Studies
in Quantitative Linguistics
1

Udo Strauss
Fengxiang Fan
Gabriel Altmann

Problems in Quantitative Linguistics 1

Second edition

RAM - Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Peter Grzybek (grzybek@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, IX +132 pp. (2nd edition).

ISBN: 978-3-9802659-4-2

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Problems in Quantitative Linguistics 1

by

Udo Strauss
Fengxiang Fan
Gabriel Altmann

Second edition

2008
RAM-Verlag

Introduction

This book is the first volume in the series "*Problems in Quantitative Linguistics*", which presents selected proposals for research, problems, questions, hypotheses, and exercises taken from various quantitative-linguistic fields. Only very few of the issues presented here have been studied in previous investigations; each of them is of serious scientific interest and can lead to findings, which may contribute to the construction of a complex linguistic theory.

The problems are of different degree of difficulty and cause different effort if tackled. Many of them can help students in the choice of themes for theses, academic teachers in finding appropriate exercises and examples for their courses, or researchers looking for new enterprises. Most of the hypotheses afford an opportunity to form an original contribution to one of the QL fields by finding a first answer to a given question, a solution to problem, a new method or approach, or an application of existing ones to new linguistic data.

The great majority of the problems concern interrelations between two or more linguistic entities. The reader is asked to set up exact definitions, quantifications and measurement methods, to collect data, perform tests, find an empirical function or derive a function from theoretical assumptions; a complete solution, however, is not always required. In the few cases where a solution or method can be found in the references the reader should feel encouraged to test it on data from other languages, text types, dictionaries etc. or to find an alternative solution.

The individual problems are presented in a unified form throughout the book as follows: (1) A *hypothesis* or a *problem* is given together with sources that should be read. These sources often provide preliminary analyses of the problem and further references. (2) A *procedure* is proposed with suggestions for the appropriate steps in the analysis. Sometimes, an in-depth analysis of the presented problem is given. (3) *References* are provided where the interested reader can find the first mention or a deeper analysis of the problem. A corresponding remark indicates if a reference is mandatory before a problem can be approached.

The instructions given with the problems do not always contain ready-made formulas; in these cases, the reader is referred to the references or to statistics text books.

The following general recommendations may help with a successful work:

1. Linguistic examples cannot be considered as evidence of a phenomenon, pattern, trend or law. The only appropriate empirical basis consists of data from complete objects (e.g. texts) or random samples.

||

2. A correlation analysis is not acceptable as a result; the same is true of a simple test of differences between objects. You should rather find at least an empirical function.
3. English or German are fine but we recommend enriching your study by at least one other language.
4. Empirical findings are often prematurely generalised. Corresponding empirical statements should be tested on several languages, text types, authors etc. depending on the kind of hypothesis.
5. Concepts, quantifications and measurements must be defined in an absolutely explicit and unequivocal way. Avoid concepts you cannot operationalise with sufficient exactness.
6. Always try a derivation of the function or distribution you assume for your data from reasonable theoretical assumptions. Often, proportionality considerations may be successful as a number of hypotheses in synergetic linguistics have shown.
7. If a function or distribution seems inadequate with respect to your data, recheck your data (sources, pre-processing, amount, artificial factors etc.), calculation, computational procedures – and your assumptions. Change or correct whatever turns out to be wrong and try once more.
8. If your mathematical model fails again: sometimes, there are some boundary conditions which affect a relation (although we think that the law of gravitation is valid we observe that some objects, e.g. birds, do not drop). Find such boundary conditions in your case and consider them as independent variables. Re-formulate your hypothesis correspondingly and start again.
9. No hypothesis should be definitively rejected or definitively accepted. Corroboration is a matter of degree.
10. To clarify your thoughts, work out a diagram of the relationship including parameters and requirements (cf. the notation in synergetic linguistics).
11. Keep in mind that data are constructs, i.e. to some extent artificial. Data collection consists in transforming facts via hypotheses (or a weaker form of assumptions or expectations) into statements. Hence, one should first set up an explicit and plausible hypothesis – then search for data.
12. If it is difficult to determine which variable is dependent and which is independent, try to integrate both variants in a larger control cycle or at least test both directions.
13. After solving several problems try to integrate all of them in a control cycle. Fill the missing vertices and edges by hypothetical ones and try to find them empirically.
14. Never give basic data in the form of percentages; always present absolute numbers.
15. When a problem is solved, do not consider it the final solution; see it as part of a greater perspective and try to describe this perspective.

16. If you think you need a classification do not just classify mechanically using a method at hand. Instead, try to set up a theory and deduce an appropriate classification from this theory.
17. Do not use functions with many parameters (e.g. polynomials) because later on these parameters will have to be interpreted (i.e. adhere to "Occam´s razor").
18. If possible, as linguist, cooperate with a programmer and a mathematician. If you are mathematician you should seek an experienced linguist, otherwise a good mathematical model may be developed – however without linguistic interpretation and hence without use.
19. Try to apply solved problems introduced in this book using new data (from other languages) so that existing theories can be corroborated or rejected.
20. Do not consider linguistic units as given a priori. Define units operationally in such a way that they can be used in hypotheses, even if their segmentation might seem somewhat artificial. Keep in mind that those linguistic units are theoretically prolific which can be used in formulating laws (not in grammatical rules).
21. Always prefer functions or distributions with a good theoretical foundation to ones which possibly displays a better fit but have no linguistic background. i.e. use empirical functions only at the beginning of a research.
22. There are nine chapters in this book. The contents of the individual chapters are not strictly homogeneous but furnish a relatively broad view of possible problems that can be solved using quantitative methods. Within each chapter, the problems are arranged alphabetically. Some problems have been analysed in more detail. Neither the chapters nor the problems need to be read successively; one can choose a problem according to one´s own preference and specialisation.

Acknowledment
We are very obliged to Reinhard Köhler who patiently read the whole book, corrected our style and our Pidgin, improved some argumentations and gave us a number of useful hints. The rest consists of our besetting sins.

Contents

Chapter 1

Phonemics and script

Accent and frequency

Hypotheses

"…words which occur most frequently are generally not preferred for accentuation." (Zipf 1935: 131)

"…the accent tends (1) to gravitate away from words of high frequency and (2) toward words in unusual usage…" (Zipf 1935: 132)

"…accent tends to settle on morphemes of the greatest average interval (´wave length´), that is, on the morphemes of the lowest relative frequency…" (Zipf 1935: 136).

Procedure

Get a text, read it aloud and classify the words into stressed words and unstressed ones. Then get the frequency of the words in these two classes from a corpus or a frequency dictionary. Within each class arrange the words in decreasing frequency and perform a nonparametric rank texts showing that the two classes of words (accented and unaccented) do not belong to the same "accent population". Try to perform this test in several languages. If there is a word which is both stressed and unstressed in different neighbourhoods, put it in both classes or eliminate it from the sample.

Reference

Zipf, G.K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. Boston: Houghton Mifflin.

Canonical word structure

Hypothesis

The relationship between syllabic and phonemic length of canonical forms is linear.

Procedure

Canonical forms are words whose phonemes have been reduced to consonant and vowel classes. Thus one gets forms like *V*, *CV*, *VC*, *CVC*, *CVV*, etc. Refer to a dictionary and transcribe all words into their canonical forms. If a computer program is used, pay attention to diphthongs and combined graphemes (e.g. E. <*sh*>, G. <*sch*>, <*ei*>, etc). Make a two-dimensional table with syllable number as the first variable and phoneme number as the second one. Show that the relationship <*syllable number, phoneme number*> is linear. Remember that *CV*

and *VC* belong to the same class (1 syllable, 2 phonemes), but *CVC* and *CVV* belong to different ones: <1 syllable, 3 phonemes> and <2 syllables, 3 phonemes> respectively.

Test hypothesis (a) without taking frequencies into account (b) taking frequencies into account. In both cases a linear relationship should result.

Reference

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Consonants and clusters

Hypothesis

A language having many consonants in the inventory also has many and long consonantal clusters (Skalička 1964b). But with increasing phoneme inventory the (relative) number of consonantal clusters decreases.

Procedure

Compute the consonant inventory size and find all clusters in a language. Use either a corpus or a dictionary. Do this for ten different languages and set up a function of dependence. Take data both from languages with small inventories and from those with large inventories.

Reference

Skalička, V. (1964). Konsonantenkombination und linguistische Typologie. *Travaux linguistiques de Prague 1, 111-114.*

Distribution of canonical forms

Problem

The canonical forms in the previous problem (using frequency) have a very regular two-dimensional distribution whose independent variables are syllabic length and phonemic length. Try to derive this distribution theoretically from reasonable assumptions.

Procedure

Use either combinatorial argumentation or a special stochastic process. Nothing is known about the form of this distribution.

References

None.

Distributional calculus

Problem

Perform the complete phonemic distributional calculus (cf. Altmann, Lehfeldt 1980) in a language which has not been studied this way as yet. Use recent literature.

Procedure

Use a dictionary or a corpus and first get all different sequences of two phonemes (not letters). Perform the classical Harary-Paper calculus using new indices. Then count the frequency of all sequences and perform the frequency phonemic distributional calculus. Compute different indicators. State whether a phoneme with high cotextuality (associativity) has also higher frequency. Find the form of this dependence (see "Cotextuality and frequency" in Chapter 4) and set up a hypothesis.

References

Altmann, G., Lehfeldt, W. (1972). Typologie der phonologischen Distributionsprofile. *Beiträge zur Linguistik und Informationsverarbeitung* 22, 8-32.

Altmann, G., Lehfeldt, W. (1980). *Einführung in die Quantitative Phonologi*e. Bochum: Brockmeyer.

Birnbaum, H. (1967). Syntagmatische und paradigmatische Phonologie. In: Hamm, J. (ed.), *Phonologie der Gegenwart: 307-352*. Graz u. a.: Böhlau.

Doležel, L., Průcha, J. (1966). A statistical law of grapheme combinations. *Prague Studies in Mathematical Linguistics 1, 33-43*

Greenberg, J. H. (1964). Nekotorye obobščenija, kasajuščiesja vozmožnych načal'nych i konečnych posledovatel'nostej soglasnych. *Voprosy jazykoznanija 4, 41-65.*

Harary, F., Paper, H.H. (1957). Toward a general calculus of phonemic distribution. *Language 33, 143-169.*

Hirsch-Wierzbicka, L. (1971). *Funktionelle Belastung und Phonemkombination*. Hamburg: Buske.

Kempgen, S. (1995). Phonemcluster und Phonemdistanzen (im Russischen). *Slavistische Linguistik 1994, 197-221*.

Kempgen, S. (1999). Modellbedingte Distributionsbeschränkungen in der Phonologie. In: K. Grünberg, W. Potthoff (eds.), *Ars Philologica. Festschrift für Baldur Panzer zum 65. Geburtstag: 179-184*. Frankfurt a. M. u. a.: Lang.

Kempgen, S. (2001). Assoziativität der Phoneme im Russischen. In: L. Uhlířová et al. (ed.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of L. Hřebíček: 124-135*. Trier: VWT.

Lehfeldt, W. (1972). Phonologische Typologie der slavischen Sprachen. *Die Welt der Slaven 17, 318-340.*

Lehfeldt, W. (2005). Phonemdistribution. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 181-*

190. Berlin: de Gruyter.
Saporta, S. (1955). Frequency of consonant clusters. *Language 31, 25-31.*
Trnka, B. (1936). General laws of phonemic combinations. *Travaux du Cercle Linguistique de Prague 6, 57-62.*
Trubetzkoy, N.S. (1939). *Grundzüge der Phonologi*e. Travaux du Cercle Linguistique de Prague 7. Prague. [Nachdruck: Nendeln: Kraus, 1968]
Vogt, H. (1942), The structure of the Norwegian monosyllable. In: *Norsk Tidsskrift for Sprogvidenskap* 12, 5-29.
Vogt, H. (1954). Phoneme classes and phoneme classification. *Word 10, 28-34.*

Distributional gaps

Hypothesis

The greater the number of phonemes in an inventory, the smaller is the proportion of possible phoneme combinations, i.e. the greater the proportion of structural gaps.

Procedure

First solve the problem “Distributional calculus”, then compute mechanically the number of interaction gaps, i.e. count the unrealized phoneme combinations. Put this number in relation to the size of phoneme inventory. Since data are available in the literature given in the preceding problem, no analysis of new data is necessary. Express the relationship formally.

References

Schulz, K.-P., Altmann, G. (1988). Lautliche Strukturierung von Spracheinheiten. *Glottometrics 9, 1-48.*

Evolution of script complexity

Problem

The symbols of two historical stages of the evolution of any script differ in their symbol complexity. Show that this change is not linear.

Procedure

Take two historical stages of the same script, e.g. Brahmi and Devanagari, Chinese iconic and modern Chinese, Japanese kanji and the respective hiragana (or katakana), Old Assyrian and newer Assyrian, Egyptian hieroglyphs and Meroitic script, etc., and measure the complexity of individual symbols. Consider the complexity of the older variant as variable x and that of the newer as y. (a) Show that the relation is not linear. (b) Try to find an adequate function.

References
Hegenbarth-Reichardt, I., Altmann, G. (2008). On the decrease of complexity from hieroglyphs to hieratic symbols. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties of characters and writing systems*:*101-110.* Berlin/New York: Mouton de Gruyter.

Exploitation of canonical forms

Problem
Find the exploitation function of canonical forms.

Procedure
Consider the phonemic length of canonical forms from the previous problem (i.e. the marginal distribution). Consider only the types, not their frequency. Since there are only two different elements (*V, C*), one can, theoretically, obtain not more than 2 elements of length 1, *V* and *C* (we admit also types *C, CC, CCC* etc. some of which exist e.g. in Slavic languages); there are theoretically $2^2 = 4$ types of length 2 (*VV, VC, CV, CC*), and in general 2^k types of length *k*. Since the observed type numbers are known from the previous problem and the theoretical ones can be computed, construct a measure of type exploitation and find the exploitation function of canonical forms. If possible compare the functions of several languages.

References
Altmann, G. (2005). Phonic word structure. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 191-198*. Berlin/New York: Mouton de Gruyter.
Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt. (p. 48 ff.)

Letter frequency

Problem
Find a common model for the rank-frequency distribution of letters.

Procedure
Rosenbaum and Fleischmann (2002, 2003) brought a number of letter and diacritic distributions from European languages.
1. Try to show that all of them follow the same theoretical distribution.
2. The authors presented also the ranking of Latin letters in the languages analysed. Use different test methods to ascertain the similarity between lan-

guages restricted to letter frequency. If successful, use more ready data and expand the investigation.
3. Draw general conclusions from the results.

References

Rosenbaum, R., Fleischmann, M. (2002). Character frequency in multilingual corpus 1 – Part 1. *Journal of Quantitative Linguistics 9(3), 233-260.*
Rosenbaum, R., Fleischmann, M. (2003). Character frequency in multilingual corpus 1 – Part 2. *Journal of Quantitative Linguistics 10(1), 1-39.*

Measurement of distinctness

Problem

Define a measure of distinctness of individual scripts.

Procedure

Take a runic script and compute its distinctness using the method of Antić, Altmann (2005). Take another runic script and compare its distinctness with the first one (cf. Mačutek 2008). Describe the difference if any. Devise a way to compute the distinctness of the Ogham script.

References

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 4, 46-53.*
Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16, 1-16.*

Measurement of ornamentality

Problem

Ornamentality is not an inherent property of script; it is a property established exclusively by our concept formation. It has no real correspondence but it can be transferred to real objects. Try to find a method for measuring ornamentality of script.

Procedure

One can proceed in three ways:
1. Set up a scale and rely on the judgements of test persons. This method has already been applied.
2. Measure ornamentality as the surplus of complexity of a given symbol over the simplest (semiotically identical) symbol.
3. Devise a new objective method taking inspiration from calligraphy or fine arts.

References
Altmann, G., Fan, F. (eds.). (2008). *Analyses of script. Properties of characters and writing systems*. Berlin/New York: Mouton de Gruyter.

Phoneme frequency and word frequency

Hypothesis
"… low frequency lexical items are composed of more rare phonemes than high frequency lexical items" (Frisch, Large, Zawaydeh, Pisoni 2001: 167)

Procedure
Since phoneme frequency is a direct function of word frequency, the hypothesis is self-evident. Try to make it more exact. Compute the frequency of phonemes and the frequency of word-forms on data from a corpus. Then for each word-form frequency get the frequencies of individual phonemes and compute their average. If the hypothesis is true, a simple function of dependence can be obtained. Try to set up this function, i.e. the dependence of mean phoneme frequency on word-form frequency. Do this for different languages if possible, and compare the results. Try to establish a general statement.

References
Frisch, S.A., Large, N.R., Zawaydeh, B., Pisoni, D.B. (2001). Emergent phonotactic generalizations in English and Arabic. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 159-179.* Amsterdam/Philadelphia: Benjamins.
Frauenfelder, U.H., Baayen, R.H., Hellwig, F.M., Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language 32, 781-804.*
Landauer, T.K., Streeter, L.A. (1973). Structural differences between common and rare words. Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior 12, 119-131*.

Phoneme inventory and word length

Hypothesis
The greater the phoneme inventory, the smaller the average word length (Nettle 1995).

Procedure
Nettle (1995) computed this relationship in 10 languages using phonemic length of words.
1. Add more languages in order to corroborate or modify the hypothesis.

2. If the hypothesis turns out to be weak, add further properties of language and set up a function with two independent variables. An additional property could be e.g. the extent of phoneme distribution (phoneme associativity, number of phonemic bigrams in language).
3. Try to test the hypothesis on texts (not a dictionary).
4. Use mean syllabic length of words as dependent variable and ascertain whether the hypothesis holds.
5. Specify which properties could have influence on word length in a language.
6. Test the hypothesis using average morph length as dependent variable.

References

Hockett, C.F. (1958). *A course in modern linguistics*. Toronto: McMillan.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics 33, 359-367.*

Weber, S. (2005). Zusammenhänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 214-226*. Berlin/New York: Mouton de Gruyter.

Power law

Problem

Try to apply Naranan and Balasubrahmanian´s modified power law and partial sums modified power series distribution to linguistic data.

Procedure

Consider as many phoneme/letter rank-frequency distributions as possible. Fit the distributions to your data. If appropriate software is not available, try to derive an estimation method for the parameters using the frequencies of lowest ranks. Try to interpret the partial sums distribution in linguistic terms.

References

Naranan, S., Balasubrahmanyan, V.K. (2005). Power laws in statistical linguistics and related systems. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 716-738.* Berlin/New York: Mouton de Gruyter.

Ranking syllable types

Problem

Find a distribution for the rank-frequency of syllables.

Procedure

Schiller et al. (1996) presented the percentages of canonical syllable types in Dutch, both types and tokens, and ranked them according to decreasing frequency. Try to find a formal relationship between rank and percentage. Use a function, not a distribution. Draw some conclusions from the result.

References

Schiller, N.O., Meyer, A.S., Baayen, R.H., Levelt, W.J.M. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics 3(1), 8-28.*

Script complexity

Problem

Up to now there are several possibilities of measuring script (symbol) complexity: the intersection method, the scaling method, Bézier curves, stroke number counting, pixel counting, fractal dimension etc. Try to define a new measure or try to compute all existing measures for the same script and compare them.

Procedure

Since some of the measures capture only isolated properties of script, try to set up a measure taking into account (a) the form of the lines, (b) the length of the lines, (c) the direction of the lines, (d) the connection of the lines. Apply the measure to the Arial fonts and compare it with the existing results.

Take the Hungarian runes, which can be found on the Internet, and compute the complexity of each symbol. Use known methods of measuring complexity. Then take other runes and compare their complexity with that of Hungarian. Consider especially the Ogham script.

References

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-73.*

Altman, G., Fan, F. (eds.). (2008). *Analyses of script. Properties of characters and writing systems.*. Berlin/New York: Mouton de Gruyter.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16, 1-16.*

Script simplification

Problem

Show that in the evolution of script the symbols were continuously simplified, and whether this simplification is linear.

Procedure

1. Use the script complexity measure proposed by Altmann (2004). Use Haarmann (1990) or Omniglot (Internet) and choose a table in which some historical epochs of a script are presented (e.g. the Arameic script p. 301). Compute the complexity of "old" form and "new" forms and the extent of simplifications.
2. Perform the same procedure concerning Japanese kanji forms and hiragana and katakana forms that developed from them.
3. Select the oldest Chinese iconic signs and compute the process of complexity change comparing them with modern Chinese symbols.
4. Take the oldest Assyrian cuneiforms and observe the change in their complexity up to the latest forms.
5. Take several runes (which can be found on the Internet, e.g. at Omniglot) and compute their complexity. State whether their mean complexity is statistically equal and give the reason. If negative, ascertain whether their age (time of first appearance) affects complexity. Try to find causal, psychological, social etc. factors causing difference in complexity. Solve the problem: is simplification of symbols linear or does it follow another trend?
6. Compare the hieroglyphs with the Meroitic script.

References

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-73.*

Haarman, H. (1990). *Universalgeschichte der Schrift.* Frankfurt: Campus.

Hegenbarth-Reichardt, I., Altmann, G. (2008). On the decrease of complexity from hieroglyphs to hieratic symbols. In: Altmann, G., Fan, F., (eds.), *Analyses of script. Properties of characters and writing systems: 105-114.* Berlin: Mouton de Gruyter.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16, 1-16.*

Syllable frequency

Hypothesis

The rank-frequency distribution of syllables behaves like the rank-frequency distribution of words.

Procedure

Use data from a corpus. (a) If possible, make a phonemic transcription of the

corpus or (b) use the written form. In both cases, partition the words in syllables and compute the frequency distribution of individual syllables (not canonical forms!). Use a syllabification program if available. Set up the rank-frequency distribution of syllables and try to fit a distribution used for word frequencies to this data.

If deviant results are obtained, what can be the cause? Try to reformulate the syllable segmentation algorithm; try to establish boundary conditions and embed them in the theoretical distribution; try to derive a theoretical distribution from combinatorial assumptions.

References

Bektaev, K.B. (1973), Alfavitno-častotnyj slovar' slogov kazachskogo jazyka. In: *Statistika kazachskogo teksta I. Trudy gruppy „Statistiko-lingvističeskoe issledovanie i avtomatizacija" III, 566-611*. Alma-Ata: Nauka.

Schiller, N.O., Meyer, A.S., Bayen, R.H., Levelt, W.J.M. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics 3(1), 8-28.*

Syllable structure

Problem

The description of syllable structure is a set of problems which must be solved stepwise.

Procedure

1. Set up the inventory of syllables in a language. (cf. "Syllable frequency")
2. Solve the problem "Syllable frequency" using a corpus.
3. Study the relationship between syllable frequency and syllable length. Since syllables are relatively short, it will be easy to find a function.
4. Syllables contain an onset and a coda. Study their symmetry and anti-symmetry and set up a symmetry indicator. Find the properties of the indicator.
5. Compare the inventory of syllables with that of phonemes (for several languages). Is there any dependence? If so, find it.
6. Try to set up an exploitation rule, i.e. compute the number of possible syllables of length x and the number of realized ones. Set up a measure of exploitation.
7. Ascertain whether the exploitation measure in (6) has some association to the phonological language type.
8. Form phonemic rules for the forming of onsets and codas.
9. Test the existence of consonant harmony between onsets and codas.

References

Berg, T. (1994). The sensitivity of phonological rimes to phonetic length. *Arbei-*

ten aus Anglistik und Amerikanistik 19, 63-81.
Booij, G. (1995). *The phonology of Dutch*. Oxford, U.K.; Clarendon
Bortoloni, U. (1976). Tipologia sillabica d´italiano. Studio statistico. In: Simone, R., Vignuzzi, U., Ruggiero, G. (eds.), *Studi di fonetica e fonologia. Atti del convegno internazionale di studi. Padova 1 e 2 ottobre 1973: 5-22.* Roma. (Pubblicazioni della Società di Linguistica Italiana 9).
Browman, C.P., Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica 45, 140-155.*
Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics 183-198.*
Derwing, B.L., Yoon, Y.B., Cho, S.W. (1993). The organization of the Korean syllable: Experimental evidence. In: O.M. Clancy (ed.), *Japanese/Korean Linguistics, Vol. 2, 223-238.* Stanford: Center for the Study of Language and Information.
Derwing, B.L., Dow, M.L., Nearey, T.M. (1988). Experimenting with syllable structure. In: J. Powers, K. de Jong (eds.), *Proceedings of the Fifth Eastern States Conference on Linguistics 83-94.* Columbus: Ohio State University.
Eisenberg, P., Ramers, K.-H., Vater, H. (eds.) (1992). *Silbenphonologie des Deutschen.* Tübingen: Narr.
Fallows, D. (1981). Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics 17, 309-317.*
Fowler, C.A., Treiman, R., Gross, J. (1993). The structure of English syllables and polysyllables. *Journal of Memory and Language 32, 115-140.*
Goldinger, S.D., Luce, P.A., Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language 28, 501-518.*
Grainger, J. (1992). Orthographic neighbourhoods and visual word recognition. In: R. Frost, L. Katz (eds.), *Orthography, phonology, morphology and meaning: 131-146.* Amsterdam: Elsevier.
Hall, T. (1962). *Syllable structure and syllable related processes in German*. Tübingen: Niemeyer.
Hamilton, P. (1995). *Constraints and markedness in the phonotactics of Australian languages.* Diss., Univ. of Toronto.
Lamontagne, G. (1993). *Syllabification and consonant cluster conditions*. Amherst, U.Mass. Diss.
Levelt, W.J.M., Wheeldon, L. (1994). Do speakers have a mental syllabary. *Cognition 50, 239-269.*
Maddieson, I. (2005). Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. *UC Berkeley Phonology Lab Annual Report (2005)*
http://www.linguistics.berkeley.edu/phonlab/annual_report/2005/OhalaConfLabReport259-268.pdf
Mohanan, T. (1989). *Syllable structure in Malayalam.* Poona, Deccan College.

