Towards the Quantitative Approach to Studying Evolution of English Verb Paradigm

Maria Glushko
Voronezh State University

1. Introduction

This paper is a part of doctoral thesis carried out in the field of lexico-semantic prognostics. There exists in linguistics a traditionally sceptic view on the possibility of forecasting the further development of language. Few isolated attempts were made in this respect, primarily in phonetics and morphology by Polivanov (1957), Shevoroshkin (1973), Whatmough (1956). The main objectives and methodology of linguistic prognostics were formulated by Kretov (1992,1993).

Linguistic prognostics employs approaches traditionally used in lexicology, lexicostatistics, mathematic linguistics, linguistic synergetics and corpus linguistics.

Kretov (1993) divides linguistic prognostics into three major subfields:

- ontognostics, i.e. explication of the unobserved fragments of the language system at its present stage;
- futurognostics, i.e. extrapolation of the revealed tendencies of the development of the system into future;
- retrognostics, i.e. reconstruction of the past stages of the system.

Both semiotic and functional approaches are employed by linguistic prognostics. Semiotic approach allows revealing theoretically semantic and grammatical peculiarities of linguistic units. The functional approach to the language, which allows tracing the dynamics of the functioning of linguistic units and then making conclusions about the future development of the language system. I employed functional approach in my work.

However, the main objectives of this paper are not predictions of the future of English verb paradigm, but the preparatory stage, i.e. tracing the general tendencies of its development. I will try to answer three questions:

1) What changes do English verb paradigms undergo?
2) How are these changes carried out, i.e. what are the mechanisms of alterations?
3) Why do these changes take place?

Nordlyd 31.1:30-45, © the author 2003
Proceedings of the 19th Scandinavian Conference of Linguistics, vol. 31.1,
Edited by Anne Dahl, Kristine Bentzen, and Peter Svenonius

2. Corpus

A few texts available on a public domain on the internet were collected for this research. A list of texts is available in the appendix to this paper. Certain constraints were imposed on the choice of texts:

a) Texts belong to different authors.
b) Texts are authentic.
c) Texts are processed to the full.
d) The length of the text exceeds 10, 000 words.
e) Texts represent the chronological scale as full as possible.

Apparently, it is difficult to meet all the requirements mentioned above, but, nonetheless, I tried to be as close to them as possible. Chronological limits of the period considered are from 1500 until 1998. Such choice is firstly and fore mostly grounded with availability of sufficient amount of e-texts, considered adequate for our purposes. Since irregular verbs are remarkable of high frequency of occurrence, texts of any genre could serve our purposes perfectly. Here I would like to acknowledge that since I am interested in the entire verb paradigm, not a particular semantic group, I don't see any need to set restrictions on the genres and define certain 'habitat' of my sample groups. Neither am I interested in personal style, social or educational background of the authors. Indeed, a text is a sample of immediate speech a linguist studies in order to obtain information about the language as an abstract construct. And I dare say that the influence of a particular writer on a language change is very close to zero. Both text's and author's characteristics will certainly comprise within figures received for each particular case. But fluctuations on their expense I consider non-determinate and irrelevant in language change studies. Deviations of this kind should rather be leveled within statistic generalization I am trying to perform.

All in all 140 books were processed. A corpus of the research is comprised with approximately 8 million words.

3. Labour saving measures

Statistic studies of the language are time consuming and painstaking primarily due to the enormous amount of calculations a scholar has to take up in order to receive the desired results. Fortunately, computer sciences at their present stage of development allow linguists to borrow methodologies and approaches for simplifying the calculation process and riding of monotonous part of work.

For these studies a text processor was produced in Delphi. It provides me with two sets of figures:

- absolute frequency of occurrence, i.e. how many times the given lexical unit occurs in the text;
- relative frequency of occurrence, i.e. absolute frequency divided by the number of words in the text.

I also made use of tables and graphic representations of Excel for summarizing and reformatting my results, driving statistic generalizations, and presenting the received suggestions in the most comprehensible form.

4. A hypothesis

The general hypothesis for this research is based on the language regularity proposed by Zipf (1935): among the words characterized with the high frequency of occurrence, the percentage of old ones is much higher then of newly coined ones. The latter are characterized with the low frequency of occurrence.

The higher frequency the word possesses, the more chances it has to secure in the given lexicon, the more resistant it is to alterations. Consequently, the lower frequency of occurrence the word has, the more probable is its loss from the lexicon.

Applying this assumption to the irregular verb, I propose, that irregularity of past indefinite and past participle forming can be put down to the fact that especially due to the high frequency of occurrence they remain unchanged from the ancient times till nowadays.

It is also assumed, that the main factor is frequency of occurrence. Hence the diminution of the functional load results in transition of the verb from the irregular into regular.

