USING QUANTITATIVE LINGUISTICS TO ASSESS PUPILS' LANGUAGE PROFICIENCY IN A BILINGUAL CONTEXT: THE CASE OF SLOVENE IN CARINTHIA

Ursula Doleschal, Gerald Robatsch
Institut für Slawistik, Celovec

UDK 81'246.2:373.5(436.5=163.6)

V članku predstavimo ovrednotenje jezikovne zmožnosti dijakov avstrijskokoroške dvojezične šole v slovenščini z metodami kvantitativnega jezikoslovja. Primerjamo šolske naloge dijakov z različnimi predznanji med seboj in z besedili iz korpusa Šolar.

dvojezična vzgoja, slovenščina, jezikovna zmožnost, kvantitativno jezikoslovje, ovrednotenje jezikovne zmožnosti

In this article we present a form of language assessment based on the methods of quantitative linguistics. We compare written class tests in Slovene from a bilingual elementary school in Austrian Carinthia with texts from the Šolar corpus.

bilingual education, Slovene, language proficiency, quantitative linguistics, language assessment

1 Introduction[1]

The measurement of linguistic competence or language proficiency has become a key topic in modern society with its focus on communication through diverse channels and in diverse languages. However, the assessment of an individual's proficiency is a complicated task, even more so in a bilingual situation with a high degree of heterogeneity as found in Carinthia (cf. Doleschal 2011: 164–166, Domej et al. 80–81). This is why virtually no studies on the proficiency of Austrian Carinthian pupils in the Slovene language have been carried out.[2]

In this article we present three applications of quantitative linguistics to Slovene language data obtained from pupils of a bilingual elementary school (Mohorjeva ljudska šola – VS Hermagoras) in Austrian Carinthia. By comparing them to texts by pupils from Slovenia selected from the Šolar corpus,[3] we investigate some aspects of their language proficiency. To this end we have chosen general quantitative indicators: type-token ratio, text coverage and frequency of word classes. The advantage of this approach is that we do not focus on errors but on the active use of language. Thus we can also describe language use, e.g., in terms of vocabulary, and contribute to the

1 We thank Emmerich Kelih, Markus Rheindorf and our reviewer for their helpful remarks.
2 Exceptions are Feinig 2000 and Danilović 2010.
3 http://www.slovenscina.eu/korpusi/solar

description of Slovene as it is acquired in Austria. Last but not least, this »descriptive assessment« should help to develop and improve didactic measures and approaches to Slovene.

In two previous, unpublished studies Robatsch (2012, 2015) investigated 155 written class tests, of which 88 were written in German and 67 in Slovene by a total of 31 fourth graders during the school years 2011/12 and 2012/13 (3 different classes). 6 children came from Slovene-speaking families, while most of the others used German at home. 10 of the latter entered school without prior knowledge of Slovene. All pupils attended the fourth grade of the *Mohorjeva Ljudska šola – Volksschule (VS) Hermagoras*, a private bilingual elementary school with daily language alternation between Slovene and German and were about 10 years old. Most of them had attended this school for 4 years.

Robatsch compared the pupils' performance in German to that in Slovene, as well as the performance of two groups (acquisition of Slovene prior to school or not). He investigated text-length, sentence length, number and frequency of verbs as well as type-token ratio. In the present study some results of these studies for the Slovene texts are compared with selected texts from the Šolar corpus. We hypothesize that 1) more exposure to Slovene before and in addition to school leads to better results in our measurements and that 2) children with Slovene as L1 and family language behave similarly to children in Slovenia.

The Slovenian corpus Šolar consists of texts produced by pupils of primary and secondary schools. Although the majority of the texts are written by pupils between the age of 13 and 19, a handful of texts composed by 6th grade pupils is also available. This subcorpus is taken as a reference for the present study. Analyses have been done with regard to *type-token-ratio*, *text coverage* and *number and frequency of conjunctions* used.