Pike, K., Pike, E. (1947). Immediate constituents of Mazateco syllables. *International Journal of American Linguistics 13, 78-91.*

Portele, T. (1995). The influence of the syllable boundary on consonant-consonant realizations. *Proceedings of the International Congress of Phonetic Sciences, Stockholm, vol 2, 594-597.*

Pulgram, E. (1970). *Syllable, word, nexus, cursus*. The Hague: Mouton.

Schiller, N.O., Meyer, A.S., Levelt, W.J.M. (1997). The syllabic structure of spoken words: evidence from the syllabification of intervocalic consonants. *Language and Speech 40(2), 103-140.*

Selkirk, E.O. (1982). The syllable. In: Hulst, H. van der, Smith, N. (eds.), *The structure of phonological representations. Part II, 337-383.* Dordrecht: Foris.

Sommer, B. (1970). An Australian language without CV syllables. *International Journal of American Linguistics 36,57-58.*

Treiman, R., Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language 27, 87-104.*

Treiman, R., Fowler, C.A., Gross, J., Berch, D., Weatherston, S. (1995). Syllable structure or word structure? Evidence for onset and rime units with disyllabic and trisyllabic stimuli. *Journal of Memory and Language 34, 132-155.*

Treiman, R., Zukowski, A. (1990). Toward an understanding of English syllabification. *Journal of Memory and Language 29, 66-85.*

Vennemann, T. (1988) *Preference laws for syllable structure and the explanation of sound change.* Berlin: de Gruyter.

Vennemann, T. (1972). Zur Silbenstruktur der deutschen Standardsprache. In: Vennemann, T. (ed.), *Silben, Segmente, Akzente*. Niemeyer, Tübingen.

Tendency towards vowel harmony

Hypothesis

In bisyllabic morphemes of a language there is a tendency towards vowel harmony.

Procedure

Collect bisyllabic word stems from a dictionary. For some languages the above-presented hypothesis holds (e.g. Indonesian languages). It differs from the usual deterministic vowel harmony holding for affixes, e.g. in Hungarian. Try to ascertain whether it holds for the language you analyze. There are two possibilities: (a) The vowel in the first syllable combines significantly with the same vowel in the second syllable. (b) Some vowels combine significantly with some other vowels but avoid combination with certain ones. Use appropriate tests to ascertain the existence of a “harmony tendency”.

References
Altmann, G. (1987). Tendenzielle Vokalharmonie. *Glottometrika 8, 104-112.*
Schulz, K.P., Altmann, G. (1988). Lautliche Strukturierung von Spracheinheiten. *Glottometrika 9, 1-48.*

Two-dimensional syllable structure

Problem
Try to find the two-dimensional structure of syllables in a European language.

Procedure
First prepare a list of all possible syllables in the given language. Count the canonical types, i.e. the number of types *V, VC, CV, CCV,…* and present their numbers in a table in which the first column contains the consonants in front of the vowel (or syllable bearer), and the first line those occurring behind the vowel, as follows:

	V VC VCC VCCC …
V CV CCV CCCV ……..	

The crossing of *CV* and *VC* means a syllable of the type *CVC*. Test the hypothesis with the numbers that the distribution is

$$P_{ij} = \frac{a^i b^j}{(i!)^k (j!)^m} P_{00}, \qquad i, j, = 0,1,...$$

where P_{ij} is the probability of syllables in line *i* and column *j*; *a, b, k, m* are parameters and P_{00} is the probability of syllables of type *V*. The estimation procedure can be found in Zörnig, Altmann (1993).

If deviation from this model is found, modify the model appropriately or develop a new model based on other assumptions.

Try to analyse several languages and find collateral phonemic properties which can be responsible for the size of the parameters.

References
Lee, Sang-Oak (1986). An explanation of syllable structure change. *Korean Language Research 22, 195-213.*

Vennemann, T. (1982) (ed.). Zur Silbenstruktur der deutschen Standardsprache. *Silben, Segmente, Akzente: 261-305.* Tübingen: Narr.
Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types. *Glottometrika 14, 190-196.*

Word length and supra-segmentals

Hypothesis

The more supra-segmental means a language has, the smaller is the average word length (Kempgen: 119).

Procedure

Investigate different languages having supra-segmentals (different tones, different accents, lengths of vowels) for word differentiation. Compute the average word lengths and find the above dependence. Compare the investigated languages with those without supra-segmentals.

References

Kempgen, S. (1990). Akzent und Wortlänge: Überlegungen zu einem typologischen Zusammenhang. *Linguistische Berichte 126, 115-134.*

Chapter 2

Grammar

Behagel´s "law"

Hypothesis

In a corpus, the greater the difference *x* between the length of two juxtaposed prepositional phrases, the greater the probability *p*(*x*) that the shorter prepositional phrase precedes the longer one (Hoffmann 1999:113). Cf. the problem "Frequency and order in freezes".

Procedure

The hypothesis seems to contradict Fenk-Oczlon´s hypothesis but this is not necessarily the case. First, operationalize the needed concept "prepositional phrase", and then try to express the hypothesis formally, solve what is necessary and use a corpus as data source. Do not restrict yourself to German or English – data from these languages are easily available – rather collect data from other languages. Refer to the attached references.

References

Allen, K. (1987). Hierarchies and the choice of left conjuncts (with particular attention to English). *Journal of Linguistics 23, 51-71.*

Bock, J.K., Warren, R.K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition 21, 47-67.*

Cooper, W.E., Ross, J.R. (1975). Word order. In: Grossman, R.E., San, L.J., Vance, T.J. et al. (eds.), *Papers from the parasession on functionalism: 63-111.* Chicago: Chicago Linguistic Society.

Edmondson, J.A. (1985). Biological foundation of language universals. In: Bailey, C.J., Harris, R. (eds.), *Developmental mechanisms of language: 109-130.* Oxford: Pergamon.

Ertel, S. (1977). Where do the subjects of sentences come from? In: *Sentence production: developments in research and theory: 141-186.* Hillsdale, N.J.: Erlbaum.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 27, 517-556.*

Fenk-Oczlon, G. (1983). Ist die SVO-Wortfolge die ´natürlichste´? *Papiere zur Linguistik 29, 23-32.*

Fenk-Oczlon, G. (1987). Frequenz und Wortfolge. Am Beispiel von ´freezes´. *Paper presented at the XIVth International Congress of Linguists.*

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448.* Amsterdam/Philadelphia: Benjamins.

Hawkins, J.A. (1983). *Word order universals*. San Diego: Academic Press.

Hawkins, J.A. (1990). A parsing theory of word order universals. *Linguistic Inquiry 21(2), 223-261.*

Hawkins, J.A. (1992). Syntactic weight versus information structure in word order variation. In: Jacobs, J. (ed.), *Informationsstruktur und Grammatik.* Opladen: Westdeutscher Verlag.

Hawkins, J.A. (1994). *A performance theory of order and constituency*. Cambridge: University Press.

Hoffmann, Ch. (1999). Word order and the principle of "Early Immediate Constituents" (EIC). *Journal of Quantitative Linguistics 682), 1999, 108-116.*

Kelly, M.H., Bock, K.J., Keil, F.C. (1986). Prototypicality in a linguistic context: effects on sentence structure. *Journal of Memory and Language 25, 59-74.*

Kuno, S. (1979). On the interaction between syntactic rules and discourse principles. In: Bedell, G., Kobayashi, E., Muraki, M. (eds.), *Explorations in linguistics: Papers in honor of Kazuko Inoue: 279-304.* Tokyo: Kenkyusha.

Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua 113-160.*

Mayerthaler, W. (1981). *Morphologische Natürlichkeit*. Wiesbaden: Akademische Verlagsgesellschaft Athenaion.

Pinker, S., Birdsong, D. (1979). Speakers´ sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior 18, 497-508.*

Ross, J.R. (1980). Ikonismus in der Phraseologie. *Zeitschrift für Semiotik 2, 39-56.*

Co-occurrence and cohesion

Hypothesis

"…syntactic cohesion is a direct result of frequency of co-occurrence: words that are used together more often tend to seem more fused and also tend to have more liaison" (Bybee 2001: 338; cf. also p. 343).

Procedure

First, define an exact measure of cohesion degrees (see also "Frequency and Cohesion", Chapter 2). Then count co-occurrences of words in a text corpus. Correlate the number of co-occurrences with the degree of cohesion. If the hypothesis does not hold, look for the boundary conditions under which it may hold.

References

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 337-359.* Amsterdam/Philadelphia: Benjamins.

Cotextuality and variation

Hypothesis

”…if they [i.e. grammatical morphemes] occur in different constructions, they move away from one another in phonological shape, meaning, and distributional properties” (Bybee 2001 346 f.).

Procedure

The hypothesis says that rich cotextuality (rich distribution) causes the rise of different allomorphs. Select 100 morphemes, both autosemantic and synsemantic ones and study their cotextuality in a corpus. Establish a direct dependence between the number of contexts and number of variants. If the hypothesis does not hold in each case, determine the boundary conditions, try to quantify them and set up the dependence *Variant forms = f*(*number of contexts, degree of other property*). Test both hypotheses. If none of them holds, define a third independent variable.

Reference

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 337-359*. Amsterdam/ Philadelphia: Benjamins.

Frequency and case

Hypothesis

”The more frequent a case in a particular language, the more it tends toward zero coding” (Fenk-Oczlon 2001: 441).

Procedure

Consider a language with overt case markers. Do not distinguish zero coding and non-zero coding but try to establish a method for scaling the magnitude of coding. Thus, Latin marks cases with suffixes such as *zero, -a, -ae, -bus, -ibus, -itis*. Determine the number of all nouns in all cases in a corpus. Try to express formally the relationship <*frequency, magnitude of coding*> by averaging the frequencies (or relative frequencies) within each class of magnitude. If that can’t be done, report on your findings and your opinion about the reason. Study a language with rich inflection and one with rich agglutination. Show the differences and try to explain them.

Reference

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistics form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam/Philadelphia: Benjamins.

Frequency and cohesion

Hypothesis

”… frequent usage of a phrase in standard usage makes the phrase cohere and become a unit” (Boyland 2001: 395).

Procedure

As cohesion in the above sense is not measurable, first give an exact definition of the concept of cohesion and make it measurable. Then collect at least 100 phrases from a corpus, measure their respective frequencies and put them in relation to their cohesion. It is to be noted that cohesion can be defined in different ways. Hence, if the hypothesis does not hold for your data, try first to redefine the measurement of cohesion. Corpus data from a language other than English would be more interesting.

References

Boyland, J.T. (2001). Hypercorrect pronoun case in English? Cognitive processes that account for pronoun usage. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 383-404*. Amsterdam/Philadelphia: Benjamins.

Frequency and derivation

Hypothesis

”The derived word has a smaller frequency of occurrence than its basic word” (Nagórko-Kufel 1984).

Procedure

Determine lemma frequencies on the basis of a frequency dictionary or a text corpus. Collect, say, 1000 nouns, and for each noun all its derivatives occurring in your source. Count the individual frequencies of nouns and their derivatives. Set up a table in which the random variable is “difference between the frequency of a noun and the frequency of its derivative”, i.e. $X = f_{noun} - f_{derivative}$. This variable will take also negative values (if the derivative occurs more frequently than the base noun).

Try to find the theoretical distribution of this difference. Show that it is not normal (test e.g. its skewness). Use Johnson´s S_U translations. Find the distribution of the variable $X = |f_{noun} - f_{derivative}|$. Try to give reasons for the form of the distribution. Try to find a discrete distribution.

Restrict your investigation to base nouns that occur in the source, but you can also consider all nouns if the base noun has zero frequency.

Continue analyzing verbs and adjectives and strive for a general theory. Refer to complexity theory or markedness theory. Compare your results with those from other languages.

References

Ginzburg, E.L. (1975). Ob odnom kriterii napravlenija derivacii. *Aktual´nye problemy russkogo slovoobrazovanija (Taškent) 372-376.*

Guiraud, P. (1960). *Problèmes et methods de la statistique linguistique*. Paris: PUF.

Harwood, F.W., Wright, A.M. (1956). Statistical study of English word formation. *Language 32, 260-273.*

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika 36, 149-176.*

Johnson, N.L., Kotz, S. (1970). *Continuous univariate distributions – Vol. 1*. Boston: Houghton Mifflin.

Nagórko-Kufel, A. (1984). Die Anwendung des Häufigkeitskriteriums bei der Wortbildung. *Glottometrika 6, 48-64.*

Frequency and irregularity

Hypothesis

"… there is a relationship between high frequency and irregularity" (Corbett, Hippisley, Brown, Marriot 2001: 202).

"The more frequently used a construction is, the greater is the likelihood that its form will be maintained, rather than being replaced by some more productive construction" (Bybee 2001: 348).

"…that which is more frequent….is more irregular." (Fenk-Oczlon 2001: 435).

Procedure

The authors consider irregularities in declination and propose a scaling procedure of irregularity. (a) Try to transfer the problem to conjugation or some other grammatical category in any language. (b) Try to generalize the problem devising a general method for scaling deviations from an expectation. (c) Use a frequency dictionary of word forms (ordered by ranks), select each 10^{th} word, determine its frequency and measure its irregularity. Then try to find a function capturing the relation <*rank, irregularity*> and analyse it. Read the discussion in the quoted article and try to generalize the concept of irregularity in language.

Set up a rank-frequency wordlist of verbs from a long text or corpus. Designate the regular verbs with *R*, the irregular ones (irregularity of any kind, without scaling) with *I*. Perform Wilcoxon´s U-test to see whether the second hypothesis holds. Then do the same for nouns. Choose a language with strong declination, and then try to generalize.

References

Corbett, G., Hippisley, A., Brown, D., Marriott, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 201-226*. Amsterdam/Philadelphia: Benjamins.

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 337-359*. Amsterdam/ Philadelphia: Benjamins.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam/Philadelphia: Benjamins.

Frequency and valency

Hypothesis

"… the more frequent a verb is, the less likely it is to have any fixed number of 'argument structures'" (Thompson, Hopper 2001: 49).

"…the more frequent a verb type, the less predictable the number of arguments; a rare verb like *to elapse* is limited to a single argument, whereas a common verb like *to get* appears in discourse with one, two, or three of the traditional arguments…" (Bybee, Hopper 2001: 5).

The hypothesis can be enlarged: frequent verbs have many prepositional (postpositional) phrases in English (g*et up, get in, get away,...*).

Procedure

Try to make this hypothesis more exact: *Number of arguments = f*(*frequency*), derive it from reasonable assumptions and test it on 100 (English) verbs of different frequencies. Consult German dictionaries of verb valency and a frequency dictionary. Try to set up the dependence function.

References

Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 1-24*. Amsterdam/Philadelphia: Benjamins.

Thompson, S.A., Hopper, P.J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 49-60*. Amsterdam/ Philadelphia: Benjamins.

Frequency of sentence patterns

Hypothesis

The rank-frequency distribution of sentence patterns abides by the Zipf-Mandelbrot law. (Köhler 2005).

Procedure

In order to test the hypothesis, analyse all the sentences of a long text ascribing them a certain general structure according to any type of grammar. Then count the number of sentences of each type in the text and set up their rank-frequency distribution. Test whether it follows the Zipf-Mandelbrot distribution. If not, what kind of distribution would be more adequate?

Perform the analysis using different grammars and comment the results. Can you draw the conclusion that the best grammar is that which most exactly follows the Zipf-Mandelbrot law?

Cf. the chapter "Textology" and try to take over some indicators which could – mutatis mutandis – express some syntactic properties.

Reference

Köhler, R. (2005). Quantitative Untersuchungen zur Valenz deutscher Verben. *Glottometrics 9, 13-20.*

Grammaticalisation

Problem

Devise a method for measuring the "grammaticalisation cline" (a) beginning from idiom down to grammatical rule, (b) beginning from lexical word down to inflectional affix, (c) beginning from phraseological expression through compounds down to blend (Hopper, Closs Traugott 2003).

Procedure

Devise a scale (or classes) for independence or cohesion and try to assign your entities to individual independence/cohesion classes. Take a large sample from a corpus and try to find some regularities or dependencies.

Reference

Hopper, P.J., Closs Traugott, E. (2003). *Grammaticalization* 2nd ed. Cambridge: Cambridge University Press.

Morph frequency

Problem

The rank-frequency distribution of morphs does not differ from that of words.

Procedure
You should base this investigation on texts in other languages than German. Partition the text in morphs and set up the rank-frequency distribution of the morphs**.** Test whether the usual distributions are adequate. Cf. the problems "Word frequency 1, 2, 3" in Chapter 4.

Reference
Best, K.H. (2005). Morphlängen. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 255-260*. Berlin/New York: Mouton de Gruyter.

Morpheme polysemy and morpheme frequency

Problem
Andrea Krott (1999) presented the dependence of the frequency of morphemes of different types on their polysemy. In two cases, the dependence can be modelled satisfactorily (with nouns and verbs); the rest displays large deviations. Comment this phenomenon.

Procedure
Evidently, polysemy alone does not explain a sufficiently large proportion of the variance. One must probably add another independent variable, which can be different with particular word classes. First try to find the answer in each case theoretically (hypothetically), then analyse a sufficiently large sample from a big dictionary and employ the frequencies in a corpus. If necessary, add further independent variables.

References
Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
Krott, A. (1999). Influence of morpheme polysemy on morpheme frequency. *Journal of Quantitative Linguistics 6(1), 58-65*.

Morphological productivity of stems

Problem
In a dictionary, stem productivity (= forming derivates and compounds) abides by a regular probability distribution.

Procedure
Use a dictionary which has derivation and compounding. Compute the number of stems forming x = 0,1,2,… derivates/compounds and set up the empirical dis-

tribution. The theoretical distribution can be set up applying a birth-and-death process (cf. Wimmer, Altmann 1995) but the birth and death rates need not be the same for all languages. Assume other birth and death rates, solve the process and generalize the problem. Find the subsidiary conditions accounting for the choice of the birth and death rates. Develop a theory.

Reference

Wimmer, G., Altmann, G. (1995). A model of morphological productivity. *Journal of Quantitative Linguistics 2(3), 212-216.*

Sequential word class frequency

Hypothesis

The cumulative sequential frequency of the main word classes (nouns, verbs) is convex, that of auxiliaries concave.

Procedure

The hypothesis has not been tested as yet. It is very general, and there will be a number of boundary conditions modifying it. Nevertheless, a pilot study could be made.

Count how many nouns occur up to position x (x = 1,2,3,…N) in a text of your choice. Obtain the cumulative positional frequency of nouns. The nouns in the above hypothesis formulation display the following sequence

Position x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Nouns cumul.	0	0	0	1	1	1	2	3	4	5	5	5	5	5	6	6

Perform the computation for all word classes. Then take the individual series and fit to them the power function

$$y = ax^b .$$

If $b > 1$, then the curve is convex. If $b < 1$, the curve is concave; if $b = y$, a straight line results.

Try to characterize texts, genres and languages by setting up the spectrum of sequential word class frequencies. Study a special word class in its historical development. Use the parameters b of individual word classes as elements of a vector and compare the vectors of individual texts. Take average b-´s for individual languages and compare their vectors. Define the word classes in individual languages in comparable terms.

Reference

Ziegler, A., Best, K.-H., Altmann, G. (2001). A contribution to text spectra. *Glottometrics 1, 97-108.*

Verb classification

Problem

Is the adequacy of fitting a distribution to a ranked series a criterion of adequacy of a classification?

Procedure

In quantitative linguistics, one frequently considers the adequacy of fitting a theoretical distribution to ranked data a sign of "correct" classification of the entities. Use the data collected by Levickij, Kiiko and Spolnicka (1996) classifying German verbs in 22 classes and giving the number of verbs in each class. Perform the ranking in this classification according to the number of verbs. Then (a) try to find a theoretical rank-frequency distribution; if that fails, (b) try to find inductively a well fitting distribution. If none is found, can you conclude that the classification is not adequate?

Reference

Levickij, V.V., Kiiko, J.J., Spolnicka, S.V. (1996). Quantitative analysis of verb polysemy in modern German. *Journal of Quantitative Linguistics 3(2), 132-135.*

Verbs and persons

Problem

Verbs can be divided into classes in many different ways. There is no "best" way. Any classification is conditioned by the aim of research. Here we shall try to test J. Scheibmann´s (2001)

Hypothesis

"..we would expect greater co-occurrence of elements whose combinations lend themselves to conveying speaker's point of view than those whose combinations do not (e.g. after Benveniste 1971, verbs of cognition would more frequently appear with a first person singular subject than with a third person singular)." (Scheibmann 2001: 65).

Scheibman classifies verbs in 10 classes following Halliday (1994) and brings frequencies of associations of these classes with grammatical persons using conversations as a data base. The data are given in Table "Scheibman". Levickij and Lučak (2005) set up 20 semantic verb subclasses. See also Jurčenko (1985), Levin (1998), Sil´nickij (1966).

Table “Scheibman”

Verb class	1s	2s	3s	1pl	2pl	3pl
Cognition	195	110	15	6	0	14
Corporeal	24	7	30	1	1	3
Existential	12	6	62	3	0	8
Feeling	19	9	10	2	0	5
Material	141	90	176	30	2	100
Perception	27	19	6	10	0	2
Perception/rel	0	0	35	0	0	4
Possessive/rel	21	31	29	5	0	16
Relational	50	41	497	6	2	45
Verbal	128	335	931	66	5	218

Here 1s = 1 person singular, etc.

Procedure

First try to perform an overall test for independence of person and verb class. Then test each cell separately (employ the test for individual cells) for significant association. Check several others of J. Scheibman´s hypotheses on her data. Then take a sample from another language and do the same. Check whether the results are identical.

Other problems: In quantitative linguistics, there is a well known hypothesis that if some entities are “adequately” ordered in classes, the rank-frequency distribution of these elements usually follows a “honest” rank-frequency distribution. Test whether J. Scheibman´s data corroborate this hypothesis.

Is it possible to consider the association of the verb with the category person as a text characteristic similar to Busemann´s Verb-Adjective Ratio? Can one scale the verb classes according to the concept of “activity” or according to the biological development of life? (Beginning with verbs of being up to verbs of psychological states, mental process…). Or can one perform an a posteriori classification for the above problem? Scaling would make the hypothesis clearer.

References

Benveniste, E. (1971). *Problems in general linguistics*. Coral Gables, Fl.: University of Miami Press.

Halliday, M.A.K. (1994). *An introduction to functional grammar*. London: Arnold.

Jurčenko, G.E. (1985). K voprosu o semantičeskoj klassifikacii glagolov anglijskogo jazyka. In: *Grammatičeskaja semantika: 45-50*. Gorkij: Gorkij University Press.

Levickij, V., Lučak, M. (2005). Category of tense and verb semantics in the English language. *Journal of Quantitative Linguistics 12, (2-3), 212-238.*

Levin, B. (1998). *English verb classes and alternations*. Chicago: University of Chicago Press.

Scheibman, J. (2001). Local patterns of subjectivity in person and verb type in American English conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 61-89.* Amsterdam/Philadelphia: Benjamins.

Silnickij, G.G. (1966). Semantičeskie klassy glagolov i ich rol´ v tipologičeskoj semasiologii. In: *Strukturno-tipologičeskie opisanie sovremennych germanskich jazykov: 244-259.*

Word class distributions

Hypothesis

The rank-frequency distributions of different word classes abide by the same probability distribution.

Procedure

Count the different word classes (noun, verbs, adjectives, adverbs,…) separately in a text. If there are ambiguous cases, decide ad hoc to which class a word belongs. Then fit the same probability distribution to all empirical distributions, e.g. Zipf´s truncated zeta distribution $P_x = C/x^a$ (x = 1,2,3,…,n), where C is the normalizing constant and $n = x_{max}$. Examine the behaviour of the parameter a. Is it equal in all cases or are there differences? Compare the results with an analogous analysis of a text in another language – even if the word classes may be different. If possible, perform an ordering of word classes according to the value of parameter a.

