Research of Morphemic Vocabulary Volume Based on the Material of German Short Stories*

Bence Nyéki[1]
nyeki.bence96@gmail.com

Vladimír Benko[2]
vladimir.benko@juls.savba.sk

[1] Saint-Petersburg State University,
Saint-Petersburg, Russian Federation
[2] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics,
Bratislava, Slovakia

Abstract

The dependence of the vocabulary volume on text sample size has been studied on the material of literary texts [Grebennikov and Assel, 2019] as well as everyday spoken language [Kosareva and Martynenko, 2015]. The present research concerns the study of the morphemic type-token ratio in samples from Franz Kafka's and Thomas Mann's short stories in German. The morphemic annotation of the samples from these texts was carried out manually and was aimed at finding the asymptote of the function "morpheme token–morpheme type." This helps to conclude whether the expansion of the list of lexemes in these authors' texts is due to the occurrence of new stems or to word formation.

Keywords: *stype-token ratio, word formation, morpheme, approximation*

1 Introduction

Our work is based on the idea that the analysis of the morphemic structure of wordsdeservesserious attention as compounding and affixation (i.e. the concatenation of morphemes) serve as productive tools for the creation of morphologically complex words with semantically transparentstructure in many languages. Consequently, the total number of morphemes in such a language is smaller than that of lexemes. As a result, it might be easier to obtain a representative sample of morphemes from a corpus than in case of words. Of course, it should be notedthat not every lexical unit is semantically compositional at the morphemic level; this feature corresponds to the problem of syntagmatic idiomaticity. Despite this fact, knowledge about the units of the morphological system of languages with rich word formation such as German and Russian can still be considered useful. Another important

aspect is that word formation is considered to play a significant role in text building. For example, Zemskaya[1992: 164] defines six ways of word formation manifestation as an activity in speech acts:

- derivation from a pretext word or syntagma during speech act production
- use of set of derivatives of the same type within a text
- formation of different derivatives from the same base
- use of words with identical derivational meaning
- juxtaposition of derivatives from homonymous words
- contrastive use of words with the same root.

These observations suggest that the study of morphemes in text can be a source of information for the description of an individual style.

This research is aimed at the modelling of the morphemic vocabulary growth (in number of morphemes) as a function of sample size on the material of German literary texts. In the next section, the research of the type-token ratio in Russian texts is briefly reviewed. In the present work, the attempt was made to adopt this basic idea to the quantitative study of texts at the morphemic level. In the third section, the data used for the present research and their annotation are discussed. Results of fitting a distribution to the datareceived are presented in the fourthsection. The last section is a summary of our conclusions and further plans.

2 Related work

In [Kosareva and Martynenko, 2015] research was done on the ORD corpus[1] ("One Day of Speech" corpus) to estimate the asymptote of the function modelling the type-token dependence in spoken Russian texts. The Weibull and Haustein functions were used for approximation, the latter of which was found to fit the data better, and the asymptotic level of about 45,000 lexemes was estimated.With the same methods, approximations by the two functions on the material of short stories of Russian writers were compared in [Grebennikov and Assel, 2019] but the authors concluded that the Haustein function is not always preferable depending on the growth stabilization. It is also worth noting that a significant difference was found between the authors. The growthof thequantity of types in Chekhov's short stories considerably slows down at a sample size of around 150,000 tokens (ca 16,000 lexemes). In case of Averchenko, however, no clear upper limit of the growth could be set.

The linguistic annotation of the data used in the present work is based on linguistic principles. [Mel'čuk, 2006: 390] should be mentioned as a theoretical framework and [Fleischer and Barz, 2018] as a description of contemporary German word formation. More details will be given in the next section.

[1] http://www.ord-corpus.spbu.ru/SocialStudies/ORD.html

3 The data

3.1 Preparing the data

In order to perform a quantitative experiment at a subword level, it is essential to have a large amount of morphemically annotated text samples. Algorithms of automatic word segmentation into smaller meaningful units are available. For example, MorfessorFlatCat is based on hidden Markov models with hidden states "stem", "prefix", "suffix" and "non-morpheme" to enable unsupervised machine learning [Grönroos et al. 2014]. However, it is obvious that such an algorithm could not be applied for data annotation in our work.

