The Average Word Length Dynamics as an Indicator of Cultural Changes in Society

Vladimir V. Bochkarev
Kazan Federal University

Anna V. Shevlyakova
Kazan Federal University

Valery D. Solovyev
Kazan Federal University

ABSTRACT

In this article we analyze the dynamics of average length of the Russian and English words belonging to the Google Books diachronic text corpus and dated back to the last two centuries. It is found that an average word length slightly increased in the nineteenth century, and then it grew rapidly over most of the twentieth century and started decreasing from the end of the twentieth to the beginning of the twenty-first century. Words which contributed most to the increase or decrease of word average length are identified, with content words and function words being analyzed separately. Long content words contribute most to the average length of words. These words reflect the main tendencies of societal development and, thus, are frequently used. The changing frequency of personal pronouns also contributes significantly to changes in average word length.

1. INTRODUCTION

Every language reflects the context in which it is spoken. While the idea that there is a direct and deterministic link between language structures and thought – the so-called Sapir-Whorf hypothesis – is debatable, there is little doubt that the vocabulary of a language

Social Evolution & History, Vol. 14 No. 2, September 2015 153–175

is a reflection of the speaker's physical and mental world, and consequently, the shifting frequencies in different words also reflect changes in the material world, society, and prevalent concerns. Earlier studies of 'world views' as embodied in language were limited to particular concepts reflecting cultural stereotypes or grammatical phenomena perceived as having an influence on speakers' behaviour. Due to the creation of large, diachronic text corpora and the development of mathematical methods for data processing now we can apply a new approach to the issue of possible connections between a language and the world (societal, physical, mental, *etc.*) of its speakers. In the present article we study changes in average world length. This is a measurable entity which can be statistically analysed, and, as we will show, there are non-random changes involving word length over time in different languages which promise to shed light on changes in the extra-linguistic world.

Much attention has been given lately to language dynamics, including data-driven or simulation-based investigations into language change viewed in the context of how speakers interact and organize themselves in different networks is (Wichmann 2008; Croft 2000; Hruschka *et al.* 2009; Trudgill 2004; Kalampokis *et al.* 2007; Michel *et al.* 2011; Nettle 1999; Blythe 2012; Fagyal *et al.* 2010). The present article focuses on such parameters as word length and frequency. Although the word length is significantly correlated with other typological parameters (Wichmann *et al.* 2011) it also varies within languages (across words belonging to different grammatical categories) and depends on the frequency, with shorter words tending to be used more frequently than longer ones. Indeed, the word frequency and length are fundamental parameters in the study of psychological processes of language acquisition and usage (Sidney *et al.* 2000; Baddeley and Scott 1971), and the frequency of language units is the basis of usage-based sociolinguistic models of language evolution (Baxter *et al.* 2009).

One observes some clear general laws of language change including the S-shaped curve of innovation diffusion (Fagyal *et al.* 2010) and the effect of word frequency on change rates whereby the less frequently a word is used, the greater the probability that it will change (Pagel *et al.* 2007). As a word drops out of use it may be compensated for by another, whereby frequencies of different words become linked. But the process of word frequency change is not only connected with the process of substitution of one form by

another one. Importantly, word frequency can vary under the influence of sociocultural factors and thus, may reflect these factors.

Our article is devoted to the processes of word frequency change under the influence of sociocultural factors. It continues recent work related to the quantitative analysis of cultural trends or 'culturomics' (Michel *et al.* 2011). The subject of our study is regularities of word average length variations and factors which cause these changes.

Data concerning average word length are available for different languages. Thus, for instance, according to the Wolfram Alpha computational knowledge engine (http://www.wolframalpha.com/input/?i=average+english%20+word+length) the average word length in English is 5.1 letters and that in Russian it is 5.28 (Sharov 2011). Nevertheless, accurate quantitative analyses of the dynamics of this parameter have not been carried out. Average word length can both decrease and increase in the course of time. Over large periods of time a language may change its morphological type (agglutinative, inflective, or isolating) (Croft 2003), and changes in the morphological type of a language are expected to influence average word length. But such changes happen very slowly. It has recently been shown (Wichmann *et al.* 2013) that the world's language families tend to have typical average world lengths which differ from family to family. This means that a characteristic average word length can persist for the amount of time that corresponds to the typical age of known language families, which is several thousand years. Language typologists and historical linguists are concerned with the diachronic study of languages over large time scales (Heggarty 2007). Here we are interested in fluctuations of word length over relatively short time spans of a few hundred years, not in the type of language-internal typological changes which happen over a larger time scale of thousands of years.

In particular, we will look at English and Russian. These languages have not changed radically during the two last centuries, so when we still observe fluctuations in average word length then they must be due to the factors other than typological change. We hypothesize that the fluctuations are due to changes in the frequencies of different kinds of words, these changes are, in turn, caused by sociocultural factors.

As described in the following section, studying the dynamics of average word length at relatively short periods of time (decades or centuries) has become possible because of the creation of big diachronic text corpora.

2. METHODS AND DATA

The creation of the digital library Google Books and the word frequency calculation enabled by the Google Books Ngram Viewer (http://books.google.com/ngrams/) has created extraordinary new opportunities for studying the evolution of the lexicon of different languages. As demonstrated in the seminal paper by Michel *et al.* (2011), these data can fruitfully be applied for analyzing cultural trends.