Type-Token-Ratio (TTR)

The TTR is here defined as the number of *distinct words* or lexemes (types) in a given text divided by the total number of words in the same text (tokens). This ratio equals 1 if no lexeme is repeated. The more repetitions of single lexemes, the lower the ratio will be. Thus, in a comparison between texts, a TTR closer to 1 for a given text indicates greater vocabulary richness.

Whether the TTR proves linguistic proficiency is a long-standing debate (cf. Berg 2014: 199; Otlogetswe 2011: 194). Following Glück (2010: 729) we assume that the TTR can shed light on the diversification of an author's lexicon.

Text coverage

Beside the TTR, text coverage can also be used as a qualifier for lexical proficiency. Fengxiang (2013: 288) defines text coverage as »the proportion of words of a text or a collection of texts covered by a given set of vocabulary.« In quantitative terms, text coverage is measured as the percentage of words from a previously defined list occurring in a given text. This parameter raises some important questions regarding language acquisition:

– What are the most frequent words of a language?
– How many of the x-frequent words does the subject know?
– How many of the x-frequent words must subjects know to express themselves reasonably well?
– What are the most frequent words in a specific genre?

Following these points, it can be assumed that high text coverage can show a subject's language proficiency regarding his or her lexicon.

Conjunctions

Conjunctions are an important means for expressing semantic relationships between propositions and are thus key to the cohesion of a text (Halliday 2002: 174; Givón 1995: 373). Conjunctions also create hierarchical relations between sentences (Toporišič 2000: 426) and are therefore an indicator of syntactic complexity: The number of different conjunctions correlates with the variety of semantic relations between clauses and their frequency with the frequency of complex sentences. Thus, higher values indicate higher language proficiency (cf. e.g., Roth, Neumann, Gogolin 2007: 9).

2 Analyses

2.1 Design of the study

The texts obtained from VS Hermagoras were digitized and their word forms lemmatized manually. The Šolar corpus also provides a lemmatization of word forms. To establish compatibility between the VS Hermagoras and the Šolar corpus, the latter has to be adjusted. Thus, a tagged line like

```
<w3 lemma="hoteti" msd="Ggnd-ez">hotela</w3>
```

has to be broken down to the pure lemma (in this case: *hoteti*). Breaking down all tagged lines to their pure lemma yields a wordlist sorted by the appearance of the observed lemma in the running text.

The relevant part of the Šolar corpus (the texts by the 6th grade pupils) consists of 26 texts of distinct authors with an average of 293 words, while the VS Hermagoras »corpus« comprises six different subcorpora which, according to the linguistic background of the pupils, will be separately compared to the Šolar subcorpus:

1. Texts from the school year 2011/12 comprise 36 written class tests (by 9 pupils) with an average word count of 97. The mother tongue and family language of these children is German.
2. Texts from the school year 2012/13 comprise 66 written class tests with an average word count of 123. These texts have to be separated into five groups according to the children's language preconditions:

   Group a: 3 pupils with Slovene as L1 and family language.

Group b: 4 pupils with German or English as L1 and family language but prior knowledge and current use of Slovene with certain relatives or in certain environments (sports club, choir).
Group c: 3 pupils with German as L1 but Carinthian-Slovene family background and/or bilingual nursery school providing some prior knowledge.
Group d: 2 pupils with a Slavic language other than Slovene as L1.
Group e: 10 pupils with German or English as L1 and family language and without prior knowledge of Slovene.

A closer look at the used words will give us more insight into the pupils' lexicon. Restrictions have to be made in order to obtain comparable data:

1. Mistakes of either orthographic or grammatical nature are not registered, as we are interested in lexical richness, and not in error analysis.
2. Titles are excluded from analysis.

2.2 Type-Token-Ratio (TTR)

The first part of the evaluation will focus on lexical richness in both corpora. The study of lexical richness has become more refined by the introduction of modifications of the original TTR like the MATTR (Moving average type token ratio, Covington 2010) or the MWTTRD (Moving window type token ratio distribution, Kubát 2013).