Firstly, the segmentations should be maximally precise and based on linguistic principles. Secondly, even the correct identification of morphs, i.e. minimal meaningful substrings of the words is unsuitable for the planned experiment. Morphs are tokens of morphemes, each of which should be represented by a single form. For instance, the morpheme *KEIT* '-ness' appears in forms *-keit* and *-heit* in the German words *Wichtigkeit* (importance) and *Dunkelheit* (darkness), respectively. Thirdly, the elaboration of a fine-grained system of supplementary morpheme tags is necessary in order to disambiguate homonymous morphemes. Due to these conditions, the data was annotated manually.

For the given experiment, the texts of two German authors were chosen. The short stories of Thomas Mann in a collection available in Project Gutenberg[2] as well as all literary works (but not diary entries and private letters) of Franz Kafka in Project Gutenberg-DE[3] were copied and saved as plain texts. The short stories only may not have contained enough tokens, so Kafka's novels were also included into the material. Thus the document containing Kafka's texts (ca 290,000 words) is much larger than the collection of Mann's short stories (ca 39,000 words) but still big enough for sampling.

The texts were first annotated by TreeTagger[4], which is a part-of-speech (POS) tagger that applies a Markov model and a decision tree [see Schmid 1994, 1995]. Apart from a relatively large number of POS-tags [5], lemmatization is also provided by the software. This information simplifies the process of morphemic annotation as a tuple consisting of a POS-tag and a lemma usually unambiguously determines the correct morphemic analysis.

Sampling was implemented as follows. From each of the two collections, 60 non-overlapping text fragments of 250 tokens each were randomly selected.

3.2 Morphemic tags

The words in the obtained samples were analyzed manually; punctuation marks were ignored. As mentioned above, contextual information was not usually needed to find the right annotation, so each POS-tag–lemma tuple was processed only once. In the case when disambiguation was necessary, the given words were analyzed in context.

The annotation includes the assignment of a string that represents a morpheme and a tag which is analogous to POS-tags to each identified word segment. The latter (for simplicity, it may be referred to as a morphemic POS-tag or MPOS-tag) is a two-level tag the first component of which takes a value from the set "ST", "PF", "SF"; the elements of this set stand

[2]https://www.gutenberg.org/files/36766/36766-0.txt

[3]https://gutenberg.spiegel.de/autor/franz-kafka-309

[4]https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[5]Documentation for German available: https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/ data/stts_guide.pdf

for “stem”, “prefix” and “suffix”, respectively (so they correspond to three of the hidden states of model applied by MorfessorFlatCat). The second component is determined by the part of speech of the whole word form when the given morpheme is added to the stem. It takes a value from the set “NN”, “VV”, “ADJ”, “ADV”, “PP” “DET”, “PPER”, “PREL”, “PD”, “PUF”, “PAV”, “PWAV”, “KOUS”, “KOUI”, “KON”, “KOKOM”, “APPR”, “PTKVZ”, “PTKZU”, “PTKNEG”, “PTKA”, “PTKANT”, “ITJ”. These values are connected to the POS-tags assigned to the words of the samples by TreeTagger. However, it should be noted that this set is smaller than the set of all TreeTagger POS-tags,so each one of the morphemic annotation units may be associated with several original TreeTagger tags. For example, attributive and substitutive/predicative roles are not reflected at the morphemic level. A one-morpheme adjective gets the same morphemic tag (ST-ADJ) in both attributive (ADJA) and predicative positions (ADJD). Of course, an attributively used adjectival word form normally consists of at least two morphemes, the last of which is an inflectional ending. Inflectional morphemes, however, were simply ignored as their main function is marking syntactic relations within a sentence and can hardly be considered informative from the aspect of textuality or individual style. Furthermore, considering inflectional morphemes in the segmentation would make it impossible to determine the correct morphemic analysis without information about the concrete word form which is to be annotated. So-called “Fugenelemente” (meaningless segments of words on the borders of constituent morphemes) were ignored as well. Examples are given in
Table 1-2.

It is worth noting that word-level POS-tags are also used in modules of word sense disambiguation systems [Wilks and Stevenson, 1997]. This means that the method of identification of morphemes by normal form and an MPOS-tag makes use of an analogy between morpheme and word level.