Some cautionary words concerning the nature of the corpus are in order. It should be taken into account, for instance, that the Ngram Viewer does not offer morphological analyses. Different inflectional forms of one and the same root (for example, English *chair* and *chairs*, or various case-forms of a particular Russian noun) are considered to be different words on a par with completely unrelated forms such as English *chair* and *table*. Thus, in what follows our definition of 'word' also does not take into account that some words are related through shared lexemes while others are not.

Using the total set of n-grams presented on the site, average word length can be counted for each year (using MatLab, for instance). Although the English corpus comprises texts dated back to 1520, great amounts of text for reliable statistic computations are only available for 1800 onwards, according to the recommendation of the creators of the Ngram Viewer.

Average word length can be calculated using the following formula:

$$L = \sum_i p_i l_i , \tag{1}$$

where l_i is the length of *i*-th word and p_i is its relative frequency (usage probability). Let us consider that a particular word frequency (*k*-th word) varies, but that correlations of other words are invariable. According to the normalization requirement $\sum_i p_i = 1$ we derive

$$L = p_k l_k + \sum_{i \neq k} p_i l_i = p_k l_k + \widetilde{L}_k \left(1 - p_k\right) \tag{2}$$

where $\widetilde{L}_k$ is an average word length without considering the *k*-th word. Suppose that the *k*-th word usage frequency varies over Δp_k. As it is seen from the formula (2), the change of average word length is

$$\Delta L = \Delta p_k \left(l_k - \widetilde{L}_k \right) \tag{3}$$

Deriving the $\widetilde{L}_k$ -value from the second formula (2), we get

$$\Delta L = \frac{\Delta p_k}{1 - p_k} \left(l_k - L \right) \tag{4}$$

The ΔL -value shows the partial contribution of a *k*-th word to the average word length change. For the majority of words, except the several most frequently used ones, $p_k << 1$ is realized as

$$\Delta L = \Delta p_k \left(l_k - L \right) \tag{5}$$

Thus, the increase of the use of words whose length is smaller than the current average word length results in a decrease of average word length and vice versa.

The value derived from the formula (5) enables us to evaluate the contribution of a given word to the average word length change. Correspondingly, we can distinguish the words which contribute most to the change of average word length.

Two different approaches to frequency calculation should be taken into consideration when comparing the further results and diagrams. First, all numbers were excluded. There are also many OCR errors and some onomatopoeia, but their contribution to overall statistical data can generally be neglected. Our corrections were also taken into account when calculating frequencies. The Ngram Viewer shows frequencies normalized by the total number of 1-grams, including numbers, but we normalize them by the number of 1-grams that can be considered true words.

The Ngram Viewer comprises texts written in seven languages, but here we analyse the data for just two of them – English and Russian.

Fig. 1 shows the average word length distribution in the English language over time according to the Ngram Viewer data.

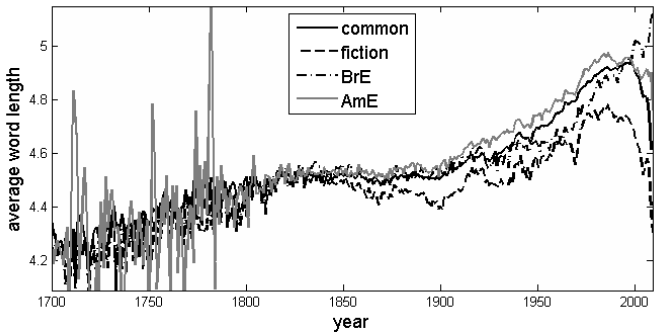

Fig. 1. Average word length over time in the English language. Each curve indicates the results for the total and fiction-only corpus, and also for British and American English separately

Three periods can be discerned in the graph: from 1800 to 1900, from 1901 to 1994, and from 1995 to 2008. In spite of fluctuations due to the relative scarcity of data, average word length is almost constant during the first period; it increases in a linear fashion during the second period, and decreases during the third period.

Words can be divided into two major groups: function words, which indicate a grammatical relationship (articles, conjunctions, particles, prepositions *etc*.), and content words (nouns, verbs, adjectives, numerals, *etc*.). Pronouns and auxiliary verbs were included into the first group. Most function words are short, consisting of only 2–3 letters (less than the average word length). Increase in the frequency of these word reduces average word length and vice versa. Almost all content words are long (longer than average word length), having the opposite influence: increase in frequency of these words lengthens average word length and vice versa.

Frequency variability of function and content words is studied separately. Fig. 2 shows changes of average word length of all the words examined and of long and short words separately.

One can see that short words, most of which are function words, almost do not change their length. At the same time, the shape of the curve which shows change of average length of long words, most of which are content words, is similar to the curve which describes word length of all long words.

