sciendo
DOI 10.2478/jazcas-2021-0048

DESIGNING A CORPUS OF CZECH MONOLOGUES: ORATOR v2

MARIE KOPŘIVOVÁ – ZUZANA LAUBEOVÁ – DAVID LUKEŠ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

KOPŘIVOVÁ, Marie – LAUBEOVÁ, Zuzana – LUKEŠ, David: Designing a corpus of Czech monologues: ORATOR v2. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 520 – 530.

Abstract: ORATOR v2 is a new 1.5M word corpus of Czech monologues, delivered to a live audience in semi-formal to formal settings. It was designed to chart the space of naturally occurring monologues which can be obtained for corpus processing. As such, it aims for diversity but does not attempt any balancing of subcategories, recognizing that some types of data are inherently easier to obtain in high volume than others. The transcription guidelines and annotation tools employed are the same as other recent spoken corpora published by the CNC, which facilitates interesting comparisons between various types of spoken Czech. The present paper sketches out three case studies, comparing ORATOR to the informal conversations of ORTOFON v2 in terms of the frequencies of demonstratives and hesitations, as well as lexical richness.

Keywords: speech, corpus, monologue, Czech

1 INTRODUCTION

With regard to spoken language, the Czech National Corpus (CNC) has historically mainly focused on collecting recordings of multi-party conversations in an informal setting, among friends and family. These interactions are thematically unspecified and unprepared; throughout the years, they have been made available to the public in a long line of corpora, culminating in the ORTOFON corpus, whose version 2 was published at the end of 2020 [1].

At the same time, after a preliminary version 1 in 2019, the full version 2 of the ORATOR corpus was also released [2]. ORATOR marks a departure from the relatively narrow focus on informal spoken Czech: it contains recordings and transcripts of mostly semi-prepared monologues of various kinds, providing a window to the opposite side of the spectrum of spoken communication. Since both corpora adhere to the same transcription guidelines[1] and were lemmatized and morphologically tagged using the same system [3], we hope they will not only enable a wealth of comparative research into various registers of spoken Czech, but they will also make interpretation of the results exceptionally straightforward and reliable.

[1] Except for the phonetic transcription layer, which is absent in ORATOR.

2 MONOLOGUE: DEFINITION AND THE EXISTING CZECH CORPORA

Spoken communication is traditionally divided into monologue and dialogue. This kind of classification is based on the number of active speakers (subjects): one only establishes a monologue, more than one a dialogue. Hoffmannová [4] defines monologue as an uninterrupted continuous activity of one subject and points out that pure monologues are very rare. Monologue is always more or less dialogical, depending on the degree of focus on the recipient, and the same is true vice versa.

Within monologues, many different genres or text types can be distinguished. Müllerová [5] introduces e.g., the following: narration of a story or memories (often with description of places or persons), introduction to a discussion, lecture, ceremonial official speech, sermons, etc.

Of course, the ORATOR corpus is hardly the first corpus of Czech to include monologues. The first spoken corpus within the CNC project – the Prague Spoken Corpus (PSC) [6] – combines two types of documents: informal, unprepared dialogue, and a structured interview with open questions. In response to the questions, speakers usually produced extensive monologues, as befits an interview. The Brno Spoken Corpus (BSC) [7] has a similar design. The formalization of the question–answer sequence led to these parts of the PSC and BSC being branded as "formal". It is however a slightly different type of formality than that in the ORATOR corpus (see below for details).

A corpus which consists entirely of pure monologues is the (aptly named) MONOLOG corpus [8]. The recordings feature a prepared and mainly read out speech by professional speakers of the Czech Radio.

3 CRITERIA FOR INCLUSION

Compared to the above-mentioned corpora, what makes ORATOR stand out is its strong emphasis on collecting naturally occurring semi-prepared monologues, e.g., university lectures, as opposed to ones that are experimentally induced and/or fully read out. Many spoken corpora focus on recordings of lectures and seminars for pragmatic reasons, because of their relative obtainability and consistent quality, which makes them well-suited for automated processing and use in NLP or ASR. By contrast, the ORATOR corpus has a broader scope: it was created as an intentional exploration of the different types of monologues which occur in communication.