5. Are irregular verbs really frequent?

To check this idea I made use of British National Corpus (BNC), available on a public domain on the internet. Let me remind the principles according to which BNC was compiled.

The first 5,000 words of all documents (=files) longer than 5,000 words in the written part of the BNC were taken. There were 2018 of these, so the sub corpus was slightly over 10M words. A frequency list was produced for each of these (truncated) documents. Then, taking the 8189 word-pos pairs occurring 100 times or more in the sample, a 2018x8189 table giving the frequency of each word in each document was produced.

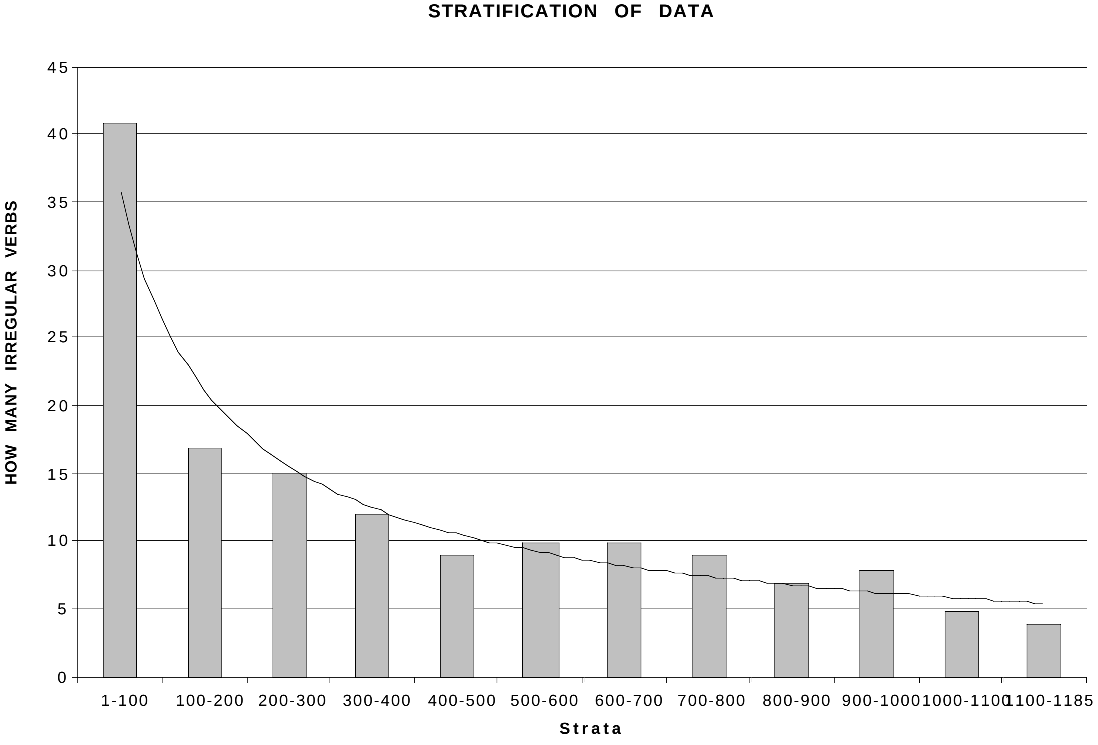
The description above justifies for BNC being a reliable source for any statistic research on the language. I was primarily interested in the distribution of the frequencies of irregular verbs compared to those for regular ones. Hence, I picked all verbs from the BNC and received a list of approximately 1200 verbs arranged according to the principle of diminution of the frequency of occurrence.

Then I conventionally subdivided this list into 12 strata, 100 verbs in each. According to my general assumption the ‘irregularity’ of verbs should be put down to their high frequency of occurrence, due to this fact they have remained unchanged throughout centuries. Applying this idea to the data described, I propose that the closer is the end of the given list of verbs, the fewer irregular ones can be spotted among them.

Plain calculations revealed the following results presented in a table:

Ordeal number of stratum	1	2	3	4	5	6	7	8	9	10	11	12
How many irregular verbs	41	17	15	12	9	10	10	9	7	8	5	4

I also presented these results as a histogram.



The dynamics of irregular verb distribution is explicitly shown on the graphic above. In order to generalize the results I set up a trend, which reveals the main tendency of distribution. I chose exponential function as approximating, as it appeared to be the most demonstrative one in this case.

According to the histogram, the first stratum contains the biggest number of irregular verbs. Comparing it to the second stratum (as well as the following ones), it can be noticed that there exists a gross gap in number of irregular verbs per stratum.

The received trend clearly justifies for the tendency of monotonous decrease of the number of irregular verbs per stratum whereas approaching the end of the list. This fact confirms the assumption about the high frequency of irregular verbs and allows proposing that the high frequency of occurrence provides the 'conservation' of irregular verbs, i.e. their possessing irregular forms of past indefinite and past participle.