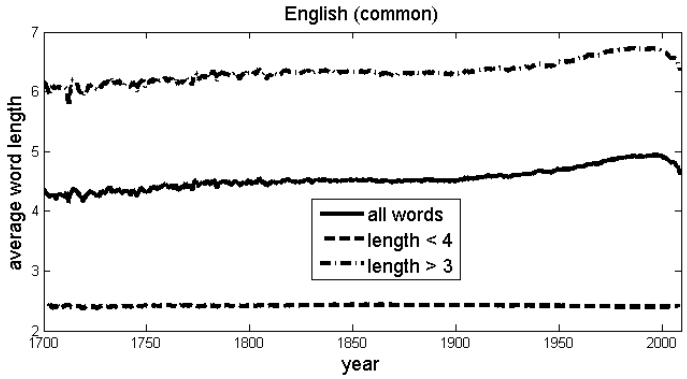

Fig. 2. Dynamics of average length of English short and long words

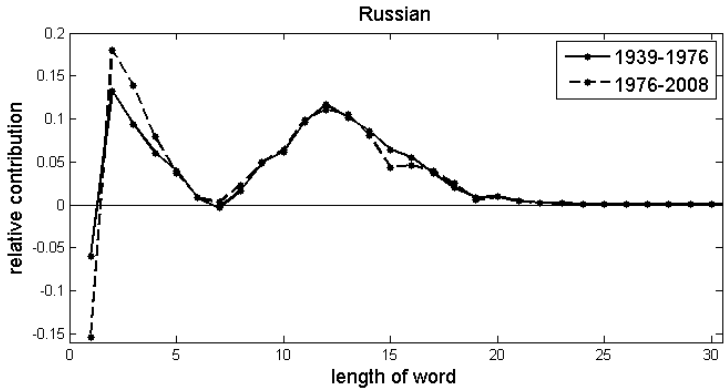

Fig. 3. Relative contribution to average word length change in Russian of words of different length for the two time intervals

Fig. 3 shows the relative contribution to average word length change of words of different length in Russian. There are peaks at two- and twelve-letter words. The period after 1939 is considered in order to avoid the influence of the orthography reform of 1918. The biggest value of average word length for the Russian language dates back to 1976. Correspondingly, two time periods are distinguished: from 1939 to 1976 when a tendency to average word length increase is observed, and 1976–2008 where the opposite tendency takes place. As the number of long words grows, their total contribution (approximately 70 per cent) becomes higher than that of short words.

3. THE INFLUENCE OF SOCIAL FACTORS ON CONTENT WORDS: A HYPOTHESIS

It can be assumed that the increase in average length is determined by vocabulary extension as new words emerge. As there are a limited number of short combinations of letters and most of them are used in any language, neologisms such as *computer, the Internet, globalization* are necessarily long. But there are other factors involved in word length variation. Fig. 4 shows the amount of words used in different periods of time.

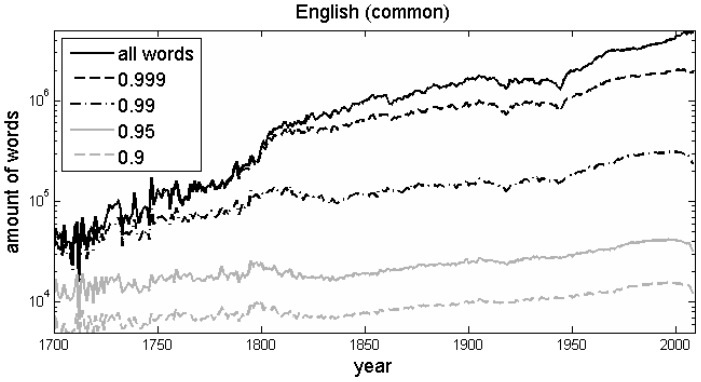

Fig. 4. The total amount of English words used in different periods (solid black line). The share of words which contributed most to the selected word group is also shown

The diagram shows that the number of words (the total number of words and the number of words which are used most frequently) grows in much the same ways in the nineteenth and twentieth centuries. Thus, this factor cannot explain why word length increases rapidly in the twentieth century, stays constant in the nineteenth century and decreases in the twenty-first century.

As a hypothesis, we suggest the following mechanism of influence of social factors on average word length. Average word length growing systematically over a large period of time reflects an active process of word formation which is connected with a certain idea that dominates in society. The actuality of the new idea, which is verbalized by a particular set of words, leads to a significant growth of the usage of these words. These concepts are usually represented by multi-syllable words. The new concepts are often complicated and the words which define them are formed through

the help of affixation, for example *socialism, globalization.* Besides, as already mentioned, the number of short words is limited in any language and cannot be used to verbalize all notions. All this leads to average word length growth.

On the other hand, decrease in the usage of content words is explained by the loss of actuality of a certain lexical field which, in turn, influences average word length on the whole. The principle of economy in language should also be mentioned as a contributor to the decrease in average word length. According to this principle, speakers will try to transmit as much information as possible using the fewest possible linguistic means. The phonetic and corresponding orthographical structure of a term may become reduced due to contractions in the pronunciation, forms may be truncated, and abbreviations may be introduced – all as a result of the accommodation of words to the needs of speakers.

We hypothesize that during the time when average word length is relatively constant and the change of average word length is sporadic there is a low influx of prevalent new ideas. In other words, a change of dominating concepts will result in a rapid change of the word stock connected with them. Graphically it is expected to incur changes in the slope of a curve or the appearance of small 'gaps' reflecting the situation when old concept are not topical anymore and the new one have not yet been formed. If average word length grows close to linearly during significantly long periods of time (*e.g.*, in English from 1900 to 2000), it can be assumed that concepts do not change rapidly at that time but rather evolve slowly.

4. EVOLUTION OF CONTENT WORDS: INVESTIGATING THE HYPOTHESIS EMPIRICALLY

We will start our survey by examining the content words in English during the period from 1990 to 2000. Below there is a list of words which contributed most to average word length increase during this period. The numbers show the level of word contribution to total average length increase. The contribution is calculated according to the formula (5).