Data collection focused on communication situations which are specifically intended to stand on their own as monologues. These may later be followed by a dialogical part (e.g., a discussion after a lecture) which is, however, not included. Some monologues may be part of more complex situations, such as meetings. Nevertheless, it is always the case that one speaker speaks without interruption,

having been allocated space and time for his or her speech, and the monologue and the dialogical part are separate. Also included were sequences of monologues linked by a moderator's commentary, such as introductory speeches at the opening of an exhibition, as well as less typical monologues, such as yoga classes or workplace fire safety trainings.

No balancing criteria were set in advance, the aim was simply to create the most diverse corpus of monologues possible and to find out which types can be obtained. Certain types of communication cannot be made public for legal or ethical reasons, and some are not appropriate because they formally intertwine short spans of monologue and dialogue in such a way that disentangling them would make the entire structure collapse.

The following criteria (cf. [9] for more details) for inclusion of a candidate recording in the corpus were determined:

1. A self-contained stretch of **monologue** by a speaker who is informed in advance about the topic, occasion, time, and location of his speech. The speaker can use different levels of preparation, such as notes, projected presentations, photographs, etc. We originally excluded speeches which were entirely or partially read out. However, this would deprive us of some types of situations, part of which requires a precisely given form, for instance because it is also a legal act (e.g., a wedding ceremony). In the case of lectures, they can contain quotations which are usually read out. The preparation of a text intended for reading also has its specificities, which is why we ended up including a small minority of these recordings to complete the picture (18 in total).
2. The context can be described as **official**, at least to a degree, and speakers were appointed either due to their expertise (public lecture, professional training, etc.), institutional role (university lecture, mayor's speech, etc.) or social status within the group (e.g., during a wedding toast). In some cases, the asymmetry of communicative roles was strengthened by the presence of a moderator. However, in smaller groups, this difference was weakened, sometimes questions were asked during the speech, especially in training sessions.
3. **Liveness**: we selected situations in which the speaker addresses a group of listeners. These ranged from smaller professional or private groups (training, wedding), to larger communities (lectures, sermons) to completely public speeches (public gatherings). We mostly wanted to avoid pre-recorded speeches which could be edited or adjusted for a particular platform. Still, as in the case of reading, we broke this rule in a few cases (9 monologues recorded specifically for an internet audience) for the sake of diversity.

The speakers were consistently anonymized and no sociolinguistic categories were identified, apart from gender. Gender information was also used to generate

nicknames for the speakers, based on randomly selected surnames supplemented by a first name initial, e.g. *Tománková, O.*[2]

4 OVERVIEW AND STATISTICS

ORATOR v2 contains over 1.5M positions in 489 recordings from 2005–2019 by 468 different speakers (some short speeches connected by the moderator form a single document and, conversely, some long lectures are divided into several parts). The ratio of recordings made specifically for this corpus vs. those acquired from external sources is about 4:3, and their length ranges from 13 seconds to 49 minutes, for a total length just shy of 149 hours. Men dominate significantly in the corpus, accounting for 71% of the number of tokens and 69% of the speakers.

As for document-level metadata, we tried to provide multiple grouping perspectives, so as to help users find their way around the corpus. Firstly, the **situational frame**: speeches were divided into official (at exhibitions, graduations, wedding ceremonies), popularizing (lectures for the public), political, professional (training sessions) and scientific (university and conference lectures). Table 1 shows the number of positions and documents in the corpus broken down by frame. Clearly the official recordings, while relatively numerous, are mostly quite short, as can be expected from the examples above. In the popularizing, professional, and scientific frames, longer lectures dominate, accounting for 49% of recordings and 78% of positions.

Secondly, 12 **situation types** provide a more fine-grained categorization of the recordings. A breakdown with examples is given in Table 2.

Thirdly, **genre** was annotated following the categories used in the latest SYN series written corpora, starting with SYN2015 [10]. While not all categories are represented, the sample is still varied and allows for interesting comparisons between written and spoken texts within a given genre.

These main divisions are complemented with information about the intended audience (public vs. restricted) and a special field identifying fringe types of monologue which technically did not meet criteria for inclusion, but were included in small amounts for diversity (cf. some examples in Section 3).

