaboration with other researchers.

## 6. Conclusion

The MultiDis web application is freely available at http://www.multidis.com.hr/ and can be used by linguists, speech and language pathologists, teachers etc., to assess the language abilities of both children and adult speakers of Croatian. It can help clinicians and educators in language sample analysis by resolving some of the main obstacles to its use. A simpler coding system fosters transcription and future development of speech-to-text could ease this process even further. Automatic lemmatisation and morphological tagging save time and enable more precise calculation of language measures. The language measures included in the application were selected based on previous research and adequately reflect the different aspects of the participants' language abilities. Therefore, the MultiDis web application supports its users by reducing both the transcription time and the linguistic knowledge required to technically perform the analysis.

## 7. Acknowledgements

This work was supported by the Croatian Science Foundation under the project entitled *Multilevel approach to spoken discourse in language development*

(UIP-2017-05-6603), by the Arts and Humanities Research Council under the project entitled *Feast and Famine Project: Confronting Overabundance and Defectivity in Language* (AH/T002859/1) and by the COST Action under the project *NexusLinguarum – European network for Web-centred linguistic data science* (CA18209). Sara Košutar was supported by the project *Young Researchers' Career Development project – Training of New Doctoral Students*. Any opinions, findings, conclusions, or recommendations presented in this manuscript are those of the author(s) and do not necessarily reflect the views of the Croatian Science Foundation.

## 8. References

Ingrida Balčiūnienė and Aleksandr N. Kornev. 2019. Evaluation of narrative skills in language-impaired children. Advantages of a dynamic approach. In: E. Aguilar-Mediavilla, L. Buil-Legaz, R. López-Penadés, V. A. Sanchez-Azanza and D. Adrover-Roig, eds., *Atypical Language Development in Romance Languages*, pages 127–414. John Benjamins Publishing Company, Amsterdam and Philadelphia.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Bastien de Clercq and Alex Housen. 2017. A Cross-Linguistic Perspective on Syntactic Complexity in L2 Development: Syntactic Elaboration and Diversity. *The Modern Language Journal*, 101(2):315–334.

Fredrik deBoer. 2014. Evaluating the comparability of two measures of lexical diversity. *System*, 47:139–145.

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2544–2547, Valletta, Malta.

Gerasimos Fergadiotis, Heather Harris Wright and Samuel B. Greenc. 2015. Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech, Language, and Hearing Research*, 58(3):840–852.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ingrida Balčiūnienė, Ute Bohnacker, and Joe Walters. 2012. MAIN: Multilingual assessment instrument for narratives. *ZAS Papers in Linguistics*, 56:1–155.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, TainaVälimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: Multilingual Assessment Instrument for Narratives – Revised. *ZAS Papers in Linguistics*, 63:1–21.

Natalia Gagarina, Daleen Klop, Ianthi M. Tsimpli, and Joel Walters. 2016. Narrative abilities in bilingual children. *Applied Psycholinguistics*, 37(1):11–17.

John J. Heilmann. 2010. Myths and Realities of Language Sample Analysis. *Perspectives on Language Learning and Education*, 17(1): 4–8.

Gordana Hržica, Sara Košutar, and Matea Kramarić. 2019. Rječnička raznolikost pisanih tekstova osoba s razvojnim jezičnim poremećajem [Lexical diversity in written texts of persons with developmental language disorder]. *Hrvatska Revija za Rehabilitacijska Istraživanja*, 55(2):14–30.

Gordana Hržica and Jelena Kuvač Kraljević. 2012. MAIN – hrvatska inačica: Višejezični instrument za ispitivanje pripovijedanja [MAIN – Croatian version: Multilingual Assessment Instrument for Narratives]. *ZAS papers in linguistics*, 56:201–218.

Gordana Hržica and Jelena Kuvač Kraljević. 2020. The Croatian adaptation of the Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics*, 64:37–44.

Gordana Hržica and Maja Roch. 2021. Lexical diversity in bilingual speakers of Croatian and Italian. In: S. Armon-Lotem and K. K. Grohmann, eds., *LITMUS in Action: Cross comparison studies across Europe*, pages 100–129. John Benjamins Publishing Company Trends in Language Acquisition Research (TILAR), Amsterdam.