3.3 Annotation principles

Although it might seem trivial to define a morpheme as the smallest segmental meaningful part of a word, in practice it can be difficult to find a theoretically supportable morphemic analysis of a particular word. For example,it needs explanation whether the following words are to be split into two morphemes or not:

a) *Mädchen* ‘girl’

b) *bekommen* ‘get, receive’

c) *entdecken* ‘discover’

d) *Augenglas* ‘glasses’

All examples above were marked as single morphemes in the samples. Word a) consist of the diminutive suffix *-chen* and a pseudostem which cannot be considered as a sign of standard German. Following the concepts in [Mel’čuk, 2006: 384], any linguistic sign should representable as a triplet of a signified, signifier and syntactics, which condition is not met by the given pseudostem due to the lack of a signified at a synchronic level. In b) both *be-* and *komm(en)* are valid linguistic signs. However, the first is an abstract verbal prefix while the second is a verb meaning ‘come’. Obviously, the meaning of the derivative cannot be composed of these semantic elements. The possible constituents of c) are *ent-,* which is a verbal prefix

Table 1: Examples of morphological analysis. Part 1

Meaning	Word form	POS	Lemma	Analysis	Comment
'war years'	Kriegsjahren	NN	Kriegsjahr	ST-NN_krieg + ST-NN_jahr	Fugenelement s is ignored
'effects'	Wirkungen	NN	Wirkung	ST-VV_wirk + SF-NN_ung	
'hands'	Hände	NN	Hand	ST-NN_hand	The same annotation is assigned to every allomorph of the same morpheme
'bitter, resentful'	bitterlich	ADJD	bitterlich	ST-ADJ_bitter + SF-ADJ_lich	
present participle from 'hold back'	zurückhaltenden	ADJA	zurückhaltend	ST-PTKVZ_zurück + ST-VV_halt + SF-ADJ_end	Participle morphemes are not considered inflection; also, they are annotated
'dilute'	verdünntem	ADJA	verdünnt	PF-VV_ver + ST-ADJ_dünn + SF-PP_t	
'(have) seen'	gesehen	VVPP	sehen	ST-VV_seh + SF-PP_t	The circumfix *ge...t* (*ge...en* is regarded as an allomorph) is marked as a single suffix which is identical to the participle morpheme *-t/-en*
'stood'	stand	VVFIN	stehen	ST-VV_steh	The vowel change within the root implies inflectional meaning so it is ignored.

Table 2: Examples of morphological analysis. Part 2

Meaning	Word form	POS	Lemma	Analysis	Comment
'is'	ist	VAFIN	sein	ST-VV_sei	In case of suppletion, the lemma is analyzed.
'one, anyone'	man	PIS	man	ST-PUF_man	ST-PUF is a special tag without an equivalent in TreeTaggertagset. It is assigned to a small set of uninflectable words
'this'	dieser	PDAT	dies	ST-DET_dies	ST-DET (determiner) is a frequent tag that is assigned to articles and demonstrative pronouns without homonyms
'(for) the'	dem	ART	die	ST-DET_die	The next examples demonstrate the disambiguating effect of MPOS-tags (in this case they disambiguate whole words)
'which' (relative pronoun)	die	PRELS	die	ST-PRELS_die	
'that' (demonstrative pronoun)	das	PDS	die	ST-PD_die	

that implies the cancellation of an action or a state as one of its senses, and *deck(en)* (cover, protect, hide). Although not hiding and discovering something can be regarded as cognitively associated senses, such a weak relation did not seem to be sufficient to split the word into two parts. The compound *Auge* (eye) + *Glas* (glass) is presented in d). In fact, it is close to be semantically compositional but still the meaning of *Glas* is too general.

In linguistics, degrees of semantic compositionality, which can also be referred to as transparency, are sometimes distinguished. In [Ransmayr et al., 2016: 267–268] eleven degrees of transparency of German derived words with the diminutive suffix -*chen* were defined. The opaquest derivatives are words without a synchronically identifiable stem, like a) above, or with a non-noun stem like *Frühchen* (premature infant).The meanings of the most transparent derivatives are simply constructed of those of their constituent morphemes. Sometimes they can be affected by pragmatic restrictions, for example, diminutives denoting clothes such as *Jäckchen* (small jacket) are usually used to refer to women's or children's clothes. Between these extremes there are instances of weakly (e.g. metaphorically) motivated compositions such as *Eichhörnchen* (squirrel) and *Hörnchen* (croissant).

