Studies
in Quantitative Linguistics
4

Reinhard Köhler
Gabriel Altmann

Problems in Quantitative Linguistics 2

RAM - Verlag

Problems
in
Quantitative Linguistics
2

by

Reinhard Köhler
Gabriel Altmann

2009
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Ján Mačutek (jmacutek@yahoo.com)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.

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

Preface

> *„It is not just that research begins with problems:*
> *research consists in dealing with problems all the way long."*
> (Mario Bunge, Philosophy of science. Vol. 1: From problem to theory. New Brunswick, London: Transaction Publishers, 2007, p. 187)

Finding a scientific problem is the first task of a young scientist. Solving it is the next one. A solution, however, does not finish a problem; on the contrary, every solution opens up a series of new problems. Thus, from time to time it would be useful for every scientific discipline to resume the topical problems, show some new ones and shed light on other aspects of old problems.

We present a collection of problems in the field of quantitative linguistics – as far as it is possible to find Ariadne´s thread in the jungle of its differently developed sub-disciplines. The whole field consists of *membra disiecta* and we try without too much violence to draw the reader´s attention to the way of unification, where theory building may begin. Today, it is not easy to imagine that in an empirical science a theory might arise without at least elementary quantification. Though in the problems presented here there is still a lot of qualitative work to be done, we try to convince the reader to form quantitative concepts, to strive for elementary quantitative solutions, to link some problems with some existing theories or to open a new field of research.

In the first volume of this series the authors presented problems concerning phonemics, script, grammar, lexicology, textology, semantics, synergetics, psycholinguistics, typology, different general problems and the relations of length and frequency to other properties. In the present volume, most of the above-mentioned domains are treated, too, but besides, a number of problems concerning pragmatics, proverbs, drama, philosophy of science, motifs, dialectology etc. are added.

If the reader decides to solve one of the problems, it is recommended to look first in "Problems Vol. 1" where a more elementary, preparatory problem concerning the same domain may be presented. If a problem has been successfully solved, one should always try to generalize it, to test the result on data from several languages or texts, to seek deviations, outliers, to enrich it with subsidiary conditions and to systematize it, i.e. to embed it in a more general framework from which it can be derived.

If one meets "hard" problems, the first step may be purely inductive, e.g. fitting a simple function to data mechanically, but in the next step, the tentatively tested function should be substantiated as to the question "why should this function be chosen?" which is nearer to a future explanation than a verbal description of the discovered phenomenon.

The problems presented here vary from classroom exercises in quantitative linguistics over take-off platforms for publications to themes for research projects.

Readers are invited to report on publications which departed from a problem in this collection or in the first volume to the editor of the *Journal of Quantitative Linguistics* (http://www.ldv.uni-trier.de/index.php?koehler) or the editor of *Glottometrics* (www.gabrielaltmann.de). Solutions to one of the problems may also be submitted for publication in one of these journals.

Readers are also invited to contribute more new problems by sending a description to one of the above-given addresses.

R.K., G.A

Contents

1. Phonology and script

1.1. Zipf's assimilation

Hypothesis

"...every assimilation points to a weakening or instability of the assimilated sound, and this weakening or instability is caused primarily by the excessive relative frequency of the assimilated sound" (Zipf 1935/68: 109). Test the hypothesis.

Procedure

Collect all phonological assimilations in a language. You may consult a textbook of phonemics (or your own knowledge) and a list of relative frequencies of sounds in the given language. Which of the following statements corresponds with your result:

(a) The hypothesis is true,
(b) on the contrary, assimilated sounds may have a rather low frequency,
(c) the sound which evoked the assimilation is very frequent,
(d) both sounds (the assimilated and the assimilating) are relatively rare.

If (a) does not hold, generalize the hypothesis and check it in several languages. Alternatively, find the conditions under which the hypothesis is true in the given language and modify it using these conditions.

If the hypothesis is true, find at least an empirical formula which can express this relationship. Systematize the hypothesis embedding it in a control cycle or show that it is a consequence of more general mechanisms.

Reference

Zipf, G.K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge, Mass: MIT.

1.2. Zipf's accent problem

Hypothesis

"The most striking feature of sentence accent is this: words which occur most frequently are generally not preferred for accentuation." (Zipf 1935/68: 131). Test the hypothesis.

Procedure

Zipf studied English and German to demonstrate this phenomenon. You might therefore want to test the hypothesis on data from other languages. Spoken language data are optimal whereas written texts must first be read aloud and transliterated. Regular word stress is to be ignored; only the main accent within a sentence is the object of study. Perform a word count and a separate count of the accentuated words in the sentences of your linguistic material. Set up a list of the vocabulary of the text with overall frequencies of the individual words and the number of their accentuated occurrences. If the data display some regularity such as a tendency, express the regularity formally, i.e. as a function (fitted to the data). If there is no clearly visible tendency find the conditions under which it can be shown, i.e. the characteristics of the texts in which an interrelation between frequency and accentuation preference appears.

Zipf himself used the modification “generally”, i.e. he saw that there are exceptions. Elaborate on these exceptions and substantiate them linguistically.

Reference

Zipf, G.K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology.* Cambridge, Mass: MIT.

1.3. Script distinctiveness

Problem

Develop a measure of sign distinctiveness.

Procedure

Distinctiveness of a sign or a letter can be defined only within a frame system. Thus, the distinctiveness of, say, the character "." differs with respect to whether it is considered as one of the characters of a typewriter or a computer font or one of the three characters ".", "-", and " " within the Morse code. Signs or letters consist of strokes (dots, straight lines, curves). First state the inventory of strokes in the writing system you wish to analyse. Then determine the properties of the strokes: length (as many categories as relevant for distinction), position (left-mid-right, top-mid-bottom), slope (horizontal, vertical, slant – as many degrees as are relevant), the aperture and widths of arcs or half-circles in as many directions as relevant (e.g. north, west, south, east), thickness (if relevant), emptiness or fullness (e.g. with squares or circles) etc. Then ascribe to each property as many degrees of complexity (integer values) as necessary. One can combine the properties e.g. top-left-short-horizontal straight line obtains value 1, etc, or top 1, left 1 short 1, horizontal 1, straight line 1 and a function of these numbers yields the necessary element of a vector. Set up a vector of these characteristics and

compare all signs with each other. Define sign distinctiveness as a function of its differences to all other signs (e.g. the mean of all the differences among the inventory). Set up a measure of global distinctiveness of the whole script system.

A variant of this kind of distinctiveness measure can be obtained if you have access to the definition of a font in terms of vectors (such as a font definition using Bézier curves). In this case a straightforward measure of character distinctiveness can be derived from the number and type of the trajectories that a character consists of and the number and (relative) co-ordinates of their control points by comparing the corresponding descriptions with those of the other characters.

A similar result can be obtained by comparing pixel (raster) definitions of each character if the given font is defined in this way.

Compare the distinctiveness of a Latin and a Greek script font.

Consider several scripts that evolved from a common predecessor and study the historical development of distinctiveness. Study the different Roman fonts your word processor offers and compare them.

Compare the distinctiveness of your handwriting with that of your colleague.

Compare distinctiveness with other properties of script and state whether there is a dependence or at least a correlation.

References

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

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 9, 46-53.*

1.4. Entropy of script system distinctiveness

Problem

With recourse to problem “1.3. Script distinctiveness” compute the entropy of this property.

Procedure

Adopt one of the definitions of a stroke in the scripts from the previous problem. Do not assign the strokes any values but state how many times a stroke with the same properties occurs in individual signs. You obtain the frequencies (representativeness) of individual strokes. Compute the entropy of the distinctiveness on the basis of these frequencies. High entropy indicates high distinctiveness. If a script has a high degree of distinctiveness all stroke types occur with equal frequency.

Compare the distinctiveness entropies of different Roman fonts. Compare them to those of the Morse and Braille scripts in terms of their relative entropies.

On the other hand, distinctiveness may be reduced if there are signs containing a long series of equal strokes. In the Morse code and in the Ogham script, the longest series of equal strokes is 5. Express distinctiveness entropy of these two scripts.

Alternatively, apply Shannon's definition of entropy.

References

None

1.5. Script complexity 2

Problem

In "Problems in Quantitative Linguistics" Vol. 1 (p. 10), complexity was treated as a purely graphical property of signs applicable to any kind of script. But since every concept can be operationalized in different ways – definitions have no truth value – some other options will be proposed here. They can be used only for alphabetic scripts as they concern the relationship between letters and graphemes on the one hand and phonemes on the other hand. The present task is the processing of more languages.

Procedure

From phoneme to grapheme: Consider the complete phoneme inventory of a language. For each phoneme state all letters or groups of letters that can represent it. For example, the English /*m*/ can be represented by letters or letter groups <*gm, m, mb, me, mm, mme, mn, mp, nm*>. The size of this orthographic set is a hint at the orthographic uncertainty of the phoneme. Process one language completely and express this kind of complexity using information theoretical measures of uncertainty (c.f. Altmann, Fan 2008) or develop new measures. Some of the works given in the references below may serve as examples of this kind of analysis.

From letter to grapheme: Bosch et al. (1974: 178) define the grapheme as "a letter or a cluster of letters that is realised in the phonological transcription as a single phoneme". Letters can occur in different graphemes. Compute the distribution of letters that can occur in 1,2,3,… graphemes. Show that it is a monotonously decreasing distribution and try to find its form. If it is not monotonous, try to find the causes and generalize your result. Fan, Altmann (2008) use the term "graphemic load of letters"; this property is at the same time a picture of writing complexity. The method presented here can be used also for those *ideographic* scripts in which the role of letters is played by different kinds of

strokes. The form of the distribution enables us to define indicators of complexity. Use the results in the given references and analyse some additional languages.

References

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

Berndt, R.S., Reggia, J.A., Mitchum, C.C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behaviour Research Methods, Instruments, & Computers 19, 1-9.*

Best, K.-H., Altmann, G. (2005). Some properties of graphemic systems. *Glottometrics 9, 29-39.*

Bosch, A. v.d., Content, A., Daelemans, W., Gelder, B. de (1974). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics 1(3), 178-188.*

Fry, E. (2004). Phonics: a large phoneme-grapheme frequency count revised. *Journal of Literacy Research 36 (1), 85-98.*

Grzybek, P., Kelih, E. (2003). Graphemhäufigkeiten (am Beispiel des Russischen). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für Slavische Philologie 31, 131-162.*

Hanna, P.R., Hanna, J.S., Hodges, R.E., Rudorf, E.H. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement*. Washington, D.C: U.S. Department of Health, Education, and Welfare.

Patterson, K.E., Morton, J. (1985). From orthography to phonology: An attempt at an old interpretation. In: Patterson, K.E., Marshall, J., Coltheart, W. (eds.), *Surface dyslexia: neuropsychological and cognitive studies of phonological reading.* London: Erlbaum.

1.6. Canonical speech segments

Problem

Find the rank-frequency and the spectrum distribution of canonically transcribed speech segments.

Procedure

Choose a text consisting of at least 1000 running words. First, if you analyse written language, the orthographic representation has to be transliterated into phonetic or allophonic symbols (by means of a software program or manually). Next, choose a classification of the sounds into at least two classes (such as V = vowel and C = consonant). More classes are possible, e.g. consonants, vowels,

glides, semivowels, reduced vowels etc. Assign the sound symbols of the text to theses classes (i.e. annotate them using corresponding symbols). Now, determine the numbers of tokens of each type (the classes) in the text and arrange the resulting numbers according to a rank-frequency distribution and the corresponding spectrum. Find the theoretical distributions that can be successfully fitted to your data

You may want to begin with Gale and Sampson's (1995) data who prepared a transcription of English sounds in three classes: V – vowel, R – reduced vowel, C – consonant and obtained types like VCV, VCCRCRCV, VRRCCV etc. The data extracted from texts are presented in Table 1.6.1, where x = number of occurrences, n_x = number of types occurring exactly x times. Then transform the spectral distribution into a rank-frequency distribution and find an appropriate theoretical distribution.

Calculate various indicators that describe these and your own data (cf. Popescu et al. 2009) and find a feature common to all of them. If possible obtain data from different languages and compare the languages.

Table 1.6.1
Spectral distribution of canonical speech segments in English
(Gale, Sampson 1995)

x	n_x	x	n_x	x	n_x	x	n_x
1	120	20	3	46	1	257	1
2	40	21	2	47	1	339	1
3	24	23	3	50	1	421	1
4	13	24	3	71	1	456	1
5	15	25	3	84	1	481	1
6	5	26	2	101	1	483	1
7	11	27	2	105	1	1140	1
8	2	28	1	121	1	1256	1
9	2	31	2	124	1	1322	1
10	1	32	2	146	1	1530	1
12	3	33	1	162	1	2131	1
14	2	34	2	193	1	2395	1
15	1	36	2	199	1	6925	1
16	1	41	3	224	1	7846	1
17	3	43	1	226	1		
19	1	45	3	254	1		

References

Gale, W.A., Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics 2(3), 217-237.*

Popescu, I.-I. et al. (2009). *Aspects of word frequencies*. Berlin-New York: Mouton de Gruyter.

1.7. Phonetic comparison of cognate languages

Problem

Express the phonetic similarity of cognate languages, or a predecessor language and a follower language by means of quantitative concepts, or measure the degree of assimilation of borrowings.

Procedure

When comparing two cognate languages with respect to their phonetic similarity (or distance), use a common phonetic system and a difference measure between the sounds at least on an ordinal scale. Such a difference measure can be based, e.g. on the number of differences of place and manner of articulation or on the number of different distinctive features of two sounds. Then obtain a random sample of cognate words or use cognate words from Swadesh's (1964) basic list. Compare the differences sound by sound in each word – evaluate quantitatively also losses or increments, epentheses, etc – and express the phonetic difference between two cognate languages as the mean of all differences.

Compare in this way two Roman languages; Latin and its followers; some Slavic languages, etc.

Show whether spatial distance between cognate languages is correlated with phonetic distance.

Evaluate quantitatively the phonetic change of borrowings or individual sounds in borrowed words in the target language.

Avoid measures that take into account only the presence of change but not weighting it phonetically, e.g. the Levenshtein distance.

References

a Campo, F., Geršić, S., Naumann, C.L., Altmann, G. (1985). Subjektive Lautähnlichkeit. *Beiträge zur Phonetik und Linguistik 50, 101-120.*

Augst, G. (1971). Über die Kombination von Phonemsequenzen bei Monemen. *Linguistische Berichte 11, 37-47.*

Austin, W.M. (1957). Criteria for phonetic similarity. *Language 33, 538-544.*

Batóg, T., Steffen-Batogowa, M. (1980). A distance function in phonetics. *Lingua Posnaniensis* 23, 47-58.

Geršić, S. (1971). *Mathematisch-statistische Untersuchungen zur phonetischen Variabilität, am Beispiel von Mundartaufnahmen aus der Batschka*. Göppingen: Kümmerle.

Grimes, J.E., Agard, F.B. (1959). Linguistic divergence in Romance. *Language* 35, 598-604.

Grotjahn, R. (1980). Zur Quantifizierung der Schwierigkeit des Sprechbewegungsablaufs. In: Grotjahn,R., Hopkins,E. (Hrsg.), *Empirical research on language teaching and language acquisition: 199-231*. Bochum: Brockmeyer..

Ladefoged, P. (1970). The measurement of phonetic similarity. *Statistical Methods in Linguistics 6, 23-32.*

Lehfeldt, W. (1978). Zur Messung der phonetischen Lautdifferenz. Eine begriffskritische Untersuchung. *Glottometrika 1, 26-45.*

Lehfeldt, W. (1980). Zur numerischen Erfassung der Schwierigkeit des Sprechbewegungsablaufs. Glottometrika 2, 44-61.

Levenštejn, V.N. (1965). Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov. *Doklady Akademii Nauk SSSR 163(4), 845–848*. [Appeared in English as: V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10* (*1966*), *707–710*]

Lindner, G. (1980). Lautfolgestrukturen im Deutschen, *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 33, 468-477*.

Peterson, G.H., Harary, F. (1961). Foundations of phonemic theory. In: Jakobson, R.(Hrsg.) *Structure of language and its mathematical aspects: 139-165.* Providence, Rhode Island: American Mathematical Society.

Swadesh, M. (1964). Linguistics as an instrument of prehistory. In: Hymes, D. (ed.), *Language in Culture and Society: A Reader in Linguistics and Anthropology: 575-584*. New York: Harper and Row.

Tolstaja, S.M. (1983). Fonologičeskoe rasstojanie i sočetaemost' soglasnych v slavjanskich jazykach. *Voprosy jazykoznanija 3, 66-81*.

1.8. Phonetic word structure

Problem

(a) Evaluate the mean phonetic difference between subsequent sounds in the first 100 most frequent words of a language;

(b) Evaluate the course of phonetic differences in each word and set up a word classification on this basis.