6. How do irregular verbs turn into regular ones?

As I've mentioned above, the general assumption is that irregularity of past indefinite and past participle forming can be put down to the fact that especially due to the high frequency of occurrence they remain unchanged from the ancient times till nowadays.

It is also assumed, that the main factor is frequency of occurrence. Hence the diminution of the functional load results in transition of the verb from the irregular into regular.

6.1 Sample group

To probe the assumption, the tendencies of functional load dynamics have been analyzed in respect of three classes of verbs:

1) irregular verbs
2) regular verbs
3) transitioning verbs, i.e. verbs characterized with having both regular and irregular forms of Past Simple and Past Participle.

A random selection of the verbs, belonging to corresponding classes has been carried out. I again made use of British National Corpora word list. The list is arranged according to decrease of relative frequency of usage principle. To make the selection representative enough, the list was initially divided into three strata, conventionally named as:

1) high frequency of occurrence stratum;
2) medium frequency of occurrence stratum;
3) low frequency of occurrence stratum.

Stratification has been carried out on condition that strata have equal number of verbs and is acquired for convenience of processing purposes.

In order to avoid erroneous interpretations, all homonymic forms of verbs, characterized with both regular and irregular forms, I excluded of the consideration. Hence, a number of verbs were initially rejected: give, put, work, leave, lead, let, send, lie, stay, ring, stick, hide, tear, wind, spell, shed, weave.

Further on, for each stratum 10 verbs of each class were selected at random. As a result, for each class, representative groups amounting at 30 verbs each, were received. Each group contained 10 verbs from each stratum. I tried to clarify it in the table below.

	high frequency of occurrence stratum	medium frequency of occurrence stratum	low frequency of occurrence stratum
Irregular verbs	10	10	10
Regular verbs	10	10	10
Verbs, possessing both regular and irregular forms.	10	10	10

Let's enumerate the verbs, included into the sample group:

Regular verb class: adjust, agree, arrive, amuse, behave, boast, celebrate, check, chop, clutch, collapse, drill, engage, exaggerate, exhaust, inspire, interrupt, join, lack, merge, occur, oppose, produce, punish, save, solve, strain, turn, want, watch.

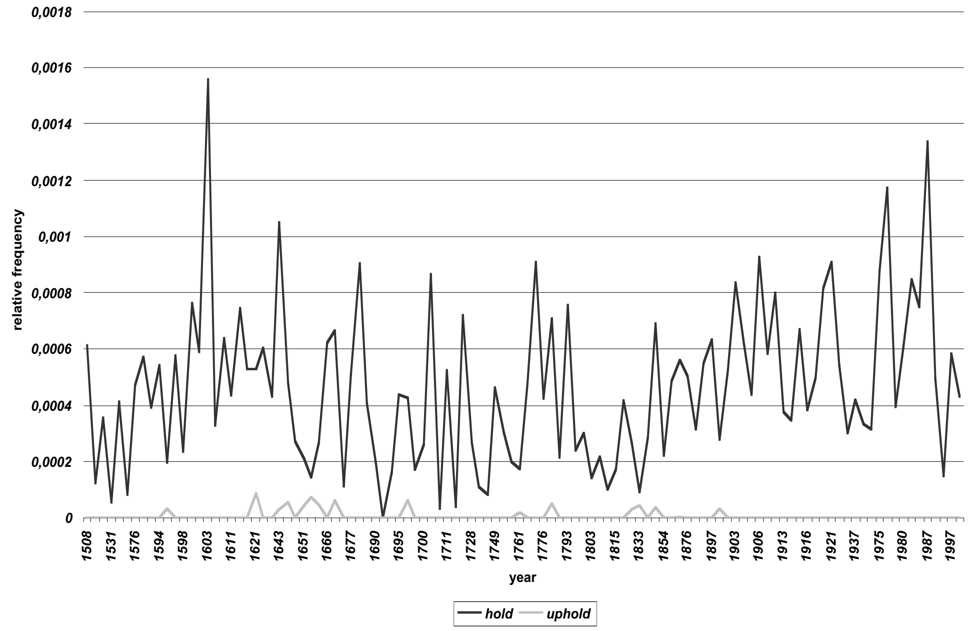
Irregular verb class: bleed, breed, cast, cling, creep, deal, draw, feel, fling, forbid, forget, freeze, hold, lend, meet, overtake, repay, shrink, speak, spin, stand, stride, sweep, swim, swing, teach, tell, uphold, weep, write.

Transitioning verb class: bind, break, burn, climb, dare, dip, dive, dream, fly, hang, kneel, lay, lean, leap, learn, melt, mix, plead, rid, show, smell, spill, spoil, strive, swell, trap, tread, wake, wrap.