1. *development*	0.00648	6. *relationship*	0.00413
2. *information*	0.00632	7. *economic*	0.00371
3. *international*	0.00596	8. *research*	0.00346
4. *university*	0.00521	9. *production*	0.00310
5. *political*	0.00415	10. *significant*	0.00304

One can see that semantically these words pertain to the key ideas of the twentieth century and reflect the priorities of the English speaking society of that time.

The following approach will help to test the hypothesis that social factors influence the frequencies of content words. The interval from 1800 to 2000 is divided into smaller ones of 25 years each (roughly corresponding to a generation), with the exception of the last interval from 2001 to 2008, which contains data for this eight-year period. The contribution of each word to the change of average word length is calculated for each interval using the formula in (5). The Appendix shows words which contributed most to average word length increase or decrease for each of these periods.

Some remarks are in order: (a) There are a lot of errors in the data during the first interval (1800–1825) due to wrong symbol recognition (OCR errors) caused by low quality of printing and the poor condition of books. For instance, the letters *s* and *t* are often recognized as *f*. This introduces some noise in the data, but not enough to invalidate overall conclusions. (b) The number of short content words is quite small, so they do not contribute much to an average word length. They were disregarded. (c) Words beginning with capital and small letters are merged and are cited in the tables in their typical orthographic forms.

For further analysis, the most 'influential' words are distinguished. These are the ten words which contributed most either to word average length decrease or increase during at least two periods. The data are summed up in Tables 1 and 2. The plus sign means that the word belongs to the top ten 'influential' words.

Let us analyse these data. One should pay special attention to the fact that the word lists at the beginning of the nineteenth and twenty-first centuries are not similar to word lists of other periods. The language development between 2000 and 2008 will be regarded later. As for the beginning of the nineteenth century, obviously, some interesting and important events took place, but the eighteenth century data are needed to survey changes taking place around the turn of the century. Unfortunately Google Books does not contain enough data, so we cannot deal further with this. The word *development* is at the top of the list of words which contributed to the average word length over one and a half centuries. This serves as evidence that this is one of the key concepts for English-speaking

society. Besides *development*, the words which are common for different periods of the nineteenth century are *one, position, government, American,* and *conditions*. They belong to different semantic groups and it is not clear that they pertain to some single idea or a set of ideas. Alongside with the word '*development*', the common words for different periods of the twentieth century from Table 1 include: *international, production, individual, education, political, university,* and *information*. It seems that these are key words in the modern English-speaking world relating to the global economic and political system based on education and information. There are some common words (marked by the plus sign) for different periods of the twentieth century presented in Table 1, which indicates that the system evolves gradually.

Table 1

Content words which contributed mostly to average word length

Word	1800–1825	1825–1850	1850–1875	1875–1900	1900–1925	1925–1950	1950–1975	1975–2000	2000–2008
one		+	+						
position		+	+						
government		+	+			+			
development		+	+	+	+	+	+		
American		+		+		+			
conditions			+	+	+				
international					+	+		+	
production					+	+			
individual					+	+			
education					+		+		
political						+	+		
university							+	+	
information							+	+	

The absolute values of the contribution to average word length are also interesting. During the nineteenth century, the contribution value is 0.003 (except for 1800–1825), and during the twentieth century the contribution value is 0.005 (except for 2000–2008). In other words, the frequency of *international, production, individual, education, political, university,* and *information* decreased in the twentieth century more rapidly than that of the *one, position, government, American,* and *conditions* in the nineteenth century. This shows that corresponding concepts were more significant for society in the twentieth than in the nineteenth century. It is interesting

that the words which contribute most to average word length increase in the twenty-first century do not belong to any single concept.

Thus, the analysis of content words contributed mostly to word frequency increase supports our hypothesis.

Let us consider the words whose frequency decreased rapidly. The data are presented in Table 2.

Table 2

Content words which contributed to average word length decrease

Word	1800–1825	1825–1850	1850–1875	1875–1900	1900–1925	1925–1950	1950–1975	1975–2000	2000–2008
particularly		+	+						
immediately		+	+	+					
God		+	+	+					
circumstances			+	+	+				
character			+	+	+	+	+	+	
Christian			+		+				
practically			+			+	+		
Mr				+	+	+			
hundred					+	+			
constitution					+		+		
little						+	+		
necessary							+	+	

The data for the early nineteenth and twenty-first centuries stand apart. It is interesting that the frequency of word *character* constantly decreases. On the whole, the words presented in this table belong to different semantic fields and do not reflect any particular idea.

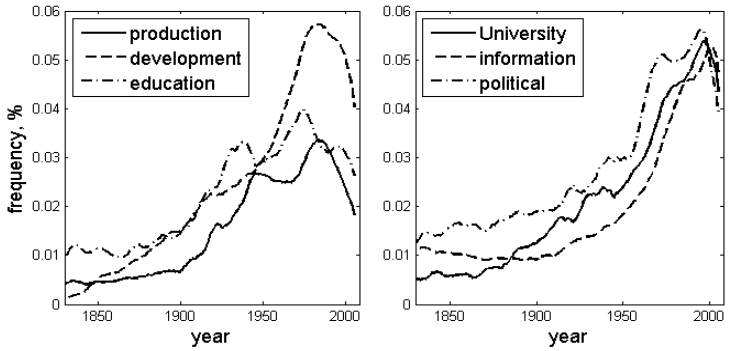

Fig. 5. Dynamics of the top ten influential words of the twentieth century

As for the absolute values of contributions to average word length of words we can make the following observations. Both for the nineteenth and the twentieth centuries, the contribution value is 0.0025 (except for 1800–1825). The value is much larger during 1800–1825 and 2000–2008, where it is 0.01. The words with significantly decreased frequency in the twenty-first century are of special interest. The less frequently used words are the following: *development*, *university*, *international*, *political*, and *information.* According to Table 1, the frequency of these words rapidly increased during the previous years. So the dominant ideas of the twentieth century do not prevail any more. Thus, the analysis of content words contributing less to word frequency supports our general hypothesis.