Analyze several languages with the same word classes, ascribe ranks to classes according to a and perform a concordance test for the equality of ordering (cf. e.g. Gibbons 1971). Try to draw some conclusions from your results. Repeat the computations fitting other distributions with one parameter and draw conclusions.

References

Gibbons, J.D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Chapter 3

Compounds and lexicology

Age and compounding propensity

Hypothesis
"The older a word, the more compounds it produces." (Altmann 1989)

Procedure
Here, word classes (specifically, parts-of-speech) must be treated separately. The compounding propensity is not the same in all word classes. In order to establish the hypothesis a historical dictionary is necessary. It should give the year or at least the century of the first appearance of a word in written documents. Draw a sample of words of the same word class, note their first appearance, and then determine the number of compounds they form in the modern language. Draw a graph by plotting an empirical curve and try to derive a theoretical function using linguistic assumptions. Then take another word class. Do the above for different languages.

References
Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107*.
Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.
Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189.* Černivcy: Ruta.

Collocations

Problem
"…the more often two elements occur in sequence the tighter will be their constituent structure" (Bybee, Hopper, p. 14). Collocations can be found by testing the cohesion of two words.

Procedure
Select any word from a corpus and seek all different words that occur immediately behind it within a clause. Then compute the significance of a collocation using the hypergeometric distribution and the Poisson distribution. Compute the conditional probability of the following word. Evaluate the collocation determining a probabilistic decision boundary. (Cf. "Association graph of the text".)

References

Bisht, R.K., Dhami, H.S., Tiwari, N. (2006). An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations. *Journal of Quantitative Linguistics 13(2-3), 161-175.*

Bybee, J., Hopper, P. (eds.) (2001). *Frequency and the emergence of linguistic Structure*. Amsterdam: Benjamins

Levickij, V. (2005), Lexikalische Kombinierbarkeit. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international Handbook: 464-470*. Berlin/New York: Mouton de Gruyter.

Levickij, V.V., Zadorožna, I. (2007). Die Stärkemessung des Zusammenhangs zwischen den Komponenten der Phraseologismen. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 399-406.* Berlin/ New York: Mouton de Gruyter.

Lin, D. (1998). Extracting collocations from text corpora. In: *First Workshop on Computational Terminology*. Montreal.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics 19(1), 143-177.*

Compound length and component length

Hypothesis

"The longer the compound, the shorter its components." (Altmann 1989)

Procedure

This hypothesis is a simple consequence of Menzerath´s law: The longer a construct, the shorter are its components. The testing is quite simple: Set up a (random) list of compounds of any length from either a dictionary or a corpus. Apply two kinds of length measurement for components: (a) in terms of the number of phonemes, (b) in terms of the number of syllables. The length of a compound is measured in terms of the number of its components. For each compound, measure its length and the average length of its components. If the hypothesis is correct, two monotonous functions for the relation <*compound length, mean component length*> will result. Unfortunately, compounds with multi-components are rare except in languages such as Hungarian or German. If necessary, consult a specialized dictionary which contains long compounds.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107.*

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Cramer, I. (2005). Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 659-688*. Berlin/New York: Mouton de Gruyter.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189.* Černivcy: Ruta.

Compound length and compound cotextuality

Hypothesis

"The longer a compound, the smaller its cotextuality." (Altmann 1989)

Procedure

Take a random sample of compounds from a corpus. Ascertain the length of individual compounds (in terms of the number of components). Then compute the cotextuality of each compound in one of the ways proposed in the previous hypothesis. Try to set up the relation <*compound length, extent of cotextuality*> by means of an empirical function as well as by means of a theoretical argument.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107.*

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189.* Černivcy: Ruta.

Compound length and polysemy

Hypothesis

"The longer a compound the fewer meanings it has (on the average)" (Altmann 1989).

Procedure

The problem is formally identical with that of word length and polysemy. The difference is that the length of a compound is measured in terms of the number of its components. Since compounding is a means to meet the specification requirement the hypothesis must hold. Try to find the relationship <*number of components, polysemy of compound*>, start from proportionality arguments and take averages, otherwise a fatal dispersion may result. Later on, when several hypotheses on compound have been established try to set up a control cycle containing the relation *polysemy of a compound = f*(*number of stems in a compound, other variable*).

References
Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107*.
Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189.* Cernivcy: Ruta.

Compound length and semantic correspondence

Hypothesis
The longer a compound, the greater its semantic correspondence with its components.

Procedure
Length of a compound means the number of its components (not its syllabic or any other length). Use the previous problem, but this time draw random samples of compounds of different lengths. Then, in analogy to the method of the previous problem, proceed in two ways.

1. Compute the mean correspondence of compounds of different length and set up the relation <*length, correspondence*>.
2. Take the minimum correspondence in each length class and set up the same relationship. If possible, study languages with longer compounds and use, if necessary, technical dictionaries.

References
http://lql.uni-trier.de

Compounds and semantic correspondence

Hypothesis
"The number of compounds in a language (having compounds) decreases proportionally to the measure of semantic correspondence of the components with the compound" (Altmann 1989).

Procedure
First set up a method for measuring the semantic correspondence between the components of a compound and the compound itself. For example, in *hangover* (German *Katzenjammer*) there is no correspondence between the meanings of the parts (*hang, over*) and the compound *hangover*. In German *Kindergarten* or *Baumschule* the correspondence is obvious; at least one component has part of the meaning of the compound. It is the same with the compound *book seller*, which has a high semantic correspondence. Differentiate the compounds also according to the degree of cohesion (or type). Take a sample of compounds and

obtain the distribution of their semantic correspondence, or try to set up the relation <*semantic correspondence*, *number of compounds*> or vice versa.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107*.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189*. Černivcy: Ruta.

http://lql.uni-trier.de

Compound forming and associations

Hypothesis

The more associations a word has, the more compounds it forms.

Procedure

Use a dictionary of word association norms which are available in many languages. Take a large random sample of basic words and note the number of associations (types). Ignore the frequency of individual associations. Then use a dictionary of compounds and for each word in the sample count the number of compounds it forms.

A monotone increasing function for <*number of associations*, *number of compounds*> is obtained if the hypothesis holds. Draw a graph, fit a theoretical curve to the empirical one using proportionality arguments.

References

None.

Compound forming and emotionality

Hypothesis

The more emotional the meaning of a word, the greater the number of compounds it produces.

Procedure

Word emotionality can be measured using the Osgood method, evaluated by test persons or using words thus evaluated from related psycholinguistic literature. Take 100 words and measure their emotionality (of whatever kind). Then use a dictionary and count the number of compounds each word produces. Plot a curve <*emotionality*, *number of compounds*> that depicts this dependence. Then try to derive the function from proportionality arguments.

References
None.

Cotextuality and compounding propensity

Hypothesis
"The greater the cotextuality of a word, the more compounds it produces." (Altmann 1989)

Procedure
Cotextuality can be ascertained in two different ways: (a) as the number of texts (of a corpus) in which a word occurs, or (b) as the number of different neighbourhoods in which it occurs, i.e. as the number of contexts. The greater the distribution of a word the more frequently it needs to be specified. Compounding is one way of specification, hence we expect a propensity in compound forming. In testing the hypothesis, distinguish word classes, take for example only nouns or verbs. Having computed the cotextuality of the words, find the number of compounds each of them forms. Then try to discover the trend <*cotextuality, number of compounds*>. Draw a graph which displays a monotone increasing tendency. Plot an empirical curve and try to analyse it using theoretical assumptions.

References
Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107.*
Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.
Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189.* Černivcy: Ruta.

Dissortativity of compounding

Hypothesis
The technique of compounding is dissortative.

Procedure
Let the number of different words with which a certain word forms compounds be its degree (term taken from the theory of graphs). If words with high degree tend to form compounds with words of high degree, the compounding is said to be assortative. If words with high degree tend to form compounds with words of low degree, the compounding is said to be dissortative. Otherwise it is neutral. Take a large sample of compounds (if possible, all the compounds of a dictionary). Compute for each component its degree. Then for all components of the

same degree compute the average degree of components forming compounds with them. Obtain the relationship <*degree of component, average degree of adjacent components*>. The result is a monotone decreasing function if the compounding is dissortative; a monotone increasing function if the compounding is assotative; a horizontal straight line if the compounding is neutral.

The problem can be tackled also with phoneme combinations (bigrams).

References

Newman, M.E.J. (2002). Assortative mixing in networks. *Physical Review Letters 89(20), 208701.*

Tamaoka, K., Meyer, P., Makioka, S., Altmann, G. (2008). On the dynamics of compounding of Japanese kanji with common and proper nouns. *Journal of Quantitative Linguistics (submitted)*

Distribution of compound length

Hypothesis

"The number of compounds decreases with their increasing length" (Altmann 1989).

Procedure

The hypothesis says that the distribution of compounds is a simply monotone decreasing distribution. Take a random sample of compounds. The length of a compound is measured in terms of the number of its components. Obtain an empirical distribution of compound length and find either the theoretical distribution or a function for <*compound length, number of compounds of that length*>. If Menzerath´s law holds, either the zeta distribution or the zeta function is obtained. Try to test the hypothesis in different languages and generalize. Remember that compounding is an expression of specification requirement.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107.*

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Fan, F., Altmann, G. (2007). Some properties of English compounds. In: Kaliuščenko, V., Köhler , R., Levickij, V. (eds), *Problems of typological and quantitative lexicology: 177-189.* Černivcy: Ruta.

Distribution of synonyms

Hypothesis

The distribution of synonyms in a dictionary has a regular form.

Procedure

Use a dictionary of synonyms. Take a systematic sample of lexical entries such as the last word on each page. Count how many entries there are having exactly $x = 1,2,3,\ldots$ synonyms. Try to fit the known models to the data.

References

Uhlířová, L. (2001). Kolik je v češtině synonym? (K dynamické stabilite v systému lexikálních synonym). In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99: 237-250*. Bratislava: Veda.

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99: 218-225*. Bratislava: Veda.

Increase of loan words

Hypothesis

The number of loan words increases in every language according to Piotrowski law.

Procedure

Use a historical dictionary of target language and try to count the number of foreign words entering the language. Note the (approximate) year of borrowing.

Another way: use a journal or newspaper and count how many new English words there are that entered the language from 1950 afterwards. Note only the first year of borrowing.

Test the hypothesis that the cumulative number of foreign words follows the Piotrowski law

$$y_t = \frac{C}{1 + ae^{-bt}},$$

where y_t is the number of foreign words at time t, C is the asymptote and a, b are parameters. Special dictionaries or catalogues of department stores can also be used for this study.

References

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung: 54-90*. Göttingen: edition herodot.

Best, K.-H. (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9-34.*

Best, K.-H. (2006). Deutsche Entlehnungen im Englischen. *Glottometrics 13, 66-72.*

Best, K.-H. (2004), Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen. *Glottometrics 8, 75-78.*

Best, K.-H. (2005). Turzismen im Deutschen. *Glottometrics 11, 56-63.*

Best, K.-H., Altmann, G. (1986). Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen. *Folia Linguistica Historica 7, 31-41.*

Körner, H. (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics 7, 25-49.*

Lexical chains

Problem

Describe aspects of the hypernymic structure of English (or another language) lexicon.

Procedure

A hypernym of a basic lexeme A is another lexeme forming a class to which A belongs. For example *furniture* is a hypernym of *chair; building* is a hypernym of *skyscraper*. The hypernym is usually contained in the definition of the meaning in a monolingual dictionary. Consider only nouns which form hypernymic chains. WordNet provides hypernymic chains for English; for other languages such chains must be built by the researcher. In forming hypernymic chains, pay attention to the following.

1. Eliminate any relation other than class inclusion; that is, do not consider "part of" relations like *head = part of the body*; *motor = part of the car* (*body* is not a hypernym to *head* and neither is *car* to *motor*).
2. Consider only the first, main meaning of the noun. If there are several meanings, form different chains.
3. Avoid circularity (which can be found also in WordNet).
4. Accept hypernyms like *entity, system, being, thing*, etc. of very high generality or abstractness but avoid definitions like *something that*.
5. Consider also abstract nouns.
6. If a noun occurs in any chain as hypernym, do not include it in the set of basic lexemes.

Having prepared the data, do the following.

(a) Try to find the distribution of the length of lexeme chains both empirically and theoretically.

(b) Order the chains so that the basic lexeme is at the first level. Compute the mean length of lexemes at the first, second, third,… level. Try to detect a trend.

(c) Consider the number (proportion) of monomorphemic words at the first, second, third …level. Try to detect a trend.

(d) Count how many different words (types) are at the first, second, third,… level and see whether there exists a regular decrease of types.
(e) Order the chains in such a way that the highest hypernym (end of the chain) stays at the first place. Perform the tasks in (b), (c) and (d).

References

Hammerl, R. (1987). Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen. *Glottometrika 8, 113-129*.
Hammerl, R. (1989). Neue Perspektiven der sprachlichen Synergetik: Begriffsstrukturen – kognitive Netze. *Glottometrika 10, 129-140.*
Hammerl, R. (1989). Untersuchung struktureller Eigenschaften von Begriffsnetzen. *Glottometrika 10, 141-154.*
Sambor, J. (2005). Lexical networks. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 447-458.* Berlin/New York: Mouton-de Gruyter.
Sambor, J., Hammerl, R. (eds.) (1991). *Definitionsfolgen und Lexemnetze*. Band I. Lüdenscheid: RAM.
Schierholz, S. (1989). Kritische Aspekte zum Martinschen Gesetz. *Glottometrika 10, 108-128.*

Lexical networks

Problem

Set up definition chains consisting of hypernyms for all meanings of a lexeme and obtain a lexical network. Study its properties.

Procedure

From a monolingual dictionary get randomly 100 nouns. If some of them are monosemic, then the lexical chain leading to the most general lexeme is simple. But for the polysemic ones there are different ways to the most general lexeme(s). A directed graph can be obtained, which has a number of properties. Evaluate at least the following properties:

1. number of terms in the graph (= number of vertices) and their distribution in language (of at least 100 nouns);
2. the width of the graphs defined by Hammerl (1989);
3. the number of branches (paths) and their distribution;
4. the number of end lexemes and their distribution;
5. average length of the paths;
6. strength of the semantic relations between the lexemes in a network;
7. lexeme productivity, etc.

Use all means of graph theory to characterize the construction of lexical networks in language.

Devise a method for performing this kind of analysis for verbs and adjectives. Compare several languages.

This discipline is not sufficiently developed; further investigations are needed.

References

Hammerl, R. (1989). Untersuchung struktureller Eigenschaften von Begriffsnetzen. *Glottometrika 10, 141-154.*

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. *Glottometrika 6, 139-151*.

Sambor, J. (2005). Lexical networks. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 447-458.* Berlin/ New York: Mouton de Gruyter.

Sambor, J., Hammerl, R. (eds.) (1991). *Definitionsfolgen und Lexemnetze*. Band 1. Lüdenscheid: RAM.

Skorochod´ko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten.* Bochum: Brockmeyer.

Stem length and compounding propensity

Hypothesis

"The shorter a word, the more frequently it occurs in compounds." (Altmann 1989)

Procedure

Take a random sample of word stems, prefereably of the same word class. Then try to find all compounds built with these stems. Note the stem positions within a compound, i.e., in the first, second, third,… position in the compound. Draw a graph to see the trend: the longer a stem the smaller the number of compounds. Then try to derive the relation <*stem length, compound number*> using first proportionality arguments. One can use also published sources.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107*.

Fan, F., Altmann, G. (2006). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 177-189.* Černivcy: Ruta.

Word length and synonymy

Hypothesis

The longer the word, the smaller the number of its synonyms.

Procedure

Use a dictionary of synonyms and take a large random sample of words. Obtain the number of their synonyms. Try to find the kind of dependence of the number of synonyms on the length of words. For each word length the average number of synonyms must be taken into consideration.

References

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99: 218-225.* Bratislava: Veda.

Chapter 4

Textology

The association graph of a text

Problem

Words in a text can be associated covertly (not by forming collocations). Set up a graph of associations and evaluate the properties of the graph.

Procedure

Compute the absolute frequencies of nouns, verbs and adjectives in a complete text. Further, count the number N of sentences in the text. Take the first two words from your list of nouns, verbs and adjectives, designate the frequency of the first as M, that of the second as n and find the number of sentences x in which both of them occur (common occurrence). In order to test the strength of the association perform the following computation: if $x > Mn/N$, compute

$$P(X \geq x) = \sum_{j=x}^{\min[n,M]} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

Choose a significance level $\alpha = 0.05$. If P is smaller than α, we conclude that an association between the two given words exists.

Perform the test for all word pairs from your list. Draw a graph of word associations and study the properties of this graph. Use the procedure for texts of different genres and languages. Set up hypotheses about the associative text structure.

References

Altmann, G. (1988). *Wiederholungen in Texten.* Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Autosemantic pace filling

Problem

Partitioning the rank frequency distribution of word frequencies in intervals of length h (h being the h-point) yields an exponential increase of the number of autosemantics in consecutive intervals. Test this statement on data from different texts.

Procedure
Determine the rank-frequency distribution of words in a text and compute the *h*-point. Partition the distribution in *h*-steps beginning from rank 1 to the highest rank. Count the autosemantics in each interval and from a table renaming the intervals in 1^{st} ,2^{nd},… i.e., rescale the distribution. An increasing function is obtained. Try to fit the function

$$y = a(1 - \exp(-kx))$$

to these data.

Popescu et al. (2008) defined two text characteristics: the *autosemantic compactness AC = ak*, where *a* and *k* are the parameters of the above fitted function, and the *autosemantic pace filling APF = a/h*. Compute these two indicators for many texts. In order to test the differences of *APF* and *AC* between two texts, use the tests given in the literature (Popescu et al. 2008).

To classify texts, use the coordinates *<1/k, a>* obtained from the above function and perform a discriminant analysis or use an up-to-date taxonomic method. Try to find justifications for your results.

References
Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter

Carroll´s vector

Problem
In an almost forgotten article, Carroll (1960) proposed a set of possible text properties. Take pairs of them and test their independence.

Procedure:
Measure the objective and subjective properties of several texts. Try to find arguments for the mutual dependence of some pairs of features. If possible, derive the dependence from a differential equation. Join the properties step by step to increasing correlated sets in order to acquire a control circuit similar to that of Köhler (1986).

Use independent properties to characterize the texts and the circuit(s) for establishing an elementary theory. Do not restrict yourself to prose. Use also the features presented in Tuldava (1995: 93-108). Set up your own scale for evaluating subjective features.

References
Carroll, J.B. (1960). Vectors of prose style. In: Sebeok, Th.A. (ed.), *Style in*

language: 283-292. Cambridge, Mass.: MIT Press
Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Tuldava, J. (1995). *Methods in quantitative linguistics.* Trier: Wissenschaftlicher Verlag Trier.
Tuldava, J. (2005). Stylistics, author identification. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.). *Quantitative linguistics. An international handbook: 368-387*. Berlin/New York: de Gruyter.

A constraint measure for texts

Problem

Ejiri and Smith (1993) proposed a "constraint measure" for texts in the form $G = \log(N/L)/\{\log(N)-1\}$. [$N$ = text length, L = text vocabulary]

Procedure

Show or simply discuss the methodological and statistical nature of this index. What does it say? Over which interval does it vary? What are its expectation and variance? How could one compare two texts by means of this index? Discuss index forming in general. Which properties must an index have?

Reference

Ejiri, K., Smith, A.E. (1993). Proposal for a new ´Constraint measure´ for text. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to quantitative linguistics: 195-39.* Dordrecht: Kluwer.
Galtung, J. (1967). *Theory and methods of social research*. Oslo: Universitets-forlaget.

Cotextuality and frequency

Hypothesis

The greater the cotextuality of an entity, the greater is its frequency of occurrence.

Procedure

Consider cotextuality in different ways. With respect to a phoneme, it can be the number of other phonemes that can occur in its neighbourhood, the number of different syllable types or the number of word types in which it occurs. With respect to a syllable it is the number of word-form types in which it occurs. With respect to a word it is the number of different texts in which it occurs. The hypothesis assumes that great cotextuality results in great frequency, although it may not always be so in reality. The hypothesis is part of Köhler´s self-regulation

cycle. Test the hypothesis in any way possible. Consider the possibility of changing the direction of dependence.

Get both the cotextuality and the frequency of words from a long text or corpus. If the hypothesis holds, it has the form of a power function. If it does not, then try to modify it.

References

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Tamaoka, K., Makioka, Sh. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, & Computers 36 (3), 531-547.*

Distances between equally long sentences

Hypothesis

Distances between equally long sentences in texts follow the Zipf-Alekseev distribution (Hřebíček 2000: 36ff.).

Procedure

Define the distance between equally long sentences in different ways in a long text. The simplest way is to count the sentences of a different length between them. Set up the distribution of distances and test whether the Zipf-Alekseev distribution is adequate.

Consider other textual units and study material from different languages in order to yield a better corroboration or rejection of the hypothesis.

If the Zipf-Alekseev distribution is not adequate, try to find another solution with corroboration.

References

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Distances between lexemes

Hypothesis

Distances between occurrences of the same lexeme follow the power law (Hřebíček 2000:32ff.).

Procedure

Define the distance between identical lexemes in a long lemmatised text as the number of intervening lexemes or as the number of intervening sentences. You

may restrict your investigation to one special lexeme. Ascertain the distances between the occurrences of the lexeme(s) and set up their distribution. It will be necessary to pool the distance classes if they are not sufficiently representative. Transform the pooled classes to x = 1,2,3,… (by simply renaming them). Then test whether <*x, number of distances of magnitude x*> is a power law.

The same can be done with respect to word forms or other well defined entities (syllables, morphemes). Try to find a connection to Skinner´s hypothesis (cf. "Phonetic aggregation" in this Chapter).

References

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Euphony

Problem

Euphony in a text can be achieved either by a special frequency of individual phonemes or by their combination or by their position in a text (e.g. rhyme). Try to develop a measure of euphony or try at least to set up an operational definition.

Procedure

Determine the relative frequencies of phonemes in a language on data from a non-poetic text sample (e.g. corpus, prosaic texts), denote p_i the relative frequency of phoneme i and ξ the random variable representing the number of occurrences of phoneme i. Let n be the number of phonemes in a verse (line) and α the chosen significance level, e.g. $\alpha = 0.05$. Consider the first phoneme in the line. If its frequency f_i is greater than np_i then compute the cumulative probability

$$P(\xi \geq f_i) = \sum_{x=f_i}^{n} \binom{n}{x} p_i^x q_i^{n-x}$$

where $q_i = 1 - p_i$. Now, the euphonic weight of the phoneme i in the line is

$$E(i) = \begin{cases} 100[\alpha - P(\xi \geq f_i)], & \text{if } \alpha > P(\xi \geq f_i) \\ 0 \quad \text{otherwise} \end{cases}$$

Let E be the set of those phonemes whose $E(i) > 0$, and $k = |E|$. Then the euphony of a line can be defined as

$$E(line) = \frac{100}{k} \sum_{i \in E} [\alpha - P(\xi \geq f_i)]$$

i.e. the mean euphony of all euphonic phonemes. Let N be the number of lines in a poem. Then the euphonic value of the whole poem can be defined as

$$E(poem) = \frac{1}{N}\sum_{j=1}^{N} E(line_j).$$

Try to do the following:
1. Analyze a poem in the described way.
2. State whether there is a special course of euphony from the beginning to the end of the poem.
3. Analyze the development of euphony with one author or in one language.
4. Develop other measures of euphony.
5. Determine whether there is a relationship between euphony and the meaning of the poem and between euphony and other properties of the text.

References

Altmann, G. (1966). The measurement of euphony. In: *Teorie verše I, 259-261*. Brno: Universita J.E. Purkyně.

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov*. Bratislava: Veda.

Hirsch-Popescu-point problems

Problem

The Hirsch-Popescu-point is that point in which $x = f(x)$ in a rank-frequency distribution or frequency spectrum. Let F_h be the cumulative frequency up to the h-point, i.e. $F(X \leq h)$. Answer the following questions:
1. Is the h-point correlated with entropy and repeat rate?
2. Which of the many indices of vocabulary richness is correlated with the h-point or with F_h?
3. Does the h-point depend on text length?
4. Is there a difference in the h-point between different genres?
5. Is the h-point characteristic for a writer?
6. Are there differences between languages as far as the Hirsch-Popescu point is concerned?

Procedure

Read "Popescu´s typological indicator a" showing the computation of the h-point, and "Repeat rate and entropy". Use a single text, compute all the characteristics mentioned above and solve the problems.

Continuation (for mathematicians)

Try to derive the h-point for some discrete probability distributions used as rank-

frequency models. If summing is difficult, give an approximation by means of integrals, series expansion etc.

References

Hirsch, J.E. (2005). An index to quantify an individual´s scientific research output. In: http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf.

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests fort he h-point and related text characteristics. *Glottometrics 15, 42-52.*

Popescu, I.-I. (2007). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 553-562*. Berlin/New York: Mouton de Gruyter .

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 13, 71-81*.