5 COMPARISON WITH ORTOFON V2

In many ways, the monologues in ORATOR are a stepping stone between the spontaneity of informal dialogues and the level of preparedness of written texts. We

[2] This is intended to remind the corpus user that the recordings were made in relatively formal/public settings. By contrast, speakers featured in the private conversations of ORTOFON v2 are identified by randomly selected first names with a surname initial, e.g., *Aleš N.*

are, therefore, convinced they will form an interesting basis for comparative research. Three simple case studies are presented in the following subsections to give an idea of the possibilities: comparisons of demonstratives, hesitations, and lexical richness between ORATOR v2 and the ORTOFON v2 corpus, where the latter consists of informal conversations. We hypothesized there would be more demonstratives in ORTOFON (as conversations are more heavily context-embedded), more hesitations in ORATOR (speakers tend to avoid long stretches of silence in monologues, leading to a higher incidence of filled pauses), and higher lexical richness in ORATOR (since the monologues are mainly expository and information-heavy).

5.1 Demonstratives

The relative frequency of demonstratives[3] in ORATOR and ORTOFON is given in Table 3. It shows that demonstratives are slightly (1.2×) more common in ORTOFON, i.e., in informal spontaneous conversations. The most frequent demonstrative lemma is *ten* 'this', which covers 92% of all demonstrative occurrences in ORTOFON, but only 86% in ORATOR (though still at the top of the frequency list).

This raises the question, where did those 6 percentage points get redistributed to? Part of the answer might be towards "long" demonstratives[4] such as *takovýhle* 'such a one' or *tenhleten* 'this one'. As Table 3 indicates, with these, the situation is reversed: they are actually about 1.3× more common in ORATOR.

In this light, our conjecture that dialogues contain more demonstratives because of their context-embeddedness might need revisiting. "Long" demonstratives, reinforced by the use of morphemes such as *hle*, are actually the ones which retain strong semantics of co(n)textual reference, and might be motivated by the frequent use of props such as photographs or slides during monologues. By contrast, the most frequent word form in ORTOFON overall is *to*, which is formally part of the paradigm of the demonstrative *ten*, but often performs more of a connective function, especially when switching speakers in dialogue (*to jo* 'yes'). As there is no speaker switching in ORATOR, the frequency list is topped instead by *a* 'and', another connective, which is arguably more useful in monologues (it is also typically the most frequent word in corpora of written Czech). So the difference in the incidence of demonstratives between monologues and dialogues might have more to do with different discourse structuring patterns rather than varying levels of context-embeddedness.

5.2 Hesitations

The transcription of both corpora notes certain non-verbal sounds, including hesitations or filled pauses, transcribed as @ (short) or @@ (long). Functionally,

[3] Retrieved via the query `[tag="PD.*"]`.

[4] Retrieved via the query `[tag="PD.*" & word=".{5,}"]`.

hesitations are connectives: speakers use them to eliminate (silent) pauses and fill the time needed to think their next utterance through [11]. As for listeners, they tend to perceive them negatively, as parasitic filler sounds [12].

Looking at the comparison in Table 4, hesitations of both types are clearly much more common in monologues. In both corpora, they tend to co-occur with pauses and other connectives such as *a* 'and' or *že* 'that'. In fact, when combined, they dominate the category of linking devices in ORATOR by a large margin, at 30,393 i.p.m.: the most common word in this category is the conjunction *a*, at 25,824 i.p.m.

Hesitations often appear when speakers attempt to convey complex notions, struggle finding the right word, or experience stress and/or high cognitive load, which would explain their increased presence in formally constrained monologues as opposed to freeform informal conversations, especially since the proportion of hesitations is highest among lectures, especially scientific ones. Intriguingly enough, they also appear in read-out speech, though at a relatively low frequency.

The lowest overall frequencies, even lower than ORTOFON, are encountered in sermons and ceremonies, though it should be noted that as with read-out speech, these are small categories with little data, so any generalizations are tentative at best. Still, a possible cause might be that speakers try to conform to a higher standard on these occasions, or that they occur repeatedly, leading to a high degree of preparation, or perhaps individual speaker proficiency.