All derived words which are not maximally transparent can be considered, using the terminology in [Mel'čuk, 2006: 390], quasimorphs, which should be stored in dictionary as separate entries. However, a typical lexicographic problem often makes this principle more difficult to apply in practice: it is not always clear how to define the meaning of constituent morphemes. Taking once again the stem of *c* as an example, it is not self-evident without thorough corpus research whether the sense 'protect, hide' should be considered a simple metaphor or a word sense which needs lexicographic description.This is a usual lexicographic problem, which inevitably occurs and must be solved more or less subjectively in each case. Despite this fact, some principles of segmentation formulated in advance can serve as considerable theoretic support. In view of the aspects discussed above, they can be summarized as follows:

1. All constituents of the analyzed word should be meaningful linguistic unitsat a synchronic level.

2. A bound morpheme is to be added to the morphemic vocabulary only if it occurs in several non-synonymous words in which it has the same sense (which although might be highly general or opaque).

3. A complex unit (quasimorph) is preferable only if its meaning is not fully transparent. Under the assumption that word meaning can be represented as a finite set of senses, which are considered to be conventionalized, this means that we expect that all senses of a morphologically complex word **w** can be represented as an element of the Cartesian product of the constituents senses. If some senses of **w** are elements of the Cartesian product, while others are not, the correspondent quasimorph should be added to vocabulary in case of the occurrence of **w** with a non-compositional sense.

These principles are simple, but they can help make consequent decisions. For example, it follows from 2) that such morphemes as the pseudostem of *Mädchen* (girl) cannot be treated as separate meaningful segments of words even if several synonymous lexemes exist in the language with the same pseudostem, e.g. *Mädchen* and *Mädel*.This idea can be generalized to any unique morpheme which occurs only as a constituent of a certain lexeme. Note that a different viewpoint is also presented in [Fleischer and Barz, 2012: 65] that is based on structural and not strictly semantic analysis.

The analysis of the noun *Aufzug* (the act of lifting, hoisting/elevator/act in theater) shows the consequences of principle 3). If it occurs as a noun derived straightly from the verb *aufziehen* (auf 'up' + ziehen 'pull'), then it is correct to segment it into two constituents. However, in a sample from Thomas Mann's texts this is not the case. It occurs in the sense 'act in theater' and is therefore added to the morphemic vocabulary as a whole unit.

It is obvious that the last principle suggests that knowledge about word senses is essential for morphemic analysis. It makes it necessary to rely on a lexicographic resource. DWDS[6] was chosen as such a resource as it ensures quick access not only to lexicographic but also corpus data if needed. For details about DWDS see [Geyken, 2007].

However, these ideas were not extended to verbs with separable verb prefixes (and words compositionally derived from them) although they often form lexical units with semantically opaque structure. Separable prefixes can take a position very far from the verbal stem in the sentence, which makes it hard to suggest an appropriate annotation method. This remained a problem to resolve.

Since morphemes are not the only elementary linguistic signs [Mel'čuk, 2006: 295–297], it is necessary to mention how non-segmental signs were handled. In German such signs are frequently applied by means of conversion and modification. Take the nouns *Schritt* (step), *Eintreten* (the act of coming in) and the verb *beenden* (finish) as examples. *Schritt* is derived from *schreiten* (to step) by vowel change, *Eintreten* is obtained by applying conversion to the verb *eintreten* (come in) and in *beenden* the noun *Ende* (i.e. its allomorph end) can be observed and as *be-* is a verbal prefix, it must be concluded that the noun is converted to a verb (there is also another verb *enden* with nearly the same meaning). As our annotation is morphemic, these non-segmental signs are simply ignored. This means that the analyses of these words are ST-VV_schreit, ST-PTKVZ_ein + ST-VV_tret and PF-VV_be + ST-NN_ende, respectively. Of course, this is a simplification: these signs are frequent in German and the morphological process of conversion is highly productive.

Now that all major problems of the data annotation are discussed, results can be presented.