Hrebs

Problem

The linguistic unit *Hreb* was named after the researcher who defined it: Luděk Hřebíček (he called it *aggregate*). A *Hreb* is a set of morphemes, words, phrases, clauses or sentences having the same meaning component.

The *Hreb* is a well definable textual unit abiding by all laws and tendencies of text forming. Test some hypotheses given below.

Procedure

1. Consider a text as a sequence of morphemes. Denote each morpheme meaning with a different number. Replace the morphemes by these numbers to obtain a sequence of numbers. Study these numbers as units and try to describe the behaviour of morpheme-*Hrebs.*
2. Consider a text as a sequence of words. Do everything with word-*Hrebs* you did with morpheme-*Hrebs.* It is advantageous to omit some word classes, e.g. conjunctions and prepositions/postpositions, some numerals and articles.
3. Consider a text as a sequence of phrases and perform the same operations as above. A long text should be used to get reliable results.
4. Consider all sentences containing the same referent as belonging to the same *Hreb*. Here a *Hreb* is a set of sentences joined by *something common* (identical entity or reference). Each sentence of the text can belong to different *Hrebs*. Hence two different sets (that of sentences and that of *Hrebs*) can be obtained.

Do the following. (a) Draw a bipartite graph whose partite sets are sentences and *Hrebs*, and compute its properties. (b) Then draw a graph of sentences joining them with edges if they contain the same referent. (c) Test the hypothesis

"the more sentences are there in a *Hreb*, the shorter they are", which is in agreement with Menzerath´s law. (d) For all kinds of *Hrebs* perform the usual denotative analysis: (i) obtain the distribution of *Hreb* sizes, (ii) compute the diffusity of *Hrebs*, (iii) compute text compactness, (iv) set up the graph of positional coincidences of *Hrebs*, (v) compute text concentration, (vi) compute text connectivity, (vii) compute the distances between *Hrebs*, (viii) determine the cliques, etc.

Try to define other kinds of *Hrebs*. Find more properties of the respective graphs and interpret them linguistically.

References

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to quantitative linguistics: 33-39.* Dordrecht: Kluwer.

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law*. Trier: Wissenschaftlicher Verlag.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Köhler, R., Naumann, S. (2007). Quantitative analysis of co-reference structure in text. In. Grzybek, P., Köhler, R. (eds). *Exact method in the study of language and text: 317-329.* Berlin/New York: Mouton de Gruyter.

West, D.B. (1006). *Introduction to graph theory*. Second edition. Upper Saddle River NJ: Prentice Hall.

Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.). *Quantitative Linguistics. An International Handbook: 423-447.* Berlin/New York: de Gruyter.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse. Ein textlinguistisches Arbeitsbuch*. Wien: Edition Prasens.

Hurst´s exponent

Problem

Characterize the behaviour of length sequences using Hurst´s exponent.

Procedure

Transcribe a text as a sequence of lengths. Use different units such as morpheme length, word length, sentence length etc. (all measured in different ways). Define the following quantities:

i = position in the sequence;

x_i = length of the unit in position i;

$R_i = max\ x_i - min\ x_i$; R_i is the range, the difference of the maximum length up to position i and the minimum length up to position i;

$\bar{x}_i = \frac{1}{i}\sum_{j=1}^{i} x_i$. $\bar{x}$ is the mean length of units up to position *i*;

$$S_i = \left[\frac{1}{i}\sum_{j=1}^{i}(x_j - \bar{x}_i)^2\right]^{1/2}$$

and compute for each step R_i/S_i. The sequence can be smoothed if one takes i = 10, 20, 30,… Fit the function $R_i/S_i = ai^H$ to the observed series and interpret the behaviour of the series using appropriate literature. Compute the correlation measure and the Hausdorff-Besicovitch dimension and draw conclusions on the text sequence. Compare different languages.

Take other quantified properties and analyze their sequence in the same way. Finally, try to determine the basis of the chaotic behaviour of linguistic sequences.

References

Çambel, A.B. (1993). *Applied chaos theory. A paradigm for complexity*. San Diego: Academic Press.

Feder, J. (1988). *Fractals*. New York: Plenum.

Hřebíček, L. (1997). Persistence and other aspects of sentence-length series. *Journal of Quantitative Linguistics 4(1-3), 103-109.*

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Hurst, H.E., Black, R.P., Simaika, Y.M. (1965). *Long term storage, an experimental study.* London: Constable.

Mandelbrot, B. (1982). *The fractal geometry of nature.* New York: Freeman.

Mandelbrot, B., Wallis, J.R. (1969a). Some long-run properties of geophysical records. *Water Resources Research 5(2), 321-340.*

Mandelbrot, B. Wallis, J.R. (1969b). Robustness of the rescaled range R/S in the measurement on noncyclic long run statistical dependence. *Water Resources Research 5(5), 967-988.*

Köhler´s word length motives 1

Hypothesis

Length motives are linguistic units behaving like other units.

Procedure

A length motive is a secondary unit consisting of a non-decreasing sequence of lengths of the primary unit, e.g. a sequence of word length. If, e.g. word length is measured in terms of syllable numbers, the sentence “Length motives are linguistic units behaving like other units” consists of the sequence

1-2-1-3-2-3-1-2-2

in which the motives are

1-2, 1-3, 2-3, 1-2-2.

The number of such motives in a text is finite, and if one computes their frequency, one can study their frequency distribution. Compute all length-motives in a long poem and examine whether their rank-frequency distribution is identical with that of word lengths themselves.

Try to set up a text typology using the parameters of the pertinent distribution as indicators. If different languages are analysed, show the differences between them.

Under the assumption that the (hypothetically) greatest word length be L and the greatest sequence length be R, how many different sequences are possible? If $L = R = 3$, the following sequences are possible:

2, 3,

1-2, 1-3, 2-2, 2-3, 3-3,

1-1-2, 1-1-3, 1-2-2, 1-2-3, 1-3-3, 2-2-2, 2-2-3, 2-3-3, 3-3-3,

…

1-1-1-…-2, 1-1-1-…-3, ….

Set up a formula for the number of possible motives. Define motive-richness as a ratio of observed motives (motive vocabulary) to possible motives. Define frequency segments in the same way (cf. also Uhlířová 2007)

References

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M (eds.), *Favete linguis. Studies in honour of Victor Krupa: 142-152*. Bratislava: Academic Press.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: *Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg* (in print)

Lua, K.T. (1990). Analysis of Chinese character stroke sequences. *Computer Processing of Chinese & Oriental Languages 6 (2).*

Uhlířová , L. (2007). Word frequency and position in sentence. *Glottometrics 14, 1-20.*

Köhler´s word length motives 2

Problem

Köhler´s length motives have a length on their own right. Find the distribution of lengths of length motives.

Procedure

In the previous problem the lengths are 2,2,2,3. Set up the distribution of lengths

of Köhler-motives in a text and study the following: is this distribution identical with those used for word length? Do texts differ in this property? Compute the moments of these distributions and display Ord´s scheme (see.“Ord´s criterion”). Try to detect some difference between texts of different genres.

References

Best, K.-H. (ed.). (1997). *The distribution of word and sentence length*. Trier: Wissenschaftlicher Verlag.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa: 142-152*. Bratislava: Academic Press.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: *Proceedings of the Jahrestagung der deutschen Gesellschaft für Klassifikation 2007 in Freiburg* (in print).

Köhler´s word length motives 3

Problem

Compute the distribution of Köhler´s word length motives using the number of morphemes as length units.

Procedure

Count the number of morphemes in the words to be studied. Then solve the tasks in the two previous problems with respect to the differently defined length motives.

If the target language is Japanese, try to perform all computations on moras as counting units. If the target language does not have clear-cut word boundaries, define them categorically. Do not forget that all results hold under this initial condition. If the word segmentation procedure is changed, the results may change, too.

Köhler´s length motives are analogous to rhythmic units which, however, have only length but no combinatorial possibilities. There are rhythmic units like 1, 1-0, 1-0-0, 1-0-0-0,…(1 meaning stressed, 0 meaning stressless syllable).

References

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M (eds.), *Favete linguis. Studies in honour of Victor Krupa: 142-152*. Bratislava: Academic Press.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: *Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg* (in print).

Köhler´s sentence length motives

Problem
Sentence lengths in text can be transformed in a sequence of numbers. Show that all problems concerning motives in this chapter can be applied to sentences.

Procedure
Compute the Hurst exponent, the Minkowski sausage and the Lyapunov coefficient for sentence length and show their differences with respect to individual authors, genres and languages. Compute the autocorrelations and compare the texts.

References
Schils, E., Haan, P.de (1993). Characteristics of sentence length in running text. *Literary and Linguistic Computing 8(1), 20-26.*

Lorenz curve

Problem
Characterize the rank-frequency distribution of words using the Lorenz curve.

Procedure
Set up the rank-frequency distribution of words in a short text. Draw the corresponding Lorenz curve. How to draw it can be found in dozens of links on the Internet. Try to use an aspect of this curve as a characterization of vocabulary richness. Do the same with Gini´s coefficient.

References
Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Lyapunov coefficient

Problem
The coefficient of Lyapunov has been introduced into linguistics by Hřebíček (1997, 2000) but its meaning is not quite clear. Perform experiments with texts of different authors, genres and languages on different units.

Procedure
Transcribe text as a sequence of values of a variable, e.g. word length, sentence

length, polysemy, length of rhythmical units, etc. Let the individual values be x_i. Estimate the Lyapunov coefficient as $\lambda = \frac{1}{k}\sum_i \ln\left|x_i - x_{i+1}\right|$, where k is the number of differences. Zero differences should be left out from the sum (because of the logarithm). The coefficient is used in the study of chaotic behavior.

Try to interpret the coefficient after consulting the relevant literature. Try to derive its variance exploiting the fact that $V(x) = \sigma^2$.

References

Çambel, A.B. (1993). *Applied chaos theory. A paradigm for complexity*. San Diego: Academic Press.

Falconer, K. (1990). *Fractal geometry. Mathematical foundations and applications*. Chichester: Wiley.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Schroeder, M. (1991). *Fractals, chaos, power laws*. New York: Freeman.

Schuster, H.G. (1995). *Deterministic chaos. An introduction*. Weinheim: VCH.

Minkowski sausage

Hypothesis

"The way in which language constructs are arranged into a positional series corresponds to such an order of growth α … which results in a relation similar to constructs and their constituents [i.e. Menzerath´s law – U.S., F.F., G.A.] when they are defined on continuity and radius of the respective Minkowski series" (Hřebíček 2000: 76).

Procedure

The hypothesis says that a "Minkowski sausage" of sequential properties of text abides by a power law. First read Chapter 4 in Hřebíček (2000: 66-76) where an application to the series of unit lengths can be found.

Transcribe a text in terms of, say, word lengths (measured in phonemes, syllables, morphemes etc.) or sentence length (measured in different ways) and obtain a sequence of numbers. Then compute a radius ε and check for all neighbors in the sequence whether the distance d between them is greater than 2ε. The distance between the neighbors x_i and x_{i-1} is given as $d_i = [(x_i - x_{i-1})^2 + 1]^{1/2}$. If $d_i > 2\varepsilon$, we have a *break*, otherwise we have a *continuity*. Add all the d_i of the continuities. Then repeat the procedure increasing ε to ten different values summing the continuities for each value separately. Set up the empirical relation $y = \log(\textit{continuity})/\log(\varepsilon)$ and fit the function $y = a\varepsilon^{-b}$ to the data. Study the values of the parameters a and b for different texts, genres, languages and try to draw both textological and typological as well as general linguistic conclusions. Find properties which yield the same parameters in different languages. Try to compute the

Minkowski-Bouligand dimension of your sequence. Show how the parameters of the power function differ for different units.

References

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.
Schroeder, M. (1991). *Fractals, chaos, power laws*. New York: Freeman.
Tricot, C. (1995). *Curves and fractal dimension*. New York: Springer.

n-Grams of length motives

Hypothesis

There are "length-motives" in text. They depend on style, genre and author. Show that their *n*-grams display regularities.

Procedure

Define a unit and a property which can be unequivocally measured within the framework of a text. A word, e.g., can be regarded as a unit and length as one of its properties. Then transcribe the text in terms of lengths of individual units. A sequence of numbers representing a time series, a Markov chain etc. is obtained. First find the frequencies of individual lengths and set up the empirical distribution. Then consider bigrams, set up the distribution and observe the difference. Finally, study trigrams, quadrigrams, pentagram,… up to decagrams. Observe the following:

1. Are there any *n*-grams that are used more often than expected?
2. How many *n*-gram types are not realized in the text? Try to express this rate of omission formally. Which kinds of *n*-grams disappear with stepwise prolongation (= increasing *n*)? Do the inventories and frequencies differ for authors, styles, genres, historically, in languages, etc.?

Define other properties and study their *n*-grams. Find some further problems concerning Köhler´s motives.

References

Brainerd, B. (1976): On the Markov nature of text. *Linguistics 176(1976), S. 5-30*
Damashek, M. (1995). Gauging similarity with N-grams: language-independent categorization of text. *Science 267, 843-848.*
Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society for Information Science 50(3), 843-848.*
Egghe, L. (2000). The distribution of N-grams. *Scientometrics 47(2), 237-252.*
Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing 9(2), 119-124.*

Köhler, R. (1983). Markov-Ketten und Autokorrelation in der Sprach- und Textanalyse. *Glottometrika 5, 134-167.*

Lua, K.T. (1995). A minimum entropy approach for Chinese text compression. *http://www.iscs.nus.sg/~luakt*

Mayzner, M.S., Tresselt, M.E., Wolin, B.R. (1965). Tables of tetragram frequency counts for various word-length and letter-position combinations. In: *Psychonomic monograph supplements 1(4), 79-143.*

Robertson, A.M., Willet, P. (1998). Applications of N-grams in textual information systems. *Journal of Documentation 54(1), 48-69.*

Runquist, W.N. (1968). Rated similarity of high m CVC trigrams and words and low m CCC trigrams. *Journal of Verbal Learning and Verbal Behavior 7, 967-968.*

Schönpflug, W. (1969). n-Gramm-Häufigkeiten in der deutschen Sprache. I. Monogramme und Digramme. *Zeitschrift für experimentelle und angewandte Psychologie XVI*: *157-183.*

Siméonoff, E. (1965). On the distributions of the "costs" of combinations of K letters in a written language. *Statistical Methods in Linguistics 4, 45-50.*

Suen, C.Y. (1979). *n*-Gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1/2, 164-172.*

Willett, P. (1979). Document retrieval experiments using indexing vocabulary of varying size II. Hashing, truncation, digram and trigram encoding of index terms. *Journal of Documentation 35(4), 296-305.*

Yannakoudakis, E.J., Tsomokos, I., Hutton, P.J. (1990). n-Grams and their implication to natural language understanding. *Pattern Recognition 23,(5), 509-528.*

Nominal style

Problem

Nominal style is sometimes contrasted to verbal style. Try to express the difference quantitatively.

Procedure

Count the number of nouns (*N*) and verbs (*V*) in a text. The rest of the words are not relevant. Perform the following test

$$X^2 = \frac{(N-V)^2}{N+V}$$

or alternatively

$$z = \left(\frac{2N}{R} - 1\right)\sqrt{R}\,,$$

where $R = N+V$. The tests are equivalent. X^2 is a chi-square test with 1 degree of freedom and the critical value is 3.84, z is a normal test, $z^2 = X^2$ and the critical value is ± 1.96. Interpret the result of the test.

Compare the style of lyrical and epical poetry, that of scientific texts and newspaper texts. Describe your observations.

References

Ziegler, A., Best, K.-H., Altmann, G. (2002). Nominalstil. *ETC – Empirical Text and Culture Research 2, 72-85*

Phonetic aggregation

Problem

According to the Skinner hypothesis, there is, within a short distance, an increased probability that a unit which was used once will be repeated. Skinner explained this phenomenon with the assumption of an increased activation level of the neurons involved. One of the consequences is, that in spontaneous speech text blocks like sentences or verses placed in short distance from one another are phonetically more similar than those lying with a long distance. This effect can be shown especially in spontaneously narrated folk-poetry. Do the following.

1. Test whether the hypothesis holds for Goethe or Shakespeare or Ovid.
2. Determine whether decreasing phonetic similarity of entities with growing distance can be considered a sign of spontaneity.

Procedure

Transcribe the selected poem phonetically (allophonemically). Devise a measure of phonetic similarity of verses. Compute the average phonetic similarity of verses in distances $x = 1,2,3,\ldots$ Determine whether there is a decreasing tendency and find a formula of this decrease, perhaps $y = ax^{-b}$. If a tendency is found, can we conclude that spontaneity is associated with phonetic similarity?

Compare folk-poetry with modern poetry. Study the speech of individual persons in a drama. Are the passages spoken by different persons phonetically more similar than passages spoken by the same person? If so, a drama can have a very complex phonetic structure.

References

Altmann, G. (1968). Some phonic features of Malay shaer. *Asian and African Studies 4, 9-16.*

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer

Skinner, B.F. (1939). The alliteration in Shakespeare´s sonnets: A study in literary behaviour. *Psychological Record 3, 186-192*.

Skinner, B.F. (1941). A quantitative estimate of certain types of sound-patterning in poetry. *The American Journal of Psychology 54, 64-79.*

Polylogue analysis

Problem

Both the frequencies and the sequences of speech acts in stage-plays furnish a number of problems which can be solved in the framework of a project.

Procedure

1. Compute the rank-frequency distribution and the spectrum for each person separately. Order the persons according to the number of words and correlate a parameter of the distributions with this order.
2. Compute the distribution of sentence lengths of individual persons and correlate the mean with the importance of the persons.
3. Classify the speech acts and compute for each person a vector whose elements are the proportions of different speech acts. Use a distance (or similarity) measure to perform a classification of persons.
4. Compute a transition probability matrix of the sequence of speakers and study its properties. Draw conclusions about the interaction of persons. Set up a weighted graph of interactions.
5. Set up the sequence of speech acts denoting them by letters and study the properties of this sequence. Are there any runs (= uniterrupted sequences of identical letters)?
6. Scale the speech acts in a certain dimension, e.g. attitude, and trace up the behaviour of this sequence. Do not use Fourier series but try to scrutinize the fractal dimension of the sequence.
7. Study formally the attitude of each person to other people on the basis of the speech act data.
8. Scaled speech acts give rise to speech act motives. Study their distribution and sequence.
9. Compare the stage-plays of one genre with those of another (e.g. drama with comedy).
10. Compare the plays of an author historically and try to find a developmental characteristic. Use different text characteristics.
11. Perform analogous computations with sentence types, which must be, beforehand, defined exactly.
12. If possible, scale the speech acts in a semantic domain by means of Osgood´s semantic differential with appropriate dimensions (Osgood et al. 1957).

References

Osgood, C.E., Suci, G.J., Tannenbaum, P.H. (1957). *The measurement of meaning.* Urbana: Univ. Illinois Press.

Snider, J.G., and Osgood, C.E. (1969) *Semantic Differential Technique: A Sourcebook.* Chicago: Aldine.

http://www.indiana.edu/~socpsy/papers/AttMeasure/attitude.htm

Popescu´s vocabulary richness

Problem

A text is the richer in its vocabulary, the more words there are with small frequencies. One of the ways of characterizing this text property is Popescu´s index R_1. Try to characterize different texts and authors.

Procedure

Since autosemantics contributing to vocabulary richness usually have higher ranks than *h* (cf. "Popescu´s typological *a*-indicator", Chapter 7), Popescu (Popescu et al. 2008) proposes the following index:

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right),$$

where $F(h)$ is the cumulative relative frequency of words having ranks smaller than or equal to *h*; *h* is the *h*-point; *N* is text length (measured in word forms). Process different texts, rank the words according to their frequencies, compute *h* and $F(h)$ and finally the above-given indicator.

If you want to compare different texts with respect to their difference in vocabulary richness, test the difference between two R_1 indices using the criterion

$$z = \frac{R_{1,1} - R_{1,2}}{\sqrt{Var(R_{1,1}) + Var(R_{1,2})}},$$

where $Var(R_{1,i}) = F(h_i)[1\text{-}F(h_i)]/N \quad (i = 1,2)$.

Try to characterize texts, authors and genres. For other indicators of vocabulary richness see the references.

References

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter

Ratios

Problem

In style analysis different ratios have been applied. The best known one is Busemann´s Verb-Adjective Ratio: (*Number of Verbs*)/(*Number of Adjectives*). Set up different ratios, normalize them, derive their sampling variance and set up an asymptotic test. Interpret the ratio.

Procedure

First set up an index that varies within the interval <0, 1>, i.e. normalize it. Busemann´s index is not normalized and in the given form it is not interpretable. The simplest way is to set up an index in the form of a proportion, in which case the variance is automatically given and one can easily set up a test.

Compute the index for different texts and compare them using your test. Interpret the index. Try to find significant differences between texts or genres and find out whether such tests can be used for stylistic analysis.

References

Altmann, G. (1978). Zur Verwendung der Quotiente in der Textanalyse. *Glottometrika 1, 91-106.*

Tuldava, J. (2005). Stylistic author identification. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 369-387*. Berlin/New York: Mouton de Gruyter.

Rhythmic units

Problem

The sequence of stressed and unstressed syllables in prosaic texts establishes a kind of rhythm displaying different properties. Find some properties and the regularities of their behaviour.

Procedure

Define a rhythmic unit as a sequence of a stressed syllable followed by unstressed syllables. (In poetry they can be defined differently.) The stressed syllable can be marked as 1, the unstressed as 0. Hence we have units such as 1, 10, 100, 1000,… The lengths of these units can be defined in terms of the number of syllables constituting them. Transcribe the text as a sequence of lengths of these units. The transcription 101001010000110, e.g., will become the sequence 2,3,2,5,1,2. Study the following:

1. Which is the distribution of lengths? Is there a general distribution holding for all prosaic texts or are there differences between texts? Test the Hyperpoisson distribution.

2. Set up a table of transition probabilities between lengths and study the sequence as a Markov chain. Determine the order of the chain. Compute the limiting state probability vector of the transition probability matrix.
3. Consider the vector in 2. as a text characteristic and compare different texts on the basis of the Euclidian distance.
4. Compute the frequencies of bigrams, trigrams, etc., set up their distributions and compute their entropies. Try to model the course of entropy from monograms to some *n*-grams (according to the length of text). Find the smallest *n* (directly or by extrapolation) at which the entropy reaches its maximum (i.e. where all *n*-grams occur exactly once).
5. Study the autocorrelation of symbols (0, 1) and separately that of lengths up to lag $k = 20$. Draw a graph of the autocorrelations.
6. Study the distances between equal lengths. Are they random or do they display a tendency? Apply Zörnig´s distribution of distances to decide the state of randomness.

References

Best, K.-H. (2002). The distribution of rhythmic units in German short prose. *Glottometrics 3, 136-142.*

Best, K.-H. (2005). Längen rhythmischer Einheiten. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.). *Quantitative Linguistics. An International Handbook: 208-204.* Berlin/New York: de Gruyter.

Eom, J. (2006). *Rhythmus im Akzent. Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhythmus im Russischen.* München: Sagner.

Marbe, K. (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker´sche Verlagsbuchhandlung.

Zörnig, P. (1984a). The distribution of distances between like elements in a sequence I. *Glottometrika 6, 1-15.*

Zörnig, P. (1984b). The distribution of distances between like elements in a sequence II. *Glottometrika 7, 1-14.*

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22.*

Text difficulty

Problem

Discuss the problem of measuring text difficulty.

Procedure

After consulting part of the voluminous literature collect all properties affecting text difficulty and analyse them step by step. Try to give reasons for the influence of a property on text difficulty (readability). You can find a survey over some properties in Kukemelk, Mikk (1993).

Collect all text difficulty (comprehensibility) formulas, present them and discuss their weak aspects. Try to find dependencies among the properties and eliminate the redundant ones if possible. If you propose a new formula, pay attention to its sampling properties and the possibility of comparing two texts with it.

References

Kukemelk, H., Mikk, J. (1993). The prognosticating effectivity of learning a text in physics. *Glottometrika 14, 82-103.*

Thematic concentration

Problem

Thematic concentration refers to the focus on a specific set of words. Try to compute thematic concentration of a poetic and a scientific text.

Procedure

Compute the frequency of words in a text. You will obtain better results if the text is lemmatized. Then set up the rank-frequency distribution of the words and compute the *h*-point (cf. "Popescu´s typological *a*-indicator", Chapter 7). Consider only autosemantic words in the pre-*h* domain (i.e. those having rank $r´ \leq h$). In prose works even proper names occur in these positions. Decide whether they belong to the theme. Popescu´s index of thematic concentration is defined as

$$TC = 2 \sum_{r´=1}^{T} \frac{(h - r´) f(r´)}{h(h-1) f(1)},$$

where

h = *h*-point
$r´$ = rank of a thematic word in the pre-*h* domain
$f(r´)$ = frequency of the thematic word at rank $r´$
$f(1)$ = frequency of the most frequent word
T = number of ranks occupied by thematic words in the pre-*h* domain.