Procedure

Use the measurement of phonetic difference as presented in the references in Problem 1.7 or develop your own measurement method. Take a large sample of words from a frequency dictionary (the most frequent ones) and test whether there is some correlation between the mean phonetic difference in the word and the frequency of the word. If Zipf's (1949) conjectures are correct, then there should be some balance or dependence between these quantities. The speaker strives for a small mean difference (production economy), the hearer for great one (decoding economy). Test whether the gamma function could express this relationship. If so, find a theoretical justification for the gamma function.

To solve problem (b), analyze as many words as possible using a dictionary and observe the course of differences. Classify the words first according to their length (in terms of sounds) and within the length classes according to the course of difference patterns. Note the number of patterns and state whether there is a regular frequency distribution (i) within a length class, (ii) in the language. If so, find the theoretical distribution and substantiate it.

References

Cf. Problem 1.7.

Zipf, G.K. (1949). *Human behaviour and the principle of least effort.* Cambridge: Addison-Wesley.

1.9. Phonetic distortion of borrowings

Problem

Borrowed words are usually phonetically modified. Express the extent of the phonetic distortion.

Procedure

First make a survey of methods treating phonetic similarity. The amount of literature on this topic is enormous (just have a look at the keyword “phonetic similarity” on the Internet using a search engine). The most frequent use of this concept can be found in dialectology (cf. e.g. Goebl 1984). Since distortion does not mean only replacement but also elimination and insertion of sounds, the most popular measure of dissimilarity is the Levenshtein distance, but many other types of distance measures are at least equivalent.

Prepare the list of borrowed words in phonetic transcription both in the source and the target languages. Then set up a table with scaled values of each property for all sounds (in both languages). An example of such a scaling for the front-back dimension is: 1. labial, 2. labio-dental, 3. alveolar, 4. palatal, 5. uvular, 6. laryngeal. Now compare each word of the source with the respective word in

the target language sound by sound. The sum of distortions in a word represents the value of the variable *D*.

(1) Set up the frequency distribution of the variable *D*. If you have computed the phonetic differences on a ratio scale, form intervals for *D* and use a continuous distribution. Is the distribution monotonously decreasing? If so, why? If not, why not?

(2) Consider an individual sound of the source language. It does not change always into the same target sound; its distortion may diversify, e.g. the sound /a/ of the source may change in /a/,/ą/,/o/ and /#/ of the target language. In that case its diversification has the value 4. For each source sound find the number of target sounds in which it can change (including elimination). Set up the distribution of the number of diversifications of sounds. Use the parameters of this distribution to measure the phonetic distance between the source and the target language.

References

a Campo, F., Geršić, S., Naumann, C.L. ,Altmann, G. (1989). Subjektive Ähnlichkeit deutscher Laute. *Glottometrika 10, 46-70***.**

Afendras, E.A., Tzannes, N.S., Trépanier, J.G. (1973). Distance, variation and change in phonology: stochastic aspects. *Folia Linguistica 6, 1-27.*

Austin, W.M. (1957). Criteria for phonetic similarity. *Language 33, 538-544.*

Batóg, T., Steffen-Batogowa, M. (1960). A distance function in phonetics. *Lingua Posnaniensis 23, 1960, 47-58.*

Benzecri, J.P. (1970). Sur l'analyse des matrices de confusion. *Revue de statistique appliquée 18, 5-63.*

Bruce, D., Murdock, B.B. Jr. (1968). Acoustic similarity effects on memory for paired associates. *Journal of Verbal Learning and Verbal Behaviour 7, 627-631*.

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study*. Nijmegen: Diss.

Geršić, S. (1971). *Mathematisch-statistische Untersuchungen zur phonetischen Variabilität, am Beispiel von Mundartaufnahmen aus der Batschka.* Göppingen, Kümmerle.

Geršić, S., Naumann, C.L., Altmann, G. (1985) Subjektive Lautähnlichkeit. *Beiträge zur Phonetik und* Linguistik 50, 101-120.

Goebl, H. (1984). *Dialektometrische Studien*. 3. vols. Tübingen: Niemeyer.

Grimes, J.E., Agard, F.B. (1959). Linguistic divergence in Romance. *Language 35, 598-604*

Grotjahn, R. (1980). Zur Quantifizierung der Schwierigkeit des Sprechbewegungsablaufs. In: Grotjahn, R., Hopkins, E. (Eds.), *Empirical research on language teaching and language acquisition: 199-231*. Bochum, Brockmeyer.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen, Diss.

Klatt, D.H. (1968). Structure of confusions in short-term memory between English consonants. *The Journal of the Acoustic Society of America 44, 401-407*

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities 37(3), 273-291.*

Lehfeldt, W. (1980). Zur numerischen Erfassung der Schwierigkeit des Sprechbewegungsablaufs. *Glottometrika 2, 44-61.*

Lindner, G. (1975). *Der Sprechbewegungsablauf. Eine phonetische Studie des Deutschen.* Berlin: Akademie-Verlag .

Lindner, G. (1980) Lautfolgestrukturen im Deutschen, *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 33, 468-477.

Łobacz, P. (1981). Classification of Polish consonantal phonemes on the basis of a subjective similarity test. *Speech Analysis and Synthesis 5, 97-120.*

Łobacz, P. (1981). Distances between Polish consonantal phonemes in tests with real and nonsense words. *Speech Analysis and Synthesis 5, 121-138.*

Miller, G.A., Nicely, P.E. (1955). An analysis of perceptual confusion among consonants. Journal of the Acoustical Society of America 27, 338-352.

Mohr, B., Wang, W.S-Y. (1968). Perceptual distance and the specification of phonological features. *Phonetica 18, 31-45*.

Nakatani, L.H. (1972). Confusion-choice model for multidimensional psychophysics. *Journal of Mathematical Psychology 9, 104-127*.

Nerbonne, J., Heeringa, W., Kleiweg, P. (1999). Edit distance and dialect proximity. In: Sankoff, D., Kruskal, J. (eds.), Time warps, string edits and macromolecules: the theory and practice of sequence comparison: V-XV. Stanford, CA: CSLI (2nd ed.)

Nerbonne, J., Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics 14(2-3),148-166.*

Peterson, G.H., Harary F. (1961). Foundations of phonemic theory. In: Jakobson, R. (Ed.), *Structure of language and its mathematical aspects: 139-145.* Providence, Rhode Island.

Singh, S. (1966). Crosslanguage study of perceptual confusion of plosive phonemes in two conditions of distortion. *Journal of the Acoustical Society of America 40, 635-656.*

Singh, S. (1971). Perceptual similarities and minimal phonemic difference. *Journal of Speech and Hearing Research 14, 113-124*.

Singh, S., Black J.W. (1966). Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *Journal of the Acoustical Society of America 39, 372-387*.

Singh, S., Frank, D.C. (1972). A distinctive feature analysis of the consonantal substitution pattern. *Language and Speech 15, 209-218.*

Thürmann, E. (1974). Phonetische Ähnlichkeit, distinktive Merkmale und auditive Dimensionen. Ein Bericht. *Hamburger Phonetische Beiträge 13, 163-192*

Tolstaja, S.M. (1968). Fonologičeskoe rasstojanie i sočetaemost' soglasnych v slavjanskich jazykach. *Voprosy jazykoznanija 66-81*

Wilson, K.V. (1963). Multidimensional analyses of confusions of English consonants. *The American Journal. of Psychology 76, 89-95*

Yokoyama, S., Itahashi, S. (1979). On the distance of Japanese words based on distinctive features and a second-order model. *Glottometrika 2, 62-81*.

Grammar

2.1. Fenk's hypothesis

Hypothesis

Within a sentence, frequent entities are positioned in the front part, rare ones in the back part. This is why "… multifunctional words tend to concentrate in the first part of the sentence …" (Fenk-Oczlon, Fenk 2002). Test the hypothesis.

Procedure

(1) Set up a word frequency list of a lemmatized text. Instead of a frequency count a frequency dictionary can be used.
(2) Partition the given text into sets of sentences of equal length, i.e. containing 2, 3, 4,… words.
(3) In each set, replace the words by their frequencies (use either the local text frequency or the frequency from a frequency dictionary). Do not eliminate punctuation marks which can turn out to be relevant.
(4) For each position in individual length sets, compute the mean frequency for each position separately
(5) Set up a hypothesis about the sequence of the mean frequencies in the course of the sentence. According to Fenk's hypothesis, the frequency should decrease monotonously. If your data corroborate it, enhance the hypothesis. If possible, propose a function expressing the decrease of the values. How do the parameters vary with sentence length?

If the hypothesis is not corroborated increase the size of your data or modify the hypothesis. Take into account the position of punctuation marks.

As the hypothesis was already tested on English data a study of a non-Indo-European language is preferable. If you can compare several languages, try to embed boundary conditions in the hypothesis, i.e. determine factors which are responsible for observed differences between the results for data from different languages.

If you obtain positive results apply the parameters of the function to typological purposes. Are the parameters apt to characterize languages?

Are there any languages which display a special course of frequency sequences? If so, have these languages/courses common features? If you observe a regularity find a mathematical function to capture it and derive it theoretically starting from some syntactic properties of the given language. How does the function change with increasing sentence lengths? Examine the data separately according to the individual sentence length classes.

References

Fenk-Oczlon, G., Fenk, A. (2002). Zipf's tool analogy and word order. *Glottometrics 5, 22-28.*

Fenk, A., Fenk-Oczlon, G. (2002a). 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.

2.2. Zipf's adverb hypothesis (1)

Hypothesis

"… adverbs of time are on the average less independent and therefore shorter than adverbs of place" (Zipf 1935/68: 242). Test the hypothesis.

Procedure

List all adverbs of place and time you find in text-books or grammars of several languages. Compute the mean lengths of these two groups and compare them using e.g. a *t*-test. Can you corroborate Zipf's hypothesis? If not, can you find some boundary conditions leading to the acceptation of this simple hypothesis? Can you find languages where the two groups possess approximately equal mean length? If so, determine the specific properties of such languages. Is it necessary to modify the hypothesis? Attach to each adverb its frequency of occurrence as counted in a corpus or from a frequency dictionary. Can you find interrelations?

Study languages which do not belong to the most studied ones such as English. Render the hypothesis more precise by measuring the extent of co-text dependence of the adverbs. This forms an independent, non-trivial problem, which is likely to be solvable only in connection with verb valency.

Reference

Zipf, G.K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology.* Cambridge, Mass: MIT.

2.3. Zipf's adverb hypothesis (2)

Hypothesis

"… adverbs of time are on the average less independent and therefore shorter than adverbs of place" (Zipf 1935/68: 242). Test the hypothesis.

Procedure

Zipf's Hypothesis has the presupposition that adverbs of time are less independ-

ent (e.g. of verbs) than adverbs of place. This assumption should be tested on its own right. As a possible measure of independence, polytextuality can be used.

Re-use the material from Problem 2.2. Classify the contexts in which the adverbs occur according to a semantic and/or pragmatic criterion. The number of classes should exceed 4. Now, determine the number of different classes in which the adverbs occur and assign to each adverb this number (polytextuality). Compare both groups using the Chi-square or *t*-test test.

References

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

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern. Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition*: *327-338.* Lengerich: Pabst Science Publishers.

Zipf, G.K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology.* Cambridge, Mass: MIT.

2.4. Auxiliary words

Problem

It is well known that auxiliary words are the most frequent words in speech and text. There is a number of different hypotheses about this fact, e.g. "… there is a striking correlation between high frequency and auxiliary status" (Krug 2001: 312). However, researchers consider mostly Indo-European languages. Elaborate on the following tasks:

Procedure

First define the class of auxiliary words in a language of your choice. Define it exactly and develop a method for scaling the degree of auxiliarity.

(1) Compare the rank-frequency distributions of auxiliary words in several texts in a highly synthetic and a highly analytic language, e.g. a Slavic and a Polynesian language. Show that there are drastic differences between the ranks (= scores on your independence scale) of auxiliaries. Perform a statistical test to show the difference.

(2) Design an index expressing the extent of the use of auxiliaries in the given language. Care for its simplicity and comparability. Characterize the languages studied. In order to make texts of different length comparable, relativize the ranks (= divide each rank by inventory size V). Perform a nonparametric test for the difference of the given languages.

References

Krug, M. G. (2001). Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J., Hopper, P. (2001), *Frequency and the emergence of linguistic structure: 309-335.* Amsterdam/Philadelphia: Benjamins

2.5. Valency and text frequency

Problem

Are verbs with valency x (x = 1, 2, 3,…, n) regularly distributed in a corpus?

Procedure

Obtain the frequencies of all verbs in a corpus. Then obtain the valency of each verb with respect to the given corpus, i.e. do not consult a valency dictionary. It can be observed that high valency verbs occur more frequently than low valency verbs. On the other hand, there are more low frequency verbs than high valency ones in the lexicon. Hence, if there is no trend, one would expect a uniform distribution. State whether the observed empirical distribution $\{P_x\}$ – where x = valency, $f(x)$ = frequency of all verbs with valency x – differs from the uniform distribution, and if so, what kind of distribution is a good model of the observed pattern? First find the distribution empirically (with the help of a statistical software package which offers a large number of theoretical probability distributions), then justify your approach, i.e. derive it from linguistic arguments.

As a by-product, determine whether the positions of the verbs in the sentences are interrelated with the degree of valency.

References

None

2.6. Valency and rank-order

Problem

The rank-frequency distribution of valency abides by a common rank-frequency distribution or function. Test the hypothesis.

Procedure

Transform the empirical distribution in the Problem “Valency and text frequency” to a rank order distribution, i.e. ignore valency but assign rank 1 to the most frequent verb, rank 2 to the second most frequent verb, etc. You obtain a

rank-frequency distribution of verbs (not valencies!). Find the form of the distribution empirically and theoretically.

Then extract from the distribution those verbs with valency 1, order the verbs according to decreasing frequency and find the distribution or a simple function for this series. Then isolate the verbs with valency 2 and perform the same. Continue up to the highest valency class with at least five verbs. After you have found all distributions or functions (sequences), compare them and formulate a new hypothesis about the form of the distribution as depending on the valency class. Is it always the same function with different parameters or do you need different functions for different valency classes?

Can this procedure be used as a criterion for the support of traditional valency attribution value to verbs?

References

None

2.7. Case diversification in Ugro-Finnic languages

Problem

Compute the diversification constant (see Problem 6.3. "Diversification constant") for the case of nouns in some Ugro-Finnic languages.

Procedure

Examine at least three Ugro-Finnic languages and 10 texts in each of them. State exactly which suffixes express case relations in nouns. Pool all allomorphs arising by vowel harmony and do not care for polysemy or polyfunctionality of individual affixes. Set up a rank-frequency sequence of affixes using absolute frequencies and show that (1) the rank frequency sequence follows the Popescu function $f(r) = 1+a*\exp(-r/b)$ where r is the rank and $f(r)$ is the absolute frequency of rank r, (2) Compute the diversification constant c,

$$c = \frac{R + f_{\max} - f_{\min} + 1 - L}{h}$$

for each language separately and show that they are very similar. Determine the category of phenomena the case belongs to using Table 1 and the formulas in Problem 6.3.

References

Best, K.-H. (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18, 26-31.*

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-35.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 2008, 97-111.*

Rothe, U. (ed.) (1991). *Diversification process in language: grammar*. Hagen: Rottmann.

Sanada, H. (2009). Diversification of postpositions in Japanese. Msc.

2.8. Valency and compounding

Hypothesis

The greater the valency of a verb, the more compounds it produces.

Procedure

Consider valency as the number of complements (alternatively as the number of all arguments, i.e. complements and circumstantials) the verb requires to form a complete sentence. For example: *I saw yesterday a good film in the cinema* has two complements (*I* and *a good film*) and two circumstantials (*yesterday* and *in the cinema*). Consult a valency dictionary in a language of your choice and obtain at least 300 verbs by random sampling. Tag each verb with its valency. Consult a common monolingual dictionary (or an Internet source) and note for each verb the number of compounds it forms. For each valency class compute the mean number of compounds. Then, in an explorative manner, find a function expressing the dependence of compositionality on valency: COMP = f(VAL). If you succeed, give a justification for the given function. Set up the differential equation which leads to the given function and substantiate it. Chart your result in form of a signal flow diagram. Incorporate it into the framework of synergetic linguistics (Köhler 1986, 2002, 2005) and show that the function can be derived from the unified theory (Wimmer, Altmann 2005).

References

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

Köhler, R. (ed.) (2002). *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279 (12.Dec. 2008)

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch*.

Quantitative Linguistics. An International Handbook: 760-774. Berlin-New York: de Gruyter.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch Quantitative Linguistics. An International Handbook: 791-807.* Berlin-New York: de Gruyter.