5.3 Lexical richness

Finally, we turn to lexical richness. A naive measure of lexical richness is the type-token ratio (TTR), which is, however, sensitive to text length, as is well-known. Therefore, we used two more sophisticated TTR-based measures: the moving-average TTR (MATTR) [13] and zTTR[5] [14]. While MATTR is calculated by sliding a fixed-size window over the text and averaging the obtained TTR values, zTTR gives the relative position of a text within a reference distribution of texts of similar length.

We ran data extraction under a variety of settings:

- window sizes of 100 or 500 tokens for MATTR
- journalistic texts or spoken dialogues as reference data for zTTR
- tallying word forms or lemmas
- per document or speaker in the document

The general shape of the results was similar across the board, so we only selected three representative examples (Figures 1–3), all word-form based, subdivided by corpus (ORATOR vs. ORTOFON) and target unit (document vs.

[5] We are grateful to Václav Cvrček for letting us use his Perl script and reference data to calculate zTTR values.

speaker in document). The left subplots show median MATTR/zTTR values with bootstrapped 99% confidence intervals computed via 10,000 iterations of Monte Carlo case resampling; the right subplots show probability density functions for the full distribution of MATTR/zTTR values in each corpus, computed via kernel density estimation.

First and foremost, what all figures clearly show is that ORATOR monologues tend to have higher lexical richness than ORTOFON dialogues, whichever way we slice them (by document or speaker). This is consistent with our expectations, based on the fact that the monologues are mostly expository – speakers are primarily trying to convey information and have a limited timeframe to do so. This makes them aim for information-dense speech, which favors increased lexical richness.

Another observation is that in the case of ORATOR, there is little difference when calculating TTR per document vs. per speaker: the density curves and confidence intervals for the medians are nearly identical, or at least overlap to a great extent (Figure 3). This makes sense: in ORATOR, documents mostly feature a single speaker, so there is little difference in the units for which TTR is calculated to begin with.

In the case of ORTOFON however, the per-speaker distributions are consistently shifted to the right, towards (slightly) greater lexical richness: a little in the case of MATTR in Figure 1 (though note that the confidence intervals for the medians do not overlap, so this looks like a reliable effect, though small), and some more with zTTR, especially in Figure 3. Since ORTOFON documents are dialogues, slicing them up by speaker actually does make a difference in the units, but the fact that it does have an effect on TTR was still somewhat surprising to us. It remains to be seen whether an underlying linguistic explanation can be uncovered, or whether this is a residual failure of both measures to compensate for different text lengths.

Turning our attention to Figure 2, we see that the per-speaker density curve for ORTOFON peaks at 0, which is a good sanity check: it shows that our sample has the same mean as the reference data extracted from another corpus of informal dialogues. The ORATOR distributions peaking above 0 is a further confirmation of the fact that semi-prepared monologues tend to be lexically richer than informal dialogues.

Finally, the journalistic zTTR was included because journalistic texts spread over a large, well-populated portion of the TTR landscape, which makes them useful as reference data for comparison across registers. Figure 3 shows that in general, the lexical richness of both dialogues and monologues is on the low end of the spectrum, with all four distribution curves squeezing almost entirely below 0, i.e., the average.

6 CONCLUSION

The ORATOR v2 corpus is freely available via the KonText search interface at https://korpus.cz/kontext; other types of access to the data can also be provided upon

request.[6] As we have sketched above, it presents many compelling research opportunities: fruitful comparisons can be drawn both within the corpus itself and with other corpora. ORTOFON v2 is an especially attractive option in this regard because the two corpora focus on opposing ends of the spoken Czech spectrum while sharing the same processing pipeline, which makes it less likely for researchers to be misled by spurious differences caused by arbitrary incompatibilities between corpora. In the previous sections, we have given a glimpse of the possible directions to explore, but these are obviously just a tip of the iceberg. We are looking forward to see what creative uses these resources will be put to by fellow linguists.

ACKNOWLEDGEMENTS

The design and compilation of CNC corpora is made possible by the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation.