Sometimes a thematic word can have several forms meaning the same, e.g. *Julia, the young girl, the love-sick*, etc. You can add the pertinent frequencies and set up the ranking in this compressed way and obtain a different picture and a different thematic concentration.

Analyze many texts and try to get the difference of thematic concentration in different genres. Then try to order the genres according to thematic concentration.

References

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V.,

Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Tokemes and the Lyapunov coefficient

Problem

In a text most of the word tokens can be replaced by a number of alternatives at the disposal of the author without changing the meaning but giving it a special nuance. The set of these alternatives and the word token itself form a tokeme M_k of size $|M_k|$, k being the position in the text. Hence a text is a sequence of tokemes. We are interested only in the *sequence of tokeme sizes* displaying the *author's information content* (freedom of choice) at a given position. Compute the Lyapunov coefficient for such a sequence.

Procedure

Up to now only one text has been studied in this way (cf. Andersen, Altmann 2006). Take the numbers in the Table in the Appendix, p. 109-115 (Tokeme size $|M_k|$) of the quoted article and compute the Lyapunov coefficient for this sequence.

For the researcher whose mother tongue is the target language, try to analyze several short texts in this way. Draw conclusions about individual texts and genres, and try to scrutinize the *author´s* information flow in general. Study also other characteristics of the sequence.

References

Andersen, S., Altmann, G. (2006). Information content of words in texts. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 91-115.* Dordrecht: Springer.

Type-token relation

Problem

Compute the fractal dimension of the Köhler-Galle type-token sequence.

Procedure

The Köhler-Galle TTR has the form:

$$TTR_x = \frac{t_x + T - \frac{xT}{N}}{N},$$

where

x = position in text (= number of words tokens up to position x)

t_x = number of types up to position x (inclusively)
T = number of types in the whole text
N = text length (= number of tokens in the whole text).

Take a text and compute the sequence as given by the formula. Then draw a graph of the sequence, which is evidently a fractal. Compute different kinds of fractal dimensions.

Perform the computation on different texts and in different languages. Observe similarities or differences.

References

Falconer, K.J. (1990). *Fractal geometry. Mathematical foundations and applications*. Chichester: Wiley.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Köhler, R., Galle, M. (1993). Dynamic aspects of text characteristics. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 46-53*. Trier: WVT.

Schroeder, M. (1991). *Fractals, chaos, power laws. Minutes from an infinite paradise*. New York: Freeman.

Tricot, C. (1993). *Curves and fractal dimensions*. New York: Springer.

Wimmer, G. (2005). The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 361-368*. Berlin/New York: Mouton de Gruyter.

Verb profile

Problem

One of the possible characterizations of a text is its verb profile, i.e. its complex picture representing its "verbal behavior".

Procedure

Set up a profile of a text concerning only its verbs. Levickij and Lučak (2005) set up the following semantic classes of verbs:

1. Exchange Verbs (*barter, buy, sell, exchange, pay, trade*).
2. Measure Verbs (*bill, charge, cost, estimate, fine, measure, price, value, weigh*).
3. Change of Ownership Verbs (*give, take, receive, borrow, lend, steal, return*).
4. Change of Position (*fall, drop, throw, slide, float, roll, fly, rotate, shift*).
5. Change of Physical State (*melt, redden, soften, freeze, harden, dry, break*).
6. Circumstance Verbs (*begin, start, stop, repeat, commence, continue, finish, halt, complete, quit, initiate, end, keep*).
7. Impact/Effect Verbs (*cut, stab, crush, smash, pierce, bite, shoot, kill*).
8. Directed Motion Verbs (*enter, come, go, arrive, descend, ascend, raise, lower, exit, rise, depart, return, leave*).

9. Verbs of Existence (*exist, live, dwell, loom, remain, reside, accumulate, aggregate, herd, gather, create, appear, disappear*).
10. Ingestion Verbs (*chew, drink, eat, gobble, ingest, munch, sip, suck, swallow*).
11. Verbs of Mental Processes (*acquire, guess, know, learn, memorize, study, think*).
12. Load/Spray Verbs (*scatter, spray, pile, pack*).
13. Manner of Motion Verbs (*bounce, dance, follow, hop, jog, jump, march, ride, sail, shuffle, stroll, track, walk, wander*).
14. Verbs of Ownership (*belong, have, hold, keep, own, possess*).
15. Verbs of Perception and Communication (*ask, communicate, feel, hear, listen, look, notice, perceive, see, shout, smell, speak, talk, tell, watch*).
16. Position Verbs (*remain, stay*).
17. Verbs of Removing (*draw, eliminate, remove, empty, scrub, sweep, peel, shell*).
18. Orientation Verbs (*aim, face, orient, point*).
19. Verbs of Psychological State (*amuse, annoy, frighten, enjoy, fancy, hate, like*).
20. Verbs of Sound Emission (*bark, chatter, roar, yelp, rumble, strike, squeak, tick*).

With all these classes the vector would have 20 elements. Try to pool the classes in different ways in order to obtain a smaller vector. Use other classifications. Try to perform different scalings of these classes e.g. in evolutionary order; from existence over moving, eating, feeling, grasping, perceiving, and so on to mental processes etc. Then set up the relevant vectors; for the above classification it would be

$$V = \{e1, e2, \ldots e20\}.$$

Then consider different texts and compute (a) the number of types (of verbs) belonging to these 20 classes; (b) the number of tokens belonging to these 20 classes. Normalize the values if necessary. Reorganize the vector according to your scaling. Compare texts and show that different genres have different verbal profiles.

In the second step do not consider clear-cut (crisp) classes but each verb belonging to a class only to a certain degree. Try to work with fuzzy sets or rough sets.

References

Halliday, M.A.K. (1994). *An introduction to functional grammar*. London: Arnold.

Levickij, V.V., Lučak, M. (2005). Category of tense and verb semantics in the English language. *Journal of Quantitative Linguistics 12(2-3), 212-238.*

Levin, B. (1999). *English verb classes and alternations*. Chicago: University of Chicago Press.

Scheibman, J. (2001). Local pattern of subjectivity in person and verb type in American English conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 60-89.* Amsterdam/Philadelphia: Benjamins.

Vocabulary richness and references

Hypothesis

Hřebíček (1985) starts from two assumptions: (1) the greater the vocabulary richness of a text, the smaller the number of references; (2) the more sentences are in the text, the more references there are. He developed the formula:

$$r = csn^b$$

where: r = number of references; s = number of sentences in text; n = text length (number of tokens, i.e. word forms), c and b are parameters.

Procedure

Define exactly what a reference is, then analyze several texts. Are there differences in parameter b for (a) individual writers, (b) individual genres, (c) individual languages. The parameters b and c in Hřebíček's formula can be estimated from the data by means of classical methods or by means of algorithms for iterative optimizisation. Compare the results with those of Hřebíček. Describe the new formula and find a foundation.

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Hřebíček, L. (1985). Text as a unit and co-references. In: Ballmer, Th.T. (ed.), *Linguistic dynamics: 190-198*. Berlin/New York: de Gruyter.

Word frequency 1

Problem

The rank-frequency distribution of words in a text conceals a number of problems that are not fully solved. Try to scrutinize some of them.

Procedure

1. Compute the rank-frequency distribution of words (i.e. lemmas, not word forms) in a text.
2. Transform it in a *cumulative* rank-frequency distribution (empirical distribution function). Find an empirical continuous function fitting well to this distribution; it may be a polynomial.

3. Compute the *curve length* using standard formulas from analysis.
4. Try to answer the question: *is this curve length in any way related to vocabulary richness?* If it does, interpret it, i.e. decide whether a greater curve length goes along with greater vocabulary richness. Study the problem on many texts (both short and long ones).
5. Try to fit all available discrete distributions to your data in 1.; do not restrict yourself to the Zipf-Mandelbrot theory.

References

Baayen, R.H. (2005). Word frequency distributions. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 397-409*. Berlin/New York: Mouton de Gruyter.

Word frequency 2

Problem

Try to elaborate on the history of the study of the rank-frequency distribution of words.

Procedure

Consult the available literature completely, restricting yourself to formulas describing rank-frequency distributions and their foundations. Show the differences in the conception of word (word form or lemma), sampling methods (random sampling, complete texts, homogeneous texts, etc.), and word classes. Begin with Estoup (1916) and do not skip Russian works. The literature concerning the problem is enormous; here only surveys are given.

References

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Chitashvili, R.J., Baayen, R.H. (1993). Word frequency distributions. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 54-135*. Trier: Wissenschaftlicher Verlag.

Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Text, Sprache, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin: Mouton de Gruyter.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 791-807.* Berlin: Mouton de Gruyter.

Word frequency 3

Problem

Usually the rank-frequency distribution of words can be captured using the Zipf (zeta) distribution. Show that the Popescu-Altmann-Köhler (PAK) curve yields better fitting results.

Procedure

Use any rank-frequency data of words in a text. Use a function-fitting program (e.g. NLREG, TableCurve, Origin, etc.) and fit the power function $y = Cx^{-a}$ to the data $y = Cx^{-a}$. Register the determination coefficient R^2. Then try the function

$$y = 1 + a*exp(-bx) + c*exp(-dx)$$

and compare the resulting determination coefficients. Look at the graphs of the functions. The power function converges to zero while the other curve converges to 1 and captures the hapax legomena better. Try to modify the power function using $y = 1 + Cx^{-a}$ and compare the determination coefficients. Take only one component of the PAK and compare the results.

References

Popescu, I.-I., Altmann, G., Köhler, G. (2008). Zipf´s law – another view (submitted).

Chapter 5

Frequency and length

Distribution of word length 1

Problem

In his study of Inuktitut word length Peter Meyer (1997, 1999) found a new distribution (a convolution of the Poisson and Thomas distributions) and established a new linguistic foundation. Analyse the distribution and try to apply it to other languages.

Procedure

Find the first moments of the distribution using the probability generating function. Then try to find estimators of the two parameters using the moments or frequency classes. Then test the distribution on any empirical distribution of word length from the literature. Consider different languages, and if the distribution is adequate, try to describe the general features of these languages.

References

Best, K.H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Meyer, P. (1997). Word length distribution in Inuktitut narratives: empirical and theoretical findings. *Journal of Quantitative Linguitics 4, 143-155*.

Meyer, P. (1999). Relating word length to morphemic structure: a morphologically motivated class of discrete probability distributions. *Journal of Quantitative Linguistics 6(1), 66-69.*

Distribution of word length 2

Problem

Find an adequate distribution of word length in several texts in a language that has not been thus studied as yet. Define word length in terms of syllable numbers. Only if the language analysed is monosyllabic, phonemes should be defined as counting units.

Procedure

Consult the available literature. Proceed inductively, i.e. fit different distributions and choose the one which is adequate for all texts. Study the parameters of the distributions and try to find a trend or differences between empirical distributions.

References
Best, K.-H. (ed.) (1997). *The distribution of word and sentence length*. Trier: Wissenschaftlicher Verlag.
Best, K.H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.
Wimmer, G., Altmann, G. (1996). The theory of word length: some results and generalizations. *Glottometrika 15, 166-180.*
Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). Towards a theory of word length distributions. *Journal of Quantitative Linguistics 1, 98-106.*
http://www.gwdg.de/~kbest/litlist.htm

Distribution of word length and Ord´s criterion

Hypothesis
In Ord´s <*I*, *S*>-scheme, all word length distributions of texts of an author are placed on a straight line.

Procedure
Use the results of the previous problem ("Distribution of word length 2") and compute for each text the functions of J. K.Ord.

Plot the computed values in an <*I*, *S*> coordinate system and compute for each author the straight line. It should be noted that in Slavic languages one sometimes takes zero-length into consideration. Compare your results with those in the literature.

References
Best, K.H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.
Best, K.-H. (2003). *Quantitative Linguistik. Eine Annäherung* (2. Auflage). Göttingen: Peust & Gutschmidt.

Frequency and compounding propensity

Hypothesis
"The more frequent a word, the more compounds it produces." (Altmann 1989)

Procedure
Draw a random sample of words of the same word class from a corpus and note their relative frequencies. Then, for each word, ascertain the number of compounds of which it is a component. To generalize, several languages should be analysed.

Order the words according to their frequency and set up the relation <*word frequency, number of compounds*>. A monotone increasing function can be obtained. Try to find an empirical formula and derive the formula from a proportionality argument.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 46-70.*

Andrukovič, P.F., Korolev, E.I. (1977). O statističeskich i leksikogrammatičeskich svojstvach slov. *Naučno-techničeskaja Informacija Serija 2, 4, 1-9*

Bertram, R., Schreuder, R., Baayen, R.H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology; Learning, Memory, and Cognition 26, 419-511.*

Hay, J. (2003). *Causes and consequences of word structure*. New York: Routledge.

Frequency and irregularity

Hypothesis

"… there is a relationship between high frequency and irregularity" (Corbett, Hippisley, Brown, Marriot 2001: 202).

"The more frequently used a construction is, the greater is the likelihood that its form will be maintained, rather than being replaced by some more productive construction" (Bybee 2001: 348).

"…that which is more frequent….is more irregular." (Fenk-Oczlon 2001: 435).

Procedure

The authors consider irregularities in declination and propose a scaling procedure of irregularity.

1. Transfer the problem to conjugation or another grammatical category in any language.
2. Try to generalize the problem devising a general method for scaling deviations from expectation.
3. Select each 10^{th} word from a frequency dictionary of word forms (ordered by ranks) and measure its irregularity. Then try to find a function capturing the relation <*rank, irregularity*> and corroborate it. Read the discussion in the quoted article and try to generalize the concept of irregularity in language.

Make a rank-frequency wordlist of only verbs out of a long text or corpus. Designate the regular verbs with *R*, the irregular ones (irregularity of any kind, without scaling) with *I*. Perform Wilcoxon´s U-test to see whether the second hypothesis holds. Then do the same for nouns. Choose a language with strong declination. Then try to generalize.

References

Corbett, G., Hippisley, A., Brown, D., Marriott, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 201-226*. Amsterdam/Philadelphia: Benjamins.

Bybee, J. (2001). Frequency effects on French liaison. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 337-359*. Amsterdam/ Philadelphia: Benjamins.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam/Philadelphia: Benjamins.

Frequency and letter utility

Hypothesis

There is a relationship between frequency of a letter and its graphemic utility. (Bernhard, Altmann 2008).

Procedure

Graphemic (positional) utility of a letter is measured as the sum of its positions in graphemes, e.g. the Italian letter <g> occurs in the graphemes <g, gl, gli, gn, gi, gg, gh, ggh> (a grapheme is the representative of a phoneme). Here <g> occurs 8 times in the first position and twice in the second position, hence PP(g) = 8(1) + 2(2) = 12. Describe exactly the phoneme-grapheme relationship (use the methods in *Analyses of Script*) in a language and compute the graphemic (positional) utility of each letter. Then use a corpus and ascertain the frequency of individual letters. Try to find at least a correlation, if possible, a function, and find a foundation for it.

References

Altmann, G., Fan, F. (eds.) (2008). *Analyses of script. Properties of characters and writing systems.* Berlin/New York: Mouton de Gruyter

Bernhard, G., Altmann, G. (2007). The phoneme-grapheme relationship in Italian. In: Altmann, Fan (eds.), *Analyses of script. Properties of characters and writing systems: 9-19.* Berlin/New York: Mouton de Gruyter.

Frequency and markedness/complexity

Hypothesis

”unmarked members of categories are more frequent than marked members“ (Bybee, Hopper 2001: 1).

"…there exists an equilibrium between the magnitude or degree of complexity of a phoneme and the relative frequency of its occurrence, in the sense that the magnitude or degree of complexity of a phoneme bears an inverse relationship to the relative frequency of its occurrence" (Zipf 1935: 49).
"…wherever the comparative magnitudes of complexity of phonemes are determinable, the magnitude of complexity bears an inverse (not necessarily proportionate) ratio to the relative frequency of occurrence." (Zipf 1935: 79).
"…it seems highly unlikely that the magnitude of complexity is the cause of the relative frequency of occurrence. It can, however, be demonstrated that the reverse is true…" (Zipf 1935: 81).
"The accent, or degree of conspicuousness, of any word, syllable, or sound is inversely proportionate to the relative frequency of that word, syllable, or sound, among its fellow words, syllables, or sounds, in the stream of spoken language. As usage becomes more frequent, form becomes less accented, or more easily pronounceable, and vice versa"(Zipf 1929: 4).
"…the term 'markedness' can easily be replaced by 'frequency'. Frequency is, moreover, a tangible empirical variable whereas markedness is a theoretical construct." (Fenk-Oczlon 2001: 435).
"Semantic unmarkedness and high frequency usually will converge" (Fenk-Oczlon 2001: 441).

Procedure

Make a list of marked vs. unmarked entities. Devise (a) a specification of the hypothesis, i.e. state in individual cases what is marked and what is unmarked, (b) a measurement procedure of markedness, (c) derive a formula, (d) perform a test.

Choose 5-10 marked – unmarked dichotomies from different language levels and perform the above procedure. Then try to compute *degrees of markedness*, because a dichotomy is an extreme reduction of information. Classes can be marked in different degrees (cf. declination, conjugation). Find for each entity a different scaling method if necessary. A good example is in Corbett, Hippisley, Brown, Marriott (2001). Compare the degrees with the frequency of units, draw a graph and set up a proportionality hypothesis. Derive curves from the hypotheses. Test them on your data.

Comment on the expression "tangible empirical variable" and "theoretical construct".

For the concept of markedness, see pp. 1, 28, 52, 54, 61, 68, 71, 82, 101, 131, 138, 140, 152-154, 185, 192, 204, 213, 215-216, 223, 226, 234, 236, 246, 292, 293, 315, 317, 330, 344, 387, 435, 439-443, 450, 457, 465, 466 in Bybee, Hopper (2001a).

References

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P. (eds.), *Frequency and the*

emergence of linguistic structure: 1-24. Amsterdam/Philadelphia: Benjamins.
Bybee, J., Hopper, P. (eds.) (2001a), *Frequency and the emergence of linguistic structure: 1-24*. Amsterdam/Philadelphia: Benjamins.
Corbett, G., Hippisley, A., Brown, D., Marriott, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 201-226*. Amsterdam/Philadelphia: Benjamins.
Fenk-Oczlon, G. (1991). Frequenz und Kognition – Frequenz und Markiertheit. *Folia Linguistica 25, 361-394.*
Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam/Philadelphia: Benjamins.
Fenk-Oczlon, G. (1990). Frequenz und Kognition – Frequenz und Markiertheit. Folia Linguistica 25, 361-394.
Greenberg, J.H. (1966). *Language universals*. The Hague: Mouton.
Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press.
Zipf, G.K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 40, 1-95.*

Frequency and order in freezes

Hypothesis

”more frequent word [occurs] before less frequent word.” (Fenk-Oczlon 2001: 437). The hypothesis concerns freezes like German “mit Kind und Kegel”.

Procedure

Collect about 500 freezes (i.e. as many as possible): from a phraseological dictionary) in a language and state whether the first word is generally more frequent than the second in the freezes. Consult a frequency dictionary or a corpus to find the frequencies. Perform a sign test on the hypothesis.

Combine the hypothesis with other ones concerning freezes and try to obtain an overall explanation. See the problem “Behagel´s ´law’” (Chapter 2).

References

Chafe, W. (1994). *Discourse, consciousness, and time*. Chicago/London: University of Chicago Press.
Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam /Philadelphia: Benjamins.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. Linguistics 27, 517-556.
Givón, T. (1984). *Syntax: a functional typological introduction. Volume 1*. Amsterdam/Philadelphia: Benjamins.
Givón, T. (1990). *Syntax: a functional typological introduction. Volume 2*. Amsterdam/ Philadelphia: Benjamins.
Siewierska, A. (1988). *Word order rules.* London/New York/Sydney: Croom Helm.
Sobkowiak, W. (1993). Unmarked-before-marked as a freezing principle. *Language and Speech 36, 393-414.*

Frequency and phoneme complexity

Hypothesis

"…there exists an equilibrium between the magnitude or degree of complexity of a phoneme and the relative frequency of its occurrence, in the sense that the magnitude or degree of complexity of a phoneme bears an inverse relationship to the relative frequency of its occurrence" (Zipf 1935: 49).
"…wherever the comparative magnitudes of complexity of phonemes are determinable, the magnitude of complexity bears an inverse (not necessarily proportionate) ratio to the relative frequency of occurrence." (Zipf 1935: 79)
"…it seems highly unlikely that the magnitude of complexity is the cause of the relative frequency of occurrence. It can, however, be demonstrated that the reverse is true…" (Zipf 1935: 81)

Procedure

Before testing this hypothesis, define the concept of phoneme complexity. Then, on the basis of a frequency count of any phoneme, try to show graphically the existence of such a dependence. Do the same for several languages beginning with those with 13 phonemes in their inventory up to about 40 phonemes. Accept, modify or reject the hypothesis in accordance with the result obtained. Compare your concept of phoneme complexity with that of Zipf. Specify the independent and the dependent variable (frequency or complexity).

References

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press. (Esp. p. 252-258).

Frequency and phoneme form

Hypothesis

”There is a clear correspondence between weak initial consonants and frequency in English: the more frequent a word, the weaker its initial consonant” (Fenk-Oczlon 2001: 439).

“…in the highest frequency class the frequency distribution of initial consonants…differs considerably from the overall distribution”. “…the share of obstruents is much lower and the share of non-obstruents much higher than in the overall distribution.” (Fenk-Oczlon 2001: 438)

Procedure

Fenk-Oczlon considers a less obstruent phoneme (e.g. glide, vowel) weaker than other ones. Define exactly the extent of obstruency, e.g. by scaling (Fenk-Oczlon’s scaling is as follows: glides, liquids, nasals, fricatives, stops; vowels are the least obstruent) and try to obtain a function for the frequency of each word in a frequency dictionary and obstruency of its first phoneme. If the hypothesis is true, there will be a great dispersion. Try to smooth the data by forming frequency classes. Test at least the first 1000 words for the difference between the frequencies of words with weak and non-weak initial consonants. If the hypothesis does not hold in your language, try to set up a different one, i.e. whether there is some relationship between the frequency of a word and its first phoneme.

References

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam/Philadelphia: Benjamins.

Frequency and production effort

Hypothesis

”…if there are two ways of saying the same thing, the one which is less ’costly’, that is, in the normal case, shorter and easier to pronounce, will win” (Dahl 2001: 475).

Procedure

First try to make precise the concept of “easier to pronounce”. It usually refers to production effort and its minimization. Second, specify the meaning of “will win” in the hypothesis. Try to make it precise: does it mean that it occurs more frequently than the “more difficult” expression, or that it will replace it? Third, state whether the hypothesise means that length and ease are the causes of frequency. Usually frequency is considered as cause of shortness and ease (see

Length and frequency). Consider this hypothesis as an example of unclear formulation and try to make it precise. Extend your argument to the testability of hypotheses. Define the concept of testability and untestability of a hypothesis.

References

Bunge, M. (1967). *Scientific research I.* Berlin/Heidelberg/New York: Springer.

Dahl, Ö. (2001). Inflationary effects in language and elsewhere. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 471-480*. Amsterdam /Philadelphia: Benjamins.

Frequency and productivity

Hypothesis

The more frequent a morpheme the greater its morphological productivity (Krott 2002).

Procedure

This is a generalization of the previous problem, extended to any type of morphological constructs (derivation, composition, reduplication). The direction of the hypothesis, i.e. what is the dependent and independent variable, is not fixed. Collect all the morphemes and their frequencies from a text corpus and obtain for each of them all the morphological constructions (types) in which they occur. Derive a theoretical function starting from Köhler´s control cycle and try to fit it to empirical data.

References

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*. http://ubt.opus.hbz-nrw.de/volltexte/2004/279/.

Frequency and reduction

Hypothesis

"*Probabilistic Reduction Hypothesis*: word forms are reduced when they have a higher probability. The probability of a word is conditioned on many aspects of its context, including neighbouring words, syntactic and lexical structure, semantic expectations, and discourse factors" (Jurafsky, Bell, Gregory, Raymond 2001: 229).