2.9. Valency and derivation

Hypothesis

The greater the valency of a verb, the greater the productivity, i.e. the more derivates are formed from it. Test the hypothesis.

Procedure

Draw a random sample of 200 simple, i.e. non-derived, verbs from a valency dictionary or corresponding electronic sources on the Internet and determine the number of their valencies. Set up a table with the verbs in the first column and the number of their valencies in the second one. Find all derivates of these verbs (take into account derivation by means of concatenation – e.g. German *Ver/**beug**/ung*, Hungarian *meg/**lep**/etés*, Slovak *roz/**del**/enie* – and by alternation such as ablaut and umlaut but not conversion or null-morpheme derivation).Write the number of derivates in the third column. Then compute first the correlation between the valency and derivation values. If you use an electronic spreadsheet program it is easy to draw a figure of the data, which may suggest the form of a dependency function. If the dispersion is too great, add further 100 verbs. If you obtain an interesting result, find a function expressing the dependence. Ideally, derive the function by means of theoretical considerations, otherwise try doing it inductively.

If this relationship exists, it surely depends on the extent of derivation in the given language. Introduce this property as an independent variable in order to render predictions more exact. Express the concept of "derivationality" using some variant of the Greenberg-Krupa indices (see references).

References

Allerton, D. (1982). *Valency and the English verb*, London/New York: Academic Press.

Emons, R. (1978). *Valenzgrammatik für das Englische. Eine Einführung,* Tübingen: Niemeyer.

Engel, U. et al. (1983). *Valenzlexikon deutsch-rumänisch*. Heidelberg.

Greenberg, J.H. (1960/1990). A quantitative approach to the morphological typology of languages. In: Denning, K., Kemmer, S. (eds.), *On Language:*

Selected Writings of Joseph H. Greenberg: 3-25. Stanford, California: Stanford University Press.
Hajič J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In: Hajičová, E. (ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*: *106-132*. Prague: Karolinum.
Helbig, G. (Hrsg.) (1971). *Beiträge zur Valenztheorie*. The Hague/Paris: Mouton.
Helbig, G. (1992). *Probleme der Valenz- und Kasustheorie,* Tübingen: Niemeyer.
Helbig, G., Schenkel, W. (1969/83). *Wörterbuch zur Valenz und Distribution deutscher Verben*. VEB Bibliographisches Institut.
Hudson, R. (1993). Recent developments in dependency theory. In: Jacobs, J., Stechow, A.v., Sternefeld, W., Vennemann, T. (eds.), *Syntax. Ein internationales Handbuch zeitgenössischer Forschung: 329-338.* Berlin/New York: de Gruyter..
Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*
Lamprecht, A. (1983). *Relationale Satzanalyse. Theorie und Praxis einer konsistenten Analyse englischer Satzstrukturen*. München: Hueber.
Nižníková, J., Sokolová, M. (1998). *Valenčný slovník slovenských slovies*. Prešov: Filozofická Fakulta Prešovskej Univerzity.
Schumacher, H**.** (1986). *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben.* Berlin-New York: de Gruyter.
Welke, K. (1988). *Einführung in die Valenz- und Kasustheorie*. Leipzig: Enzyklopädie.

2.10. Valency and synonymy

Hypothesis

The greater the valency of a verb, the more synonyms it has. Test the hypothesis.

Procedure

For many verbs, high valency may be connected with a penetration of a verb in the semantic domain of another verb, and the number of their synonyms may thereby increase. Again, prepare a table (electronic spreadsheet) containing the valency and the number of synonyms of individual verbs and compute first the correlation, then show that the dependence is a straight line but its slope is not 1. Only few valency dictionaries give also the synonyms. It is, anyway, better to consult a synonym dictionary in addition to the valency dictionary.

State whether the relationship remains the same if you examine several other languages. If not, modify the hypothesis and add suitable boundary conditions.

Since synonymy is associated with polysemy, demonstrate the impact of both variables (valency and polysemy) on synonymy and derive appropriate formulas.

References

Cf. Problem 2.9.

2.11. Valency and length

Hypothesis

The greater the valency of a verb, the shorter it is.

Procedure

Frequency "shortens" the words and at the same time gives the verbs chances to enlarge their valency. Hence, at least a correlation between verb length and verb valency should be observable.

Prepare data and examine the behaviour of the two variables. Word length should be measured in a canonical form, i.e. after lemmatization, and not in terms of the number of phonemes or graphemes but rather in terms of the number of syllables. We do not expect a linear relation and we cannot predict the direction of their dependence. The inverse variant of the dependence, viz. "the shorter a verb, the greater its valence" is plausible, too.

References

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 760-774.* Berlin-New York: de Gruyter.

2.12. The control cycle of valency

Problem

Integrate valency, synonymy, polysemy, frequency and length in a control cycle.

Procedure

Recall the notation as used in synergetic linguistics and draw a corresponding diagram. Find requirements and order parameters which govern the processes associated with the resulting structure.

References

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 760-774.* Berlin-New York: de Gruyter.

2.13. Valency of nouns and adjectives

Problem

Which of the relations of valency with other properties discussed above hold also for nouns and adjectives?

Procedure

Collect data in analogy to the descriptions above for nouns and adjectives instead of verbs. Examine the corresponding results and determine which of the relations show clear tendencies and describe the differences between the three groups. Next, set up a hypothesis about the properties of the parts-of-speech that determine the form of their valency relations and formulate them as boundary conditions.

References

None

2.14. Valency: the distribution of variants

Hypothesis

The number of variants of verbs is distributed according to the positive negative binomial distribution.

Procedure

Count the number of variants of the verbs in a valency dictionary or an online source. As many verbs as possible should be taken into account, ideally the complete dictionary. A variant of a verb can be operationalized as an entry in the valency dictionary with an individual complementation pattern and mostly a special meaning nuance. Hence, each verb has $x = 1, 2, \ldots$ variants with

$$P_x = \frac{\binom{k+x-1}{x}}{1-p^k} p^k q^x, \quad x = 1, 2, \ldots.$$

Test the hypothesis, i.e. fit the distribution to the data and perform a Chi-square test.

Substantiate the hypothesis or develop a different one if the distribution does not fit with the data. Find the reasons for the rejection of the hypothesis, i.e. study the derivation of the negative binomial distribution.

Reference

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

2.15. Valency and complementation patterns

Hypothesis

The complementation patterns of verbs are distributed according to the Zipf-Mandelbrot distribution.

Procedure

Examine a suitable valency dictionary and count the number of verbs which possess a given pattern in their description. An example of a complementation pattern is "N_s+N_d/N_i/CL_{that}/P_{in}N", which describes the valency of a verb with an obligatory subject, and obligatorily a direct or indirect object or alternatively a clause introduced with *that* or a prepositional object with the preposition *in* (such as *believe*: I believe this / I believe him / I believe that it is true / I believe in the power of algebra). Facultative (non-obligatory) elements are also possible.

There will be a large number of patterns which are found with only a single verb (variant), a somewhat smaller number of patterns which describe the behaviour of two verbs etc. The pattern with the largest number of verbs has rank 1. The resulting rank-frequency distribution is expected to follow the Zipf-Mandelbrot distribution. Test the hypothesis, i.e. fit the distribution to the data and perform a Chi-square test.

Reference

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

2.16. Distribution of the semantic subcategories of arguments

Hypothesis

The number of admissible semantic subcategories of an argument in a complementation pattern is distributed according to the positive Poisson distribution.

Procedure

Some valency dictionaries provide information not only about number and type of the arguments in a pattern but also about semantic subcategories for a finer selection of the lexical instances for an argument type. Examples of such semantic subcategories are +anim (only animated arguments are possible), -anim (only non-animated arguments), similarly +/- human, +/-abstract, +/-collective, etc. Count the number of arguments with x = 1, 2, … possible subcategories in a given complementation pattern. Each verb contributes to the count as many times as it has complements, i.e. a verb with two complements adds two independent counts. The variable x is expected to follow the distribution

$$P_x = \frac{\lambda^x}{x!\ (e^\lambda - 1)} \quad P_{x-1}, \qquad x = 1,\ 2,\ 3,\ \ldots.$$

Test the hypothesis, i.e. fit the distribution to the data and perform a Chi-square test.

Reference

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

2.17. Number of arguments and number of semantic subcategories

Hypothesis

There is a linear functional dependency of the number of admissible semantic subcategories on the number of arguments in a complementation pattern.

Procedure

It goes without saying that the more arguments (complements) a complementation pattern has the more semantic subcategories are admissible. Determine the form of the actual dependency. It is expected to be a linear one.

Test the hypothesis, i.e. fit the linear function $y = ax + b$ to the data and calculate the coefficient of determination.

Reference

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

2.18. Frequency and allomorphy

Hypothesis

The number of allomorphs of a morpheme is an increasing function of morpheme frequency. The more frequent a linguistic unit the easier it is to remember. Therefore, frequent units can be irregular, which again can be exploited in order to form economical, short allomorphs. The verb *to be* is a good example of frequent lexemes; the set of allomorphs occurring in its word-forms is {*be, am, are, is, was, were*} whereas a rather rare morpheme such as *simul* comes as a singleton morph in all its derivative and inflectional contexts (*simul-ate, -ation, -taneous*, …). Test the hypothesis.

Procedure

Collect data on morpheme frequency and the number of allomorphs of the corresponding morphemes on textual data or on data from a frequency dictionary. Draw a graph of the data points (morpheme frequency, number of allomorphs) for each morpheme of your sample and check optically whether the points roughly form a smooth line or show, at least, a clear tendency. If possible, fit a function to the data.

Collect the data randomly; do not take selected morphemes. Since the domain of allomorph numbers is smaller than that of frequencies, reverse the hypothesis and present the dependence as Frequency = f(allomorph number). Consider, of course, mean frequencies.

References

None

2.19. Semantic relevance of affixes (1)

Hypothesis

The more relevant an affix (category) is with respect to the meaning of the word the closer it is to the stem. Test the hypothesis.

Procedure

First, define an independent measure of semantic relevance, which takes into account to what degree the meaning of the complete word-form differs from the meaning of the stem. Position closeness is easier to define: the number of morphs between the stem and the affix under study. Determine the closeness and the relevance values of all the words with affixes in textual data and try to find a corresponding function.

Consider the construction of the measure of semantic relevance as a separate problem. Develop a scaling procedure. Set up the distribution of this measure with all word forms in a text. Find a probability distribution using theoretical argumentation. Perform this procedure on different languages. Define the difference between languages using this criterion.

Reference

Bybee, J.L. (1985). *Morphology: a study of the relation between meaning and form.* Amsterdam, Philadelphia: J. Benjamins. (= Typological Studies 9)

2.20. Semantic relevance of affixes (2)

Hypothesis

The greater the semantic cohesion within a word (i.e. the more relevant the affixes with respect to the stem) the greater the probability of a morpho-phonemic effect. Test the hypothesis.

Procedure

Derive a measure of semantic coherence within a word on the basis of the measure of relevance as defined above (cf. Semantic relevance of affixes (1)).

Determine the coherence values of all the words with affixes in textual data from an inflectional or agglutinative language. For each of these words note whether a morpho-phonemic effect can be observed. Form appropriate intervals on your coherence scale and pool the words in groups according to these intervals. For each of the groups calculate the relative frequency of morpho-phonemic effects

$$M_i = E_i / S_i$$

where E_i is the number of words in group i showing a morpho-phonemic effect and S_i the number of all words in group i. The groups can now be represented by data pairs (C_i,M_i) where C_i is the mean value of coherence of the words in group *i*. Show whether there is a tendency or even a functional dependency in the data which supports the hypothesis.

Reference

Bybee, J.L. (1985). *Morphology: a study of the relation between meaning and form.* Amsterdam, Philadelphia: J. Benjamins (= Typological Studies 9).

2.21. Word order and topic assignment

Hypothesis

There are various observations concerning topic coding, and various hypotheses have been set up. Moreover, different interpretations exist as to the reasons for certain patterns such as the motivations for the ranking of the word order patterns within a sentence.

COMMENT > COMMENT-TOPIC > TOPIC-COMMENT > TOPIC(REPETITION)

(zero topic) (zero comment)

One interpretation is based on a continuity or predictability scale (when the topic is easy to induce the effort of coding or emphasizing is reduced), another on the psycholinguistic principle "Attend first the most urgent task". Apparently, economy, iconicity, and processing factors are involved.

Test the hypothesis that the frequencies of the four word order patterns in texts follow a probability distribution from the group of diversification distributions.

Procedure

Determine the frequencies of the four word order patterns on data from texts in at least two languages and fit an appropriate probability distribution to the data. Interpret the result. Set up a theoretical model of the mechanism which controls the word order patterns.

References

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 646-658.* Berlin/New York: de Gruyter.

Givón, T. (1985). Iconicity, isomorphism and non-arbitrary coding in syntax. In: Haiman, John (ed.), *Iconicity in syntax*. Amsterdam, Philadelphia: Benjamins.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

2.22. Syntactic properties

Problem

The quantitative investigations into properties and interrelations of syntactic structures (cf. Köhler 1999 and Köhler, Altmann 2000) have been tested only on corpora from two languages (English and German) so far. Extend this empirical basis.

Procedure

Find a syntactically annotated corpus in a language other than German and English. Alternative: Compile such a corpus, even if it is a small one. Conduct investigations in analogy to the above-cited ones and compare your results to the ones obtained by Köhler and Altmann..

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*

Köhler, R., Altmann, G. (2000). Probability dstributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1), 189-200.*

2.23. Efficiency of the POS system

Problem

In Vulanović (2008), the theoretical efficiency of various typologically attested part-of-speech systems is calculated. Find the efficiency of existing POS systems. Take account of the frequency of the corresponding phenomena as observed in communication (texts).

Procedure

The efficiency measure as proposed by Vulanović is based on system properties. Taking into account with which frequency the words belonging to a POS class occur and how often a linguistic means of marking a propositional function is needed may change the picture of efficiency in actual usage.

Annotate texts according to grammatical and lexical descriptions of some languages with respect to POS classes of the words and with respect to the application of propositional functions. Formulate an efficiency measure which takes frequency into account and apply it to the annotated texts.

Reference

Vulanović, R. (2008). A mathematical analysis of parts-of-speech systems. *Glottometrics 17, 51-65.*

2.24. Length and complexity of syntactic structures

Problem

Quantitative studies of properties such as complexity and length of syntactic structures, information content, position in the mother constituent etc. (cf. Köhler 1999 and Köhler, Altmann 2000) are defined and operationalized with respect to a phrase structure grammatical analysis of sentences. Define properties and interrelations for sentences analyzed according to a dependency grammar.

Procedure

Find at least two properties which can be ascribed to parts of a stemma. Operationalise them and measure the properties accordingly using a syntactically annotated text corpus such as the Czech National Corpus. Determine the frequency distributions of the properties and fit theoretical probability distributions to the data. If the two (or more) properties display a functional dependency set up a hypothesis and fit the corresponding function to the data.

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*

Köhler, R., Altmann, G. (2000). Probability dstributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1), 189-200.*

2.25. Grammar, text, corpus, language

Problem

The quantitative studies in Köhler (1999), Köhler/Altmann (2000) are based on a phrase structure grammatical analysis. Moreover, the particular grammar as used by the Lancaster annotators has some specific properties. To what extent are results as presented in the studies cited above descriptions of the individual texts/ then corpus, of the language, or of the properties of the selected grammar?

Procedure

Scrutinize the syntactic analysis which led to the annotation in the Susanne corpus (Sampson 1995). Examine how the results would change depending on the choice of the specific properties of variations of the grammatical principles.

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*

Köhler, R., Altmann, G. (2000). Probability dstributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1), 189-200.*

Sampson, G. (1995). *English for the Computer*. Oxford: Clarendon Press.

2.26. Functional dependencies in syntax

Problem

In Köhler (1999), several hypotheses on functional dependences between syntactic properties have been set up and tested empirically.
Some of them have the form $y = Ax^b e^{cx}$:

- average constituent frequency as a function of constituent complexity,
- the relation between complexity and length,
- the dependence of depth of embedding on constituent position.

Provide a theoretical derivation or justification of this formula with respect to the specific hypotheses.

Procedure

Determine the hypothetical mechanisms behind the links between the variable pairs under study. You might want to follow the modelling procedure as described in (Köhler 2006) and make use of the extension of the synergetic-linguistic tools presented here.

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern - Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler: 327-338.* Lengerich: Pabst Science Publishers

2.27. Distribution of complexity

Hypothesis

The syntactic complexity of clauses follows – as the complexity of syntactic constructions in general – the hyper-Pascal distribution.

Procedure

In Köhler/Altmann (2000), the distribution of the complexity of syntactic constructions was analysed and modelled with the hyper-Pascal distribution. In this paper, the totality of all the constructions in the corpora was used as data. The theoretical considerations which lead to the model should be valid in particular for individual kinds of constructions.