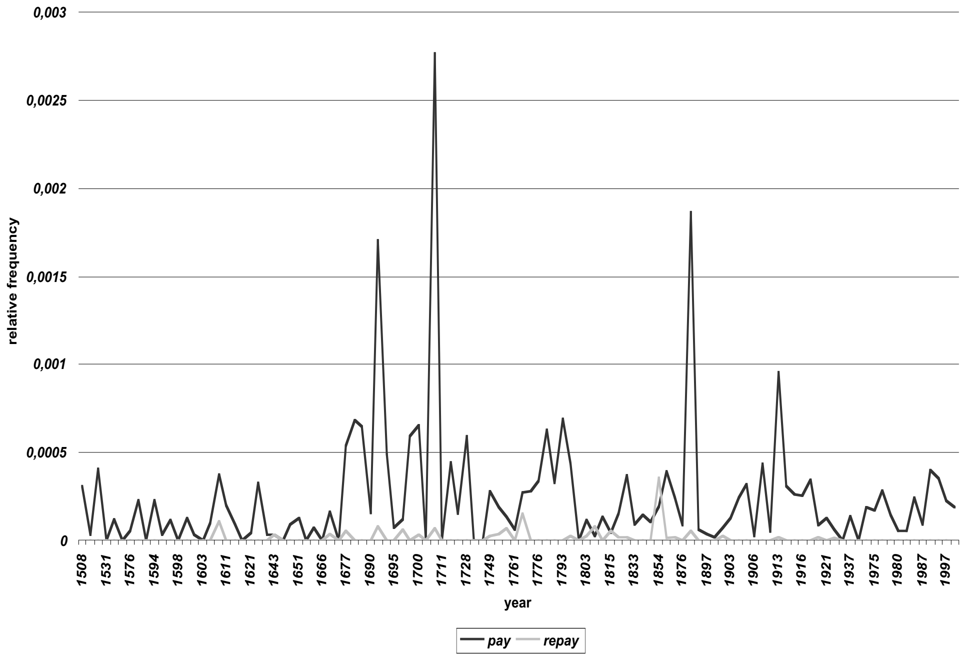
Hence, our sample group is comprised with 90 verbs.

Three derivative verbs from low-frequency stratum were included into the representative sample, i.e. uphold, repay, overtake. First task was to check whether their inclusion as of independent lexical units was relevant or whether the dynamics of evolution of functional load on derivatives would repeat those of producing verbs. For this purpose graphics of functional loads evolution were received and compared. I obtained relative frequencies of occurrence for verb pairs, i.e. hold/uphold, pay/repay, take/overtake. Below (graphs 1, 2, 3) one can see the curves representing the interposition of the dynamics of these verbs.

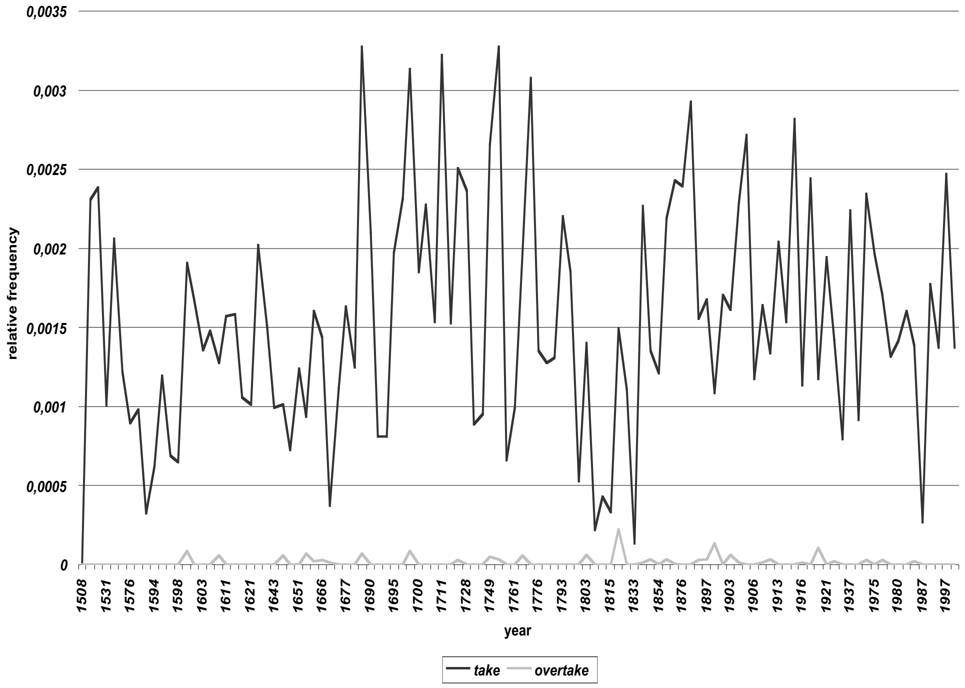
GRAPH 1
Hold & uphold



GRAPH 2
Pay & Repay



GRAPH 3
Take & Overtake



Apparently, no mutual dependence of oscillations can be spotted, which proves the inclusion of derivatives into the sample group is relevant.