We gain the following results by simply counting the number of types and tokens in the different corpora:

Table 1: Types and tokens in the subcorpora

ŠOLAR		1293 types	7620 tokens
VS Hermagoras – texts 2012/13	group 2a[4]	577 types	1989 tokens
VS Hermagoras – texts 2012/13	group 2b	571 types	2313 tokens
VS Hermagoras – texts 2012/13	group 2c	449 types	1635 tokens
VS Hermagoras – texts 2012/13	group 2d	317 types	1042 tokens
VS Hermagoras – texts 2012/13	group 2e	306 types	1331 tokens
VS Hermagoras – texts 2011/12		696 types	3219 tokens

The original formula of the Type-Token-Ratio is heavily affected by the length of the associated text – the longer the text, the more word repetitions will occur (cf. Covington 2010: 94; Kettunen 2014) – and makes the corpus with the longest texts – the Šolar subcorpus – the one with the lowest TTR-values. Hence, the previously mentioned MATTR is useful since it eliminates this bias by creating $N - W_x + 1$ individual TTRs (N = number of tokens; W = text window with a length of x tokens). Subsequently, the mean of all the individual TTRs results in the final MATTR-value

4 Subsequently VS Hermagoras *12/13a* or *VS 12/13a*. This abbreviation procedure is also applied to the other subcorpora.

(cf. Covington 2010: 96). Because of the moderate text length in our data we choose a window size of 100. The output[5] shows the following ratios:

Table 2: MATTR of the subcorpora. Window size = 100[6]

Subcorpus	Tokens	MATTR
Šolar	7620	0.603
VS Hermagoras 2012/13a	1989	0.630
VS Hermagoras 2012/13b	2313	0.621
VS Hermagoras 2012/13c	1635	0.597
VS Hermagoras 2012/13d	1042	0.566
VS Hermagoras 2012/13e	1331	0.550
VS Hermagoras 2011/12	3219	0.633

The highest values in the TTR and thus the least repetitions of single words are shown by three Austrian-Carinthian groups, whereas the pupils from Slovenia are ranked fourth. How is this possible? The reason is to be sought in the distribution of high frequency lemmata. Let us take verbs as an example.[7] A closer look at the most frequent verbs reveals why the VS 11/12 texts show the highest value – and not the more advanced Šolar subcorpus:

Table 3: Verbs in the subcorpora

Verbs Šolar	count	Verbs VS 12/13a	count	Verbs VS 12/13b	count	Verbs VS 12/13c	count
biti	1319	biti	322	biti	406	biti	285
iti	46	iti	20	imeti	40	iti	32
priti	46	imeti	14	iti	29	imeti	16
povedati	33	priti	12	peljati	19	videti	14
dati	33	videti	11	hoteti	12	narediti	12
imeti	30	peljati	10	dati	10	priti	9
oditi	30	igrati	9	videti	10	igrati	8
začeti	27	reči	8	reči	9	peljati	7
vprašati	25	jesti	6	pasti	9	deževati	5
dobiti	23	začeti	6	vzeti	8	teči	5

5 Generated by Covington & Fall's MATTR – A CASPR project. http://ai1.ai.uga.edu/caspr/ MATTR2.zip

6 There is no consensus as to the margin of what the MATTR of a given text should reach. At a Window size of 500, Kettunen (2014) analysed the EU Constitution and discovered a MATTR between 0.39 and 0.60 depending on language (Slovene: 0.53). He did the same for parts of the Leipzig corpus (which contains randomly selected sentences from newspapers and web pages of different languages, http://corpora.uni-leipzig.de/download.html) and obtained a MATTR between 0.61 and 0.86 (Slovene: 0.73). Furthermore, the authors of the MATTR considered a value of 2 W-0.02 for any English text.

7 Like conjunctions, verbs are indicators of proficiency, since the formation of sentences depends on the use of verbs. A good command of verb lexemes is thus the basis for expressing a variety of states.