Collect complexity data separated according to the different kinds of constructions, i.e. phrases and clauses and test the hypothesis. If the hypothesis is confirmed on your data, test whether the parameters of the distributions differ with respect to construction kind.

References

Köhler, R., Altmann, G. (2000). Probability dstributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(1), 189-200.*

2.28. Information structure (1)

Hypothesis

The information of the elements of a syntactic structure decreases with increasing position of the element in the structure.

Procedure

Information (in the information theoretical sense) is the degree of uncertainty - or a measure of the information elements (bits) needed for coding. The information of an element of a syntactic structure can be measured in terms of the number of alternatives which could be used instead of the given element (paradigmatic relation).

The easiest way to test the hypothesis is using a syntactically annotated corpus. For each structure type in a long text or in a corpus, count the number of alternatives that can be used at a given position in the structure: begin with the first position in a structure type (e.g. in a NP) and determine the number of element types (phrase types and word classes) that can start the construction (e.g., determiner, proper noun, pronoun, ...). Now, for any element type you find,

determine the number of followers, i.e. element types at position 2, etc. The logarithm of such a count is a measure of the information of an element at the given position. If you take the logarithm to base 2 you get information in terms of bits. Study the dependence of information on position on data from as many languages as possible.

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*

Köhler, R**.** (2000). A study on the informational content of sequences of syntactic units. In: Kuz'min, L.A. (ed.), *Jazyk, glagol, predloženie. K 70-letiju G. G. Sil'nitskogo: 51-61*. Smolensk.

2.29. Information structure (2)

Problem

Study the information of linguistic elements in paradigmatic relations.

Procedure

The individual elements or features which constitute a paradigm or a category do not occur with the same frequency. Therefore, the information associated with a given category or element type (cf. the problem Information structure (1)) can be measured using the concept of entropy. In this way, the probability distribution of the elements is taken into account. Calculate the entropies of the elements in a syntactic construction using their frequencies at their positions in the structure and compare the result to the findings of the above problem. What do you conclude?

References

Altmann, G. (1980). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*

Köhler, R. (2000). A study on the informational content of sequences of syntactic units. In: Kuz'min, L.A. (ed.), *Jazyk, glagol, predloženie. K 70-letiju G. G. Sil'nitskogo: 51-61*. Smolensk.

2.30. Diversification of aspect

Problem

Some languages express the verbal aspect using morphological means, e.g. Slavic languages; other ones use analytic constructs, phrases, etc. Show that if the means diversify, their frequency is linked with their length (cf. also the more general Problem 9.12)

Procedure

Consult a grammar text-book of the given language and extract all forms expressing aspect. Then check the occurrences of the individual forms in a corpus. Measure the length of individual means in different ways, e.g. in terms of word numbers, syllable numbers, morpheme numbers, etc. First compute the correlation between length and frequency, then find a function expressing the dependence. Embed the problem in the system of synergetic linguistics.

References

Bache, C. (1982). Aspect and Aktionsart: Towards a semantic distinction. *Journal of Linguistics, 18(1), 57-72.*

Binnick, R.I. (1991). *Time and the verb: A guide to tense and aspect.* New York: Oxford University Press.

Binnick, R.I. (2006). Aspect and Aspectuality. In: Aarts, B., McMahon, M.S. (eds.), *The Handbook of English Linguistics: 244–268*. Malden, MA: Blackwell Publishing.

Chertkova, M.Y. (2004). Vid or Aspect? On the Typology of a Slavic and Romance Category [Using Russian and Spanish Material]. *Vestnik Moskovskogo Universiteta, Filologia, 58(9-1), 97-122.*

Comrie, B. (1976). *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge-New York: Cambridge University Press.

Dahl, Ö. (ed.) (2000). *Tense and Aspect in the Languages of Europe*. Berlin: Mouton de Gruyter.

Gautier, L., Haberkorn, D. (eds.) (2004). *Aspekt und Aktionsarten im heutigen Deutsch.* Tübingen: Stauffenburg.

Herweg, M. (1990). *Zeitaspekte*. Wiesbaden: Westdeutscher Verlag.

Kortmann, B. (1991). The Triad "Tense-Aspect-Aktionsart". *Belgian Journal of Linguistics, 6, 9-30.*

Löbner, S. (2002). Is the German Perfekt a perfect perfect? In: Kaufmann, I., Stiebels, B. (eds.), *More than Words: A Festschrift for Dieter Wunderlich*: *369-391*. Berlin: Akademie Verlag.

MacDonald, J.E. (2008). *The syntactic nature of inner aspect: A minimalist perspective.* Amsterdam-Philadelphia: John Benjamins Pub. Co.

Richardson, K. (2007). *Case and aspect in Slavic*. Oxford-New York: Oxford University Press.

Sasse, H.-J. (2002). Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state? *Linguistic Typology, 6(2), 199-271.*

Sasse, H.-J. (2006). Aspect and Aktionsart. In: Brown, E.K. (ed.), *Encyclopedia of language and linguistics (Vol. 1, 535–538)*. Boston: Elsevier.

Smith, C.S. (1991). *The parameter of aspect*. Dordrecht-Boston: Kluwer Academic Publishers.

Tatevosov, S. (2002). The parameter of actionality. *Linguistic Typology, 6(3), 317-401.*

Hollebrandse, B., Hout, A.v., Vet, C. (eds.) (2005). *Crosslinguistic views on tense, aspect and modality.* Amsterdam: Rodopi.

Zalizniak, A.A., Šmelev, A.D. (2000). *Vvedenie v russkuju aspektologiju* [Introduction to Russian aspectology]. Moskva: Jazyki russkoj kul'tury.

2.31. Case control

Problem

(1) Is there a hierarchy of cases? (2) Is there a hierarchy of case markers? Examine the problems and try to set up hypotheses.

Procedure

Consider a language with a well developed case system, e.g. Latin, German, a Slavic language, etc. Count both the number of individual cases and the occurrence of individual endings in a text. One can use a tagger or pencil and paper. The result is a table of the following form:

	Nom	Gen	Dat	Acc	Voc	Loc	Instr	Abl	…
-a									
-ae									
-am									
-arum									
-as									
……									
zero									

No other grammatical category is here relevant.

(a) Solve the first problem showing that the cases are not distributed uniformly. This can be done by testing the column sums for homogeneity.

(b) Show that if the column sums are ranked according to frequency, one obtains a regular distribution. Trace down this distribution empirically (e.g. using a software) and finally substantiate the distribution. Give arguments why it *must* be so. For example, the nominative occurs in almost all sentences because…The next case occurs in a smaller number of sentences because… etc. Base your justification concerning the mathematical form on proportionality argumentation.
(c) Solve the second problem testing the sums of rows for homogeneity.
(d) Find the rank-frequency distribution of endings and substantiate it linguistically.
(e) Can one state that the shorter the ending the more frequent it is?
(f) Study the diversification of individual endings and solve some pertinent problems. Read the diversification problems in this volume.

References

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 66-78.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 94-108.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

3. Semantics

3.1. Verb and noun polysemy

Hypothesis

According to D. Gentner (1981) "Common verbs have greater breadth of meaning than common nouns." Test the hypothesis.

Procedure

Take samples of nouns and verbs of different frequency ranges from frequency dictionaries. Find the numbers of their meanings in a monolingual dictionary and compare them. Instead of English, the most frequently investigated language, an analysis of a less examined language would be welcome. Select among languages without explicit (morphological) marking of the difference between noun and verb. Does the hypothesis hold?

Now take systematic samples of nouns and verbs from a monolingual dictionary, e.g. the last verb and the last noun on each page. Count their meanings and compare their averages.

Select 100 nouns which can form verbs by derivation from a monolingual dictionary. Count the number of noun meanings and verb meanings and compare them.

Do the same as above but take verbs from which nouns may be derived.

Perform the same test for English to examine the case of conversion.

Extract a list of nouns and verbs from a text and consult a monolingual dictionary to determine the number of their meanings. Compute the average number of meanings of nouns and verbs respectively and compare them using a *t*-test. Can you corroborate the hypothesis?

Perform the same procedure with different texts. Is the difference – if there is any – between the averages caused by the text type? If so, set up a new hypothesis and test it in other languages.

Can you see some other factors having influence on the diversification of meaning?

Can you give some psychological, developmental or epistemological grounds for this difference – if it exists?

References

Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory 4(2), 161-178.*

Oguy, O. (2005). Aproksimativni metodi v semasiologičnich poslidženijach: rezuľtati ta perspektivi zastosuvanija. In: Altmann, G., Levickij, V., Perebij-

nis, V. (ed.), *Problemi kvantitativnoï lingvistiki – Problems of Quantitative Linguistics: 134-148*. Černivci: RUTA.
http://wordnet.princeton.edu/
http://www.sfs.uni-tuebingen.de/GermaNet/

3.2. Polysemy of parts-of-speech

Problem

Different parts-of-speech tend to display different polysemy behaviour. Test the hypothesis.

Procedure

This is a generalization of Problem 3.1.

From a monolingual dictionary obtain by random or systematic sampling samples of nouns, verbs, adjectives and adverbs. Obtain their polysemy values (i.e. number of meanings) and set up its frequency distribution for each part-of-speech separately. The difference in polysemy can be shown by comparing the means of the distributions or performing a homogeneity test.

Try to show that smaller skewness of the distribution is associated with greater polysemy. Use also the entropy and the repeat rate for characterizing polysemy (cf. Problems Vol. 1, p. 113)

Compare the distributions using Ord's scheme (Cf. Problems Vol. 1, p. 111). Plot the results and set up a hypothesis on the formation of polysemy with different parts-of-speech. Can you give grounds for the differences – if they exist?

Set up the word list of a text and sort the words according to their parts-of-speech. Assign the words their polysemy values by determining their numbers of meanings using a monolingual dictionary. Then set up the distribution of meanings for each part-of-speech (X = number of meanings, Y = number of words with x meanings) separately. Compare the distributions using Ord's scheme. Analyze several texts and, comparing the locations of individual parts-of-speech in Ord's scheme, set up a new hypothesis.

Test whether there is a difference between the dictionary samples and text samples of the same parts of speech.

References

Oguy, O. (2005). Aproksimativni metodi v semasiologičnich poslidženijach: rezuľtati ta perspektivi zastosuvanija. In: Altmann, G., Levickij, V., Perebijnis, V. (ed.), *Problemi kvantitativnoï lingvistiki – Problems of Quantitative Linguistics: 134-148*. Černivci: RUTA.

.

3.3. Synonymy and morphological productivity

Hypothesis

The greater the morphological productivity of a word (i.e. the more derivatives and compounds it produces) the greater is its synonymy. Test the hypothesis.

Procedure

Take randomly 100 words from a dictionary and calculate the number of derivatives and compounds that are formed with a given word as base. Then obtain the number of synonyms of the given bases from a dictionary of synonyms. Show that the dependence <morphological productivity, synonymy>, i.e. $Syn = f(MP)$, is a monotonously increasing function. Find an appropriate function and give reasons for its adequacy, or vice versa, give reasons for the necessity of this relation and embed them in a differential equation whose solution yield the given dependence.

Integrate the dependence into a synergetic control cycle; find other factors which influence morphological productivity, synonymy, or both.

References

None

3.4. Synonymy and postpositional phrases

Hypothesis

The more postpositional phrases formed with a verb, the greater the synonymy of the verb. Test the hypothesis.

Procedure

Take randomly 50 verbs from an English monolingual dictionary and get all postpositional phrases like *get in, get out, get around, get off, get out of, get from under, get through,…* Then count the number of synonyms of the given word (here *get*) in a dictionary of synonyms. Determine the direct dependence of synonymy on the number of postpositional phrases, i.e. study the relationship <No. of different postpositional phrases, No. of synonyms>, i.e. $Syn = f(PF)$, find an appropriate function and give reasons for its adequacy.

If possible, analyze also other languages and compare the results with those of English. See also Problem 2.9 (Valency and synonymy) and 3.3 (Synonymy and morphological productivity).

If the result is positive, embed it in the control cycle of Problem 3.3

References

None

3.5. Semantic partitioning of space

Problem

In every language the space is partitioned semantically by different word classes and morphemes, e.g. prepositions, postpositions, prefixes, affixes, adverbs of place. Set up the semantic space for each class separately and evaluate some of its properties.

Procedure

Collect all words (morphemes) of a given class representing space (location, direction, transition). Using the definitions and analytic means in the references, define and evaluate some of the following properties: (1) for the location system: fineness, orientation, symmetry, efficiency; (2) for the directional system: discrimination, discrimination entropy, symmetry, location-direction syncretism.

Compare the results with those from languages analyzed in the references. Generalize the results. Consider another class of units, compute the properties of this system and compare them with those of the class analyzed first. Is the representation of space in both classes equal?

Give a complete analysis of a language (i.e. all classes expressing location, direction or transition) and set up the spatial world-view of that language. If possible, join the results with psychological, ethno-historical or geographic background.

References

Altmann, G., Dömötör, Z., Riška, A. (1968). The partition of space in Nimboran. *Beiträge zur Linguistik und Informationsverarbeitung 12, 56-71.*

Altmann, G., Dömötör, Z., Riška, A. (1968). Reprezentácia priestoru v systéme slovenských predložiek. *Jazykovedný časopis 19, 25-40.* [German translation available]

3.6. Synonymy and morphological status of the word

Hypothesis

The simpler the word morphologically, the more synonyms it has. Test the hypothesis.

Procedure

According to the morphological status words can be classified or scaled as simple, derived, reduplicated, compound, reduplicated compound, compound with derivation. Take a random sample of words from a monolingual dictionary and sort them into the above classes, which can be scaled ordinally. Then consult a dictionary of synonyms and count the number of synonyms of each word in the list. For each class, compute the average number of synonyms and test the hypothesis that the number of synonyms of a word is a function of its morphological complexity, i.e. *No. of synonyms = f*(*morphological complexity*). If the trend is not linear, find an appropriate function and substantiate it. Use the concept of specification to obtain a synergetic-linguistic background (Köhler 2005).

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: de Gruyter.

3.7. Word senses (1)

Problem

Determine the word sense distributions in texts.

Procedure

Compile word lists for a number of texts in a lexically disambiguated corpus with word sense annotation. Determine the probability distribution of the senses with respect to each word and each text.

(a) Do your findings confirm those reported in the literature about linguistic diversification (cf. the references)?

(b) Can you find interrelations between the distributions you obtained and

i. text length,
ii. text type,
iii. part-of-speech of the word,
iv. word frequency,
v. word age, or
vi. word length?

If you find an interrelation, express it by means of a simple function.

Compare your results with those in Problem 3.2 concerning the dictionary.

References

Best, K.-H. (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18, 26-31*.

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-35.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 97-111.*

Rothe, U. (ed.) (1991). *Diversification process in language: grammar*. Hagen: Rottmann.

Sanada, H. (2009). *Diversification of postpositions in Japanese*. Msc.

3.8. Word senses (2)

Problem

Determine the number of senses of the words as realised in texts and thematic domains.

Procedure

Collect data on word polysemy from a semantically annotated text corpus, i.e. the number of word senses used in the texts under study (as opposed to dictionary-based polysemy). Determine frequency and probability distributions as far as you find sufficiently ambiguous words. Compare your findings to the polysemy of the words as given in a dictionary with respect to individual texts or to thematic domains. In case of disagreement, give grounds for this fact.

References

Arapov, M.V. (1987). Upotrebiteľnosť i mnogoznačnosť slova. *Učenye Zapiski Tartuskogo Universiteta 774, 15-28.*

Levickij, V. (2005). Polysemy. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 456-464.* Berlin-New York: de Gruyter.

3.9. Distribution of word synonymy

Problem

Determine the distribution of word synonymy.

Procedure

Take a random sample of words from a dictionary and look up the words in a synonym dictionary or use Wordnet, Germanet or another electronic source – de-

pending on the language you study and on availability. Count the number of synonyms for each word and arrange these numbers according to their size.

Would you expect a specific probability distribution to fit to the data? Which one? First find an empirical distribution (or simple function) and perform a fit and a goodness-of-fit test. If the result is positive, substantiate the function, set up an appropriate differential or difference equation and formulate the hypothesis. This is a usual inductive-deductive approach, very helpful in creating theories.

References

Wimmer, G., Altmann, G. .(2001). Two hypotheses on synonymy. In: Ondrejovič, S., Považaj, N. (eds.), *Lexicographica ´99: 218-225*. Bratislava: Veda.
http://wordnet.princeton.edu/
http://www.sfs.uni-tuebingen.de/GermaNet/

3.10. Synonymy and polysemy

Problem

Find the interrelation between word synonymy and polysemy.