Laura M. Justice, Ryan P. Bowles, Joan N. Kaderavek, Teresa A. Ukrainetz, Sarita L. Eisenberg, and Ronald B. Gillam. 2006. The Index of Narrative Microstructure: A Clinical Tool for Analyzing School-Age Children's Narrative Performances. *American Journal of Speech-Language Pathology*, 15(2):177–191.

Maria Kapantzoglou, Gerasimos Fergadiotis, and Alejandra Auza Buenavides. 2019. Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, 62(1):70–83.

Inge S. Klatte, Vera van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language Sample Analysis in Clinical Practice: Speech-Language Pathologists' Barriers, Facilitators, and Needs. *Language, speech, and hearing services in schools*, 53(1):1–16.

William Labov and Joshua Waletzky. 1967. Narrative analysis: Oral versions of personal experience. In: J. Helm, ed., *Essays on the verbal and visual arts*, pages 3–38. University of Washington Press, Seattle and London.

Barbara J. Leadholm and Jon F. Miller. 1992. *Language sample analysis: The Wisconsin guide*. Wisconsin State Department of Public Instruction, Madison.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs (3rd ed.)*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

David Malvern and Brian Richards. 1997. A new measure of lexical diversity. In: A. Ryan and A. Wray, eds., *Evolving models of language*, pages 58–71. Multilingual Matters, Clevedon.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language*

*Development. Quantification and Assessment*. Palgrave Macmillan, London.

Mercer Mayer (1969). *Frog, where are you?* Dial Press, New York.

Phillip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, University of Memphis.

Danielle S. McNamara, Arthur C. Graesser, Phillip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, New York.

Jon M. Miller. 1981. *Assessing language production in children: experimental procedures*. University Park Press, Baltimore.

Marilyn A. Nippold, Laura M. Vigeland, Megan W. Frantz-Kaspar, and Jeannene M. Ward-Lonergan. 2017. Language Sampling With Adolescents: Building a Normative Database With Fables. *American Journal of Speech-Language Pathology*, 26(3):908–920.

John M. Norris and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4):555–578.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied Linguistics*, 24(4): 492–518.

Mollee J. Pezold, Caitlin M. Imgrund, and Holly L. Storkel. 2020. Using Computer Programs for Language Sample Analysis. *Language, Speech, and Hearing Services in Schools*, 51(1):103–114.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Stroudsburg, PA. Association for Computational Linguistics.

Mabel L. Rice, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53(2):333–349.

Tanja Samardžić and Nikola Ljubešić. 2021. Data Collection and Representation for Similar Languages, Varieties and Dialects. In: M. Zampieri and P. Nakov, eds., *Similar Languages, Varieties, and Dialects: A Computational Perspective, Studies in Natural Language Processing*, pages 121–137, Cambridge University Press, Cambridge.

Cheryl M. Scott and Sharon L. Stokes. 1995. Measures of syntax in school-age children and adolescents. *Language, Speech, and Hearing Services in Schools*, 26(4):309–319.

Ida J. Stockman. 2010. Listener reliability in assigning utterance boundaries in children's spontaneous speech. *Applied Psycholinguistics*, 31(3):363–395.

Mildred C. Templin. 1957. *Certain language skills in children; their development and interrelationships*. University of Minnesota Press, Minneapolis.

Text Inspector. 2018. *Online lexis analysis tool at textinspector.com*

Martin Tutek, Filip Boltužić, Ivan Smoković, Mario Šaško, Silvije Škudar, Domagoj Pluščec, Marin Kačan, Dunja Vesinger, Mate Mijolović, and Jan Šnajder. 2021. *Podium: a framework-agnostic NLP preprocessing toolkit. GitHub repository*. https://github.com/TakeLab/podium

Guido Van Rossum. 2020. The Python Library Reference, release 3.8.2. Python Software Foundation. https://py.mit.edu/_static/spring21/library.pdf

Marleen F. Westerveld, Gail Gillon, and Jon F. Miller. 2004. Spoken language samples of New Zealand children in conversation and narration. *Advances in Speech Language Pathology*, 6(4):195–208.

Danielle Wetherell, Nicola Botting, and Gina Conti-Ramsden. 2007. Narrative in adolescent specific language impairment (SLI): a comparison with peers across two different narrative genres. *International journal of language & communication disorders*, 42(5):583–605.