Verbs VS 12/13d	count
biti	185
iti	22
imeti	9
vzeti	8
igrati	7
jesti	7
peljati	7
videti	6
hoteti	6
priti	5

Verbs VS 12/13e	count
biti	181
imeti	48
rasti	12
zrasti	6
piti	6
misliti	5
cveteti	5
začeti	5
igrati	4
reči	4

Verbs VS 11/12	count
biti	358
imeti	77
iti	44
peljati	24
jesti	15
rasti	12
vedeti	9
pozdraviti	9
veseliti	8
priti	8

Table 3 shows clearly that one verb, *biti*, appears above average and basic arithmetic operations prove the crucial role of *biti* – as a full and an auxiliary verb, as can be seen in Table 4:

Table 4: Percentages of *biti* in the subcorpora

Subcorpus	*biti* in relation to Σ verbs	*biti* in relation to Σ words
Šolar	51.1 % (1319:2583)	17.3 % (1319:7620)
VS Hermagoras 12/13a	48.2 % (322:667)	16.2 % (322:1986)
VS Hermagoras 12/13b	50.2 % (406:809)	17.6 % (406:2313)
VS Hermagoras 12/13c	50.7 % (285:562)	17.5 % (285:1629)
VS Hermagoras 12/13d	49.9 % (185:371)	17.8 % (185:1042)
VS Hermagoras 12/13e	55.7 % (181:325)	13.6 % (181:1327)
VS Hermagoras 11/12	43.9 % (358:816)	11.1 % (358:3214)

However, the use of *biti* differs from text to text – stories written in the past or future tense, e.g., necessitate the extended use of *biti* as an auxiliary verb. Eliminating these auxiliary verbs results in adjusted outputs, as shown in Table 5:

Table 5: Occurrences of *biti* in the subcorpora

Subcorpus	*biti* (all)	*biti* (full verb)	ratio
Šolar	1319	156	0.12
VS Hermagoras 12/13a	322	52	0.16
VS Hermagoras 12/13b	406	96	0.23
VS Hermagoras 12/13c	285	68	0.24
VS Hermagoras 12/13d	185	34	0.18
VS Hermagoras 12/13e	181	164	0.91
VS Hermagoras 11/12	358	183	0.51

By eliminating all tokens of *biti* as an auxiliary verb, we ultimately obtain different MATTR-values as well:

Table 6: MATTR of the subcorpora without *biti* as an auxiliary verb. Window size = 100

Subcorpus	Tokens	MATTR
Šolar	6459	0.686
VS Hermagoras 2012/13a	1718	0.708
VS Hermagoras 2012/13b	2003	0.699
VS Hermagoras 2012/13c	1418	0.667
VS Hermagoras 2012/13d	890	0.640
VS Hermagoras 2012/13e	1314	0.555
VS Hermagoras 2011/12	3044	0.633

In almost all subcorpora the MATTR rises significantly upon eliminating *biti* as an auxiliary verb.[8] Only in the case of the VS 12/13e texts does the MATTR remain nearly unchanged. As can be seen in Table 5, this group used *biti* mostly as a full verb. Otherwise the values for the first three groups are now closer to one another, and the group VS 11/12 (with German as family language) drops to rank 5.

Thus, the results show that the groups without use of Slovene as family language or in other environments (VS 11/12 and VS 12/13e) use *biti* as a full verb more often than the other children. It follows that they use the past and future tense less and we can also infer a lower variety of verbs.

2.3 Text coverage

Applying the Šolar subcorpus as reference, we consider lexemes with a frequency of 3 or higher as reference lexemes for a text coverage analysis. This frequency is found for 403 lexemes. The comparison with the different VS Hermagoras subcorpora shows notable inequalities:

Table 7: Text coverage for the Hermagoras subcorpora

Corpus	text coverage
VS Hermagoras 12/13a	0.46
VS Hermagoras 12/13b	0.44
VS Hermagoras 12/13c	0.37
VS Hermagoras 12/13d	0.31
VS Hermagoras 12/13e	0.21
VS Hermagoras 11/12	0.41

The descending text coverage values correspond to the exposure to Slovene: Group VS 12/13a shows the highest text coverage, whereas VS 12/13e shows the lowest. As an interesting exception, group VS 11/12 does quite well and has the 3rd highest text coverage ratio.