Let us consider the diagrams of the top ten influential words in the twentieth century: *development*, *information*, *international*, *university*, *political*, *relationship*, *economic*, *research*, *production*, and *significant*. It is clearly seen that these diagrams reflect the common rules of average word length change in the nineteenth – twenty-first centuries. During the nineteenth century, the majority of these words (except for *development*) do not increase their frequency and even decrease in frequency. Their frequency increases in the twentieth century. In the twenty-first century the frequency of these words decreases. Further, the decrease in frequency of some words (*development*, *production*) begins earlier, approximately in 1980.

5. THE EVOLUTION OF FUNCTION WORDS

There are several dozen function words. The function words which we distinguish (see the Appendix) are the following: articles, prepositions, conjunctions, pronouns, auxiliary verbs, particles and some adverbs.

Below is a list of words which contributed most to word length increase during the twentieth century.

1. *the*	0.04110	6. *his*	0.01307
2. *of*	0.03947	7. *to*	0.01136
3. *he*	0.02033	8. *was*	0.00932
4. *it*	0.01848	9. *at*	0.00923
5. *I*	0.01688	10. *and*	0.00886

This list of function words is dominated by personal pronouns (four items). The dynamics of personal pronouns is rather interesting. It can clearly be seen that the frequency of the majority of these words is stable or decreases rapidly until the end of the twentieth century, after which the frequency of all the pronouns increases rapidly. As all of them are short, the decrease in their frequency contributes to average word length increase and the increase in their frequency in the beginning of the twenty-first century contributes to average word length decrease. This is why this factor has the same influence as the evolution of content words which was described earlier. But it is not clear whether these two factors are mutually connected.

The majority of other function words are used with a near constant frequency. Some of them show the same rules of frequency change as personal pronouns, but this tendency is much weaker.

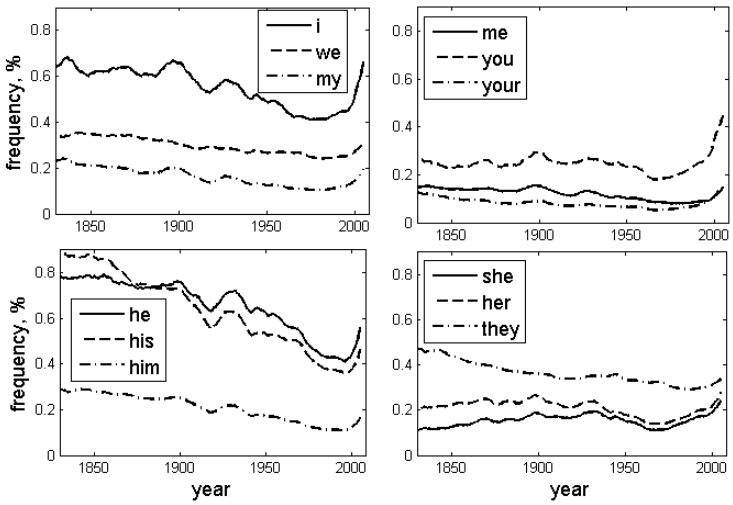

Fig. 6. Dynamics of personal pronouns in English

6. AVERAGE WORD LENGTH DYNAMICS IN THE RUSSIAN LANGUAGE

We now turn to the Russian words in order to test our hypothesis about the relationship between changes in average word length and societal changes. It is interesting to compare English with Russian

because Russian society differs greatly from the English-speaking society and has had completely different views and values for almost a century. The average word length dynamics in Russian is shown in Fig. 7. As can be seen, the overall situation is generally the same as in the English language: word length is near constant over the nineteenth century, grows in the twentieth century, but the decrease starts earlier – around 1975.

Let us consider the most interesting period in the twentieth and the early twenty-first century. It should be noted that the chosen time periods for both languages (1900–1925 *etc*.) are not relevant for the Russian language since several important events in Russian history as the socialist revolution (1917), World War II and the collapse of the Soviet Union took place in the middle of the intervals, which leads to some vocabulary mixing during the corresponding periods. The results for the Russian languages are, nevertheless, rather clear.