4 Results

Having annotated the text fragments, it was necessary to find a distribution which can model the growth of morphemic vocabulary as a function of sample size. As mentioned above, the Weibull and Haustein functions were applied for similar goals [Kosareva and Martynenko, 2015; Grebennikov and Assel, 2019]. Our paper is not aimed at comparing how different functions fit the data. Only cumulative Weibull distribution was chosen for modelling, which is usually defined as follows:

$$y = 1 - e^{-(\frac{x}{\eta})^{\beta}}$$

In related work, a slightly different equation is given as the Weibull function [Grebennikov and Assel, 2019: 380]:

$$y = N_{max} - N_{max}e^{-cx^{d}}$$

Apart from the exponent, the difference is that the right side of the last equation is

[6]Digitales Wörterbuch der deutschen Sprache (Digital Dictionary of the German Language): https://www.dwds.de/

multiplied by N_{max} which is the asymptote of the function, i.e. the theoretical maximal volume of the morphemic vocabulary.

To simplify computation, data was manipulated to enable the use of the former (more standard) formula. Firstly, empirical values (the registered number of lexemes at a given sample size) were divided by a hypothetical value of maximal volume. Then the equation was linearized (analogously to [Kosareva and Martynenko, 2015]) to estimate the parameters of Weibull distribution using the slope and intersect of the linear regression. Now theoretical values of y could be calculated, which were multiplied by N_{max}. The most appropriate value of N_{max} was estimated by the least-squares method: it was determined as the highest observed vocabulary volume plus 50 (it was clear from the data that the asymptote was significantly higher than any observed value). Then 50 was added to N_{max} again in each following step until the least sum of quadratic deviations between the theoretical and observed values of y was reached (some steps could be omitted adding immediately more than 50 to the upper limit of vocabulary volume). Of course, obtained values are approximate and they can be defined more precisely; still the results clearly show the difference between the texts of the two German authors.

Tables 3-4 show some hypothetical values of N_{max} and the corresponding sum of quadratic deviations. The data necessary for calculating theoretical values of y, given the N_{max} which has eventually proved best, are presented in Tables 5-6. These tables serve as illustrations and they contain only every third observed value of the morphemic type-token function (of course, the whole sets of observations were used to find distribution parameters). The curves of cumulative Weibull distributions determined by the calculated parameters and N_{max} are depicted in Figure 1.

Table 3: Hypothetical values of N_{max} and the corresponding sums of quadratic deviations for samples from Franz Kafka's texts

N_{max}	**Sum of quadratic deviations**
1793	148624
2293	22794,88
2493	16697,63
2543	16109,97
2593	15775,62
2643	15652,69
2693	15706,58

Figure 1 shows that the growth of the functions is nearly identical until vocabulary size reaches approximately 1000 morphemes. After that the growth of Kafka's vocabulary slows down and stabilizes at a sample size of about 60 thousand morpheme tokens. The other function has a considerably higher asymptote and stabilizes at about 80 thousand morpheme tokens.

Table 4: Hypothetical values of N_{max} and the corresponding sums of quadratic deviations for samples from Thomas Mann's texts

N_{max}	**Sum of quadratic deviations**
2154	264363,87
2654	64892,02
3154	30309,74
3354	26620,6
3554	25473,1
3604	25460,79
3654	25534,69

Table 5: Calculations for finding distribution parameters and theoretical values for the samples from Franz Kafka's texts. The highest observed volume of vocabulary is 1743 morphemes. $N_max = 2643$, $\beta = 0.69$, $\eta = 13573.52$, $\Sigma(y_i - y_j)^2 = 15652.69$.

Sample size x	**Observed value** y_i	$ln(x)$	y_i/N_{max}	$ln(-ln(1 - y_i/N_{max}))$	**Weibull value**	**Theoretic value** y_j	**Quadratic deviations**
726	315	6,59	0,12	-2,06	0,12	327,12	146,98
1453	496	7,28	0,19	-1,57	0,19	507,97	143,23
2213	675	7,7	0,26	-1,22	0,25	656,35	347,91
2936	803	7,98	0,3	-1,02	0,29	775,03	782,11
3669	908	8,21	0,34	-0,87	0,33	879,94	787,22
4389	990	8,39	0,37	-0,76	0,37	971,58	339,14
5109	1062	8,54	0,4	-0,67	0,4	1054,26	59,95
5796	1129	8,66	0,43	-0,58	0,43	1126,25	7,54
6516	1196	8,78	0,45	-0,51	0,45	1195,62	0,14
7232	1246	8,89	0,47	-0,45	0,48	1259,3	176,94
7973	1302	8,98	0,49	-0,39	0,5	1320,37	337,4
8724	1359	9,07	0,51	-0,33	0,52	1377,86	355,72
9476	1417	9,16	0,54	-0,26	0,54	1431,51	210,51
10209	1467	9,23	0,56	-0,21	0,56	1480,44	180,52
10949	1511	9,3	0,57	-0,16	0,58	1526,8	249,71
11698	1551	9,37	0,59	-0,12	0,59	1570,93	397,28
12434	1609	9,43	0,61	-0,06	0,61	1611,8	7,85
13127	1659	9,48	0,63	-0,01	0,62	1648,22	116,27
13842	1700	9,54	0,64	0,03	0,64	1683,86	260,59
14607	1743	9,59	0,66	0,07	0,65	1719,99	529,59