Procedure

Obtain a random sample of words belonging to the same part-of-speech from a monolingual dictionary and obtain the number of their meanings (polysemy). Adhere to the way in which polysemy is marked in the dictionary. Then take each of these words and find the number of its synonyms in a synonymy dictionary or in Wordnet. Then show that the mean number of synonyms is a power function of the number of meanings. This relationship follows from Köhler's synergetic control cycle (1986, 2005) and has been corroborated only once, in Italian.

Study different languages to give the hypothesis a more general validity.

References

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

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin-New York: de Gruyter.

Ziegler, A., Altmann, G. (2001). Beziehung zwischen Synonymie und Polysemie. In: Ondrejovič, S., Považaj, M. (eds.) (2001). *Lexicographica ´99: 226-229.* Bratislava: Veda

3.11. Synonymy, length and frequency of words

Problem

Find the interrelation between synonymy and length, and synonymy and frequency of words.

Procedure

First, set up hypotheses about the possible interrelation between (a) frequency and number of synonyms of a word, (b) length and number of synonyms of a word. Formulate these hypotheses in terms of mathematical functions. Then take a random sample of words from a dictionary and determine their lengths. The frequencies of words can be obtained in a text corpus. Synonymy can either be taken from a dictionary of synonyms or automatically by consulting an electronic source such as Wordnet or Germanet, etc. Fit the functions to the data and calculate the determination coefficients R^2 to assess the goodness-of-fit.

Interpret the results. Try to embed the result in the control cycle of previous problems.

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: de Gruyter.

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

4. Lexicology

4.1. Definition chains (verbs and adjectives)

Hypothesis

Verbs and adjectives have shorter hypernym chains than nouns. Test the hypothesis.

Procedure

In Problems Vol. 1, "Lexical chains", there was a task to set up hypernym chains of nouns and to measure the length of the chains. Thereby, a frequency distribution was obtained. Take a sample of 100 verbs and of 100 adjectives from a monolingual dictionary, set up their respective distributions of hypernym chain lengths, and for each of them separately show whether

(a) the tails of the distributions are as long as those of nouns. For German and Polish nouns, the basic data can be found in Sambor, Hammerl (1991).

(b) Find a model of the form of the chain length distributions. Start from the assumption that the longer a chain, the smaller the probability that another hypernym will be added because every next (more general) hypernym is rather part of a technical vocabulary and therefore increases the encoding and memory effort (cf. Köhler 2005). If possible, base your derivation on the Wimmer-Altmann's Unified Theory.

(c) Scrutinize the problem why nouns, verbs and adjectives have different lengths of definition chains. Give not only linguistic arguments; take recourse to other sciences (e.g. biology, physics) too.

References

Ballmer, T.T, Brennenstuhl, W. (1986). *Deutsche Verben: eine sprachanalytische Untersuchung des deutschen Verbwortschatzes.* Tübingen: Narr.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 760-774.* Berlin/New York: de Gruyter.

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

Wimmer, G. Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 648-659*. Berlin/New York: de Gruyter.

4.2. Survival of word classes

Problem

The members of some parts-of-speech classes (e.g. pronouns) have a stronger survival persistence than those of other classes. Using the Romance data published by M.A. Kapitan (1994) set up (1) an indicator of survival persistence, (2) an indicator of survival homogeneity.

Procedure

The data of M.A. Kapitan are presented in reduced form in Table 4.2.1

Table 4.2.1
Survival of parts of speech in Romance languages (the first 1000)
(Kapitan 1994)

Parts of speech	Latin	Romanian	Italian	French	Spanish	Portuguese
N	355	102	196	167	177	183
V	323	85	136	108	138	138
ADJ	178	49	84	70	77	79
ADV	87	11	19	15	19	18
NUM	5	5	4	4	5	5
PRON	6	5	5	4	5	5
PREP	13	7	6	7	8	8
CONJ	26	6	7	6	5	5
INTERJ	3	1	1	1	1	1

The survival persistence is given by the mean of inherited words; survival homogeneity is given by the extent of equality of survival proportions. Propose some indicators, if possible, also their variances, and show the order of parts-of-speech according to (1) and (2). Study also other language families. If there are other parts of speech, modify the classification. Interpret all results and substantiate them.

References

Kapitan, M.E. (1994). Influence of various system features of Romance words on their survival. *Journal of Quantitative Linguistics 1(3), 237-250.*

4.3. Frequency and survival of words

Hypothesis

"The more frequent a word is, the more chances of survival it has" (Kapitan 1994: 242). Test the hypothesis.

Procedure

The simplest way to obtain data is the examination of texts in two historically different stages of the same language. Another possibility is to analyze texts in a classical language and its descendants, e.g. Latin and some present-day Romance languages. The latter kind of data was furnished by Kapitan (1994) who coded the frequencies in intervals $(2^n, 2^{n+1}\rangle$, the most frequent words in the interval $<2^0, 2^5>$. Thus in Table 4.3.1 we find intervals of rank-order which can be presented as ordinal numbers. The table gives the number of words surviving in individual classes in five Romance languages.

Table 4.3.1
Number of survived Latin words in individual frequency classes up to 1000 in five Romance languages (Kapitan 1994: 242)

Frequency range		Latin		Romanian	Italian	French	Spanish	Portuguese
<1-32>	1	32		21	25	22	25	25
<33-64>	2	32		18	18	18	17	17
<65-128>	3	64		33	45	38	41	41
<129-256>	4	128		42	64	60	65	64
<257-512>	5	256		71	132	103	119	119
<513-995>	6	483		88	175	142	169	177

(1) Find a function describing the decrease of the *relative* number of survivals and try to give reasons for its form. Pay attention to the third class which deviates from the decreasing trend.
(2) Scrutinize the homogeneity of survivals in the five languages.
(3) Does one obtain the same function if one determines different frequency intervals? Generalize the problem.

References

Arapov, M.V., Cherc, M.M. (1974). *Matematičeskie metody v istoričeskoj lingvistike*. Moskva: Nauka [*Matematische Methoden in der historischen Linguistik*. Bochum: Brockmeyer 1983]

Kapitan, M.E. (1994). Influence of various system features of Romance words on their survival. *Journal of Quantitative Linguistics 1(3), 237-250.*

4.4. Word class distributions 2[1]

Hypothesis

Word classes (parts-of-speech etc.) display a regular rank-frequency sequence. Test the hypothesis of goodness-of-fit with some function and generalize to any kind of word classes.

Procedure

In Problems Vol. 1, p.28 the rank-frequency sequence was presented as a distribution; here we propose to find the "best" plain or probabilistic function expressing the regular representation of word classes. The simplest way is to consider parts-of-speech. The first problem is to collect all published data. For parts-of-speech cf. Best (1994, 1997, 2000, 2001), Hammerl (1990), Judt (1995), Schweers, Zhu (1991), Tuzzi, Popescu, Altmann (2009), Zhu, Best (1992), Ziegler (1998, 2001), for inflection classes Belonogov (1964), for verb forms Bull (1947), Robbins (1926), for tenses Hills, Anderson (1929, 1930), for personal pronouns Hills, Anderson (1931), for affixes Pierce (1961, 1962), Veenker (1968, 1969, 1973, 1975, 1976), etc. The amount of literature on this topic is very extensive.

Find the best function for all classes. If necessary, modify or generalize a function. The following main functions are commonly applied:

Zipf's zeta function: $f(r) = \frac{C}{r^a}$, $r = 1,2,3,\ldots$

Mandelbrot's function: $f(r) = \frac{C}{(r+a)^b}$, $r = 1,2,3,\ldots$

Zipf-Alekseev function: $f(r) = ar^{-b-c\ln r}$, $r = 1,2,3,\ldots$

Altmann function: $f(r) = \frac{\binom{b+r}{r-1}}{\binom{a+r}{r-1}} f(1)$, $r = 1,2,3,\ldots$

[1] Cf. the same problem in Problems Vol. 1, p. 28. Here, it will be generalized.

Negative hypergeometric distribution:

$$P(r) = \frac{\binom{M-r-2}{r-2}\binom{K-M+n-r}{n-r+1}}{\binom{K+n-1}{n}}, \quad r = 1,2,...,n+1$$

Various other functions used as models of phoneme/letter and word frequencies can be tested, too. Find the best empirical result, substantiate linguistically the given function and connect it with Wimmer-Altmann's (2005) general theory.

Consider the first four functions as simple sequences (i.e. non-normalized); the last one is a regular distribution.

References

Belonogov, G.G. (1964). Raspredelenie častot pojavlenija flektivnych klassov russkich slov. *Problemy kibernetiki 11, 189-198.*

Best, K.-H. (1994). Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1(2), 144-147.*

Best, K.-H. (1994). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. *Glottometrika 16, 276-285.*

Best, K.-H. (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge tur Sprachwissenschaft 4, 37-51.*

Best, K.-H. (2001). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1, 1-26.*

Bull, W.E. (1947). Modern Spanish verb-form frequencies. *Hispania 30, 451-466.*

Hammerl, R. (1990). Untersuchung zur Verteilung der Wortarten im Text. *Glottometrika 11, 142-156.*

Hills, E.C., Anderson, J.O. (1929). The frequency of moods and tenses of verbs in recent Spanish plays. *Hispania 12, 604-606.*

Hills, E.C., Anderson, J.O. (1930). The frequency of verbs and tenses in recent Spanish plays. *Hispania 13, 413-416.*

Hills, E.C., Anderson, J.O. (1931). The relative frequency of Spanish personal pronouns. *Hispania 14, 335-337.*

Judt, B. (1995). *Wortartenhäufigkeiten im Deutschen und Französischen.* Göttingen: Staatsexamensarbeit.

Pierce, J.E. (1961). A frequency count of Turkish affixes. *Anthropological Linguistics 3, 31-42.*

Pierce, J.E. (1962). Frequencies of occurrence of affixes in French. *Anthropological Linguistics 6, 30-41.*

Robbins, F.E. (1926). Statistics of Greek verb-forms. *Classical Journal 15, 101-108.*

Schweers, A., Zhu, J. (1991). Wortartenklassifizierung in Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (ed.), *Diversification processes in language: grammar: 157-165.* Hagen: Rottmann.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, (in print).*

Veenker, W. (1975). *Verzeichnis der ungarischen Suffixe und Suffixkombinationen.* Hamburg: Societas Uralo-Altaica.

Veenker, W. (1969). *Vogul suffixes and pronouns. An index a tergo*. The Hague: Mouton.

Veenker, W. (1973). *Verzeichnis der ostostjakischen (Vach) Suffixe und Suffixkombinationen (unter Einschluß der wichtigsten Pronomina).* Hamburg: Societas Uralo-Altaica.

Veenker, W. (1975). *Verzeichnis der čeremissischen Suffixe und Suffixkombinationen.* Hamburg: Finnisch-Ugrisches Seminar.

Veenker, W. (1976). *Verzeichnis der votjakischen Suffixe und Suffixkombinationen.* Hamburg: Finnisch-Ugrisches Seminar.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 791-907.* Berlin-New York: de Gruyter.

Zhu, J., Best, K.-H. (1992). Zum Wort im modernen Chnesisch. *Oriens Extremus 35, 45-60.*

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5, 269-280.*

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 295-312*. Trier: Wissenschaftlicher Verlag.

4.5. Vocabulary comparisons

Problem

Collect all methods used in comparing the vocabularies of two texts and evaluate them.

Procedure

The first part of the work is quite mechanical. Consult Köhler's Bibliography of Quantitative Linguistics (1995) and find all works touching the theme. The second stage is the Internet search for various keywords (authorship attribution, intertextual distance,…). Some recent works are Rudman (1998), Merriam (2003), Labbé (2007), where further references can be found.

When you dispose of a list of methods and formulas, elaborate on the mathematical properties of the proposed formulas, e.g. determine the domains of coefficients, derive their variances, and study their behaviour with increasing sample size, especially if you work with asymptotic quantities..

Compare two texts by means of all collected similarity indicators. Evaluate their effectiveness. If you compare more than two texts, do not perform any classifications or attribution studies. If you have comparable texts in time succession, e.g. Nobel-Prize lectures in literature, end-of-year speeches of presidents etc., show that similarity is linked with time distance. Which index can express this dependence in the best way?

Analyze all existing argumentations concerning authorship attribution, show their weak points, collect existing criticisms and systematize this domain.

References

Brunet, E. (1988). Une mesure de la distance intertextuelle: la connexion lexicale. Le nombre et le texte. *Revue informatique et statistique dans les sciences humaines 24(1-4), 81-116.*

Kļaviņa, S.P. (1977). *Sopostavlenie funkcionaľnych stilej latyšskogo jazyka (lingvostatističeskoe issledovanie).* Vilnius: Diss.

Köhler, R. (1995). *Bibliography of quantitative linguistics*. Amsterdam-Philadelphia: Benjamins.

Labbé, C. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics 14(1), 33-80.*

Labbé, C., Labbé, D. (2001). Inter-textual distance and authorship attribution Corneille and Molière. *Journal of Quantitative Linguistics 8, 212-231.*

Merriam, T. (2003). An application of authorship attribution by intertextual distance in English. *Corpus 2, 167-182.*

Muller, Ch. (1968). *Initiation à la statistique linguistisque.* Paris: Larousse.

Muller, Ch. (1977). *Principes et méthods de statistique lexicale*. Paris: Hachette université.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities 351-365.*

Tuldava, J. (1971). Statističeskij metod sravnenija leksičeskogo sostava dvuch tekstov. *Linguistica 4, 199-220.*

Tuldava, J. (1998). *Probleme und Methoden der quantitative-systemischen Lexikologie.* Trier: Wissenschaftlicher Verlag.

Viprey, J.-M., Ledoux, C.N. (2006). About Labbé's “Intertextual Distance”. *Journal of Quantitative Linguistics 13(2-3),265-283.*

4.6. Word commonness

Problem

In the Czech corpus-linguistic tradition, the concept of word commonness has been introduced (cf. especially Savický, Hlaváčová 2002). Show that this is the same as polytextuality operationalized in a different way.

Procedure

Savický and Hlaváčová consider the corpus an uninterrupted sequence of words. They partition this sequence in segments of equal length and observe the occurrence of a given word in these segments. This is exactly the setting of Frumkina's law (cf. Problems Vol. 1, Chapter 9 and the references therein), here generalized from a single text to an entire corpus. However, in this new form it conforms even better with the concept of polytextuality. The authors use the empirical mean of occurrences (ARF) to measure word commonness. Show that this measure is identical with the expectation (mean) of the negative hypergeometric distribution representing Frumkina's law.

Another measure of commonness introduced by Savický and Hlaváčová (2002) is the average logarithmic distance

$$ALD = \frac{1}{N}\sum_{i=1}^{f} d_i \log_{10} d_i ,$$

where the d_i are the distances between occurrences of one and the same word, f is the word's frequency and N is the sum of the distances. Consider d_i a geometrically distributed variable and derive the expectation of ALD.

References

Altmann, G., Burdinski, V. (1982). Toward a law of word repetitions in text-blocks**.** *Glottometrika 4, 147-167.*

Frumkina, R.M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T.N. (ed.), *Strukturno-tipologičeskie issledovanija: 124-133*. Moskva: ANSSSR.

Savický, P., Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics 9(3), 215-231.*

4.7. Indicator of association

Problem

In grammar, lexicology and textology one often uses an association index originating from information theory, viz.

$$I(w_1, w_2) = \log\left(\frac{N \times f(w_1, w_2)}{f(w_1) \times f(w_2)}\right)$$

where w_1 and w_2 are two different words (or other entities), $f(w_i)$ (i = 1,2) is the frequency of the words in a corpus of size N (which can be skipped in the formula) and $f(w_1,w_2)$ is the common occurrence of these words. Derive the variance of this indicator.

Procedure

If you retain N, then first define exactly whether N is the number of words or sentences in the text. Then define the term *co-occurrence*: are you going to examine immediate neighbourhood or common occurrence in a sentence – indirect or direct neighbourhood. Then derive the variance using the Taylor expansion method.

Set up an asymptotic test for the significance of the indicator on the basis of the variance and test all associations of one word. Order them according to the strength of association (quantile of the normal distribution) and determine a boundary at which the co-occurring elements can be considered a compound.

References

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

Church, K., Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics 16, 22-29.*

Cramér, H. (1946*). Mathematical methods in statistics*. Princeton: Princeton University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19, 61-74.*

Ferrer i Cancho, R., Reina, F. (2002). Quantifying the semantic contribution of particles. *Journal of Quantitative Linguistics 9(1), 35-47.*

Han, D., Ito, T., Furugori, T. (2002). Structural analysis of compound words in Japanese using semantic dependency relations. *Journal of Quantitative Linguistics 9(1), 1-17.*

Han, D., Kato, K., Furugori, T. (2001). Automatic segmentation of compound words in Japanese using contextual information. *Technical Report of IEICE, NLC 2001-05, 29-34* (in Japanese).