8 This occurs because this operation reduces the number of *biti*-tokens drops dramatically but affects text size relatively little.

Thus, the group of pupils with Slovene as family language most closely approximates the vocabulary used by Slovenian pupils, as we would expect. Note that the reference group is about two years older than the pupils we investigated.

2.4 Conjunctions

Let us now turn to the frequency of conjunctions. As for the Šolar corpus, we can easily extract the tagged conjunctions, yielding a total of 855 tokens. The conjunctions used are:

Table 8: Conjunctions in the Šolar subcorpus

a	dokler	ki	nato	saj	torej
ali	drugače	kjer	niti	tedaj	vendar
ampak	in	ko	oziroma	temveč	zakaj
če	kako	kot	pa	ter	zato
čeprav	ker	naj	preden	toda	

We are able to compare how often conjunctions are used in each subcorpus. The following table is sorted by the frequency of each conjunction in the Šolar subcorpus. Conjunctions printed in bold appear only in the Šolar texts:

Table 9: Frequency of conjunctions in all subcorpora

KONJ_DIS	Šolar	VS 12/13a	VS 12/13b	VS 12/13c	VS 12/13d	VS 12/13e	VS 11/12
=in	253	57	60	59	42	66	158
=da	138	28	27	18	8	12	28
=ko	109	16	13	11	12	0	9
=pa	86	11	6	4	4	1	1
=ki	36	8	7	0	4	0	5
=a	29	0	0	0	0	0	1
=ker	28	11	19	15	10	8	10
=saj	27	0	0	0	0	0	0
=zato	27	6	5	5	6	2	0
=če	18	7	5	1	1	0	2
=ampak	15	6	12	6	2	2	5
=ali	15	5	3	4	0	1	4
=vendar	12	0	0	0	0	0	0
=kot	11	2	5	4	1	4	7
=kjer	10	1	0	0	0	1	0
=kako	10	2	2	0	0	0	0
=nato	5	5	12	6	4	0	1
=preden	4	0	2	0	0	0	0
=dokler	4	0	0	0	0	0	0

=niti	3	0	0	1	0	0	0
=ter	3	0	0	0	0	0	0
=toda	2	0	0	0	0	0	0
=drugače	2	0	0	0	0	0	0
=oziroma	2	0	0	0	0	0	0
=torej	1	0	0	0	0	0	0
=temveč	1	0	0	0	0	0	0
=čeprav	1	1	0	0	0	0	0
=tedaj	1	0	0	0	0	0	0
=naj	1	0	0	0	0	0	0
=zakaj	1	0	2	0	0	0	0
txtLength	7620	1986	2313	1629	1042	1327	3214
conj. ratio	11.2	8.4	7.8	8.2	9.0	7.3	7.7

The following graph demonstrates the crucial role of the conjunction *in* in all subcorpora. Only the Šolar texts show a larger variety of conjunctions.