6.2 Counts

The first step was the input of the verbs from the sample group into the text analyzer. Since I was interested in the functional load on each verb, I also had to take into consideration its forms, i.e. past simple form, past participle form, gerund, and third person singular form. Absolute frequencies of occurrence of these forms were summed to form the data on each of the given verbs.

The second step was processing the 140 texts described above. All results received were registered in Excel tables, which I am not including into my article due to their being far too voluminous to be presented anywhere but for the thesis itself.

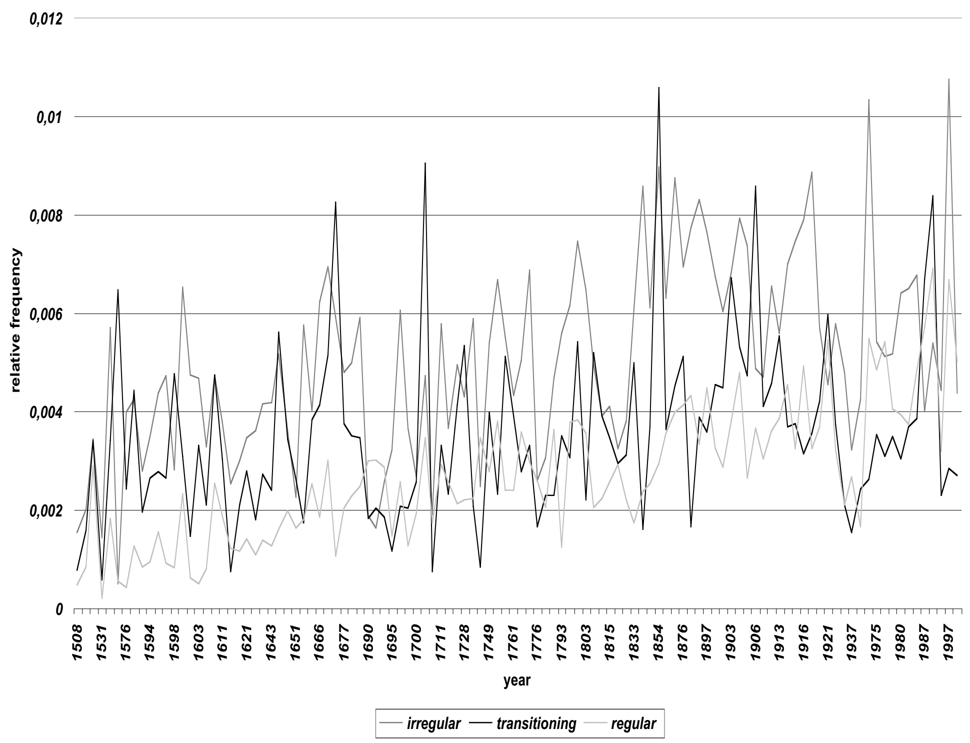
The emphasis of the third step was generalization. As a result of processing the texts I received the data on three groups of verbs:

- irregular;
- regular;
- transitioning, i.e. ones, possessing both regular and irregular forms.

Hence, I divided all frequency of occurrence data into three classes and summed figures that referred to each year. Three rows of figures were received – dynamics of frequency of occurrence evolution for irregular, regular, and transitioning verbs consequently.

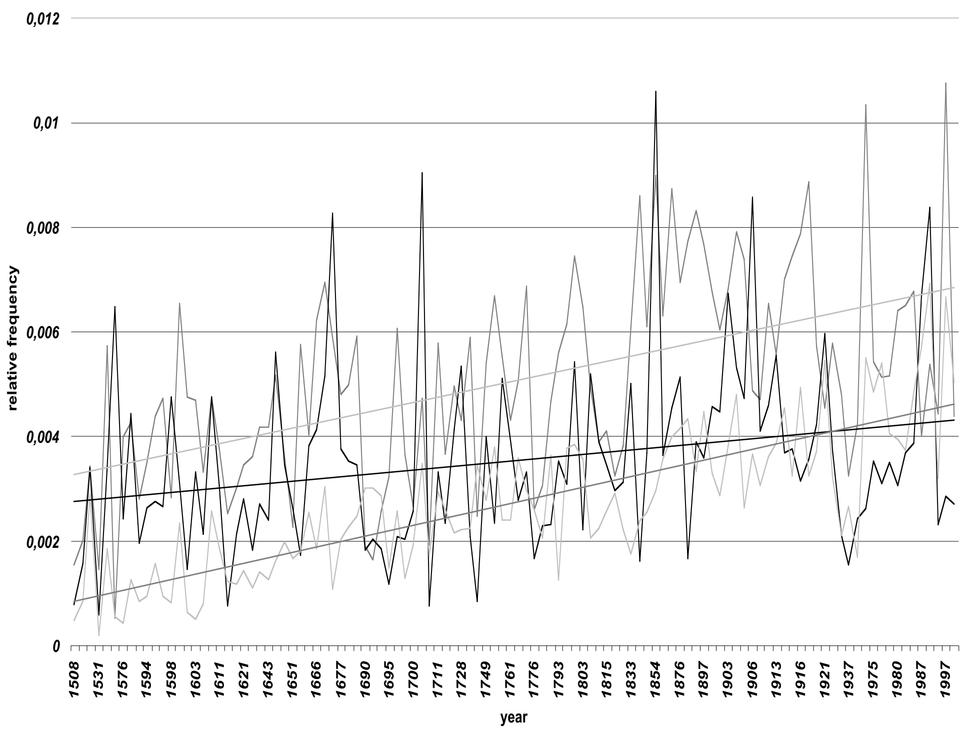
The fourth step was producing a diagram with the curves demonstrating the dynamics of the oscillations that the functional load on the verb paradigms undergo (graph 4).