Table 6: Calculations for finding distribution parameters and theoretical values for the samples from Thomas Mann's texts. The highest observed volume of vocabulary is 2104 morphemes. $N_{max} = 3604$, $\beta = 0.72$, $\eta = 17676.11$, $\Sigma(y_i - y_j)^2 = 25460.79$.

Sample size x	**Observed value** y_i	$ln(x)$	y_i/N_{max}	$ln(-ln(1 - y_i/N_{max}))$	**Weibull value**	**Theoretic value** y_j	**Quadratic deviations**
700	332	6,55	0,09	-2,34	0,09	334,84	8,08
1367	533	7,22	0,15	-1,83	0,15	526,67	40,04
2129	739	7,66	0,21	-1,47	0,2	704,28	1205,4
2826	867	7,95	0,24	-1,29	0,23	843,63	546,03
3537	990	8,17	0,27	-1,14	0,27	969,93	403
4228	1107	8,35	0,31	-1	0,3	1080,9	681,4
4897	1184	8,5	0,33	-0,92	0,33	1179,42	21
5575	1283	8,63	0,36	-0,82	0,35	1271,79	125,72
6308	1369	8,75	0,38	-0,74	0,38	1364,44	20,82
6911	1425	8,84	0,4	-0,69	0,4	1435,76	115,86
7629	1489	8,94	0,41	-0,63	0,42	1515,63	709,07
8328	1566	9,03	0,43	-0,56	0,44	1588,66	513,7
9005	1634	9,11	0,45	-0,5	0,46	1655,43	459,37
9705	1698	9,18	0,47	-0,45	0,48	1720,76	517,99
10414	1747	9,25	0,48	-0,41	0,49	1783,43	1327,29
11161	1824	9,32	0,51	-0,35	0,51	1846	483,93
11883	1900	9,38	0,53	-0,29	0,53	1903,38	11,42
12588	1964	9,44	0,54	-0,24	0,54	1956,71	53,09
13351	2025	9,5	0,56	-0,19	0,56	2011,67	177,77
14085	2104	9,55	0,58	-0,13	0,57	2062,02	1762,4

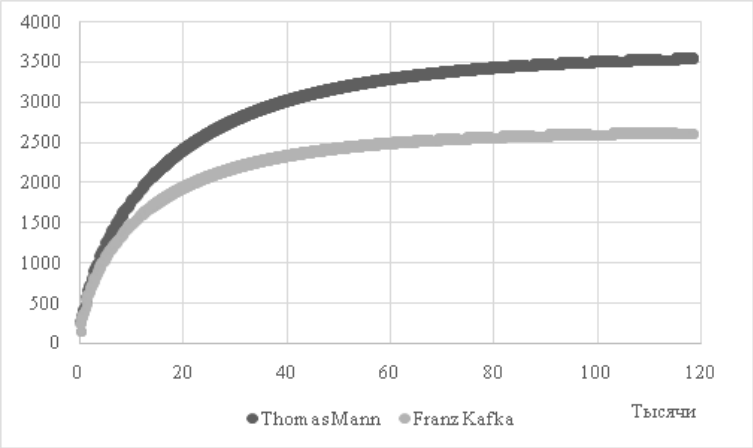

Figure 1: Type-token ratio functions for morphemes.