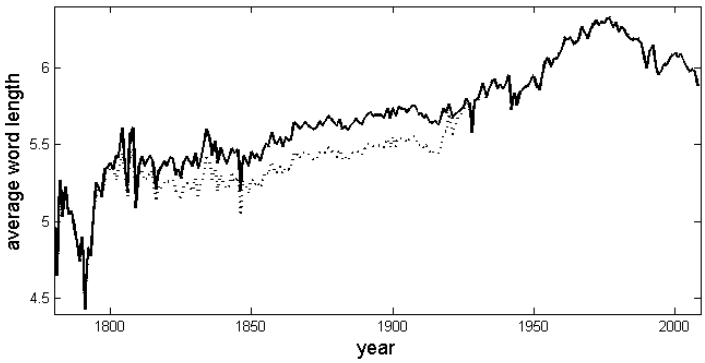

Fig. 7. The black solid line shows an average word length in the Russian language with the account of the orthographical rules of that time, the dotted line shows it excluding the letter ъ at the end of words (neutralizing the effects of one of the main rules of the orthographical reform of 1918)

Unlike what we observe in the English language, the top ten influential words for the chosen periods, that is the ones that contributed most to average word length, do not coincide much. Different concepts are present at each interval. From 1900 to 1925, the concepts related to economic issues dating back to the nineteenth

century mix with revolutionary ideas. From 1925 to 1950, the most influential words relate to the ideas of the socialist regime and Soviet power and key issues of the time, such as construction and agriculture. The priorities change in the period between 1950 and 1975, now starting to revolve around manufacturing and management. The leading concepts changed again between 1975 and 2000, now relating to the idea of nationhood. Legal issues dominate between 2000 and 2008. It is typical that between 1975 and 2000, the frequency of words connected with manufacturing and socialism decreased sharply, something which contributed greatly to a decrease of average length of words during this period. The concept changes between 1925–1950 and 1950–1975 can be considered as a small gap on the diagram a bit before 1950.

Fig. 8 shows diagrams for the top ten words. The lines for the Russian language in Fig. 8 are not as smooth as the corresponding ones for English. Many of the words were used widely only after the revolution. The increase in frequency was perturbed by World War II. It should be noted that priorities changed during the 1950s: the problems of effectiveness became acute and the term *revolution* was left behind. It is interesting that the frequency of influential Russian words started decreasing in 1980 just as in the English.

Fig. 8. Dynamics of the top ten influential Russian words of the 20th century. Translation of the Russian words in the left window: organisations, revolution, socialist, and soviet; in the right window: production, efficiency, Russian, and state

Thus, there is a direct dependency between fluctuations of average word length and the formation of word groups which are semantically connected with a dominating concept.

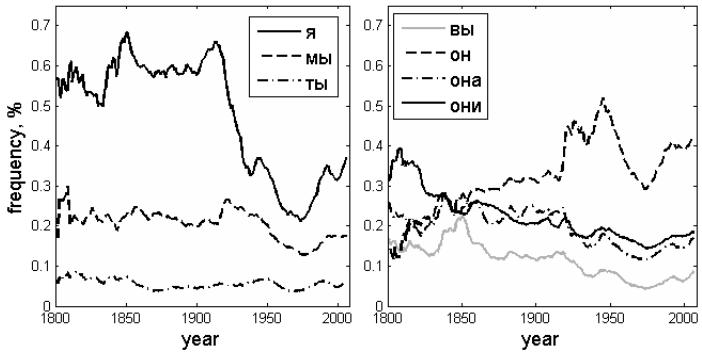

Fig. 9. Dynamics of personal pronouns in Russian. Translation of the Russian words in the left window: I, we, and you (singular); in the right window: you (plural), he, she, and they

As for function words, Russian personal pronouns generally behave as English personal pronouns (Fig. 6), and a decrease in frequency takes place in 1975. Nevertheless, there is a conspicuous special behavior of the Russian first person pronoun. The rapid decrease in frequency of *я* ('I') in the Russian language after 1917, which continues up to 1975, can be explained by detraction of the role of personality during the time of socialism and the imposition of collectivist ideology.

7. CONCLUSION

The present article introduces data concerning the change of average word length in the English and Russian languages over the last two centuries based on Google Books e-library and the Ngram Viewer. It is shown that the change of word length is a result of change in word frequency. The change in frequency of content words and function words has been analysed separately.

We have shown that a decrease in average word length correlates with two socio-cultural factors: the existence of some fundamental idea or several interconnected ideas and variation of frequency of personal pronouns.

Decrease in average word length correlates with the retreat from ideas which have been dominating (the number of words

which verbalize this concept decreases rapidly), and the frequency of personal pronouns increases. When there are no prevailing concepts, average word length tends not to change.

It is interesting to observe that there are some similarities in the data for the English and Russian languages in spite of great social, cultural, economic and political differences between their speakers. This shows that the regularities traced are not accidental, and also that they reflect some global, international trends exerting their influence in parallel with national trends. The most conspicuous similarity is a steeper increase in word length for both languages in the twentieth as compared with the nineteenth century (compare Fig. 1 and Fig. 7). We interpret this as an evidence of an overall increase in the speed of change affecting both the Russian- and English-speaking societies throughout the twentieth century, probably due to factors such as industrialization, urbanization, and migration, constantly driving changes in the world as perceived by the speakers whose dynamics of word usage we have studied in this paper.

The Ngram Viewer system enables the study of language evolution at the micro-level through frequency data that ultimately reflect extra-linguistic factors contingent upon cultural and historical factors.

ACKNOWLEDGEMENTS

The authors wish to thank Søren Wichmann for help and useful advice which contributed much to the improvement of the quality of the article.

This work was supported by the Russian Foundation for Basic Research through project RFBR № 12-06-00404-a.

REFERENCES

Baddeley, A. D., and Scott, D. 1971. Word Frequency and the Unit Sequence Interference Hypothesis in Short-Term Memory. *Journal of Verbal Learning and Verbal Behavior* 10: 35–40. DOI: 10.1016/S0022-5371(71)80090-7.

Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J. 2009. Modeling Language Change: An Evaluation of Trudgill's Theory of the Emergence of New Zealand English. *Language Variation and Change* 21: 257–296.