"… words which are strongly related to or predictable from neighboring words, such as collocations (sequences of commonly cooccurring words), are more likely to be phonologically reduced" (Jurafsky, Bell, Gregory, Raymond 2001: 230).
"…, predictability not only affects vowel duration, but has an additional independent non-categorical effect on word duration" (Jurafsky, Bell, Gregory, Raymond 2001: 239).
"The higher the probability of the word given its neighbor, the shorter the word" (Jurafsky, Bell, Gregory, Raymond 2001: 240).
"The different formal words of any vocabulary are then, it seems, the residues of specific past acts of abbreviatory process, and are the names of experiential categories which are frequently referred to in the stream of speech." (Zipf 1935: 271).
"…as a meaningful configuration becomes relatively more frequent, it becomes simultaneously less articulated and more integrated." (Zipf 1935: 272).
"deletion […] is more prevalent in high-frequency words than in low-frequency words" (Pierrehumbert 2001: 138).
"… high-frequency discourse items tend to reduce faster than lower-frequency items"(Bush 2001: 257).
"... frequent words reduce faster than infrequent words" (Fenk-Oczlon 2001: 436).
"…abbreviatory acts of truncation seem to arise on the whole as a consequence of the increased frequency in usage of a word, whether within the entire speech-community or within certain minor groups thereof." (Zipf 1935: 33).
"…where frequency and abbreviatory substitution are connected, the frequency is the cause of the abbreviatory substitution;" (Zipf 1935: 36).
"…the accumulated effect of acts of durable abbreviatory substitution during the evolution of language is in part reflected by the frequency-magnitude relationship of words today." (Zipf 1935: 36)
"…with temporary abbreviatory substitutions one cannot prove statistically that frequency is the inevitable cause of all substitutions of shorter forms." (Zipf 1935: 37)

Procedure

For each word try to separate all factors (neighboring words *N*, syntactic structure *Sy*, lexical structure *L*, semantic expectation *Se*, discourse factors *D*). Perform the necessary measurements and try to express the Extent of Reduction as $R = f(N, Sy, L, Se, D)$. Begin with a linear relationship and add complexity stepwise, e.g. $R = f(N)$, $R = f(Sy)$, etc. and combine them. Accept a relationship only if it reduces the variance. Express the shortening at least as a function of neighborhood probability.

Set up more hypotheses on this relation and show that this is a very rich branch of research.

References

Bush, N. (2001). Frequency effects and word-boundary palatalization in English. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 255-280*. Amsterdam/ Philadelphia: Benjamins.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 431-448*. Amsterdam/ Philadelphia: Benjamins.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W.D. (2001). Probabilistic relations between words: evidence from reduction in lexical production. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 229-254*. Amsterdam/ Philadelphia: Benjamins.

Pierrehumbert, J.B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 137-157*. Amsterdam/ Philadelphia: Benjamins.

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press. (Esp. p. 252-258).

Frequency and variety

Hypothesis

"…the number of different words (i.e. variety) seems to be ever larger as the frequency of occurrence becomes ever smaller." (Zipf 1935: 26)

Procedure

The hypothesis is very simple. It says that the frequency distribution (frequency spectrum) of words decreases monotonously. The hypothesis is not specified. Collect as many word frequency distributions as possible and try to find a common distribution or show that there are many distributions. The most common cases are the Zipf (zeta) distribution, the Zipf-Mandelbrot distribution, the Waring distribution etc. Try to find the conditions under which a special distribution holds. See also "Word frequency 1,2,3", Chapter 4.

References

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter (to appear).

Word frequency in: http://lql.uni-trier.de

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press. (Esp. p. 252-258).

Length and frequency

Hypothesis

"…the larger a word is in length, the less likely it is to be used." (Zipf 1935: 22).
"…the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences;" (Zipf 1935: 25).
"…high frequency is the cause of small magnitude." (Zipf 1935: 29).
"It is found that a simple stochastic model gives a rough prediction for the results obtained when all words are combined, but not when words are classified as function and content words. Function words are short and their frequency of occurrence is a decreasing function of their length; content words are longer and their probability is relatively independent of length." (Abstract) (Miller, Newman, Friedman 1958).
"…the greater the number of strokes [in a Japanese *kanji*], the smaller the number of occurrences of a word." (Sanada 2007)
"…the length of a morpheme tends to bear an inverse ratio to its relative frequency of occurrence." (Zipf 1935: 173)
"The magnitude of complexity of a morpheme bears an inverse (not necessarily proportionate; possibly some non-linear mathematical function) relationship to its relative frequency." (Zipf 1935: 176)

Opinions on the above differ; hence large scale research into these hypotheses is necessary.

Procedure

Measure measure the length of each word (form) in a frequency dictionary. Define length in one of the three ways: (a) as the number of phonemes in the word, (b) as the number of syllables in the word, (c) as the number of morphemes in the word. Determining the boundaries between syllables or morphemes is not necessary; their number in the word is sufficient. Then begin with words with frequency 1 and compute their mean length; continue with words with frequency 2 and compute their mean length, etc. Words with high frequencies can be pooled. If the hypothesis is true, a monotonously decreasing function can be obtained in all cases. Show the difference between the curves graphically; try to derive the functions from proportionality arguments. Compare several languages; study a strongly agglutinating language if possible, and state whether strong agglutination has an influence on the parameters of the functions. If an oscillating function results, read the corresponding references below.

References

Baker, S.J. (1951). A linguistic law of constancy: II. *The Journal of General Psychology 44, 113-120.*

Baker, S.J. (1951). Ontogenetic evidence of a correlation between the form and frequency of use of words. *The Journal of General Psychology 44, 235-251.*

Belonogov, G.G. (1962). O nekotorych statističekich zakonomernostjach ruskoj pis´mennoj reči. *Voprosy jazykoznanija 11/1, 100-101.*

Breiter, M.A. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1, 224-231.*

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 387 – 433.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Grzybek, P., Altmann, G. (2002). Oscillation in the frequency-length relationship. *Glottometrics 5, 97-107.*

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie.* Paris : P.U.F.

Guiter, H. (1977). Les relations /frequence-longeuer-sens/ des mots (langue romanes et anglais). In: *XVI Congresso Internazionale di Linguistica e Filologia Romanza, Napoli, 15-20 Aprile 1974, 373-381*. Napoli: Macchiaroli/ Amsterdam: Benjamins.

Hammerl, R. (1990). Länge-Frequenz, Länge-Rangnummer. Überprüfung von zwei lexikalischen Modellen. *Glottometrika 12, 1-24.*

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier, WVT.

Herdan, G. (1966). *The advanced theory of language as choice and chance.* Berlin: Springer.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R., Zörnig, P., Brinkmöller. (1990). Differential equation models for the oscillation of the word length as a function of the frequency. *Glottometrika 12, 25-40.*

Kornai, A. (2002). How many words are there? *Glottometrics 4, 61-86.*

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen in die quantitative und systemtheoretische Linguistik: 75-126.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Leopold, E. (1997). Frequency spectra within word length classes. In: *Third International Conference on Quantitative Linguistics, August 26-29, 1997, Helsinki, Finland: 156.* Helsinki: Monila.

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse*. Hamburg: Kovač.

Leopold, E. (2000). Length-distribution of words with coinciding frequency. In: *Proceedings of the fourth conference of the International Quantitative Linguistic Association, Prague, August 24-26: 76-77.*

Miller, G.A., Newman, E.B., Friedman, E.A. (1958). Length-frequency statistics for written English. *Information and Control 1, 370-389.*

Miyajima, T. (1992). Relationship in the length, age and frequency of Classical Japanese words. *Glottometrika 13, 219-229.*

Sanada, H. (1999). Analysis of Japanese vocabulary by the theory of synergetic linguistics. *Journal of Quantitative Linguistics 6, 239-251.*

Sanada, H. (2006). The selection of scholarly terms in basic vocabulary lists. *Goi Kenkyu (Studies on vocabulary) 4, 21-42.*

Sigurd, B., Eeg-Olofsson, M., van de Weijer, J. (2004). Word length, sentence length and frequency Zipf revisited. *Studia Linguistica 58 (1), 37-52.*

Tuldava, J. (1995*). Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag.

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, Mass.: Harvard University Press.

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press.

Length and polysemy

Hypothesis

The longer a word the smaller the number of its meanings (Zipf).

Procedure

The hypothesis is very old. Usually one considers the *average* number of meanings for a certain length. The result is the usual power function, well known from the literature, well corroborated and very general. However, there remains an unsolved problem.

Draw a large sample from a dictionary – if possible, incorporate the entire dictionary. Measure the syllabic length (x) and the number of its meanings (y).of each of the words. Define exactly how the second variable is measured. Then set up a two-dimensional distribution of the number of words (z) depending on (x, y), i.e. $P(z) = f(x,y)$. The problem is difficult; it is not easy even to obtain a sample. If you do not succeed in collecting corresponding data use those in Altmann et al. (2002: 88) concerning the Indonesian lexicon. If necessary, pool some classes.

Since the lengthening of a word (by derivation or compounding) is caused by the requirement of specification, reflect on the possibility of taking the number of meanings as the independent variable and length as dependent variable. In that case it will suffice to evaluate only a large sample from a dictionary and fit a continuous function. Proceed as is usual in synergetic linguistics.

References

Altmann, G., Bagheri, D., Goebl, H., Köhler, R., Prün, C. (2002). *Einführung in die quantitative Lexikologie.* Göttingen: Peust & Gutschmidt.

Altmann, G., Beőthy, E., Best, K.-H. (1982). Die Bedeutungsmenge und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543.*

Baker, S.J. (1950). The pattern of language. *Journal of General Psychology 42, 25-66.*

Fickermann, I., Markner-Jäger, B., Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. *Glottometrika 6, 115-126.*

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 387- 433.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Hoffmann, Ch. (2001). Polylexie lexikalischer Einheiten in Texten. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: 76-97.* Trier: Wissenschaftlicher Verlag.

Levickij, V. (2005). Polysemie. In: Altmann, G., Köhler, R., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 458-464*. Berlin/ New York: Mouton de Gruyter.

Sambor, J. (1984). Menzerath´s law and the polysemy of words. *Glottometrika 6, 94-114.*

Length and word classes 1

Hypothesis

"…adverbs of time are on the average less independent and therefore shorter than adverbs of place." (Zipf 1935: 242)

"It is found that a simple stochastic model gives a rough prediction for the results obtained when all words are combined, but not when words are classified as function and content words. Function words are short and their frequency of occurrence is a decreasing function of their length; content words are longer and their probability is relatively independent of length." (Abstract) (Miller, Newman, Friedman 1958).

Procedure

Consider all temporal and spatial adverbs of a language with the help of a thorough grammar. Define length exactly. Compute the average lengths of the two classes and compare them using a statistical test. Is Zipf right? If the sampling is very difficult, try to take only simple adverbs, ignore complex ones.

References

Miller, G.A., Newman, E.B., Friedman, E.A. (1958). Length-frequency statistics for written English. *Information and Control 1, 370-389.*

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press.

Length and word classes 2

Problem
Synsemantic words occur more frequently than autosemantic ones. Hence, they are shorter than autosemantics. Thus, some word classes are on the average shorter than other ones.

Procedure
Arrange the words in a frequency dictionary (or a corpus), in ascending order according to length and assign to each word its class membership (in languages where it is possible). Then assign to each word a rank according to its length. Perform a (nonparametric) rank test to show that the word classes have different ranks (the classes differ). Another alternative: compute the mean length of all words in a class and test against the mean length of another class.

References
None

Sentence length and clause length

Problem
Test Sherman´s law and Menzerath´s law concerning sentence and clause length.

Procedure
Count the sentence lengths in several texts in terms of clause numbers, and clause lengths in terms of word numbers. Set up their frequency distributions (the random variable is length).

1. Show that both distributions follow the negative binomial distribution. Compare the parameters of the distributions in different genres and languages.
2. Show that there is a dependence between the parameters of the negative binomial.
3. Test Menzerath´s hypothesis and state that the longer a sentence, the shorter are the clauses. The dependence has the form of a power function.
4. Compare different texts and try to find the possible differences between texts and languages. If possible, investigate especially strongly agglutinating languages. Try to find some divergences and give an explanation.
5. Study the development of sentence length in a special class of texts, e.g. newspaper texts, in the course of several decades.

References
Altmann, G. (1988). Verteilungen der Satzlängen. *Glottometrika 9, 147-170.*
Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2002). Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft 7, 7-13.*

Best, K.-H. (2005). Satzlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook*: *298-304*. Berlin/ New York: de Gruyter.

Dzhurjuk, T. (2006). Sentence length as a feature of style. *Glottometrics 12, 55-62.*

Heeschen, V. (1994). How long are clauses and sentences in a Papuan language like Eipo? *Semaian 10, 50-70.*

Heups, G. (1983). Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. *Glottometrika 5, 113-133.*

Kaßel, A., Livesey, E. (2001). Untersuchungen zur Satzlängenhäufigkeit im Englischen: Am Beispiel von Texten aus Presse und Literatur (Belletristik). *Glottometrics 1, 27-50.*

Kelih, E., Grzybek, P. (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (Am Beispiel slowenischer Texte). *Glottometrics 8, 23-41.*

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. *Glottometrika 16, 213-275.*

Teupenhayn, R., Altmann, G. (1984). Clause length and Menzerath´s law. *Glottometrika 6, 127-138.*

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (Eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Ludek Hřebiček: 266-282.* Trier: Wissenschaftlicher Verlag.

Word length and polytextuality

Hypothesis

According to a hypothesis originating from Köhler´s control cycle, the longer a word the smaller its polytextuality, i.e. the longer a word the smaller the number of different texts in which it occurs; or alternatively, the smaller the number of neighbourhoods in which it occurs.

Procedure

Set up some sets of words of different lengths. In each set place only words of the same word class. Take 5 words of different lengths from one of the classes. Count their occurrences in individual texts of a corpus. Show graphically the relationship <*length, polytextuality*> for each set. Then try to apply theoretical reasoning in order to derive the appropriate function. Show whether the individual sets (containing different word classes) display different parameters of the functions.

References

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: R. Köhler (Hg.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 387 – 433.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Rothe, U. (1983). Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. *Glottometrika 5, 101-112.*

Word length and position in sentence

Problem

Several researchers stated that word length differs in different positions of the main clause of a sentence. Try to find a general formula for the change of word length in individual positions considering clause length as a boundary condition.

Procedure

Separate the sentences/clauses in a long text according to their length and compute for each position the mean word length. Draw a graph of the curve and try to find a formal expression of the curve. Generalize the result. Do not study only Indo-European languages, but avoid monosyllabic languages.

References

Behagel, O. (1930). Von deutscher Wortstellung. *Zeitschrift für Deutschkunde 44, 81-89.*

Croft, B. (1981). *Language universals and linguistic typology. Syntax and morphology*. Oxford: Blackwell.

Fenk, A., Fenk-Oczlon, G. (2005). Within-sentence distribution and retention of content words and function words. In: Grzybek, P. (ed.), *Word length studies and related issues: 157-170.* Dordrecht: Springer.

Greenberg, J. H. (1963/1969). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. H. (ed.), *Universals of language. Report of a conference held at Dobbs Ferry, New York, April 13-15, 1961*. 2nd ed.. Cambridge, Mass.: MIT.

Hawkins, J.S. (1983/1988). *Word order universals*. San Diego: Academic Press.

Hawkins, J.S. (1990). A parsing theory of word order universals. *Linguistic Inquiry 21(2), 223-261.*

Hawkins, J. S. (1992). Syntactic weight versus information structure in word order variation. In: Jacobs, J. (ed.), *Informationsstruktur und Grammatik: 196-219.* Opladen: Westdeutscher Verlag.

Hawkins, J. S. (1994). *A performance theory of order and constituency*. Cambridge: University Press.

Hoffmann, C. (2002). "Early immediate constitutents" – ein kognitiv-funktionales Prinzip der Wortstellung(svariation). In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen in die quantitative und systemtheoretische Linguistik: 31-74.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6, 46-57.*

Niemikorpi, A. (1997). Equilibrium of words in the Finnish frequency dictionary. *Journal of Quantitative Linguistics 4(1-3), 190-196.*

Siewierska, A. (1993). Syntactic weight vs. information structure and word order variation in Polish. *Journal of Linguistics 29, 233-265.*

Uhlířová, L. (1997). Length vs. order: word length and clause length from the perspective of word order. *Journal of Quantitative Linguistics 4(1-3), 266-275.*

Uhlířová, L. (1997a). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost 58, 174-184.*

Uhlířová, L. (1997b). Length vs. order. Word length and clause length from the perspective of word order. *Journal of Quantitative Linguistics 4, 266-275.*

Word/Morph length and composition

Hypothesis

"The shorter a word, the more frequently it occurs in compounds." (Altmann 1989: 104)

Procedure

Divide a sufficiently large number of randomly sampled nouns from a dictionary in classes according to their length (syllabic with words, phonemic with morphemes). Then, for each word/morpheme, find all compounds in which it occurs. Compute averages and set up the function *Mean compound activity = f*(*mean frequency*). Repeat the procedure with other word classes.

References

Altmann, G. (1989). Hypotheses about compounds. *Glottometrika 10, 100-107.*

Krott, A., Schreuder, R., Baayen, R.H. (1999). Complex words in complex words. *Linguistics 37, 905-926.*

Prün, C. (2005). Quantitative Morphologie: Eigenschaften der morphologischen Einheiten und Systeme. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 227-242.* Berlin/New York: de Gruyter.

Chapter 6

Semantics, synergetics, and psycholinguistics

Abstractness

Problem

Devise a measure of *text abstractness*.

Procedure

First consider only nouns. Try to scale their abstractness (not confounding it with generality) using (a) the abstractness of affixes, (b) the kind of definition in a monolingual dictionary, (c) the possibility of the perception of their denotates. Ask test persons to perform a scaling in a defined interval. Do not use contextual constraints. Try to perform an analogous procedure with adjectives and finally with verbs. Describe your scaling procedure exactly.

Set up an overall abstractness index on the basis of your scale of abstractness for the processed word classes,. Take a poetical and a scientific text and compute the extent of their abstractness.

For the researcher trained in statistics: try to derive the expected value and the variance of the index and set up an asymptotic significance test for the difference of two texts. Show that scientific texts are more abstract than poetic ones. Evaluate many texts and try to ascribe abstractness to individual genres.

References

Altarriba, J., Bauer, L.M., Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers 31,* 578-602.

Gilhooly, K.J., Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation 12, 395-427.*

Groeben, N. (1982). *Leserpsychologie: Textverständniss – Textverständlichkeit*. Münster: Aschendorf.

Paivio, A., Yuille, J.C., Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values of 925 words. *Journal of Experimental Psychology Monograph Supplement 76.*

Wiemer-Hastings, K., Graesser, A.C. (1998). Abstract noun classification: A neural network approach. *Proceedings of the 20th Annual Conference of the Cognitive Science Society: 1036-1042.* Hillsdale, NJ: Erlbaum.

Wiemer-Hastings, K., Krug, J., Xu, X. (2006). Imagery, context availability, contextual constraint and abstractness. www.hcrc.ed.ac.uk/cogsci2001/pdf-files/1106.pdf

Distribution of polysemy

Problem

It is assumed that polysemy has a law-like distribution in the dictionary. There are different models; generalizing one speaks of Krylov´s law. Test the different forms of the law or develop new models.

Procedure

Consider the distribution problem in isolation after consultation of the relevant literature. Test the models individually on the data collected by P. Steiner (1995) processing the complete German dictionary of Wahrig (distinguishing word classes; pool the classes if necessary). Find the model that displays the best fit. Try to find arguments for its foundation.

In the article by Levickij, Drebet, Kiiko (1999) one can find data on the distribution of polysemy in German (Table 1,2,3). Try to find a theoretical distribution common to all these data.

References

Krylov, J.K. (1982). Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf´s law: 234-255*. Bochum: Brockmeyer.

Levickij, V. (2005), Polysemie. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 458-464.* Berlin/ New York: Mouton de Gruyter.

Levickij, V.V., Drebet, V.V., Kiiko, S.V. (1999), Some quantitative characteristics of polysemy of verbs, nouns and adjectives in German. *Journal of Quantitative Linguistics 6(2), 172-187.*

Steiner, P. (1995). Effects of polylexy on compounding. *Journal of Quantitative Linguistics* 2(2), 133-140.

Familiarity and frequency

Hypothesis

"…word occurrences are observed by human perception, and […] the frequencies of words are stored in memory." (Köhler, Rapp 2007).

"…the familiarity of a word shows a relative increase as the frequency of perception of that word grows…" (Köhler, Rapp 2007).

Procedure

Conduct an investigation of the relation between frequency and familiarity on data from a language other than English. Elicit judgements on familiarity from test persons. The frequencies of the individual words can be determined using a

frequency dictionary or by counting in a corpus Then test the Köhler-Rapp hypothesis

$$y = \frac{V}{1 + Ax^B},$$

where *y* is the degree of familiarity, *V* is the maximal value of familiarity in your data (i.e. a kind of empirical limit), *x* is the frequency, and *A* and *B* are parameters to be estimated from your data. *B* is negative.

Fit the function to your data and compute the determination coefficient. Compare your results with those from English. If necessary, smooth the data.

References

Kacinik, N., Shears, C., Chiarello, C. (2000). Familiarity for nouns and verbs: not the same as, and better than, frequency. In: Gleitman, L.R., Joshi, A.S.K. (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society: 1035*. Hilsdale, NJ: Erlbaum.

Köhler, R., Rapp, R. (2007). Familiarity and frequency: a psycholinguistic application of synergetic linguistics. *Glottometrics 15, 62-70*.

Kreuz, R.J. (1987). The subjective familiarity of English homophones. *Memory & Cognition 15, 154-168.*

Familiarity of slang words

Problem

Find a scaling procedure for the familiarity of a slang word, its semantic variability and uncertainty.

Procedure

There are three possible outcomes to the question "do you know the meaning of…?" A correct answer, a false answer and "I don´t know". Set up a measure of familiarity of slang words and ask *n* test persons. Study the polysemy of slang words using the answers of test persons. Set up a rank-frequency distribution of meanings of each slang word and try to find an adequate probability distribution.

Compute the entropy of this distribution and try to find the relationship between familiarity and entropy of polysemy.

References

Altmann, G. (2005). Der Diversifikationsprozess. In: Köhler, R., Altmann, G., Piotrowski, R,.G. (eds.). *Quantitative linguistics. An international handbook: 648-659*. Berlin/New York: de Gruyter.

Köhler, R., Rapp, R. (2007). Familiarity and frequency: a psycholinguistic application of synergetic linguistics. *Glottometrics 15, 62-70*.

Serdelová, K. (2005). Some properties of slang words. *Glottometrics 9, 40-45.*

Kanji frequency

Hypothesis

The more frequent a word is, the earlier its *kanji* are learned (Sanada 2006). *Kanji* is a Chinese character in Japanese.

Procedure

Check the reading/writing plans for children in primary schools (for the languages: Chinese, Japanese or Korean). Decide on the scaling of time, e.g. 1,2,3,… (first learned, second learned, third learned) or first month learned, second month learned,… or first year learned, second year learned,…Then count the frequency of individual signs from a frequency dictionary. If the time intervals of learning are determined, determine the average of frequencies of characters. State whether the order of learning or time of learning is a function of frequency. Set up a proportionality relationship. Combine the previous problem with the present one, namely consider

Time of learning = f(*number of strokes, frequency*).

Try to find a two-dimensional dependence.

References

Hall, J.E. (1954). Learning as a function of word-frequency. *American Journal of Psychology 67, 138-140.*

Sanada, H. (2006). The selection of scholarly terms in basic vocabulary lists. *Goi Kenkyu (Studies on vocabulary) 4, 21-42.*

Learning and complexity

Hypothesis

The greater the number of strokes in a Japanese *kanji* (sign), the later a *kanji* is learned (Sanada 2006).

Procedure

Consider Chinese characters in Chinese, Japanese and Korean. At school, children learn every month a certain number of new characters every month. Make a list of characters in the order of learning and compute their complexity. Apply two complexity measures: (a) number of strokes leaning against the way of writing Chinese symbols, (b) another measure of complexity e.g. Altmann (2004). Try to solve the following three problems.

1. State whether the above hypothesis can be expressed by an adequate function in these languages, i.e. <*number of strokes, learning order*> or <c*omplexity*,

learning order>. If no clear-cut function follows, smooth the series in different ways. Try to obtain an adequate function.
2. You can perhaps observe that the empirical order is not quite smooth, hence try to take into account means of school levels. Compute also the variance of the levels and try to ascertain whether there is a relationship <*dispersion of complexity, learning order*>. Most probably this dependence will be smoother than (1).
3. Compare the three languages and state whether the dependencies are similar.

References

Sanada, H. (2006). The selection of scholarly terms in basic vocabulary lists. *Goi Kenkyu (Studies on vocabulary) 4, 21-42.*

Learning with children

Problem

The learning of a language by children is a regular process which can be captured by a function. Find the function(s).

Procedure

Children learn different entities of language very consistently. Try to obtain observations concerning

1. the learning of vowels and consonants from the 1st to the 30th month;
2. the learning of new words in the first ten years;
3. the prolongation of word length (not only lemmas but also forms);
4. the development of sentence length;
5. the development of text length;
6. the development of the rank-frequency distribution of word classes.

Try to connect your data in different networks and observe their changes.

References

This is an independent discipline with enormous literature. The best available is

Ke, J., Yao, Y. (2008). Analysing language development from a network approach. *Journal of Quantitative Linguistics 15(1), 70-99.*

Meaning and frequency

Hypothesis

"The primary meaning of a word is then its statistically most frequently occurring meaning in the group for which one wishes to establish the primary meaning." (Zipf 1935: 276)

Procedure

Generalize the problem in the following way. Show that the individual meanings of any word follow a (ranking) probability distribution, i.e. the frequencies of individual meanings are statistically ordered. Select randomly some words with many meanings from a dictionary and print out all sentences containing them in a corpus, in order to identify the meaning in the given sentence. This is a simple problem of diversification, sometimes called Beőthy-law.