Hurt, J. (1976). Asymptotic expansion of functions in statistics. *Aplikace Matematiky 21, 444-456.*

Kabayashi, Y., Tolunaga, T., Tanaka, H. (1994). Analysis of Japanese compound nouns using collocational information. In: *Proceedings of the 15ht International Conference on Computational Linguistics (COLING´94): 865-869.* Kyoto.

Kendall, M.G., Stuart, A. (1969). *The advanced theory of statistics I,II.* London: Griffin.

Li, W. (1989). *Mutual information functions of natural language texts*. Santa Fe Institute Working Papers 008.

Manning, Chr.D., Schütze, H. (1999). *Foundations of statistical naural language processing*. Cambridge (Mass.), London: MIT Press.

Oehlert, G.W. (1992). A note on the delta method. *The American Statistician 46(1), 27-29.*

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

4.8. Word stability

Problem

A concept akin to word commonness (cf. Problem 4.6) is that of statistical word stability. The main difference is that word commonness is based on text passages of equal lengths whereas word stability refers to paragraphs. Compute the stability of words in a single text and in an entire corpus.

Procedure

The coefficient of statistical word stability was introduced by Marusenko (1983) in the form

$$WS = \frac{F_w \times m_w}{N \times n},$$

where

F_w = frequency of a given word *w* in the text
m_w = number of paragraphs containing word *w*
N = number of words in the text
n = number of paragraphs in the text.

This coefficient can be used for estimating the importance of a word (cf. Zubov 2004) and text content.

Determine the interval of *WS*, derive its expected value and in any case its variance. Consider *N* and *n* constant characteristics of a given text, i.e., you need to conjecture the distributions of F_w and m_w.

Even if the attempt to find the theoretical background does not succeed, perform the following analyses: (a) Compute *WS* for each word of the text, order the words according to decreasing *WS*. (b) Make statements about the positioning of different parts of speech. (c) Derive the distribution of *WS* theoretically and compare it with your empirical results. (d) Compare *WS* with other coefficients of this kind.

What is the difference between *WS* and Frumkina's law?

References

Abracos, J., Lopes, G.P. (1997). Statistical methods for retrieving most significant paragraphs in newspaper articles. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization: 51-57.*

Akišina, O.V. (2001). Formaľnoje vyraženie osnovnogo soderžanija anglojazyčnogo reklamnogo teksta. In: Zubov, A.V. (ed.), *Materialy ežegodnoj naučnoj konferencii studentov i magistrantov universiteta, 5-6 aprelja 2000g. Časť vtoraja: 3-8*. Minsk: MGLU.

Alavi Džafar, A. (2000). Formaľnoe predstavlenie osnovnogo soderžanija anglijskich korotkich raskazov. In: Zubov, A.V. (ed.), *Materialy ežegodnoj naučnoj konferencii prepodavaelej i aspirantov universiteta, 5-6 aprelja 2000g. Časť treťja: 3-8*. Minsk: MGLU.

Bolšakova, Ju.G. (2000). Statistika v ocenke osnovnogo soderžanija tekstov-opisanij finansovych operacij. In: Zubov, A.V. (ed.), *Materialy ežegodnoj naučnoj konferencii prepodavatelej i aspirantov universiteta, 5-6 aprelja 2000g. Časť vtoraja: 6-7*. Minsk: MGLU.

Brandow, R., Mitze, K., & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management, 31(5), 675–685.*

Čaplja, A.I., Čaplja, S.G., Zubov, A.V. (1973). Avtomatičeskij otbor ključevych i poliključevych slov. In: Čaplja, A.I. (ed.), *Sbornik naučnych soobščenij fakulteta inostrannych jazykov: 76-93.* Machačkala: DGU.

Marusenko, M.A. (1983). O formirovanii slovnika slovarja statističeski ustojčivych naučno-techničeskich terminov. In: *Strukturnaja i prikladnaja lingvistika. Mežvuzovskij sbornik. Vypusk 2, 82-89.*

Zechner, K. (1996). Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In: *Proceedings of the International Conference on Computational Linguistics, Copenhagen: 986-989.*

Zubov, A. (2004). Formalization of the procedure of singling out the basic text contents. *Journal of Quantitative Linguistics 11(1-2), 33-48.*

Zubov, A.V., Čaplja, A.I., Čaplja, S.G.(1978). Avtomatičeskij otbor ključevych slov. In: *Strukturnaja i prikladnaja lingvistika. Vypusk I: 198-205.* Leningrad: LGU.

4.9. Word length and meaning generality

Problem

Is there a relationship between length and meaning generality of a word?

Procedure

Prepare about 100 hypernym chains using a monolingual dictionary according to the following instructions:

A hypernym of a basic lexeme A is another, more general lexeme forming a class to which A belongs. E.g., *furniture* is a hypernym of *chair; building* is a hypernym of *skyscraper*. A hypernym is usually part of the definition of the meaning in a monolingual dictionary. Consider only nouns which form hypernymic chains. WordNet and similar electronic sources provide hypernymic chains for some languages; for other languages such chains have to be formed by the researcher. The general procedure we recommend is as follows.

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

The degree of generality can be estimated as the mean of the levels on which the noun was found. E.g. in the chain *hammer – instrument – equipment – thing*, the noun *instrument* would obtain value 2 of generality. If a word occurs in more than one chain the mean of its generality values may be used.

Compute the mean generality (x) and length (y) of each word and find a clear relationship. If your data oscillate heavily, increase the size of the sample. Conjecture the form of the tendency if there is any.

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 Linguistik. Ein internationales Handbuch. 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.*

5. Textology

5.1. Belza-Skorochod´ko´s chaining coefficient

Hypothesis
In technical texts, the chaining connection of sentences is greater than in poetic texts. Test the hypothesis.

Procedure
The chaining coefficient C_T reflects the tendency to form uninterrupted sequences of coherent, i.e., semantically connected sentences. This concept is operationelized as the mean length of such sentences in a text or text collection. It is, therefore, defined as

$$C_T = \frac{1}{s} \sum_{i=1}^{s} k_i ,$$

where k_i represents the length of the i-th chain (there may be chains of length 1) and s is the number of all chains in the text.

Whether consecutive sentences are semantically connected or not can be operationalized in different ways, on the basis of co-reference in general, anaphorical co-reference, etc.

To test the hypothesis, define and operationalize the notion of semantic connection (you may try several variants) and collect a number of pragmatically homogeneous texts, i.e. texts of the same text sort, genre, thematic field, etc. Then calculate the C_T values of the texts in your collection.

According to Belza (1971), Russian technical texts have C_T = 7.4, popular scientific ones 6.6, newspaper or belletristic texts 5.3. That means, in Russian technical texts, a coherent chain is formed, on the average, by 7.4 sentences.

Compare similar text types in different languages and perform tests of equality. The C_T coefficient is a simple mean, hence a t-test can be applied for comparisons.

Alternative methods to measure coherence can be found in the references.

References
Bateman, J., Rondhuis, K. (1994). *Coherence relations: Analysis and specifications.* Darmstadt: GMD-IPSI (Technical report)

Beaugrande, R.-A. de, Dressler, W.U. (1981). *Introduction to text linguistics*. London/New York: Longman.

Belza, M.I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury svjaznych tekstov. In: *Semantičeskie problemy avtomatizacii informacionnogo potoka: 58-73.* Kiev.

Dijk, T.A. van, Kintsch, W. (Eds.). (1983). *Strategies of Discourse Comprehension*. New York/London et al.: Academic Press.

Foltz, P.W. (1996). Comprehension, Coherence, and Strategies in Hypertext and Linear Text. In: J.-F. Rouet, J.J. Levonen, A. Dillon, R.J. Spiro (Eds.), *Hypertext and Cognition: 109-136*. Mahwah/New Jersey: Lawrence Erlbaum Associates Publishers.

Fritz, G. (1982). *Kohärenz: Grundfragen der Dialoganalyse.* Tübingen: Narr.

Fritz, G. (1999). Coherence in Hypertext. In: Bublitz, W., Lenk, U., Eija, V. (Eds.), *Coherence in Spoken and Written Discourse.* (pp. 221-232). Amsterdam/Philadelphia: John Benjamins.

Haliday, M.A., Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hobbs, J.R. (1979). Coherence and coreference. *Cognitive Science 3, 67-90.*

Hobbs, J.R. (1985). *On the coherence and structure of discourse.* Stanford, CA: Center for the Study of Language and Information Technical Report 85-37.

Rickheit, G., Schade, U. (2000). Kohärenz und Kohäsion. In: Brinker, K. Antos, G., Heinemann, W., Sager, S.F. (Eds.), *Text- und Gesprächslinguistik -- ein internationales Handbuch zeitgenössischer Forschung. 1. Halbband: 275-282*. Berlin/ New York: de Gruyter.

Schade, U., Langer, H., Rutz, H., Sichelschmidt, L. (1991). Kohärenz als Prozeß. In: G. Rickheit (Ed.), *Kohärenzprozesse. Modellierung von Sprachverarbeitung in Texten und Diskursen*: 7-58. Opladen: Westdeutscher Verlag.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung bei Wissenserwerb mit Texten*. Weinheim: Beltz.

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

Strohner, H., Rickheit, G. (1990). Kognitive, kommunikative und sprachliche Zusammenhänge: Eine systemtheoretische Konzeption linguistischer Kohärenz. *Linguistische Berichte 3-23.*

Stutterheim, C. von (1997). *Einige Prinzipien des Textaufbaus. Empirische Untersuchungen zur Produktion mündlicher Texte*. Tübingen: Niemeyer.

Wolf, F., Gibson, E. (2005). Representing discourse coherence: A corpus-based analysis. *Computational Linguistics 31(2), 249-287.*

Wolf, F., Gibson, E. (2005). *Coherence in natural language. Data structures and applications*. Cambridge, Mass: MIT Press.

5.2. Crowding of autosemantics

Problem

Show that the autosemantic words display a special kind of crowding in fixed rank intervals in the rank-frequency distribution of word forms in a text.

Procedure

Determine the frequencies of occurrence of word forms (or lemmas) in a text and set up the rank-frequency distribution. Compute the *h*-point and partition the text into intervals of length *h*. Then count the number of autosemantics in the individual intervals. A monotone increasing sequence is obtained, which can be captured by the function

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

(Popescu et al. 2009). In short texts, the sequence is not very smooth. Now, having *h* and *a*, define the indicator

$$APF = a/h \quad (APF = \textit{autosemantic pace filling})$$

and compute it for several texts. This index characterizes the usage of autosemantics in a given text. At the same time, the indicator

$$AC = ak$$

yields the *autosemantic compactness* of the text (Popescu et al. 2009).

By application of these indicators, the linguistic evolution of a writer, of genres, styles, even differences between languages can be studied. A test for *APF* can be easily set up.

Using the above formulas continue developing an appropriate theory of positioning and frequency of autosemantics in texts.

Reference

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gryuter.

5.3. Semantic reduction in texts

Project

A certain proportion of the words in a text are polysemantic. The extent of polysemy can be obtained by means of a good dictionary, e.g. by counting the individual meanings marked by Arabic or Roman numbers or by letters. Other measurement procedures are not excluded, e.g. using a mega-corpus and studying the meaning of a word in each of its contexts.

Words occuring in a text are not isolated but embedded in their specific co-text and context, by which the words are disambiguated; their sematic potential being reduced and specified. The co-textual disambiguating effect may come from close (affixes, compounds, reduplications), or more distant (phrasal) elements; contextual effects are due to co-referents or the referred situation. The reduction of polysemy causes in most cases a sequence of words to become more or less monosemic; i.e., the polysemy values of subsequent words in a text is usually a sequence 1,1,1,…. Contrasting such a sequence against a polysemy sequence whose values are taken from the polysemy of the words in a dictionary, shows the extent of the reduction. It can be expected, due to fixed terminology, that the reduction is smaller in scientific texts, and greater in poetry where often words evoking strong imagination are used. Hence, a text can be presented in form of a sequence of semantic reductions.

The resulting sequences can be called *semantic sequences*. Such sequences have many global properties, but they can also be partitioned into *P*-segments in different ways. Such a segment can be defined in the following way: A *P*-segment is a non-decreasing sequence of numbers, e.g. 1,1,1,2,8. It is analogous to other sequences of this kind e.g. length sequences, frequency sequences, polytextuality sequences, etc. introduced by R. Köhler and L. Uhlířová (cf. the references). Such sequences have some interesting properties: they partition each text in an unambiguous and exhaustive way, they have a granularity between words and syntactic contructions, and they capture syntagmatic test structures.

This state of affairs generates four problems:

(1) What is the mean extent of monosemization in text and how is monosemization distributed?

(2) What are the properties of the complete sequence of reduction degrees in text?

(3) What are the properties of *P*-sequences, e.g. frequencies, lengths, combinations, distributions?

(4) Can we mechanically distinguish genres using these properties? Evidently, the solution of all these problems is a task for a team of researchers, hence we shall discuss them only theoretically.

The second package of problems arises from the fact that each word in a dictionary belongs to at least one class of words sharing some semantic compon-

ents with it. However, if they are used in a text, some (word-class identifying) components may become lost because words in texts are used either generally, or they may be specified by deixis, articles, context, prepositions, etc. Languages differ in using different means. Again, the same package of problems arises but this time concerning the loss or increase of the number of components.

Practically all properties of linguistic units that can be quantified – or at least nominally classified – can be used to set up symbolic or numerical sequences. If these sequences can by a definition be partitioned in shorter subsequences, called segments, we obtain abstract linguistic units, void of form and meaning but expressing a sequence of quantified properties. Since all properties of units can be captured in this way, a new discipline can be initiated.

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: 145-152*. Bratislava: Slovak Academic Press.

Köhler, R. (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe: 416-421*. Bratislava: Veda.

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory 1(1), 115-119.*

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications: 637-646*. Berlin, Heidelberg: Springer,.

Uhlířová, L. (2009). Word frequency and position in sentence (To appear).

5.4. Rank-frequency distribution and arc length

Problem

Show that the arc length of rank-frequency distributions is correlated with entropy.

Procedure

Compute the rank-frequency distribution of word forms in several texts. Compute the relative frequencies and using them compute the Shannon-entropy

$$H = -\sum_{x=1}^{V} p_x \log_2 p_x$$

where V is the number of different word forms and p_x the relative frequencies. Then compute the arc length of the distribution using

$$L = \sum_{x=1}^{V-1} [(f_x - f_{x+1})^2 + 1]^{1/2},$$

where f_x are the absolute frequencies. Show that there is at least a correlation between H and L, and show the form of the relation.

References

None.

5.5. Popescu´s vocabulary richness

Problem

Evaluate the word frequencies in different texts and compute the vocabulary richness using Popescu´s indicators (Popescu et al. 2009).

Procedure

Vocabulary richness can be evaluated in many different ways. First study the available literature and prepare a survey of indicators and procedures. Evaluate their advantages and disadvantages. Then use all indicators presented in Popescu et al. (2009) and compare your results with those presented in the book.

Perform tests of equality, classify your texts according to richness in different classes and interpret the results.

References

Brunet, É. (1978). *Le vocabulaire de Jean Giraudoux. Structure et evolution*. Genève: Slatkine.

Cossette, A. (1994). *La richesse lexicale et sa mesure*. Paris: Champion.

Dugast, D. (1980). La mesure de la richesse lexicale: une esquisse historique. *Verbum 3(1), 115-134.*

Honore, T. (1979). Some simple measures of richness of vocabulary. *ALLC Journal 7, 172-177.*

Kuraszkiewicz, W. (1963). *La richesse du vacabulaire dans quelques grands texts polonaise en vers.* Wroclaw: Ossolineum.

Ménard, N. (1983). *Mesure de la richesse lexicale*. Paris: Slatkine.

Ménard, N., Santerre, L. (1979). La richesse lexicale individuelle comme marquer sociolinguistique. *Cahiers der linguistique 1, 165-188.*

Muller, Ch. (1968). Mesure de la richesse lexicale. *Travaux de linguistique et de littérature 6, 73-84.*

Muller, Ch. (1971). Sur la mesure de la richesse lexicale. Théorie et experience, homage à René Michéa. *Études de linguistique appliqué 1971, 74-87.*

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Ratkowsky, D.A., Hantrais, L. (1975). Tables for comparing the richness and structure of vocabulary in texts of different lengths. *Computers and Humanities 9, 69-75.*

Ratkowsky, D.A., Halstead, M.H., Hantrais, L. (1980). Measuring vocabulary richness in literary works. A new proposal and re-assessment of some earlier measures. *Glottometrika 2, 125-147.*

Tešitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics 3, 103-120.*

Thoiron, P. (1986). Indice de diversité et mesure de la richesse lexicale. In: Muller, Ch. (Festschrift), *Méthodes quantitatives et informatiques dans l´étude des texts. Colloque international de CNRS à l´Université de Nice, 5-8 juin 1985: 831-840.* Paris: Champion.