Blythe, R. A. 2012. Neutral Evolution: A Null Model for Language Dynamics. *Advances in Complex* Systems 15(3–4): 1150015. DOI: 10.1142/S0219525911003414.

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. Harlow: Pearson Education.

Croft, W. 2003. *Typology and Universals*. Cambridge: Cambridge University Press.

Fagyal, Z., Swarup, S., Escobar, A. M. *et al.* 2010. Centers and Peripheries: Network Roles in Language Change. *Lingua* 120: 2061–2079.

Heggarty, P. 2007. Linguistics for Archaeologists: Principles, Methods and the Case of the Incas. *Cambridge Archaeological Journal* 17: 311–340.

Hruschka, D. J., Christiansen, M. H., Blythe, R. A. *et al*. 2009. Building Social Cognitive Models of Language Change. *Trends in Cognitive Sciences* 13: 464–469.

Kalampokis, A., Kosmidis, K., and Argyrakis, P. 2007. Evolution of Vocabulary in Scale-Free and Random Networks. *Physica A* 379: 665–671.

Michel, J.-B., Shen, Y. K., and Aiden, A. P. *et al.* 2011. Quantitative Analysis of Culture using Millions of Digitized Books. *Science* 331: 176–182. DOI: 10.1126/science.1199644.

Nettle, D. 1999. Using Social Impact Theory to Simulate Language Change. *Lingua* 108: 95–117.

Pagel, M., Atkinson, Q. D., and Meade, A. 2007. Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History. *Nature* 449. DOI:10.1038/nature06176.

Segalowitz, S. J., and Lane, K. C. 2000. Lexical Access of Function versus Content Words. *Brain and Language* 75: 376–389. DOI: 10.1006/brln. 2000.2361.

Sharov, S. A. 2011. The Statistics of the Words in the Russian Language. *In Russian* (Шаров С. А. Статистика слов в русском языке). URL: http://www.lingvisto.org/artikoloj/ru_stat.html.

Trudgill, P. 2004. *New Dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.

Wichmann, S. 2008. The Emerging Field of Language Dynamics. *Language and Linguistics Compass* 2.3: 442–455.

Wichmann, S., Rama, T., and Holman, E. W. 2011. Phonological Diversity, Word Length, and Population Sizes across Languages: The ASJP Evidence. *Linguistic Typology* 15: 177–197.

Wichmann, S., and Holman, E. W. 2013. Languages with Longer Words have More Lexical Change. In Borin, L., and Saxena, A. (eds.), *Approaches to Measuring Linguistic Differences* (pp. 249–284). Berlin: de Gruyter Mouton.

APPENDIX

In the following tables the numbers indicate a word's contribution to the variation of the average words length. Positive numbers correspond to an increase of the average words length, negative ones to a decrease of the average words length.

Table 1

Function words in American English with decreasing frequency

1825–1849		1850–1874		1875–1899		1900–1924	
'he'	4.025	'of'	5.725	'and'	2.651	'I'	3.339
'to'	3.846	'to'	3.217	'of'	1.885	'he'	2.966
'be'	2.222	'and'	2.491	'the'	1.617	'and'	2.665
'his'	1.544	'his'	1.670	'by'	0.986	'his'	2.647
'by'	0.967	'be'	1.175	'to'	0.980	'to'	2.209
'him'	0.845	'we'	0.992	'we'	0.896	'my'	1.399
'is'	0.603	'my'	0.717	'i'	0.604	'was'	1.183
'who'	0.480	'our'	0.696	'or'	0.437	'de'	1.110
'that'	0.429	'it'	0.625	'are'	0.356	'by'	1.061
'it'	0.427	'all'	0.477	'our'	0.340	'him'	0.879

1925–1949		1950–1974		1975–1999		2000–2008	
'i'	4.086	'the'	8.128	'of'	20.340	'of'	5.339
'it'	2.273	'of'	4.386	'the'	11.844	'the'	3.920
'the'	2.221	'he'	3.395	'in'	5.730	'is'	2.049
'he'	2.073	'it'	3.151	'by'	3.384	'in'	1.440
'of'	2.056	'i'	2.366	'be'	3.389	'by'	1.352
'and'	1.685	'his'	1.652	'is'	3.044	'be'	0.886
'to'	1.383	'be'	1.154	'may'	0.829	'or'	0.568
'at'	1.328	'so'	1.143	'no'	0.638	'a'	0.450
'was'	1.138	'by'	1.132	'not'	0.608	'may'	0.414
'her'	1.106	'was'	1.099	'has'	0.576	'are'	0.300

Table 2

Function words in American English with increasing frequency

1825–1849		1850–1874		1875–1899		1900–1924	
'her'	–0.430	'up'	–0.383	'everything'	0.415	'per'	–0.569
'p'	–0.439	'are'	–0.4204	'was'	–0.511	'be'	–0.640
'in'	–0.644	'v'	–0.503	'her'	–0.571	'on'	–0.666
'its'	–0.816	'was'	–0.675	'she'	–0.664	'of'	–0.770
'we'	–0.921	'the'	–0.877	'for'	–0.701	'or'	–0.864
'on'	–1.038	'at'	–1.002	'he'	–0.790	'are'	–0.974
'and'	–1.381	'in'	–1.105	'you'	–0.949	'for'	–1.331
'of'	–2.774	'is'	–1.519	'de'	–1.010	'a'	–1.755
'the'	–3.040	'i'	–1.763	'in'	–1.412	'in'	–2.260
'a'	–3.886	'a'	–1.768	'a'	–2.791	'is'	–2.705