References

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 646-658*. Berlin/New York: Mouton de Gruyter.

Baayen, R.H. (2005). Morphological productivity. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 243-255*. Berlin/New York: Mouton de Gruyter.

Paivio, A., Yuille, J.C., Madigan, S. (1968). Concreteness, imagery, and meaningness values for 925 nouns. *Journal of Experimental Psychology Monograph.*

Hay, J. (2003). *Causes and consequences of word structure.* New York: Routledge.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (ed.), *Diversification processes in language: grammar: 47-55*. Hagen: Rottmann.

Reder, L.M., Anderson, J.R., Bjork, R.A. (1974). A semantic interpretation of encoding specificity. *Journal of Experimental Psychology 102, 648-656.*

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press. (Esp. p. 252-258)

Morpheme inventory and morpheme polysemy

Hypothesis

The greater the average polysemy of morphemes the smaller the inventory of morphemes in a language (Krott 2002: 77f).

Procedure

The hypothesis results from the Zipfian hypothesis of equilibrium between full polysemy and no polysemy. Both extremes are impossible. The hypothesis is not easy to test: data from at least 10 languages are needed in order to see the course

of the function. The dispersion will be most probably very great. One should begin with languages living in isolation up to languages in technical civilizations. One must have at his disposal complete morpheme/morph lists of the languages and a team of specialists.

References

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quatitativen und system-theoretischen Linguistik.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Prün, C., Steiner, P. (2005). Quantitative Morphologie: Eigenschaften der morphologischen Einheiten und Systeme. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 227-242*. Berlin/New York: Mouton de Gruyter.

Morphology vs. phonemics

Hypothesis

Since grammatical endings are usually short (they frequently developed from independent words) some of their phonemes were eliminated, usually vowels. Hence one can ask: *Is there an interrelation between (consonant) cluster forming and the extent of inflection (agglutination) in language?* (Skalička 1964).

Procedure

Set up a table of individual consonant clusters in texts from at least 10 languages from different families. Apply an association measure (e.g. Harary, Paper) to express the "clustering tendency". Measure the extent of inflection or agglutination (affix forming). Apply e.g. Greenberg´s/ Krupa´s indices and compare the results with respect to the languages.

1. Find the relationship between clustering and inflection/agglutination.
2. Define a new index of inflection/agglutination.
3. Try to find the relation of cluster forming with other language properties.

References

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

Harary, F., Paper, H.H. (1957). Toward a general calculus of phonemic distribution. *Language 33, 143-169.*

Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*

Skalička, V. (1964). Konsonantenkombination und linguistische Typologie. *Travaux linguistiques de Prague 1, 111-114.*

Phoneme inventory vs. morpheme length

Hypothesis
The greater the inventory of phonemes in a language, the shorter are its morphemes (cf. Hockett 1968: 93).

Procedure
Take at least three languages with very different phoneme inventory size and use ready sources (e.g. Karpilovs´ka 2002 for Ukrainian where one finds all roots of this language). Compute the mean morpheme length in these languages using random samples of about 500 morphemes. State whether they are equal or change with increase of inventory. Set up a hypothesis, add more languages and test the hypothesis.

References
Hockett, Ch. F. (1968[3]). *A course in modern linguistics*. New York: McMillan.
Karpilovs´ka, E.A. (2002). *Korenevij gnizdovij slovnik ukrainskoi movi.* Kiiv: Ukrains´ka enciklopedija.

Polysemy and compounding

Hypothesis
"The greater the polylexy of a word the more compounds it forms." (Rothe 1988).

Procedure
Test the hypothesis using a monolingual dictionary and sampling systematically about 1000 words. Define exactly polysemy and the concept of a compound for your language. For each selected word find the number of its meanings (given in the dictionary) and the number of compounds it forms. Show that the dependence *Number of Compounds = f*(*number of meanings*) increases monotonously. Find the appropriate function and test its adequateness.

References
Hammerl, R. (1990). Überprüfung einer Hypothese zur Kompositabildung (am polnischen Sprachmaterial). *Glottometrika 12, 73-83.*
Rothe, U. (1988). Polylexy and compounding. *Glottometrika 9, 121-134.*

Semantic classes

Problem
Does semantic classification of words conform to a rank-frequency distribution?

Procedure

A hypothesis in QL states: if a linguistic class is constructed "naturally", then its elements abide by a proper rank-frequency distribution of the Zipf type. Test this hypothesis on the data of Levickij and Lučak (2005, Table 7, p. 223, last two columns). The authors establish 20 verb classes and give their frequencies in English.

As a continuation, reorder the columns of this table according to the ranks in the last but one line of the table. Try to find a two-dimensional rank-frequency distribution for this class/tense classification. If not successful, find at least the correlation between the two classifications.

Reference

Levickij, V., Lučak, M. (2005). Category of tense and verb semantics in the English language. *Journal of Quantitative Linguistics 12(2-3), 212-238.*

Semantic diversification

Hypothesis

The individual meanings of any word occur with different frequencies. The frequencies arranged in decreasing order follow a proper rank-frequency distribution.

Procedure

The fact itself is known from many publications. The task here is to show the distribution of individual word classes. Take a sample of words consisting of say 5 nouns, 5 verbs, 5 adjectives, 5 prepositions etc. and obtain the frequencies of individual meanings of all these words. Set up empirical rank-frequency distributions and try to find appropriate models. Besides, show that prepositions are more diversified than nouns, etc. Show that languages differ drastically in semantic diversification, but abide by the given models. Devise new models starting from theoretical arguments.

Analyse the complete set of conjunctions in your language and evaluate the results.

References

Altmann, G. (1985). Semantische Diversifikation. *Folia Linguistica 19, 177-200.*

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 646-657*. Berlin/New York: de Gruyter.

Altmann, G., Best, K.-H., Kind, B. (1987). Eien Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8, 130-139.*

Beőthy, E., Altmann, G. (1984). Semantic diversification of Hungarian verbal prefixes. III. "föl-", "el-", "be-". *Glottometrika 7, 45-56.*

Rothe, U. (1986). Die Semantik des textuellen *et*. Frankfurt: Lang.
Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

Chapter 7

Typology

Entropy and synthetism

Problem
Does the entropy of word frequencies have something common with language synthetism?

Procedure
Compute the word-form frequency spectrum in several texts (the frequency f_x is the number of words in a text occurring exactly x times). Then compute the entropy according to the formula in "Repeat Rate and Entropy" (Chapter 8). Compute the mean of the entropies of all texts and compare the resulting value with the following table.

Mean entropies of word frequency spectra in 20 languages

Language	mean H: spectra
Hungarian	0.9577
Latin	1.2203
German	1.2980
Romanian	1.3252
Bulgarian	1.3279
Czech	1.3510
Russian	1.5145
Italian	1.5325
Tagalog	1.5721
Indonesian	1.5823
Slovenian	1.6344
Marathi	1.6532
Lakota	1.8002
Kannada	1.8683
English	2.2791
Rarotongan	2.6337
Samoan	2.7099
Maori	2.7696
Marquesan	2.8490
Hawaiian	2.8946

As can be seen in the table, the more analytic a language the greater the entropy of word frequency spectra. Where in the interval of the values is your result placed? Try to explain this phenomenon. Try to compute also the Repeat Rate for your texts and consult the reference. Compare your results with those in the problem "Synthetism in language" in this chapter.

Reference

Popescu, I.-I., Altmann, G. (2007). On diversity of word frequencies and language typology. *Göttinger Beiträge zur Sprachwissenschaft 14, 81-91*.

Homonymy and synonymy of affixes 1

Problem

If there is a declination system in a language (e.g. Latin, Slavic languages etc.), some of the declination affixes are homonymous (same form but different meanings/functions); other ones are synonymous (different forms for the same meaning/category).

Procedure

1. Try to express quantitatively the extent of homonymy and synonymy in a language.
2. Show that the homonyms are not represented with the uniform frequency (e.g. test for homogeneity).
3. Show that the synonyms within a category are not represented with the same frequency (e.g. test for homogeneity).
4. Set up the empirical frequency distribution of some categories and scrutinize the relation between case frequency and mean length of affixes.
5. See continuation in the next problem.

Reference

Skalička, V. (2005-2006). *Souborné dílo I-III* [*Collected Works I-III*]. Praha: Nakladatelství Karolinum.

Homonymy and synonymy of affixes 2

Hypothesis

Synonymous affixes follow a proper rank-frequency distribution.

Procedure

Set up a list of all affixes (both derivative and inflectional) of a language with the corresponding categories or class meanings, e.g. English *–s* expresses genitive of nouns, plural of nouns, third person singular of verbs; English *–ity* and other

affixes express abstractness, etc. Ignore the fact that the genitive itself can have a great number of different meanings; consider only the categories. The list should be prepared in form of a table with affixes in the first column and the categories (meanings) in the other ones. Mark the cells where the affix (row) has the given meaning (column). Add the frequency of individual affixes on the basis of a corpus or a frequency dictionary.

Arrange the rows according to the frequencies of the affixes. Now count the number of marks in each category (column) and obtain the number of affixes with which the given category is associated. Reorder the columns according to these numbers.

For each row and column find its rank-order distribution separately. Use a distribution with not more than two parameters because the rows and columns are short. Observe the behaviour of the parameters. Finally try to find a two-dimensional distribution for the whole table. Interpret the results. Consider the rank-frequency order as the only criterion of the "correctness" of your affix list and the ascription of categories. If you do not obtain satisfactory results, consult other grammatical descriptions of the language. On the other hand, deviation from your order can be a sign of beginning self-organisation (leaving the equilibrium) or a sign of the impact of self-regulation (restoring equilibrium). In every language there will be some "exceptions" showing the dynamics of language.

References
None.

Inflection in general

Problem
Devise different measures for the degree of inflectivity in language. Set up different measures for grammar (*langue*) and for text (*parole*). Try to compute the difference in inflectivity between written and spoken French using only texts. Compare individual cases and compute the development of inflection loss in spoken French. Compare Old English with modern English; Latin with Spanish; Old Russian with modern Russsian. Preliminarily, apply the Greenberg/Krupa indices but try also to define new one(s).

Hypothesis
"…the greater the number of different inflectional affixes a language possesses, the smaller proportionately will be the number of different roots occurring in the stream of speech compared to the number of the different words (which are really inflected roots) made up from these roots." (Zipf 1935: 252-253).

Procedure (as applied to Zipf´s hypothesis)

1. Define exactly the concept of inflection. With regard to Greenberg´s index, draw a large sample from a corpus, count the number of words and the number of inflected words. Their ratio gives a measure of inflectivity. Consider it a proportion which can be manipulated statistically. Compare several languages.
2. Try to obtain a function representing Zipf´s hypothesis. According to Zipf it should be $y = k/x^2$, a consequence of his theory of word distribution. Check its goodness or develop a new theory if necessary.

References

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press. (Esp. p. 252-258).

Morph length

Hypothesis

According to Skalička (2006: 988, 1054), one of the indicators of typological techniques is morph length. That is, "in languages with high degree of polysynthesis the morphs are short."

Procedure

Check this assumption on texts from different languages. First define the way of measuring morph length (in phonemes or syllables), decide whether a zero morpheme should be postulated; then transcribe the text morph(em)ically and obtain the distribution of lengths. According to Best (2005), the data should be distributed according to the hyper-Poisson distribution. Test this hypothesis and show the differences between the parameters in particular languages. Study several texts in the languages examined testing the distributions for homogeneity.

However, the shortest class can display a characteristic behaviour in individual languages. Consider the relative frequency of the smallest class ($x = 0$ or 1 depending on the way of measurement) as a characteristic of language and try to find another measurable feature associated with it.

References

Best, K.-H. (2005). Morphlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handboook: 255-260.* Berlin/New York: Mouton de Gruyter.

Skalička, V. (2006). *Souborné dílo III [Collected Works III]*. Praha: Nakladatelství Karolinum. [especially the articles: "Zum Problem des Donausprachbundes". *Ural-altaische Jahrbücher 40 (1-2), 1968, 1-9* and "K voprosu o tipologii". *Voprosy jazykoznanija 1966, 22-30*.]

Popescu´s typological indicator *a*

Problem

Popescu´s indicator *a* expresses the degree of synthetism of a language. Compare a language of your choice with the table below and the usual indices of synthetism.

Procedure

Compute the frequency of different words (not lemmas but word forms) in a text. Ascertain the *h*-point (cf. Textology) in the following way: (a) The *h*-point is that point at which *rank = frequency*. (b) If such a point cannot be found, apply the formula

$$C = \frac{1}{f_r - r},$$

where f_r is the frequency at rank r, and r is the given rank. C increases up to a point where there is a break; it becomes negative and increases again. Join the greatest positive C with the smallest negative C by a straight line. The intersection of this straight line with the x-axis is the *h*-point (Popescu et al 2008).

Consider the text length N and set up the indicator a as follows

$$a = \frac{N}{h^2}.$$

Analyze several texts (as many as possible) from the given language and compute the average *a*. Insert this value in the table below.

Compare your *average a* with that of the neighbouring languages in the table using the test

$$t = \frac{|\bar{a}_1 - \bar{a}_2|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s^2 = \frac{\sum_{i=1}^{n_1}(a_{i1} - \bar{a}_1)^2 + \sum_{i=1}^{n_1}(a_{i2} - \bar{a}_2)^2}{n_1 + n_2 - 2}$$

Mean values of quantity *a* in 20 languages (from Popescu et al. 2008)

Language	**Mean a**	**n**		**Language**	**Mean a**	**n**
Samoan	4.56	5		Italian	8.41	5
Rarotongan	5.02	5		Romanian	9.15	6
Hawaiian	5.37	6		Slovenian	9.19	5
Maori	5.53	5		Indonesian	9.58	5
Lakota	5.69	4		Russian	10.10	5
Marquesan	5.69	3		Czech	10.33	10
Tagalog	7.24	3		Marathi	11.82	50
English	7.65	13		Kannada	16.58	47
Bulgarian	7.81	10		Hungarian	18.02	5
German	8.39	17		Latin	19.56	6

and t has $n_1 + n_2 - 2$ degrees of freedom. The values of individual *a´*s can be taken from Popescu et al. (2008) (Table 3.1.1).

For another test for difference between two individual texts cf. Popescu et al. (2008).

References

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., A., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Root length and extent of derivation

Hypothesis

Looking at Skalička´s system above we can assume that if a language has short roots (on the average), it has ample derivations.

Procedure

First define the measurement of root length (e.g. in terms of phoneme number), then define the extent of derivation. One can use a Greenberg-Krupa index. The problem can be solved on the basis of a dictionary or a corpus. In order to obtain the empirical form of the dependence, several languages must be analysed or ready results can be taken from the typological literature. At a higher level the dependence should be derived theoretically. If it does not hold, restrict the ceteris

paribus condition and search for a third variable or try to find boundary conditions.

References

Skalička, V. (2005-2006). *Souborné dílo I-III [Collected Works I-III]*. Praha: Nakladatelství Karolinum.

Synthetism in language

Problem

Languages using many affixes and inflections are highly synthetic. Try to develop a measure of synthetism and apply it to several languages.

Procedure

Begin with defining such a measure in terms of non-root morphemes found in a corpus (only types). Then try to define it in terms of the distribution of words with 0,1,2,… affixes, which include prefixes, infixes, suffixes and circumfixes. Try to develop other measures of synthetism and finally check whether a high degree of synthetism is associated with average word length.

Begin with four definitions:

(1) *W*/*M* (*W = number of words, M = number of morphemes*).

(2) *R*/*M* (*R = number of root morphemes*)

(3) *S*/*W* (*S = number of sentences*)

(4) *L*/*V* (*L = number of lexemes/lemmas, V = number of word forms*).

Try to study whether they give the same value. Analyze several short texts in each language.

If possible, develop some statistical properties of these or new indices. Find other properties associated with synthetism. Test whether $L = aV^b$ (Tuldava 1995: 154).

References

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

Kelemen, J. (1970). Sprachtypologie und Sprachstatistik. In: Dezsö, L., Hajdú, P. (eds.), *Theoretical problems of typology and the Northern Eurasian languages: 53-63.* Amsterdam.

Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*

Slavíčková, E. (1968). Toward a typological evaluation of related languages. *Travaux linguistiques de Prague 3, 281-298.*

Tuldava, J. (1995). The ratio of word forms and lexemes in texts. In: Tuldava, J. (1995), *Methods in quantitative linguistics: 151-159.* Trier: WVT.

Vocalic language

Problem

There are different opinions concerning the number of vowels and consonants in a language. Try to give a clear definition of vocalicness.

Procedure

Consider several possible measures of "vocalicness". Test some of them in different languages using both inventories and corpora and try to find a relationship between one of the measures with some other properties of language, e.g. degree of inflection.

Reference

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Word length and agreement

Hypothesis

If there is much agreement in text, the words must be longer on the average (Skalička 2005-2006).

Procedure

This is one of the possible hypotheses from Skalička´s system. Agreement is usually represented by special (derivational or inflectional) affixes. These affixes lengthen the word, especially in languages with high agglutination.

1. Study 10–20 texts from one language, computing the average word length and the proportion of words having any kind of agreement. Count only explicit agreement; e.g. in German *zu diesen schönen Häusern* three words are joined by agreement, but in English *to these nice houses* only two words; in Indonesian (*kepada rumah-rumah bagus ini*) there is none. A word can have more instances of agreement at the same time. Define exactly the presence of such an instance. Then try to show whether there is a relation <*agreement, word length*>.
2. Perform the same procedure with the data from 10 different languages (not only Indo-European ones) and observe whether the above association is present. If the hypothesis is correct, try to set up a function. If there is no trend, try to search for further intermediate variables.

References

Skalička, V. (1966). Ein "typologisches Konstrukt". *Travaux linguistiques de Prague 2, 157-163.*

Skalička, V. (2005-2006). *Souborné dílo. I-III [Collected Works I-III]*. Praha : Nakladatelství Karolinum.

Word order and inflection

Hypothesis

"The more highly inflected a language is, the greater is the liberty which it may take in positional arrangement. The presence or absence of inflectional devices, and the degree of their usage, modifies the scope of syntactical arrangement" (Zipf 1935: 246).

Procedure

Devise a method for measuring the freedom of word order/word arrangement in the sentence. Apply one of the Greenberg or Krupa indices (or propose a new one) as measure of inflectionality. Try to express *Freedom of Word Order = f*(*degree of inflection*) in the form of a function. Test the adequateness of the function.

References

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*

Skalička, V. (1966). Ein "typologisches Konstrukt". *Travaux linguistiques de Prague 2, 157-163.*

Zipf, G.K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton Mifflin; 1968[2]: Cambridge, Mass.: The M.I.T. Press.

Chapter 8

General problems

Distributions

Problem

Find empirically appropriate distributions for different entities that can be found in a text.

Procedure

R = rank-frequency distribution; *F = frequency spectrum* or *frequency of frequencies*

Consider the following entities: phonemes (*R*), letters (*R*), graphemes (*R*), syllables (*R*), syllable length (*F*), words (*F*), word length (*F*), clause length (*F*), sentence length (*F*), word classes (*R*), number of meanings of words (taken from a dictionary) (*F*). Find the "best" distributions, study their forms and parameters and, if possible, try to find a common origin for all *R*'s and for all *F*'s. Characterize the text as a vector of distribution properties.

References

Grzybek, P. (2006). History and methodology of word length studies. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 15-90.* Dordrecht: Springer.

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 329-337*. Dordrecht: Springer.

Entropy and inventory size

Problem

Since linguistic entities have very characteristic rank-frequency distributions or frequency spectra, it must hold that the entropy of these distributions directly depends on the inventory size. In phonemics (phoneme or letter distribution) this dependency has already been shown.

Procedure

Try to show that this hypothesis has a broader scope. Prepare data about phoneme/letter frequencies from languages with different inventory sizes and from individual texts of different lengths, compute the word frequency distributions, the inventory sizes (type or token inventory), and the entropies and scrutinize the dependence of the entropy on the size of inventory. In the first step, restrict your

analysis to a purely empirical investigation. In the second step, try to find a theoretical distribution fitting to the data and derive the entropy theoretically. Then try to show that empirical entropies follow the theoretical function.

Reference

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Fitting a distribution

Problem

Explorative fitting of a distribution to data is merely the first step, not the last.

Procedure

In the appendix to the article by Gale and Sampson (1995: 237) there is a complete frequency spectrum of canonical forms.

1. Use a fitting program to compute all distributions which are empirically adequate for this data.
2. Construct the empirical rank-frequency distribution of this data and find the rank-frequency distribution.
3. Explain why the concept "unseen species" in linguistics is wrong.

If an adequate formula for the above data is obtained, try to derive it from theoretical assumptions. Show that there are always several "good" distributions for the given set of data. The preference must be given to the one that can be derived from a theory.

Reference

Gale, W.A., Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics 2(3), 217-227.*

Setting up hypotheses by means of factor analysis

Problem

Factor analysis can help to obtain sets of properties which are in some way associated with one another. These associations within a factor are a source for hypotheses. Try to find such hypotheses.

Procedure

Consider first J. Tuldava´s article (1995a: 84) where one can find some associated properties. Since the data are not presented, perform text analyses in your language measuring 12 properties as given by Tuldava. Then try to set up hypotheses concerning the mutual dependences of these properties. For each factor try

to draw a Köhlerian self-regulation cycle for each factor, consider each property in its log-linear form and set up the formulas. First combine pairs of properties, then three, four etc. at once, i.e. capture the factor as a multidimensional structure. Finally compare the results in several languages, i.e. check whether your control cycle is valid. Tuldava used the following properties: number of nouns, adjectives, pronouns, verbs, adverbs, pre-/postpositions, conjunctions, content words, concentration of function words, entropy, frequent word forms, and rare words. Consider more properties.

References

Tuldava, J. (1995a). An attempt at quantitative analysis of the style of fiction. In: Tuldava, J. (1995), *Methods in quantitative linguistics: 73-92*. Trier: WVT.

Tuldava, J. (1995b). A comparison of subjective and objective characteristics of style. In: Tuldava, J., *Methods in quantitative linguistics: 93-108*. Trier: WVT.

Iconicity

Problem

There is an immense number of works on icon, index and symbol. In all languages one finds enormous number of all these signs in all languages; there are books describing individual languages and the iconicity in them. Unfortunately, there is no method for measuring the extent of iconicity, indexality and symbolicity of individual signs. Hence, it would be of utmost importance for for semiotics to develop a quantification of these properties, which would make it easier to find their relation to other properties of language.

There are words of iconic origin which today have the status of a symbol. But the way from icon to symbol is not abrupt; it has a history in which the extent of iconicity decreases and that of symbolicity increases. Try to establish such a method.

References

None.

Index formation

Problem

Index formation is a complex procedure. Setting up a ratio of some quantities does not suffice. The properties of the index must be given.

Procedure

Mikk (1997) set up an *index of word class complicatedness* used in text compre-

hension studies, defined as *WCC = (N + Adj)/(V + Adj)*. Try to interpret this index and find the range of *WCC*. If you cannot find it, transform the index in another one which varies within the interval <0,1> and interpret the new index. Find the expected value and the variance of the new index and set up an asymptotic test enabling you to compare two texts.

Alternatively, Tuldava and Villup (1976: 94) defined the Index of Substantivity = *N/V*. Perform the same analysis as for *WCC*.

Cf. also "Ratios" in Chapter 4.

References

Altmann, G. (1978). Zur Anwendung der Quotiente in der Textanalyse. *Glottometrika 1, 91-106.*

Mikk, J. (1997). Parts of speech in predicting reading comprehension. *Journal of Quantitative Linguistics 4(1-3), 156-163.*

Tuldava, J., Villup, A. (1976). Sõnaliikide sagedusest ilukirjandusproosa autorikõnes. *Töid keelestatistika alalt 1, 61-102* (with English summary).

Menzerath´s law

Problem

According to Menzerath´s law, the following statement holds: the longer a construct, the shorter its components.

Procedure

Check this hypothesis on data from a language in which it has not been studied as yet. Consider (a) sentence length (measured in clauses) vs. clause length (measured in words); (b) word length (measured in syllables) vs. syllable length (measured in phonemes); (c) word length (measured in syllables) vs. morph length (measured in phonemes); (d) word length (measured in syllables) vs. syllable duration (measured in milliseconds).

Try to analyse a strongly agglutinating language. If the results are not corroborative, explain why.

Study the relationship sentence length vs. word length, known as Arens´ law and explain the results.

References

Cramer, I.M. (2005). Das Menzerathsche Gesetz. In Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 659-688*. Berlin/New York: Mouton de Gruyter.

Grzybek, P., Stadlober, E. (2007). Do we have problems with the Arens´ law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 205-217.* Berlin/New York: Mouton de Gruyter.