GRAPH 4
Genetalization of Tendencies



Unfortunately, these curves do not provide one with a clear idea on the changes of the functional load on each of the verb groups. Consequently, on receiving graphic representation of dynamics, three linear trends were set that represent maximally generalized tendencies of evolution of verb classes we are interested in. On the graph 5 their interrelation and co position are demonstrated.

GRAPH 5
Genetalization of Tendencies



_____ linear trend, irregular verb group

_____ linear trend, transitioning verb group

_____ linear trend, regular verb group

Apparently, irregular verbs as a whole possess the highest frequency of occurrence. Linear trend, characterizing this class, is placed in the zone of highest frequency of occurrence. Strictly parallel to it, in the zone of relatively low occurrence, one can see a linear trend that characterizes irregular verbs. As it should have been expected, zones of frequencies of irregular and regular verbs don't overlap.

Finally, the linear trend, that characterizes evolution of 'transitioning' verbs frequency of occurrence is situated between the two described above. Besides, it is easy to notice, that in the beginning of the considered period the so-called 'transitioning' verbs are closer to irregular ones I terms of relative frequency of occurrence. By the beginning of the 20th century, the averaged relative frequency of occurrence of this verb class equals with that of regular verbs.

Further on, the tendency of diminution of the averaged relative frequency of this verb class remains. Thus, our assumption that transition of English verbs from the category of irregular ones to the category of regular ones, is stipulated by their frequency of occurrence, is supported with the results of this experiment.

7. New horizons

Although I consider the interposition of the linear trends to be highly satisfactory, there still are some points left to be explained later. As it can be easily noticed, according to my results the entire English verb paradigm is directed upwards, which in terms of frequency means that the functional load on all verb classes is growing. This, in its turn, hypothetically means that each existent verb these days is used more often then it was used in the 16th century.

I don't have ready consistent explanations of this fact yet. However, the plausible idea is that this process can be put down to the analytic development of English. I'm not currently aware of any experimental studies of parts of speech in general with regard to English. But theoretical premises are a few. Melnikov (1988) gives a detailed account of analytic/synthetic development of the language from the position of systematic linguistics. Polikarpov (1997) provides an exciting overview of the works that refer to this problem, as well as presents an interesting experiment on the development of Russian. Research has been has been carried with regard to Roman languages, i.e. Kapitan (1994) and Titov (2001). Again, they don't study the dynamics of the parts of speech, but their representation in the dictionary. However, for English, as well as for Roman languages it has been spotted that the share of verbs in the dictionaries reduces significantly over time on the expense of the noun class. Which might probably form the starting point for explaining my results on the increase of the functional load on the verb paradigm – but this is not more then the hypothesis I expect to be the most productive so far. However, any constructive critic in this respect is highly welcome.

8. Language change and linguistic variation

After studying evolution of verb classes in general, the focus of our interests was shifted to the group of verbs possessing both regular and irregular forms, i.e. verbs on the stage of variation.

Language change, as any process, comes in several stages. Baudouin de Courtenay (1963) defined four phases of the language change:

Phase A	snik
Phase B	snik:snek
Phase C	snek:snik
Phase D	snek

Apparently, between the phases B and C equilibrium distorts in favour of the new characteristics.

Rastorgueva (1989) defines three stages of the language change:

- appearance of the new features;
- their co-existence and competing with the old ones;
- final acceptance of new features and disappearance of the old ones.

The first stage corresponds phase A by Baudouin de Courtenay. Characterizing the second stage Rastorgueva (ibid.) unites phases B and C into one, implying that it can be subdivided into any number of synchronic stages with different correlation of old and new features. The third stage by Rastorgueva is identical to phase D by Baudouin de Courtenay.