Frame	Positions	Documents
popularizing	812,671	188
scientific	361,770	75
professional	178,229	52
official	164,013	164
political	18,966	10

Tab. 1. Number of positions and documents in ORATOR v2, broken down by situational frame

Situation type	Positions	Documents
lecture (academic, general public)	1,204,668	240
public assembly	63,891	24
meeting	49,451	19
tour (e.g. castle tour)	43,073	31
opening speech	34,644	64
introduction of a work of art	33,931	35
training (e.g. workplace safety)	32,438	11
instructions (e.g. yoga class)	26,176	12
celebratory address	17,483	20
ceremony (e.g. wedding)	12,658	14
sermon	9,087	8
closing speech	8,149	11

Tab. 2. Number of positions and documents in ORATOR v2, broken down by situation types

[6] Please use the form at https://korpus.cz/clarin/helpdesk to submit your request.

Type of demonstratives	i.p.m. in ORATOR v2	i.p.m. in ORTOFON v2
all	65,156.17	78,221.82
“long”	7,782.38	6,025.17

Tab. 3. Comparison of the relative frequency of different types of demonstratives (in instances per million)

(Sub)corpus	i.p.m. of @	i.p.m. of @@
ORTOFON v2	7,613	1,606
ORATOR v2	24,797	5,596
- read	5,939	2,124
- lectures	26,284	5,890
- scientific	35,329	8,986
- sermons	6,933	550
- ceremonies	6,320	1,185

Tab. 4. Relative frequency of short @ and long @@ hesitations in various (sub)corpora (in instances per million)