1925–1949		1950–1974		1975–1999		2000–2008	
'n'	–0.443	'pp'	–0.267	'up'	–0.970	'me'	–0.570
'l'	–0.480	'c'	–0.280	'can'	–0.995	'we'	–0.589
'c'	–0.500	'p'	–0.296	'me'	–1.016	'he'	–0.670
'b'	–0.540	'l'	–0.302	'my'	–1.530	'my'	–0.692
'v'	–0.561	'on'	–0.367	'her'	–1.906	'she'	–0.895
'p'	–0.588	'al'	–0.410	'she'	–1.927	'her'	–0.951
'in'	–0.640	'j'	–0.509	'to'	–1.983	'and'	–0.953
'for'	–0.655	'in'	–0.605	'a'	–2.561	'you'	–1.485
'f'	–0.770	'an'	–0.673	'you'	–4.072	'to'	–1.789
'm'	–0.777	'b'	–0.815	'i'	–6.011	'i'	–3.537

Table 3

Content words in American English with increasing frequency

1825–1849		1850–1874		1875–1899	
'one'	–0.527	'mr'	–0.517	'conditions'	0.494
'beautiful'	0.359	'president'	0.486	'monsieur'	0.482
'intellectual'	0.356	'general'	0.329	'american'	0.411
'inhabitants'	0.287	'massachusetts'	0.295	'university'	0.380
'connection'	0.285	'conditions'	0.253	'government'	0.358
'influence'	0.271	'development'	0.252	'literature'	0.321
'neighborhood'	0.261	'railroad'	0.224	'mademoiselle'	0.318
'constantinople'	0.255	'signified'	0.223	'development'	0.305
'spiritual'	0.251	'position'	0.222	'political'	0.258
'development'	0.243	'especially'	0.213	'temperature'	0.258

1900–1924		1925–1949		1950–1974	
'organization'	0.842	'international'	0.526	'development'	1.390
'conditions'	0.701	'production'	0.485	'information'	1.177
'department'	0.633	'administration'	0.435	'university'	1.013
'individual'	0.577	'research'	0.435	'relationship'	0.845
'school'	0.567	'government'	0.431	'behavior'	0.759
'development'	0.562	'temperature'	0.428	'patients'	0.721
'production'	0.550	'individual'	0.408	'environmental'	0.698
'business'	0.540	'development'	0.403	'children'	0.683
'american'	0.537	'relationship'	0.397	'research'	0.671
'education'	0.531	'problems'	0.391	'significant'	0.649

1975–1999		2000–2008	
'information'	1.223	'international'	0.581
'management'	0.617	'students'	0.502
'understanding'	0.418	'learning'	0.416
'technology'	0.409	'research'	0.410
'internet'	0.378	'university'	0.265
'assessment'	0.376	'political'	0.265
'including'	0.351	'understanding'	0.261
'performance'	0.335	'chapter'	0.260
'perspective'	0.334	'knowledge'	0.239
'understand'	0.310	'organizations'	0.219

Table 4

Content words in American English with decreasing frequency

1825–1849		1850–1874		1875–1899	
'mentioned'	–0.328	'country'	–0.264	'heaven'	–0.221
'concerning'	–0.328	'particular'	–0.269	'inhabitants'	–0.225
'celebrated'	–0.330	'happiness'	–0.275	'circumstances'	–0.244
'immediately'	–0.334	'knowledge'	–0.321	'received'	–0.244
'necessary'	–0.340	'christian'	–0.347	'signifies'	–0.258
'distinguished'	–0.374	'character'	–0.349	'general'	–0.275
'considered'	–0.389	'happiness'	–0.275	'signified'	–0.283
'particularly'	–0.413	'circumstances'	–0.426	'spiritual'	–0.295
'scriptures'	–0.449	'god'	0.580	'according'	–0.298
'god'	0.669	'inhabitants'	–0.835	'immediately'	–0.312

1900–1924		1925–1949		1950–1974	
'replied'	–0.215	'hundred'	–0.195	'satisfactory'	–0.225
'hundred'	–0.216	'interesting'	–0.208	'character'	–0.232
'constitution'	–0.236	'character'	–0.219	'company'	–0.233
'england'	–0.243	'country'	–0.223	'corporation'	–0.262
'circumstances'	–0.250	'temperance'	–0.234	'consideration'	–0.273
'character'	–0.257	'beautiful'	–0.235	'corresponding'	–0.278
'christian'	–0.257	'intelligence'	–0.255	'advertising'	–0.289
'mademoiselle'	–0.340	'practically'	–0.258	'business'	–0.297
'thousand'	–0.364	'business'	–0.268	'constitution'	–0.325
'monsieur'	–0.479	'little'	–0.345	'temperature'	–0.374

1975–1999		2000–2008	
'considerable'	–0.387	'elizabeth'	–0.118
'distribution'	–0.391	'family'	–0.126
'individual'	–0.398	'treatment'	–0.130
'behavior'	–0.430	'problems'	–0.132
'population'	–0.431	'william'	–0.157
'general'	–0.454	'married'	–0.165
'development'	–0.487	'william'	–0.157
'administration'	–0.490	'temperature'	–0.197
'education'	–0.543	'said'	–0.233
'government'	–0.630	'children'	–0.262