Thoiron, P., Labbé, D., Serant, D. (eds.) (1988*). Études sur la richesse et la structure lexicale.* Paris-Genève: Champion-Slatkine.

Wimmer, G., Altmann, G. (1999). Review article: On vocabulary richness. *Journal of Quantitative Linguistics 6(1),1-9.*

5.6. Alliteration

Hypothesis

Each poem displays some degree of non-random alliteration.

Procedure

Alliteration is the repetition of equal phonemes (sounds, letters,…) at the beginning of words in a verse line. In order to test the above hypothesis, determine the relative frequencies of the phonemes (sounds, letters,…) in the language under study, or, if not possible, use those from the given poem. Let the relative frequency of the phoneme *i* be p_i; let the number of words in a verse be *n*; let the number of words beginning with phoneme *i* be *r*. Then the probability of finding exactly *r* words beginning with *i* is given by

$$(1) \qquad P(X_i = r) = \binom{n}{r} p_i^r q_i^{n-r}$$

and the probability that *r* or more words begin with *i* is given by

$$(2) \qquad P(X_i \geq r) = \sum_{x=r}^{n} \binom{n}{x} p_i^x q_i^{n-x},$$

where q_i = 1 - p_i. If (2) is smaller than 0.05, accept the hypothesis that alliteration of the given phoneme in the given verse takes place. Test all verses for all phonemes occurring at least twice in the verse. Set up an index of alliteration. Compare individual poems, authors, lyrical and epical poems, try to discover a historical development.

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.

5.7. Alliteration structure

Hypothesis

Every poem has a non-random alliteration structure. Test the hypothesis.

Procedure

If there is more than one repeated phoneme at the beginning of words in a verse, e.g. phoneme *i* occurs at the beginning of two words and phoneme *j* at the beginning of three words then the procedure in "5.6. Alliteration" becomes a little more complex. With two repeated phonemes, we have to use the trinomial distribution, and in general we use the multinomial distribution. Let p_i be the probability of phoneme *i*, p_j the probability of phoneme *j* and $1-p_i-p_j$ the probability of the other ones. Then the probability that in a verse with *n* words there are exactly k_i words beginning with phoneme *i*, k_j words with phoneme *j* and $n- k_i - k_j$ words beginning with any other (non repeated) phoneme is given by

(1)

$$P(X_i = k_i, X_j = k_j, X_{n-k_i-k_j} = n - k_i - k_j) = \frac{n!}{k_i!\,k_j!\,(n-k_i-k_j)!}\, p_i^{k_i}\, p_j^{k_j} (1 - p_i - p_j)^{n-k_i-k_j}$$

To compute the given and the more extreme probability we add all probabilities to obtain

$$P(X_i \geq k_i, X_j \geq k_j, X_{n-k_i-k_j} = n - k_i - k_j) =$$

$$= \sum_{\substack{x_i \geq k_i \\ x_j \geq k_j}} \frac{n!}{x_i! x_j! (n - x_i - x_j)!} p_i^{x_i} p_j^{x_j} (1 - p_i - p_j)^{n - x_i - x_j} \quad (2)$$

Having computed the extent of alliteration for individual verses, study the course of alliteration in the poem and express the extent of alliteration using an appropriate indicator. Formula (2) can be used to compute the extent of alliteration at the beginning of verses, too.

Reference

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

5.8. Autosemantic dissortativity

Hypothesis

The associative graph of autosemantic words in a text is dissortative. Test the hypothesis.

Procedure

First replace all pronouns in a text by the words to which they refer to. Then eliminate from the text all auxiliaries, leave only autosemantics. Compute the coincidence of autosemantics within the sentence using the method in the problem "Association graph of the text" (Problems 1: 41) For each vertex (autosemantic word) compute its degree (= number of coincidences with other autosemantics). Determine whether words with high degree are coincident rather with words of high degree than with those of low degree. In the first case the graph is assortative; in the second case it is dissortative. Perform the test simply by computing the correlation of degress.

Show that texts of a special kind (e.g. scientific ones) are more assortative than poetic texts (or vice versa). Try to find the extent of assortativity for different text kinds and study the problem in a diachronic perspective.

References

Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. *arXiv:con-mat/0104209 v1, 11. Apr.2001.*

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

Newman, M.E.J. (2003). Mixing patterns in networks. *Physical Review E, 67, 026126*.
Newman, M.E.J., Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E, 68, 036122.*

5.9. Superhreb

Hypothesis

There are superunits between the levels of text and *Hreb* containing *Hrebs* as components.

Procedure

If we consider *Hrebs* as aggregates of sentences containing the same word or the same symbol or the same meaningful entity (also synonyms), etc., then according to Menzerath´s law there may exist a level consisting of *Superhrebs*, i.e. units which contains *Hrebs* as components. Establish units of this kind and find the respective level.

References

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.
Schwarz, C. (1996). The distribution of aggregates in texts. *ZET-Zeitschirft für Empirische Textforschung 2, 62-66. .*

5.10. Golden section (1)

Hypothesis

The radians of the "writer´s view" angle of the rank-frequency distribution are never smaller than the golden section 1.618.

Procedure

Compute the rank-frequency distribution of word forms in a text. Let r = rank, $f(r)$ = frequency at rank r. Compute the Hirsch-Popescu h-point according to the formula

$$h = \begin{cases} r, & \text{if there is an } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases}$$

i.e. $h = r$ if there is a rank r whose frequency $f(r)$ is equal to the rank; if not, take two neighbouring ranks r_i and r_j such that $r_i < f(r_i)$ and $r_j = r_i + 1 > f(r_j)$ and compute the second part of the formula. Join this point $P(h,h)$ using a straight line with the points $P(1,f(1))$, i.e. with the highest frequency, and $P(V,1)$, i.e. with the frequency of the highest rank (V = vocabulary or inventory). The angle α associated with the h-point is called "writer´s view" (cf. Popescu, Altmann 2007). Compute first cos α given as

$$\cos \alpha = \frac{-\left[(h-1)\left(f(1) - h\right) + (h-1)\left(V - h\right)\right]}{\left[(h-1)^2 + \left(f(1) - h\right)^2\right]^{1/2}\left[(h-1)^2 + \left(V - h\right)^2\right]^{1/2}}$$

then α and finally the radians

$$\alpha \text{ rad} = 2\pi\alpha/360.$$

α rad must be at least $\pi/2 = 1.57$ but it is always greater than 1.618, i.e. its lower limit for texts is the golden section.

Perform this investigation on many texts and test the hypothesis. At the same time, study the maximum of α rad. Theoretically, it is equal to π but its value is not known empirically.

References

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glotto-metrics 15, 2007, 71-81.*

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gryuter.

5.11. Strange attractor of the writer´s view

Hypothesis

The α radians of the "writer´s view" lie in a strange attractor.

Procedure

First solve the problem "5.10. Golden section" and collect results from many texts. Consider also the text lengths N. Enter the points <N, α radian> in a Cartesian coordinate system. One obtains a "boomerang"-like area, probably with several outliers. Find at least a function capturing the course of the points. If

possible, find the approximate area of the points by means of a system of two differential equations.

References

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 15, 2007, 71-81.*

5.12. Aristotle´s Categories

Problem

Analyze a text in terms of Aristotle´s categories.

Procedure

Consider the Aristotelian categories:

substance	- what is something? (desk, girl)
quantity	- how great is something? (two meters)
quality	- how is something? (able, green)
relation	- in which relation is it to something? (greater)
place	- where is it? (in the school)
time	- when is it? (today, tomorrow)
posture	- in which position is it? (it sits, it hangs)
possession/habit	- what does it have? (it is armed, it has a hat)
action	- what does it do? (it runs, it cuts)
passion (receiving)	- what does it suffer? (it is cut, burnt)

"Nowadays, these categories are commonly seen as having a value that is merely historical, in part because Aristotle's notion of substance is commonly rejected. This rejection often stems from a misunderstanding of his real meaning, which was that substance is that which exists of itself and not in another" http://en.wikipedia.org/wiki/Category_(philosophy) (Dec. 1, 2008).

Using these categories, a text is analyzed not into words but many times in syntagms, phrases etc. Register the number of individual categories and show the differences between texts. If the number of categories is not sufficient, establish further ones.

References

Aristotle (1953). *Metaphysics*. Ross, W. D., trans. Oxford University. Press.

Aristotle (2004). *Categories* Edghill, E. M., trans. University. of Adelaide Library.

5.13. The Skinner effect

Problem

If the appearance of linguistic entities in text is also self-stimulated in the sense that if a unit occurs, the probability of its occurrence in its close vicinity increases (Skinner 1939, 1941, 1957), then automatically the following hypothesis arises: Verses of a poem situated closer to each another are phonetically more similar than distant verses. Test the hypothesis.

Procedure

Perform either a phonetic transcription line by line, or a morphemic transcription of a long poem in a language of your choice. Define a measure of similarity between verses, compute the mean similarities for the distances 1,2,3,… and scrutinize whether the course of similarities decreases with increasing distance.

Perform the analysis using both artistic texts and folk poetry.

Can one conclude that a text following this regularity was created more spontaneously than a text not displaying this regularity?

Try to follow this phenomenon historically. Is alliteration a consequence of this phenomenon?

References

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

Bunde, A., Eichner, J.F., Kantelhardt, J.W., Havlin, S. (2005). Long-term memory: a natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Physical Review Letters 94, 048701*

Corral, A., Ferrer-i-Cancho, R., Diaz-Guilera, A. (2009). Universal complex structures in written language. *http://arxic.org/abs/0901.2924 (07,01,2009)*

Skinner, B.F. (1939). The alliteration in Shakespeare´s sonnets: A study in literary behavior. *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.*

Skinner, B.F. (1957). *Verbal Behaviour.* Acton: Copley Publishing Group.

5.14. The <I,J>- scheme

Problem

Obtain rank-frequency distributions of word forms from many texts. Plot the *I* and *J* indicators in a Cartesian coordinate system <*I,J*>.. Study the positions of

texts in different genres and also their position in historical succession. Describe your observations.

Procedure

The positioning of the *I* and *S* indicators of Ord (1972) for different distributional data is well known in linguistics (cf. also Problems Vol. 1, p. 111-112). Popescu, Mačutek, Altmann (2009) introduced a new possibility using the entropy of the distribution. Define

$$I = \frac{m_2}{m_1'} = \frac{s^2}{\overline{x}},$$

which is identical with that of Ord´s *I*. Since entropy in rank-frequency distributions is also an indicator of skewness, the coordinate *J* is defined as

$$J = \frac{H}{s_{\overline{x}}},$$

i.e. the entropy defined as $H = -\sum_{i=1}^{V} p_i \log_2 p_i$, where *V* is vocabulary size (and the highest rank), p_i are the relative frequencies and $s_{\overline{x}}$ is the standard deviation of the mean.

(1) Can you distinguish highly analytic from highly synthetic languages using this scheme?

(2) Can you clearly distinguish English texts from German or Slavic ones?

(3) Is there a development in the work of individual author?

(4) Can you demonstrate the development of German toward analytism on data from a historical corpus?

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: 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., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies.* Lüdenscheid: RAM.

5.15. Text Cohesion (1)

Problem

Determine the block distribution of anaphora and cataphora in texts of various kinds.

Procedure

In analogy to the block-wise distribution of function words (shown by Frumkina 1962 and others given in the references) and of syntactic constructions/functions (cf. Köhler 2001), determine the number of text blocks (try block sizes of 10, 30, 50, 100, words), in which there are 0,1,2,... occurrences of anaphora and cataphora. Fit the negative hypergeometric distribution ("Frumkina's Law") to the data. Observe dependences of the parameter values on block length, number of blocks, category type.

The negative hypergeometric distribution is defined as

$$P_x = \frac{\binom{M+x-1}{x}\binom{N-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}, \qquad x = 0,1,\ldots,n$$

where *K*, *M* and *n* are parameters.

Find out whether the parameters can be considered as text characteristics.

Under which special circumstances would it be possible to use the limiting cases of the negative hypergeometric, namely Poisson, binomial and negative binomial distributions?

If some parameters or their functions seem to be appropriate to express text cohesion, set up an indicator and show its properties.

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, 146-167.*

Bektaev, K.B., Luk´janenkov, K.F. (1971). O zakonach raspredelenija edinic pis´mennoj reči. In: Piotrowski, R.H. (ed.), *Statistika reči i avtomatičeskij analiz teksta: 47-112.* Leningrad: Nauka.

Brainerd, B. (1972). Article use as an indirect indicator of style among English-language authors. In: Jäger, S. (ed.), *Linguistik und Statistik: 11-32.* Braunschweig: Vieweg.

Francis, I.S. (1966). An exposition of a statistical approach to Federalist dispute. In: Leed, J. (ed.), *The computer and literary style: 38-78.* Kent, Ohio: Kent State University Press.

Frumkina, R.M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T.N. (ed.), *Strukturno-tipologičekie issledovanija; 124-133*. Moskva: Akademija Nauk SSSR.

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.). *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček*: *136-148.* Trier: Wissenschaftlicher Verlag.

Maškina, L.E. (1968). *O statističeskich metodach issledovanija leksiko-gramatičeskoj distribucii.* Minsk: Diss.

Mosteller, F., Wallace, D.L. (1964). *Inference and disputed authorship: The Federalist*. Reading, Mass.: Addison-Wesley.

Paškovskij, V.E., Srebrjanskaja, I.I. (1971). Statističeskie ocenki pis´mennoj reči boľnych šizofreniej. In: *Inženernaja lingvistika*. Leningrad.

Piotrowski, R.G. (1984). *Text, Computer, Mensch.* Bochum: Brockmeyer.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

5.16. Text Cohesion (2)

Hypothesis

The distances between anaphora and cataphora represent a monotonously decreasing sequence.

Procedure

Determine the distances between all the anaphora (cataphora) in a text in terms of mumbers of the number of intervening words and set up the frequency distribution of the distances (i.e. the number of occurrences of 0,1,2… word distances). Approximate this sequence by means of the Zipf-Alekseev function

$$y = ax^{-b-c \ln x}$$

where x is the distance (x = 0,1,2,… or x = 1,2,3,… depending on the distance definition you apply), y is the number of occurrences of that distance and a,b,c are parameters. If the above function is not adequate, find a more adequate one. (Do not forget that $x^0 = 1$)

Is a function of the distance (e.g. its mean) linked in some way to the parameters of the negative hypergeometric distribution in Problem 5.14? If so, determine the kind of linkage and express it formally.

References

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.*

Skinner, B.F. (1957). *Verbal behaviour*. Acton: Copley

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

5.17. Text Cohesion (3)

Hypothesis

The frequency distribution of the grammatical functions of anaphora and cataphora follows a probability distribution from the class of diversification distributions.

Procedure

Determine for each of the anaphora (cataphora) in a text its grammatical function and count the number of occurrences of the individual functions. Fit an appropriate probability distribution to the data. Substantiate the distribution.

References

Alekseev, P.M. (1978). O nelinejnych formulirovkach zakona Cipfa. In: Piotrovskij, R.G. (ed.), *Statistika reči avtomatičeskij analiz teksta: 53-65.* Moskva-Leningrad: Naučnyj sovet po kompleksnoj probleme „Kibernetika“ AN SSSR.

Altmann, G. (1985a). Semantische Diversifikation. *Folia Linguistica 19, 177-200.*

Altmann, G. (1985b). Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4, 139-155.*

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 33-46.* Hagen: Rottmann.

Altmann, G. (1996). Diversification processes of the word. *Glottometrika 15, 102-111.*

Altmann, G., Best, K.-H., Kind, B. (1987), Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8, 130-139.*

Beöthy, E., Altmann, G. (1984a). The diversification of meaning of Hungarian verbal prefixes. II. ki-. *Finnisch-Ugrische Mitteilungen 8, 29-37.*

Beöthy, E., Altmann, G. (1984b). Semantic diversification of Hungarian verbal prefixes. III. „föl-", „el-", „be-". *Glottometrika 7, 45-56.*

Best, K-H. (1994). Word class frequency in contemporary German short prose texts. *Journal of Quantitative Linguistics 1, 144-147.*

Best, K.-H. (2009). Diversifikation des Phonems /r/ im deutschen. *Glottometrics 18, 26-31.*

Haight, F.A. (1966). Some statistical problems in connection with word association data. *Journal of Mathematical Psychology 3, 217-233.*

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

Hoffmann, L. (2000). Anapher im Text. In: Brinker, K., Antos, G., Heinemann, W. (eds.), *Text- und Gesprächslinguistik. Linguistics of text and conversation: 295-304*. Berlin-New York: de Gruyter.