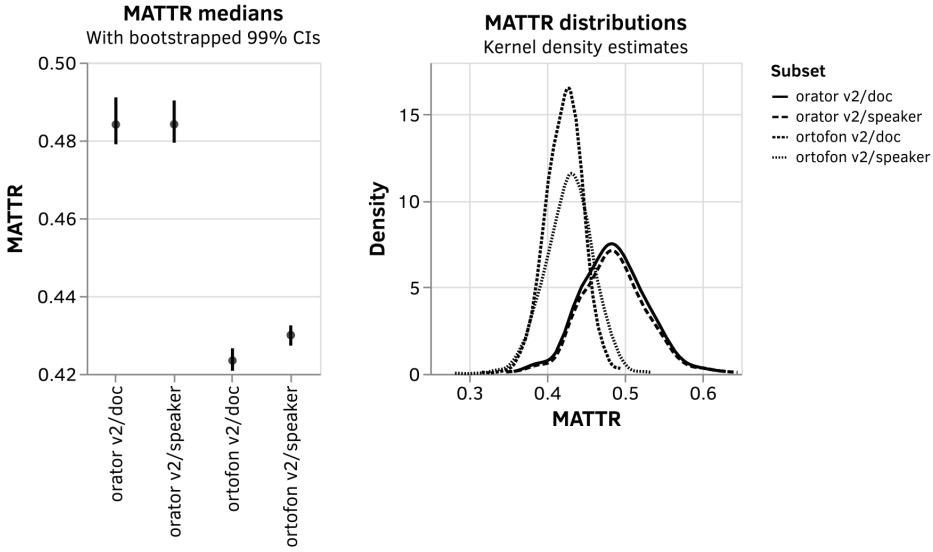

Fig. 1. MATTR computed on word forms, with a window of 500 tokens

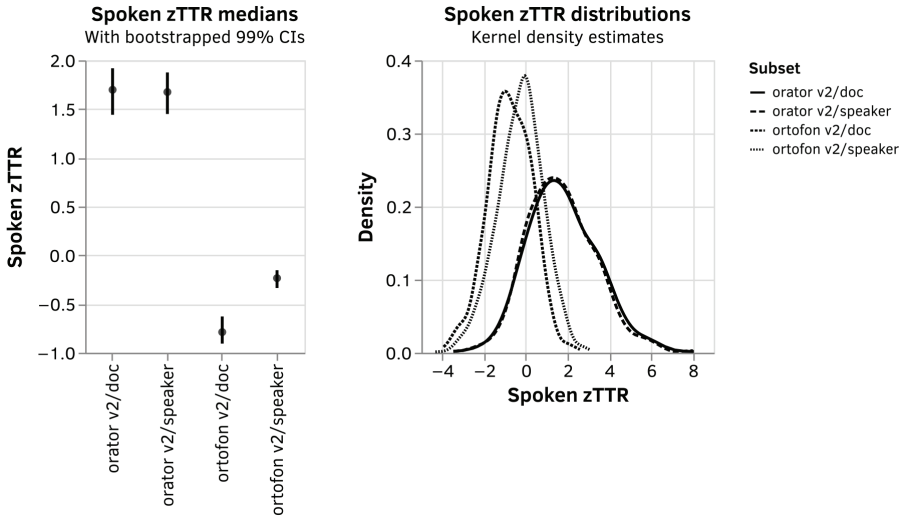

Fig. 2. zTTR computed on word forms, against spoken dialogue reference data

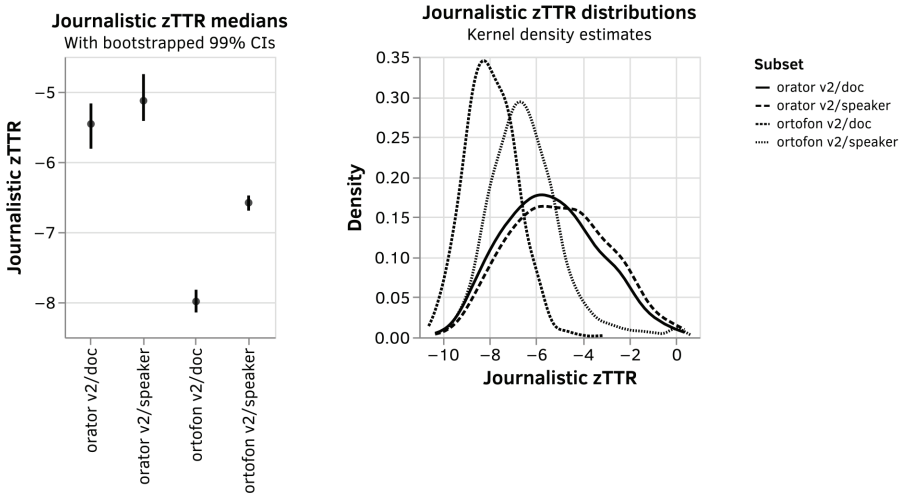

Fig. 3. zTTR computed on word forms, against journalistic reference data

References

[1] Kopřivová, M., Laubeová, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2020). ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. ÚČNK FF UK: Prague. Accessible at: https://korpus.cz.

[2] Kopřivová, M., Laubeová, Z., Lukeš, D., and Poukarová, P. (2020). ORATOR v2: Korpus monologů. ÚČNK FF UK: Prague. Accessible at: https://korpus.cz.

[3] Kopřivová, M., Komrsková, Z., Lukeš, D., and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. Korpus – Gramatika – Axiologie, 15, pages 47–67.

[4] Hoffmannová, J. (2017). Monolog. CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org.

[5] Müllerová, O. (2000). Žánry a syntaktické rysy mluvených projevů. In Tváře češtiny, pages 21–54, Ostrava. Ostravská univerzita.

[6] Čermák, F., Adamovičová, A., and Pešička, J. (2001). PMK: Pražský mluvený korpus. ÚČNK FF UK, Praha.

[7] Hladká, Z. (2002). BMK: Brněnský mluvený korpus. ÚČNK FF UK, Praha.

[8] Štěpánová, V. (2016). Korpus Monolog 1.1. Accessible at: http://monolog.dialogy.org.

[9] Kopřivová, M., Komrsková, Z., Poukarová, P., and Lukeš, D. (2019). Relevant criteria for selection of spoken data: theory meets practice. Jazykovedný časopis, 70(2), pages 324–335.

[10] Křen, M. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In Proceedings of LREC, pages 2522–2528, Portoroz. ELRA.

[11] Čermáková, A., Jílková, L., Komrsková, Z., Kopřivová, M., and Poukarová, P. (2019). Diskurzní markery. In Syntax mluvené češtiny, pages 244–351, Prague. Academia.

[12] Skarnitzl, R., and Machač, P. (2012). Míra rušivosti parazitních zvuků v řeči mediálních mluvčích. Naše řeč, 95, pages 3–14.

[13] Kubát, M., and Milička, J. (2013). Vocabulary Richness Measure in Genres. Journal of Quantitative Linguistics, 20(4), pages 339–349.

[14] Cvrček, V., and Chlumská, L. (2015). Simplification in translated Czech: a new approach to type-token ratio. Russian Linguistics, 39(3), pages 309–325.