5 Conclusions

A logical interpretation of the results is that there are more foreign roots in Thomas Mann's short stories than in Franz Kafka's texts. It can be noticed that Mann uses more foreign proper names (e.g. *Florentinum*, *Fontana*). However, it seems that the type-token ratio should be considered at word level as well in order to justify this hypothesis. For example, an author whose morphemic vocabulary grows slowly but uses many different lexemes could rely on word formation to support expressivity.

In the present paper, it has been showed that quantitative aspects of derivational morphology can be regarded as a feature of individual style. A significant difference has been found between the growth of morphemic vocabulary in two German authors' texts. In future research, larger samples need to be taken in order to compare the dependence of the number of lexeme and morpheme types on sample size. As word formation is a productive linguistic and cognitive process, this may considerably contribute to the quantitative research of style.

References

[Zemskaya, 1992] Zemskaya E. (1992) Word formation as an activity. (In Rus.) = Slovoobrazovaniekakdeyatel'nost'. Nauka, Moscow, Russia. – 221p.

[Kosareva and Martynenko, 2015] Kosareva E. O., Martynenko G. Ya. (2015) The Type–TokenRatio in Everyday Spoken Russian. Structural and Applied Linguistics, Vol. 11. (In Rus.) = Otnoshenietekst – slovarv povsednevnoyustnoyrechi. Strukturnaya I prikladnayalingvistika, Vol 11. Saint Petersburg, Russia. Pp. 220–228

[Grebennikov and Assel, 2019] Grebennikov A. O., Assel A. N. (2019) XIX–XX Centuries' Russian Short Stories Corpus. ApproximationModels. Proceeding of the International Conference «Corpus Linguistics–2019». Saint Petersburg University Press. (In Rus.) = Bazarusskogorasskaza XIX–XX vekov. Modeliapproksimatsii. Trudy mezhdunarodnoykonferentsii«Korpusnayalingvistika – 2019».Izdatel'stvo Sankt-PeterburgskogoUniversiteta, Saint Petersburg, Russia.Pp. 379–386

[Mel'čuk, 2006] Mel'čuk I. A. (2006): Aspects of the Theory of Morphology. Trends in Linguistics. Studies and Monographs 146. Mouton de Gruyter, Berlin, Germany. – 616p. Available at https://anekawarnapendidikan.files.wordpress.com/2014/04/aspects-of-the-theory-of-morphology-by-igor-melcuk.pdf

[Fleischer and Barz, 2012] Fleischer W., Barz I. (2012) Word Formation of Contemporary German (In Ger.) = Wortbildung der deutschen Gegenwartssprache. Fourth, revisededition. De Gruyter, Berlin, Germany – 481p.

[Grönroos et al., 2014] Grönroos S.-A., Virpioja S., Smit P., Kurimo M. (2014)MorfessorFlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. Proceedings of the 25th International Conference on Computational Linguistics. Association for Computational Linguistics, August 2014, Dublin, Ireland. Pp.1177-1185. Available at https://www.aclweb.org/anthology/C14-1111.pdf

[Schmid, 1994] Schmid, H (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK. Pp. 44–49. Available at https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/tree-tagger1.pdf

[Schmid, 1995] Schmid, H. (1995) Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. Pp. 1–9. Available at https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/tree-tagger2.pdf

[Wilks and Stevenson, 1997] Wilks Y. and Stevenson M. (1997) Sense Tagging: Semantic Tagging with a Lexicon. ACL SIGLEX workshop, Washington, DC. Pp. 74–78

[Ransmayr et al., 2016] Ransmayr J., Schwaiger S., Durco M., Pirker H., Dressler W. U. (2016)Grading of the Transparency of Diminutives Ending in -chen: a corpus linguistic research. German Language. Journal for Theory, Practice, Documentation 44(3). (In Ger.) = Gradierung der Transparenz vonDiminutiven auf -chen: Eine korpuslinguistische Untersuchung. Deutsche Sprache. ZeitschriftfürTheorie, Praxis, Dokumentation 44(3). Pp. 261–286

[Geyken, 2007] GeykenA. (2007) The DWDS corpus: A reference corpus for the German language of the 20th century. In: Collocations and Idioms: Linguistic, lexicographic, and computational aspects. Ed. by Fellbaum C. London, UK. Pp. 23–41 Draft available at https://www.dwds.de/dwds_static/publications/text/DWDS-Corpus_Desc4_draft.pdf