Meyer, P. (2007). Two semi-mathematical asides on Menzerath-Altmann´s law. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 448-460*. Berlin/New York: Mouton de Gruyter.

Naranan-Balasubrahmanyan distribution

Problem

Fit the Naranan-Balasubrahmanyan distribution to rank-frequencies of phonemes/letters. Apply it also to rank-frequencies of words. Show that the model can be derived from the Unified Theory.

Procedure

The Naranan-Balsubrahmanyan distribution is defined as $P_x = Ce^{-a/x}x^{-b}$, x = 1,2,3,… where a and b are parameters and C is the normalizing constant. Draw a large sample of phonemes or letters (graphemes) from a text (or from available literature) and try to fit the above distribution to the empirical rank-frequency distribution. Derive some simple estimators and test the fit by means of the chi-square criterion.

References

Krylov, Ju.K. (1987). Stacionarnaja model´ porožděnija svjaznogo teksta. *Acta et Commentationes Universitatis Tartuensis 774, 81-102.*

Balasubrahmanyan, V.K., Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics 3(3), 177-228*.

Naranan, S, Balasubrahmanyan, V.K. (2007). Statistical analogs in DNA sequences and Tamil language texts: rank frequency distribution of symbols and their application to evolutionary genetics and historical linguistics. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 484-497*. Berlin/New York: Mouton de Gruyter.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics 3(1), 38-50.*

Wimmer, G., Altmann, G. (2006). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the science of language. Word length studies and related issues: 93-117*. Boston: Kluwer.

Ord´s criterion

Problem

Take a group of texts, compute the distribution of a certain variable and try to characterize them using Ord´s criterion.

Procedure

If the frequencies are obtained, compute the first three moments

$$m_1' = \frac{1}{N} \sum_{x=x_{\min}}^{x_{\max}} x f_x , \qquad \text{(mean, average)}$$

$$m_2 = \frac{1}{N} \sum_{x=x_{\min}}^{x_{\max}} (x - m_1')^2 f_x \qquad \text{(variance, second central moment)}$$

$$m_3 = \frac{1}{N} \sum_{x=x_{\min}}^{x_{\max}} (x - m_1')^3 f_x \qquad \text{(asymmetry, third central moment)}$$

and set up the indicators:

$$I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2}.$$

Then plot the points <*I*, *S*> of individual texts into the Cartesian coordinate system, in which you can see the position and distance between texts.

The plot is a kind of elementary classification which enables us to set up hypotheses about the status and development of the given property. Try to perform *sentence length* investigations for different genres and for the same genre in a historical perspective.

Expand the binomials in the central moments and simplify them. Consider 10 poetic and 10 scientific texts in your language. Compute the distribution of sentence lengths in each of them. Then use Ord´s criterion to display the difference between these two genres.

If you want to compare word frequency distributions in *different languages*, state that there are differences in their location: they lie on different straight lines.

References

Best, K.H. (2005), Wortlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 260-273*. Berlin/New York: Muton de Gruyter.

Oakes, M.P. (2007). Ord´s criterion with word length spectra for the discrimination of texts, music and computer programs. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 508-519*. Berlin/New York: Mouton de Gruyter.

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Repeat rate and entropy

Problem

The *repeat rate* is defined as

$$R = \sum_x p_x^2 = \frac{1}{N^2} \sum f_x^2$$

and *entropy* is defined as

$$H = -\sum_x p_x \, ld \; p_x \; = \; ld \; N - \frac{1}{N} \sum_x f_x \, ld \; f_x \; .$$

where N is the sample size, p_x is the probability of the given entity, f_x is the absolute frequency of the entity, and *ld* is the logarithm with base 2. Examine whether these indicators depend on the size of inventory of entities.

Procedure

Perform phoneme counts (in different languages) and word counts. For word counts, set up the rank-frequency and the spectrum distributions. Compute the above indicators and study their relation to the size of phoneme inventory and to the vocabulary size of individual texts. Try to find a dependence.

References

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Sample size

Problem

Sample size is a crucial problem in all quantitative linguistic investigations. For every test there must be a sufficient number of sampled cases, otherwise the test would be weak. Frequency distributions must have a sufficiently large number of cases in each *x*-class. The points for a frequency curve are usually means of a sufficiently large number of cases. For phoneme frequencies Kubáček´s method is to be used. Try to find an inductive method yielding a adequate sample size which can be used generally for any linguistic entity.

Procedure

Let us exemplify a method for phoneme frequencies. If you do not know how many phonemes are to be sampled for a phoneme count, try the following empirical method. Extract 1000 phonemes (letters) from a text and write the frequencies in a column. Extract the next 1000 ones and add them to the previous frequencies using a new column. Repeat this until 10000 phonemes have been counted. Then compute the relative frequencies for each column in this table. Compute the sum of absolute differences between each pair of two neighbouring columns and observe the decrease of this sum. Try to fit the function $y = a10^{-bx}$ (iteratively) to this decreasing series, where x is the order number of the column, and find the point x at which $y < \delta$, setting $\delta = 1/10K$ (K = number of phonemes in the inventory). Using this x, set the necessary sample size at $N = 1000(x+1)$.

Generalize the method and use it for determining the required sample size of syllables, morphs, even words. Compare the results of your computation with the classical methods of statistics.

References

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Kubáček, L. (1994). Confidence limits for proportions of linguistic entities. *Journal of Quantitative Linguistics 1(1), 56-61.*

The problem of infinity

Problem

Try to solve or discuss the problem of infinity in language. Some mathematical models show that a linguistic entity can be infinite, e.g. the number of sentences in a language, the size of the word stock, the length of a sentence, even the length of a word etc. We know that there are some limits in language, e.g. to the number of phonemes, the number of syllables, and the number of words an individual can store in his memory (take polyglots and specialists into account!) – yet, is there a limit to the number of meanings? If so, try to find a foundation. If not, try to explain this fact.

Procedure

Start from Zipf´s unification and diversification forces and consider the consequences of the hypothetical existence of a word with infinitely many meanings. In the case of sentence length, take into account Köhler´s interpretation of Menzerath´s law. For phoneme inventories, consider the problem of an effectual distinctivity. For the number of different syllables start from the necessity of adequate redundancy, etc. Generalize the problem taking into account the Köhlerian “requirements”

References

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 760-774*. Berlin/New York: Mouton de Gruyter.

Köhler, R. (1989). Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. Chapter 7 in: Altmann, G., Schwibbe, M.H., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen: 108-112*. Hildesheim: Olms .

Zipf, G.K. (1935). *The psycho-biology of language. An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G.K. (1949). *Human behaviour and the principle of least effort.* Reading, Mass.: Addison-Wesley.

Tightness/Cohesion

Hypothesis

"…the more often two elements occur in sequence the tighter will be their constituent structure" (Bybee, Hopper 2001: 14; Bybee, Scheibmann 1999). Clear examples are cases in which two words have fused together because of their frequent co-occurrence and now behave essentially as single words, e.g. *want to > wanna, going to > gonna, I am > I´m, can not > can´t, do not > don´t, I don´t know > I dunno, would have >would´ve.*

"Pairs of words that are frequently used together, whatever their apparent constituency and status as lexical or grammatical (*don´t you, told you, that you, last year*), are more likely to show effects of coarticulation than words that are used together less often" (Bybee, Hopper 2001: 7).

Procedure

Try to devise a measure for tightness/cohesion of constituents. Then test pairs that occur significantly frequently together in a corpus. Try to corroborate the hypothesis *Tightness = f(Frequency of co-occurrence*); collect 100 pairs and test the hypothesis. Then try to show that the more frequently a pair occurs, the greater its tightness, i.e. *Cohesion = f(Frequency of occurrence).*

References

Boyland, J.T. (1996). *Morphosyntactic change in progress: a psycholinguistic approach.* Diss: Linguistics Department, University of California.

Bybee, J. (2000). Lexicalization of sound change and alternating environment. In: Broe, M., Pierrehumbert, J. (eds.), *Laboratory V: Language acquisition and the lexicon: 250-268*. Cambridge: Cambridge University Press.

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P. (eds.), *Frequency and the*

emergence of linguistic structure: 1-24. Amsterdam/Philadelphia: Benjamins.
Bybee, J., Scheibman, J. (1999). The effect of usage on degree of constituency: the reduction of don´t in American English. *Linguistics 37, 575-596.*
Fan, F., Altmann, G. (2007). Measuring the cohesion of compounds. In: Kaliuščenko, V., Köhler , R., Levickij, V. (eds), *Problems of typological and quantitative lexicology: 177-189. Č*ernivcy: Ruta.
Krug, M. (1998). String frequency: a cognitive motivating factor in coalescence, language processing and linguistic change. *Journal of English Linguistics 26, 286-320.*
Krug, M. (2001). Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 310-335*. Amsterdam/Philadelphia: Benjamins.

Zipf´s and Zipf-Mandelbrot´s law

Problem

Review the history of Zipf´s and Zipf-Mandelbrot´s law.

Procedure

Begin with collecting literature. Prepare a comprehensive bibliography of these laws – if possible at all. Do not omit Russian works. Then review all formulas that were developed for this purpose. Review all derivations leading to the individual formulas. Try to classify them into families with different background. Separate linguistic argumentations from general argumentations but show them all. Refer to the use of this law in other sciences. Zipf´s law is the famous power law, Mandelbrot´s version a generalisation. Show all generalisations you can find in the literature. Show also divergent cases.

References

Glottometrics 3-5 (2002) [= a collection of articles to honor G.K. Zipf]
Guiter, H., Arapov, M.V. (eds.) (1982). *Studies on Zipf´s law.* Bochum: Brockmeyer.
http://www.nslij-genetics.org/wli/zipf

Chapter 9

Research projects

Frumkina´s law (Word occurrence in passages)

Problem

Divide a long text into passages of constant length *n* (e.g. 50, 100, 150, 200,… words). Then choose a (not too rare) word and count how many times this word occurs in the individual passages, i.e. compute the number of passages containing the given word that occurs $x = 0,1,2,\ldots$ times. The "distribution of word occurrence" in passages will turn out to be the negative hypergeometric distribution or one of its limiting cases (binomial, negative binomial, geometric, Poisson d.).

Procedure

Use the FITTER or another appropriate software to find the distribution.
The problem has several aspects:

1. If the length of the passages increases, either the parameters of the distribution change or the distribution converges to a limiting form. Study the problem for different words – perhaps check all word classes and show whether the length of the passage has an influence on a parameter or whether different (limiting) changes occur in different word classes.
2. Fix the passage length and compute the distribution for many words of the same word class. Obtain the frequency of the given word in the whole text and try to set up a relation between the (relative) frequency of the word and one of the parameters of the distribution.
3. Try to draw conclusions from the form of the distribution (or from the values of the parameters) on the word class of the given word could belong.
4. State the empirical conditions under which the basic negative hypergeometric distribution converges to its limiting cases (which words, which word classes, what frequency, what length of words, etc.).
5. Do the following:
   (a) draw conclusions on the semantic relevance of the word in a text from the kind of distribution (or its parameters);
   (b) draw conclusions on the psychic/emotional state of the given author from the deviant form of the distribution of some words.
   (c) Try to find the differences between languages, genres, styles, authors concerning the distribution of "word occurrence" in passages.

   This problem is a theme for a research project involving a team of linguists, psychologists and programmers.

References

Altmann, G. (1988). *Wiederholungen in Texten.* Bochum, Brockmeyer.

Altmann, G., Burdinski, V. (1982). Towards a law of word repetitions in text-blocks. *Glottometrika 4, 147-167.*

Bektaev, K.B., Lukjanenkov, K.F. (1971). O zakonach raspredelenija edinic pis'-mennoj reči. In: Piotrowski, R.G. (ed.), *Statistika reči i avtomatičeskij analiz teksta: 47-112.* Leningrad: Nauka.

Best, K.-H. (2001/2003). *Quantitative Linguistik. Eine Annäherung.* 2., überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2005). Sprachliche Einheiten in Textblöcken. *Glottometrics 9, 1-12.*

Billmeier, G. (1968). Über die Signifikanz von Auswahltexten. Untersuchung auf der Grundlage von Zeitungstexten. In: Moser, Hugo u.a. (Hrsg.), *Forschungsberichte des Instituts für deutsche Sprache 2, 126-171.*

Brainerd, B. (1972a). Article use as an indirect indicator of style among English-language authors. In: Jäger, S. (ed.), *Linguistik und Statistik: 11-32.* Braunschweig, Vieweg.

Frumkina, R.M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T.N. (ed.), *Strukturno-tipologičeskie issledovanija: 124-133.* Moskva: ANSSSR.

Herdan, G. (1956). *Language as choice and chance.* Groningen: Nordhoff.

Knauer, K. (1955). Grundfragen einer mathematischen Stilistik. *Forschungen und Fortschritte 29, 140-149.*

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In Uhlířova, L., Wimmer, G., Altmann, G., Köhler, R. (Eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Ludek Hřebíček: 136-148.* Trier: WVT.

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse*. Hamburg: Kovač.

Maškina, L.E. (1968). *O statističeskich metodach issledovanija leksiko-grammatičeskoj distribucii.* Minsk, Diss.

Morton, A.Q., Levison, M. (1966). Some indicators of authorship in Greek prose. In: Leed, J. (ed.), *The computer and literary style: 141-179.* Kent, Ohio: Kent State UP.

Mosteller, F., Wallace, D.L. (1964). *Inference and disputed authorship: The Federalist*. Reading, Mass, Addison-Wesley.

Muller, Ch. (1972). *Einführung in die Sprachstatistik.* München: Hueber.

Paškovskij, V.E., Srebrjanskaja, I.I. (1971). Statističeskie ocenki pis'mennoj reči bol'nych šizofreniej. In: *Inženernaja lingvistika.* Leningrad.

Piotrowski, R.G. (1984). *Text – Computer – Mensch.* Bochum: Brockmeyer.

Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. (1985). *Mathematische Linguistik.* Bochum: Brockmeyer.

Suhren, S. (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen „De lütte Prinz“*. Staatsexamensarbeit, Göttingen.

Zwirner, E., Ezawa, K. (Hrsg.) (1966, 1968, 1969). *Phonometrie, Erster-Dritter Teil.* Basel/ New York: Karger.

Zwirner, E., Zwirner, K. (1935). Lauthäufigkeit und Zufallsgesetz. *Forschungen und Fortschritte 11, Nr. 4*: 43-45. (Also in: Zwirner & Ezawa (Hrsg.), Dritter Teil: 55-59.)

Zwirner, E., Zwirner, K. (1938). Lauthäufigkeit und Sprachvergleichung. *Monatsschrift für höhere Schulen 37: 246-253.* (Also in: Zwirner & Ezawa (Hrsg.), Dritter Teil, 68-74.)

Skalička´s typological system

Hypothesis

The system takes into account the following language properties:

1. Root length
2. Word length
3. Distinguishing word classes
4. Word complexity
5. Conversion
6. Number of affixes
7. Extent of derivation
8. Number of synsemantics
9. Affix length
10. Forming of compounds
11. Number of preposition and postpositions
12. Homonymy of affixes
13. Synonymy of affixes
14. Fixedness of word order
15. Inflection
16. Internal inflexion
17. Number of clauses
18. Number of endings in the word
19. Morpheme discontinuity
20. Existence of infinitives, participles, verbal nouns
21. Number of vowels and consonants
22. Extent of agreement
23. Differentiation of root and auxiliary elements
24. Differentiation of inflection and derivation
25. Sentence markedness
26. Vowel harmony
27. Suppletivism
28. Article formation
29. Possessivity
30. Extent of declination

All these properties are interrelated. Try to corroborate the interrelations both empirically and theoretically.

Procedure

Quantify at least some of the above properties. Measure them in texts or a corpus. For a property with many possible values, one language is sufficient to partially corroborate a hypothesis; for binary features, at least 10 languages are necessary. The reading of some works by Skalička is necessary, e.g. Skalička (1966). Do not forget that before quantifying, measuring and testing, a clear hypothesis must be set up. Begin with any two properties. If you solve the problem involving more than two properties, draw a relationship diagram and extend it stepwise (cf. Köhler 1986). The very extensive register of properties in Skalička´s collected works (2005-2006) can be used to find his respective articles written in German, English, French, Russian or Hungarian.

References

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26, 178-194.*

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Skalička, V. (1964a), Konsonantenkombinationen und linguistische Typologie. *Travaux linguistiques de Prague 1, 111-114.*

Skalička, V. (1964b). Typologie a konfrontační lingvistika. *Československá rusistika 7, 1962-1964, 210-212*

Skalička, V. (1966). Ein "typologisches Konstrukt". *Travaux linguistiques de Prague 2, 157-163.*

Skalička, V. (2005-2006). *Souborné dílo. I-III [Collected Works I-III]*. Praha : Nakladatelství Karolinum**.**

Synonymy

Problem

Synonymy is part of several control cycles whose other elements are various word properties. Each point below shows a potential word property, i.e. a potential hypothesis concerning synonymy. The properties are as follows:

1. Length in terms of phoneme, syllable, morpheme and mora numbers.
2. Frequency of the word in a corpus.
3. Polysemy as the number of meanings in a dictionary (or senses in WordNet).
4. Polytextualiy as the number of texts or even neighbourhoods of a word (contexts) (e.g. collocations, see also point 11).
5. Morphological status: simple, derived, reduplicated, compound (the simpler the more synonyms).
6. Class attribution: the greater the number of word classes to which a word belongs, the more synonyms it has (cf. direct conversion in WordNet).

7. Morphological productivity: the more derivates, compounds, reduplications are possible, the more synonyms the word has (a data base for German is on the Internet).
8. Age of the word in terms of centuries from its first appearance in writing. The older a word the more synonyms (it depends on the word class). Not easy to ascertain.
9. Provenance: number of the historical stations of the words, e.g. from Latin to French to Russsian; from Arabic through English to German etc. The longer the way, the more synonyms.
10. Verb valency: number of actants with which a verb can be joined. Valence increases polysemy which in turn increases synonymy.
11. Special case for some languages: number of prepositions with which the verb can form a phrase (*get in*, *get out*, *get around*, *get off*, *get out of*, *get from under*, *get through*,…). The more prepositional phrases there are, the more synonyms, because many prepositional/postpositional phrases can be replaced by a unique word.
12. Number of grammatical categories a word has (case, number, tense,…). The more categories the greater the synonymy. The categories allow a word to occur in different contexts, ergo increase of polytextualiy → polysemy.
13. Emotionality vs. Notionality (e.g. mother vs. bank). Must be doen on test subjects. A scale must be devised. Hypothesis not known. (It can be in both directions but it is assumed that the greater the emotionality, the more synonyms there are, e.g. *you swine*!)
14. Pollyanna: the position of the word on the good-bad scale. (Test subjects)
15. Abstractness vs. Concreteness, e.g. beauty vs. revolver. (A special scaling procedure must be devised.)
16. Specificity vs. Generality (e.g. revolver vs. instrument). Measurement according to "definition chains".
17. Dogmatism of the word (e.g. must vs. can; all vs. some; always vs. sometimes).
18. Number of associations (connotative potency). Use dictionaries of associations. The more associations, the more synonyms.
19. Number of possible functions in the sentence (e.g. the word can be subject, predicate, object, complement,…).
20. Diatopic variation of the word: the more forms there are in the dialects, the more synonyms are formed. (Can be measured as the number of competitors in a dialect atlas.)
21. Discourse properties: does a word indicate an association with a social group?
22. The degree of standardization (high, middle class, city dialect, slang…).
23. Diversification: in how many word classes can a word be transferred by means of affixes (not by conversion!), e.g. German: *Bild* (noun), *bildhaft* (adjective/adverb), *bilden* (verb).
24. Originality: (a) Genuine word, (b) calque, (c) borrowing.

25. Number of fixed phrases a word forms (special case of polytextuality, say polytextuality I).

Each property supports or disfavours synonymy forming (or it is to be eliminated if it behaves neutrally). Hypotheses should be set up and tested. Stepwise cycles are to be formed; finally a complex synergetic control cycle should be constructed.

Procedure

Collect randomly about 500 words from a synonymy dictionary and for each of them count the number of its synonyms. Then study one of the above mentioned word properties and try to show that *Synonymy = f*(*Property*). Test step by step all the hypotheses; from case to case you will have to devise a measurement procedure for the individual properties. If there is a dependency, represent this finding in a diagram where you connect synonymy to the given property with an arch. Continue until all properties are checked for an interrelation with synonymy. Then try to find dependencies between the individual properties.

Continuation

Synonymy can come into existence in different ways:

1. Under special circumstances a given word does not contain the needed sense and is replaced by another one (e.g. irony, sentiment, slang,… e.g. Latin *caput, testa*).
2. The special circumstance can be represented by the environment in which it occurs (the rest of the sentence). This case connects the synonymy to polytextuality. Every environment changes slightly the meaning of the word. In order to specify the intended meaning, one chooses a more adequate word.
3. Every word has the tendency to increase its polysemy, but some of the meanings are dropped and expressed by other words because of the necessity of specification at the given occasion.
4. Find several other motives leading to the rise of synonymy.

References

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Wimmer, G., Altmann, G. (2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99. Zborník na počest Kláry Buzássyovej: 218-225*. Bratislava: Veda.

Ziegler, A. (2001). Zum Gesetz der Synonymie. Modellanpassungen im Deutschen und Englischen. In: Slavomír Ondrejovič, Matej Považaj (Hrsg.): *Lexicographica '99. Zborník na počest Kláry Buzássyovej: 230–236*. Bratislava: Veda.

Ziegler, A., Altmann, G. (2001). Beziehung zwischen Synonymie und Polysemie. In: Slavomír Ondrejovič, Matej Považaj (Hrsg.): *Lexicographica '99. Zborník na počest Kláry Buzássyovej: 226–229*. Bratislava: Veda.

Word frequency and collateral properties

Hypothesis

Many properties of the word are associated with its frequency.

Procedure

This hypothesis is very voluminous and can be tested only stepwise. Collateral properties are those that can be established for one and the same linguistic unit, e.g. the word. Then all its properties taken into account are operationalized (quantified). Random samples of words are taken from a dictionary and their properties are measured. The frequency of the given words is obtained and its relationship to the properties measured can be studied.

A "reversed" way may be taken: first count the frequencies of all lemmas in a text. Then consider some properties of these lemmas and test their association with frequency.

Some of the properties can be measured only on the nominal scale, nevertheless try to find a way to establish an ordinal scale (rank them). To facilitate the solution to this problem, we repeat the list of 27 word properties taken from Popescu et al. (2008). The list is in no way complete; in the course of research new properties can be established (see the list of properties in "Synonymy").

1. Length: measured in terms of phoneme, letter, syllable, mora or morpheme numbers. Sometimes one calls this property material complexity.
2. Polysemy: number of meanings in a dictionary.
3. Morphological status: simple word, reduplicated word, derivation, compound.
4. Class membership: the number of word classes to which it belongs, e.g. by conversion (*the hand*, *to hand*).
5. Polytextuality: the number of texts in which it occurs or the number of contexts (direct neighbours in text (collocation))
6. Productivity: the number of derivatives, compounds, reduplications that can be formed with the word. Data can be obtained on the Internet.
7. Age: the number of years or centuries from the first appearance of the word in texts
8. Provenance: through how many languages did it come into the language under study.
9. Valency with verbs: the number of cases or prepositions with which it can co-occur.

10. The number of its grammatical categories: case, number, gender, tense, person, mode, etc. or the number of affixes it can combine with (e.g. not all verbs can be combined with all prefixes).
11. Degree of emotionality vs. notionality. Compare for example the emotionality of the words "mother" and "pencil".
12. Pollyanna: the degree of the word on the "good – bad" scale.
13. Degree of abstractness vs. concreteness of the word, e.g. "beauty" vs. "pencil".
14. Specificity vs. generality, e.g. "pencil" vs. "instrument".
15. Degree of dogmatism, e.g. "can" vs. "must", "all" vs. "some", "always" vs. "sometimes".
16. Number of associations (= connotative potential) that can be built upon hearing or seeing a word. There are dictionaries of word associations.
17. Synonymy: number of synonyms in a dictionary.
18. Number of different functions in sentence, e.g. a word can be subject, object, predicate etc.
19. Diatopic variation: in how many places in a dialect atlas can the word be found?
20. Dialectal competition: how many competitors of the word are there in a dialect atlas?
21. Discourse properties: in what degree does a word indicate the attribution to a social group?
22. The degree of standardization: standard language, social idiolect, argot etc.
23. Diversity: in how many word classes can the word enter by way of derivation, e.g. German *Bild* (N) -> *bildhaft* (Adj), *bilden* (V), *bildlich* (Adj, Adv).
24. Originality: genuine, borrowing, calque, folk etymology, substrate, etc.
25. Phraseology: in how many idioms can the word be found?
26. Degree of verb activity, e.g. *sleep* vs. *run*.
27. Degree of expression of a property by an adjective, e.g. *nice*, *pretty*, *beautiful*.

Try to establish more properties and find their relations with frequency.

References

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Author Index

Subject Index