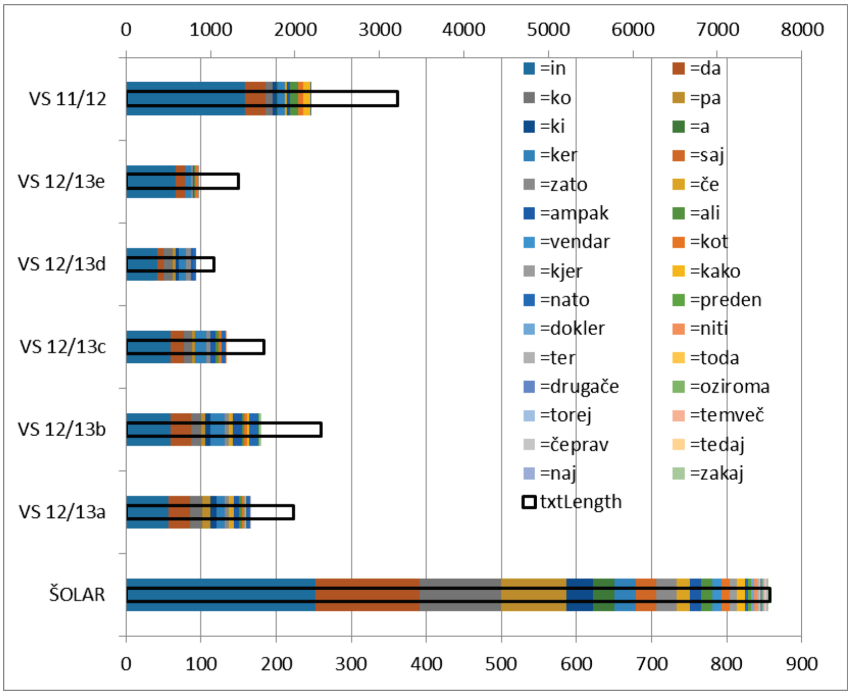

Chart 1: Frequency of conjunctions in all subcorpora – graph. Segments sorted by the most frequent conjunctions in the Šolar corpus from left to right (in – da – ko – pa – ki etc.)

Table 9 and Chart 1 allow us to conclude the following: The highest percentage of conjunctions is found in the Šolar subcorpus, as shown by the conjunction ratio (percentage of conjunctions in the word count). As for the VS Hermagoras sub-

corpora, the conjunction ratio is quite stable (between 7.7 and 9.0 %). However, *in* is relatively more frequent in the texts produced by pupils without a Slovenian environment (VS 11/12 and 12/13e), and both the ratio and variety of conjunctions are least for group 12/13e.

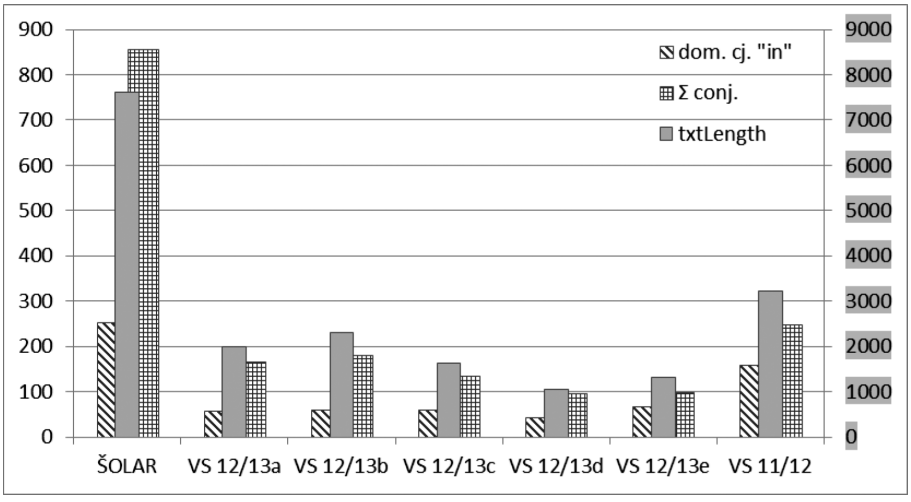

Chart 2: Text length vs. conjunctions

A graphical reproduction of the data from Table 9 reveals the dominant role of the conjunction *in* in all corpora. Chart 2 moreover clearly shows the relationship between text size and the overall amount of conjunctions.

While all corpora show a high *in*-ratio, only the Šolar corpus also shows a high number of different conjunctions, as seen in Table 9.

3 Conclusion

This paper describes three applications of quantitative linguistics to the assessment of language proficiency, highlighting the use of corpora for comparative studies: a Type-Token-Ratio operation using the rather new MATTR-formula, an investigation of text coverage and a quantitative approach concerning conjunctions.