According to Rastorgueva (ibid.) language change goes the following way:

Initial stage	Stage of variation	Final stage
_1	_1	---
---	_2	_2
appearance of parallel variants	coexistence of parallel variants	selection

It is important to keep in mind that not only one, but a few features can be changing simultaneously and each of them has competing parallel variants. Besides, in the schemes above it is not shown as a rule that there are parallel variants at any stage, although at the stage of variation they are the most numerous.

9. From stage 2 to stage 3

In this part of the paper I will demonstrate the study of the verbs on the stage of variation, i.e. the modern verbs that have the biggest number of parallel variants. For compiling examples of such verbs I made use of the encyclopedia by Russakovsky (1998) which the author claims to be the most full edition on English verb forms ever.

The main objective of this part of the research is to understand the mechanism of the selection. Applying the initial assumption, it is proposed that the major role in the selection process is played by the frequency of the verb occurrence. Hence, I was interested in how the functional load on the verb itself influences the correlation of the amount of its regular and irregular forms.

9.1 Sample group

I selected 77 verbs from Russakovsky (ibid.) that registered in the dictionaries as possessing both regular and irregular forms.

9.2 Corpus

The corpus is the same as described above.

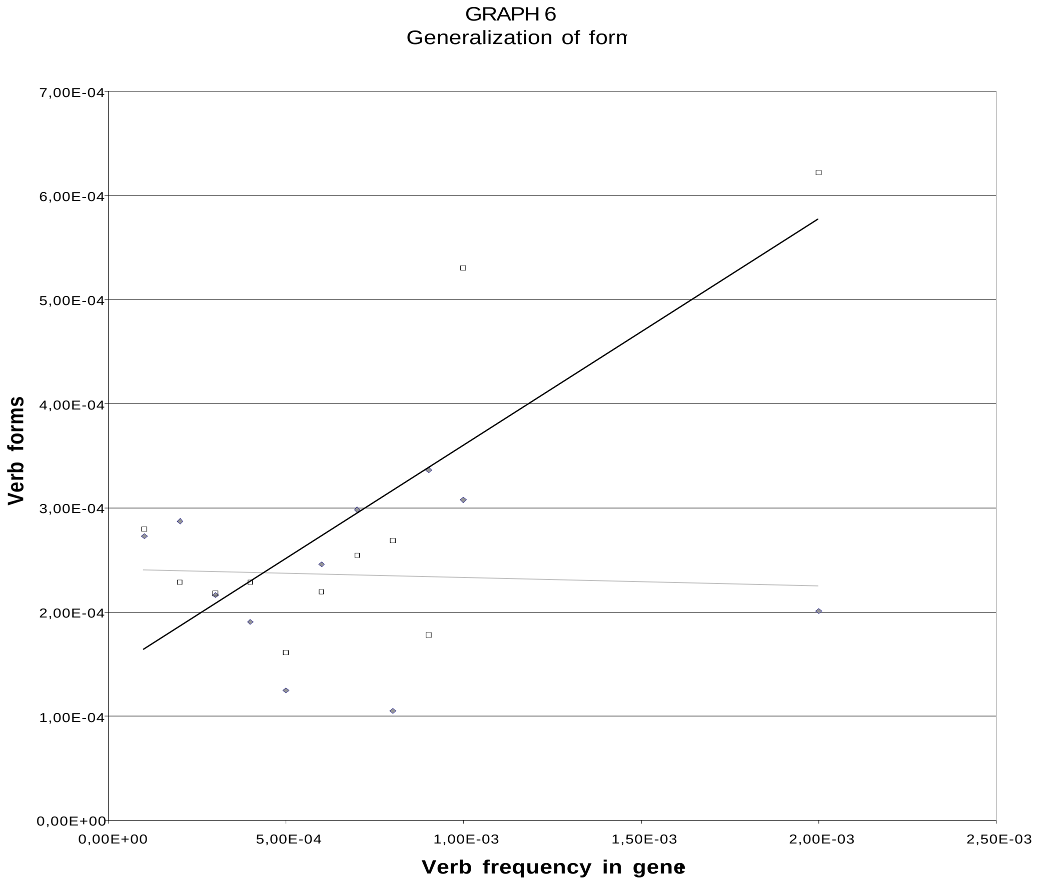
9.3 Counts

For each of these verbs several counts have been made. I was interested in relative frequency of occurrence of the verbs in general (i.e. the sum of absolute frequency of all verb forms, found in the given text, divided into number of words in the text). There have also been received results on the relative frequency of occurrence of irregular and regular forms of these verbs.

After receiving data on all the three properties described above it was found out how frequency of the verbs correlates with that of its forms (regular/irregular). The frequency interval of verb occurrence, starting at 0 and ending at 0,006230704 was divided into 11 equal pieces – from 0 to 0,001 and higher. Then I studied the correlation of verb frequency with that of its forms. I'll try to make my further actions more comprehensible, providing an example.