We were able to show some general aspects as well as specific details of the lexical competence of 10-year old pupils in relation to their sociolinguistic profile. The results confirm our hypotheses and are in line with the findings of Daniłović (2010: VIII; 88–103): Children who have acquired Slovene at home or before school and use the language also outside school achieve higher results and are closer to the Slovenian reference group. But we have also shown that groups with similar characteristics (VS 11/12, 1213e) exhibit different degrees of proficiency. It is thus necessary to follow up on our results and especially build a Slovenian reference corpus of younger pupils' texts.

References

BERG, Thomas, 2014: On the Relationship between Type and Token Frequency. *Journal of Quantitative Linguistics* 21/3. 199–222.

BUSCH, Brigitta, DOLESCHAL, Ursula, 2008: Mehrsprachigkeit in Kärnten heute. *Wiener Slavistisches Jahrbuch* 54/2008. 7–20.

COVINGTON, Michael, McFALL, Joe, 2010: Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17/2. 94–100.

DANILOVIĆ, Mirijam, 2010: *Jezikovna zmožnost dijakov višje šole za gospodarske poklice na avstrijskem Koroškem. Diplomsko delo.* Maribor: Filozofska fakulteta.

DOLESCHAL, Ursula, 2011: Bilingual Education in Austria: The Case of Slovene in Carinthia. *Uporabno jezikoslovje* 9–10. 158–175.

DOMEJ, Teodor et al., without year: *Jahresbericht über das Schuljahr 2012/13*. Klagenfurt: Landesschulrat für Kärnten. http://www.2sprachigebildung.at/jahresbericht2013.pdf

FEINIG, Tatjana, 2000: Slowenisch an Kärntner Schulen als Erst-, Zweit- oder Fremdsprache? Erhebungen und Anmerkungen zur Sprachkompetenz im Slowenischen in Bezug auf dessen soziolinguistische Situation. Allan James (ed.): *Aktuelle Themen im Zweitspracherwerb.* Wien: Edition Praesens. 143–168.

FENGXIANG, Fan, 2013: Text Length, Vocabulary Size and Text Coverage Constancy. *Journal of Quantitative Linguistics* 20/4. 288–300.

GIVÓN, Talmy, 1995: *Functionalism and Grammar.* Amsterdam, Philadelphia: John Benjamins Publishing Company.

GLÜCK, Helmut (ed.), 2010: *Metzler Lexikon Sprache.* 4. Auflage. Stuttgart, Weimar: Metzler.

HALLIDAY, Michael, 2002: Linguistic Studies of Text and Discourse. Jonathan Webster (ed.): *Collected Works of M.A.K. Halliday.* London, New York: Continuum.

KETTUNEN, Kimmo, 2014: Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics* 21/3. 223–245.

KUBÁT, Miroslav, MILIČKA, Jirí, 2013: Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20/4. 339–349.

OTLOGETSWE, Thapelo, 2011: *Variability Measures in Corpus Design for Setswana Lexicography.* Cambridge: Cambridge Scholars Publishing.

ROBATSCH, Gerald, 2012: *Untersuchung des Leistungsstandes von Schülerinnen und Schülern der VS Hermagoras*. Unpublished seminar paper. Universy of Klagenfurt.

ROBATSCH, Gerald, 2015: *VS Hermagoras: Schularbeitenanalyse nach quantitativen Kriterien. Korpus 2012/13. Stichwortauswertung*. Unpublished project account, University of Klagenfurt.

ROTH, Hans-Joachim, NEUMANN, Ursula, GOGOLIN, Ingrid, 2007: *Abschlussbericht über die italienisch-deutschen, portugiesisch-deutschen und spanisch-deutschen Modellklassen*. Hamburg: Univ. Hamburg.

Šolar: http://www.slovenscina.eu/korpusi/solar (Tadeja Rozman, Mojca Stritar Kučuk, Iztok Kosem, Simon Krek, Irena Krapš Vodopivec, Špela Arhar Holdt, Marko Stabej, 2013: *Learners' corpus Šolar 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1036)