The verb **learn** in the text dated 1516 has a relative frequency of occurrence **0,00075561**, which goes into interval 8 out of 11 described above. Then I check what frequencies regular and irregular forms have. Irregular form **learnt** does not occur in this text. Hence, this time for irregular forms in interval 8 I write 0. Regular form **learned** has a relative frequency of occurrence **0,000307.** This figure I write down to regular forms interval 8.

The same operation I carried out for each of the 77 verbs throughout 140 texts. As a result, for each interval I receive two groups of figures: for irregular and regular forms. Then I summed frequencies for regular forms in each interval, as well as for irregular ones. Eventually, I received 11 values for irregular forms and 11 values for regular ones and set up two linear trends that demonstrate the maximum generalized correlation of the verb frequency with the distribution of its forms.



_____ regular forms
_____ irregular forms

The graph shows that there does not exist uniform tendency for regular and irregular forms. The linear trend, characterizing generalized tendency of irregular verb forms has an almost constant character on the whole interval. Decrease of functional load, registered on the considered corpus, I consider insignificant. At the same time the linear trend, characterizing generalized tendency of functional load on regular verb form bears strongly pronounced increasing character.

Taking into consideration the described above assumption that the functional load on verb paradigm is growing steadily, it is proposed that increase of frequency of occurrence of the verbs possessing both regular and irregular forms is provided with growth of regular verbs occurrence, whereas functional load on irregular forms remains almost constant.

10. Conclusion

This paper agues for several major findings. Firstly, the process of regularization of the English verb paradigm is primarily facilitated and predetermined by the change of the functional load factor. I.e. the high frequency of occurrence acts as a conservation force preventing a verb from the transition into the regular category. The decrease of the functional load on a verb results in its converting into regular. Secondly, an important fundamental insight I consider the fact that the functional load on the entire English verb paradigm increases over centuries. Thirdly, for the transitioning verb group this uniform increase is carried out on the expense of regular conjugation forms, i.e. certain amount of irregular ones survive but these days 'strong' past forms are used more frequently then they used to in XVI-th century.

References:

Baudouin de Courtenay, J. 1963. *Ob obshih prichinah yazikovih izmenenii. Izbrannie trudi po yazikoznaniu,* Moskva.

BNC http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

Kapitan, M.E. 1994. 'Influence of Various System Features of Romance Words on their Survival,' *Journal of Quantitative Linguistics*, 1:3, 237-250.

Karpilovska, E.A. 1990. *Konstruvannya skladnih slovotvorcheskih edinits,* Naukova dumka, Kiiv.

Kretov, A.A. 1992. 'Nauchnii prognoz v leksicheskoi semantike,' *Funktsionalnaya semantika slov,* Sverdlovsk, pp. 99-110.

Kretov, A.A. 1993. *Osnovi leksiko-semanticheskoi prognostiki,* Dissertatsia na soiskanie uchenoi stepeni doktora filologicheskih nauk, Voronezh.

Melnikov, G.P. 1988. *Systemology and Cybernetic Problems in Linguistics*, Gordon and Breach, L.- Sidney.

Otsaluk, S.I. 1993. 'Prognozuvannya semantichnoi derivatsii: postanovka problemi,' *Materiali mezhdunarodnoi naukovoi konferentsii "Semantika movi tekstu",* Ivano-Frankivsk, pp. 231-232.

Palmaitis, M.L. Toporov, V.N. 1984. 'Ot rekonstruktsii staroprusskogo do rekriatsii novoprusskogo,' *Balto-slavyanskie issledovaniya*, Nauka, Moskva, pp. 36-63.

Polikarpov, A.A. 1997. 'Some Factors and Regularities of Analytic Synthetic Development of Language System,' http://lexigraph.nm.ru/

Polivanov, E.D. 1957. 'Foneticheskie konvergentsii,' *Voprosi yazikoznaniya* 3, 77-83.

Rastorgueva, T.A. 1989. *Ocherki po istoricheskoi grammatike angliiskogo yazika,* Visshaya shkola, Moskva.

Russakovsky, E.M. 1998. *Entsiklopedia form angliiskih glagolov,* Prestizh, Moskva.

Shevoroshkin, V.V. 1973. 'Distributivnaya fonetika russkogo yazika v sravnitel'no tipologicheskom aspekte,' *Problemi strukturnoi lingvistiki*, Nauka, Moskva, pp. 575-585.

Titov, V.T. 2001. 'Kvantitativnaya harakteristika chastei rechi v romanskih yazikah,' *Vestnik VGU,* Voronezh, pp. 74-95.

Whatmough, Joshua. 1956. *Language. A modern Synthesis,* Secher and Warburg, London.

Zipf, George Kingsley. 1935 *The Psycho-Biology of Language,* Cambridge, Mass.