Horvath, W.J. (1963). A stochastic model for word association tests. *Psychological Review 70, 361-364.*

Hřebíček, L. (1996). Word associations and text. *Glottometrika 15, 96-101.*

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

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics 14, 241-257.*

Köhler, R. (1989). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

Köhler, R. (1990). Elemente der synergetischen Linguistik. *Glottometrika 12, 179-17.*

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

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-25.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Raether, A., Rothe, U. (1991). Diversifikation der deutschen Komposita 'Substantiv plus Substantiv'. In: Rothe, U. (ed.) (1991): 85-91.

Rothe, U. (1986), *Die Semantik des textuellen et*. Frankfurt: Lang.

Rothe, U. (ed.) (1991). *Diversification processes in language: Grammar*. Hagen: Rottmann.

Ziegler, A. (1998), Word class frequencies in Brasilian-Portuguese press texts. *Journal of Quantitative Linguistics 4, 269-280.*

5.18. Hapax legomena and Markov chains

Hypothesis

The distance between hapax legomena in a text is a Markov chain of first order (Popescu et al. 2009: 227ff.). Test the hypothesis.

Procedure

Compute the frequencies of the words in a text. Then replace the hapax legomena by a symbol, say 1, and the other words by 0. The distance between two hapax legomena is the number of intervening other words (= the number of zeroes between two ones). If this sequence is a Markov chain, then the distances (Y) are distributed according to the modified geometric distribution

$$P(Y=k)=\begin{cases}1-\alpha, & for\ k=0\\ \alpha\, pq^{k-1}, & for\ k=1,2,3,...\end{cases}$$

Fit the distribution to the frequencies of distances. Iterative fitting is possible with the Altmann-Fitter, point estimators are shown in Strauß et al. (1984). Compare the parameters α and p (q = 1-p) in different texts and characterize texts and genres by means of parameter intervals.

Examine texts in different languages and state whether the hypothesis holds true. If so, associate the parameters with some other property of the given language, e.g. synthetism/analytism.

If the hypothesis can be rejected, do not pass to a higher order of the Markov chain because this leads only to further modifications of the geometric distribution; look for another solution (cf. Problem 5.19).

Cf. also “Distances between equally long sentences” in Problems, Vol. 1, p. 44.

References

Brainerd, B. (1976). On the Markov nature of the texts. *Linguistics 76, 5-30.*

Popescu, I.- I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin/New York: Mouton de Gruyter.

Strauß, U., Sappok, Ch, Diller, H.J., Altmann, G. (1984). Zur Theorie der Klumpung von Textentitäten. *Glottometrika 7, 73-100.*

5.19. The frequency sequence of words

Problem

The sequence of word frequencies displays a kind of regularity. Find the regularity and use it for comparisons.

Procedure

Compute the frequencies of words in a text (lemmas or word forms). You may want to apply one of the freely available word count programs. Then replace the words of the text by their frequencies. Ignore interpunction. One obtains a kind of time series consisting of a sequence of frequencies.

(a) Use Fourier analysis to express the oscillation of the frequencies.
(b) Compute Hurst´s exponent for the series (cf. Problems Vol 1, "Hurst´s exponent").
(c) Compare texts using the results of (a) and (b).
(d) Set up the distribution of differences between neighbouring frequencies and find the theoretical discrete distribution. Substantiate the given distribution by grammatical, typological, psychological etc arguments.
(e) Compute some properties of the empirical distribution and compare them with other texts.
(f) Draw some typological consequences from the distribution in (c).

References

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., 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.*

Spiegel, M.R. (1974). *Fourier Analysis*. New York: McGraw-Hill.

5.20. Golden section (2)

Problem

Show the asymptotic existence of the golden section in texts using the Popescu-Altmann (2009) method.

Procedure

Take texts of different length, up to complete novels, compute the rank-frequency distribution of lemmas or word forms and for each text compute the *h*-point (see Problem 5.10, "Golden section 1"), the arc length *L* (see Problem 6.1 "Arc length and typology"), the maximum arc length according to

$$L_{max} = V - 1 + f(1) - 1$$

where *V* is the vocabulary size (i.e. the greatest rank) and *f*(1) is the greatest frequency. Using these quantities compute two text indicators *p* and *q* defined as

$$p = \frac{L_{\max} - L}{h - 1}$$

and

$$q = \frac{L_{\max} - L}{\sqrt{N}}.$$

Adding them, i.e. computing

$$p + q = (L_{\max} - L)\left(\frac{1}{\sqrt{N}} + \frac{1}{h-1}\right)$$

show that this quantity converges to the golden section 1.618… Scrutinize the behaviour of both *p* and *q* and study the way of convergence of *p*+*q*. If possible, support this peculiar phenomenon by linguistic arguments, find the mathematical background and show the simple functions or intervals for *p*, *q* and *p*+*q* for your data. Compare your results with those of Tuzzi et al. (2009).

Then analyze the rank-frequency sequence of another entity covering the whole text, e.g. parts of speech, compute the above indicators and examine whether here *p*+*q* converges to the golden section, too.

Finally, examine individual parts-of-speech. Differentiate appropriate sub-categories, e.g. in the case of pronouns, personal, deictic, relative, interrogative,… ones, would form such a group of sub-categories. Set up the rank-frequency sequence of these sub-categories and analyze the change of *p*, *q*, *p*+*q* on this level. Can you see a tendency or a kind of self-similarity, etc?

References

Popescu, I.-I., Altmann, G. (2009). A modified text indicator. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Problems of quantitative text analysis: 13-39.* Černivci: ČNU.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). The golden section in texts. *ETC – Empirical Text and Culture Research 4 (submitted)*

6. Typology and universals

6.1. Arc length and typology

Problem

Is the relation of arc length of the rank-frequency distribution of word-forms to the greatest frequency a typological indicator? Scrutinize the problem.

Procedure

Set up the rank-frequency sequences of the first 50 word-forms of several texts, separately for each text. Determine the individual $f(1)$, i.e. the frequencies of the words at rank 1, and compute the arc lengths according to

$$L = \sum_{r=1}^{V-1} [(f(r) - f(r+1))^2 + 1]^{1/2}$$

where $f(r)$ is the frequency at rank r and V is the vocabulary (= number of word-form types). Note L and $f(1)$. Then repeat the procedure but take the first 100 word forms; again, note the $f(1)$ and L. Continue to maximally 1000 words and obtain finally 20 points $< L_i, f(1)_i >$. When you have these quantities for several texts, compute the function

$$L = af_1^b.$$

Perform the same procedure on texts from several languages. The more synthetic a language, the steeper is the function, i.e. the greater is parameter b. Study the behaviour of parameter b and explain typologically its role. Analyze texts from a family of related languages e.g. Roman.

References

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18-44.*

6.2. Length of morphs

Hypothesis

Morph length is distributed regularly (Saporta 1966; Best 2001). Test the hypothesis.

Procedure

Sol Saporta noticed that the length of Spanish morphs measured in terms of phoneme numbers displays a very regular pattern and asked whether this phenomenon is universal. His data are as follows

Number of phonemes	Number of different morphs
0	6
1	59
2	97
3	307
4	387
5	327
6	261
7	143
8	64
9	19
10	4
11	1
12	2
13	1
14	1

Find a discrete distribution expressing the frequencies of morphs of the individual lengths and examine the problem "… what factor other than chance might be operating to produce such a distribution." (Saporta 1969: 69).

If the fitting was successful, study another language or a group of languages, draw a random sample of ca. 1500 morphs and generalize the above result. Formally and methodically, the problem does not differ from that of word length distribution. Compare your results with those of Best (2001), compare the phoneme inventories of the languages investigated and conjecture the Saporta factors.

References

Best, K.-H. (2001). Zur Länge von Morphen in deutschen Texten. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 1-14.* Göttingen: Peust & Gutschmidt Verlag.

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J.H. (ed), *Universals of language: 61-72.* Cambridge, Mass.: The M.I.T. Press.

6.3. Diversification constant

Hypothesis

The diversification of a given phenomenon can be characterized by the same constant in all languages (Popescu, Altmann 2008).

Procedure

Consider the diversifications studied in Problem 7.5 (Diversification Distribution). According to the hypothesis, all cases of the same phenomenon must diversify in a very similar way, so that a property of the rank-frequency distribution capturing this diversification can be used for characterization. Popescu proposed the coefficient

$$c = \frac{R + f_{\max} - f_{\min} + 1 - L}{h},$$

where R is the number of diversified classes, i.e. the maximum rank; f_{max} and f_{min} are the maximum and the minimum frequencies of the distribution respectively; L is the empirical arc length of the distribution computed as

$$L = \sum_{r=1}^{R-1} [(f_r - f_{r+1})^2 + 1]^{1/2},$$

representing the sum of the Euclidean distances between neighbouring frequencies; and finally, h is the h-point computed as

$$h = \begin{cases} r & \text{if there is an } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \text{if there is no } r = f_r \end{cases}$$

where the subscripts 1 and 2 indicate any two neighbouring classes such that $f_1 > r_1$ and $f_2 < r_1 + 1$. If $f_{min} > R$, i.e. the smallest frequency is greater than the greatest rank, one can subtract from each frequency ($f_{min} - 1$). Thereby the arc length L does not change but the h-point can be computed more easily.

Compare your results with the Popescu-Altmann table below and determine whether your result lies in the given interval. The intervals for individual c-s are in the fourth column, the intervals for mean c-s in the fifth column of the table. If your results diverge, find a possible factor causing it.

Table 1
Comparison of mean $\overline{c}$
(Popescu, Altmann 2008)

Category	$\overline{c}$	$\mathbf{s_c}$	**Int. c**	**Int.** $\overline{c}$
Sounds, phonemes, letters	1.05	0.02	<1.00, 1.10>	<1.04, 1.06>
Word classes (parts of speech)	1.10	0.02	<1.06, 1.15>	<1.09, 1.12>
Rhythmic patterns	1.14	0.11	<0.92, 1.36>	<1.10, 1.18>
Paradigmatic classes	1.15	0.05	<1.04, 1.26>	<1.09, 1.20>
Colour classes	1.18	0.07	<1.05, 1.32>	<1.15, 1.22>
Prepositions, postpositions, conjunctions	1.24	0.11	<1.03, 1.46>	<1.17, 1.32>
Case diversification	1.33	-	-	-
Allomorphs of plural	1.37	0.21	<0.97, 1.77>	<1.31, 1.43>
Affixes (Meaning diversification)	1.39	0.16	<1.06, 1.71>	<1.32, 1,44>
Words (Meaning diversification)	1.47	0.21	<1.06, 1.88>	<1.44, 1.50>

Study especially the diversification of functions and meanings of nominal cases for which Popescu and Altmann had only one specimen, and find a preliminary interval.

References

Best, K.-H. (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18, 26-31*.

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS *Glottometrics 18, 13-25.*

Popescu, I.-I. Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 2008, 97-111*.

Rothe, U. (ed.) (1991). *Diversification process in language: grammar*. Hagen: Rottmann.

Sanada, H. (2009). *Diversification of postpositions in Japanese*. Msc.

6.4. Synthetism – analytism

Problem

Show the locations of a language of your choice on the synthetism-analytism scale using only word frequencies from various texts.

Procedure

Since synthetic languages have many word-forms, the rank-frequency sequence of word forms is very long; in analytic languages it is short. Hence, compute the rank-frequency distribution of word forms in a text (or several texts of the same language – not a mixture of texts) and compute the Popescu indicator

$$q = \frac{L_{\max} - L}{N^{1/2}}$$

where L is the arc length between the greatest and the smallest frequency computed in form of a sum of Euclidean distances, L_{max} is the maximum arc length and N is text length (number of word-form tokens). The exact definitions can be found in Popescu, Mačutek, Altmann (2009a,b).

Having analyzed several texts from the given language, compute the mean q for these texts and find the place of your language in the table below. If you have analyzed at least 20 texts in a language contained in the table and obtained a very deviating value of mean q, correct the table by computing the unweighted mean of your value and that in the table. If you analyzed a new language, insert your language and its q-value simply in the table. The aim is to obtain as many languages as possible.

The mean indicators $\overline{q}$
(From Popescu, Mačutek, Altmann 2009)

Language	$\overline{q}$	**Language**	$\overline{q}$
Kannada	0.273	Russian	0.382
Latin	0.278	Italian	0.412
Hungarian	0.281	English	0.435
Indonesian	0.312	Tagalog	0.446
Marathi	0.324	Lakota	0.449
German	0.334	Maori	0.479
Czech	0.336	Marquesan	0.504
Romanian	0.356	Rarotongan	0.520
Bulgarian	0.370	Hawaiian	0.542
Slovenian	0.376	Samoan	0.565

References

Popescu, I.-I., Mačutek, J., Altmann, G. (2009a). A modified text indicator. In: Kelih, E., Levickij, V., Altmann, G. (2009). *Methods of text analysis: 208-229.*. Černivci: ČNU.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009b). *Aspects of word frequencies*. Lüdenscheid: RAM.

6.5. Methodological problems

Problems

1. Is the aim of typology classification? If so, is it purposeful to consider only grammar or only phonemics or only both together as the basis of classification?
2. Is it possible to draw any consequences from a typological language classification you know? If so, consider some of these consequences as hypotheses and test them empirically. If not, what is the very aim of the classification?
3. Is it possible to use in typology only categorical (nominal) concepts in typology or is it more purposeful (more exact, more prolific) to apply quantitative ones? If you prefer the former, are you sure that no ambiguities remain? Are all languages unequivocally ascribed to classes? If you prefer the latter, collect all existing typological indicators beginning with the work of Greenberg up to now.
4. Normalize all indicators, i.e. transform them in such a way that they vary in the interval <0,1>. Explain why an index whose right boundary is infinity does not give any reasonable description of the linguistic reality. Do properties with infinite values exist in language?
5. Interpret each indicator. Is it possible to interpret an index in the interval <0, ∞)? If so, does a value of 1000 correspond to a high or to a low degree of a property?
6. Find the sampling distribution of the indicator. Consult statisticians if necessary. If there are difficulties, solve in any case the following problem:
7. Derive the theoretical expectation and the variance of the indicators and use them for setting up an asymptotic normal test. Set up confidence intervals around the mean and classify preliminarily all languages you have at your disposal according to all indicators. Do the indicators always yield the same classification?

References

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Anreiter, P. (1989). Transformierte sprachtypologische Profilvektoren. *Glotto-metrika 10, 32-45.*

Cysouw, M. (2005). Quantitative methods in typology. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics: an international handbook: 554-557.* Berlin: Mouton de Gruyter.

Fronzaroli, P. (1975). Problemi di classificazione delle lingue su base quantitative. In: *Colloquio sul tema: le tecniche di classificazione e loro applicazione linguistica: 123-141.* Roma: Academia Nazionale dei Lincei.

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26(3), 178-194.*

Greenberg, J.H. (1974). *Language typology: a historical and analytical overview*. The Hague, Paris: Mouton.

Kasevič, V.S., Jachontov, S. (eds.) (1982). *Quantitative typology of Afro-Asiatic languages.* St. Petersburg: University Press (in Russian).

Krámský, J. (1959). Quantitative typology of languages. *Language and Speech 2, 72-85.*

Kroeber, A.L. (1960). On typological indices. 1. Ranking of languages. *International Journal of American Linguistics 26, 171-177.*

Kroeber, A.L. (1960). Three quantitative classifications of Romance. *Romance Philology 14, 189-195.*

Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*

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

Lekomceva, M.I. (1963). K tipologii fonologičeskich struktur slova v slavjanskich jazykach. *Slavjanskoe jazykoznanie 1963, 277-295.*

Lekomceva, M.I. (1963). Tipologija fonologičeskich sistem. *Issledovanija po strukturnoj tipologii 1963, 42-51.*

Mejlach, M, (1973). Indeksy morfologičeskoj tipologii. In: *Problemy grammatičeskogo modelirovanija: 155-170*. Moskva: Nauka.

Sankaran, C.R., Taskar, A.D., Ganeshsundaruam, P.C. (1950). Quantitative classification of languages. *Bulletin of the Deccan College Institute 10, 85-111.*

Stepanov, A.V. (1995). Automatic typological analysis of Semitic morphology. *Journal of Quantitative Linguistics 2(2), 141-150.*

Winter, W. (1969-70). Some basic difficulties in the application of quantifying techniques to morphological typology. *Actes du X-e Congrès International des Linguistes: 3, 545-549.* Bucarest.

6.6. Word order (1)

Hypothesis

In language universals research, formulations can be found such as: "With overwhelmingly more than chance frequency, languages with dominant order VSO

have the adjective after the noun." (Greenberg 1966: 85). Replace this fuzzy statement by an exact one.

Procedure

Assume that a language with dominant VSO order has a probability of $p = 0.5$ of having the adjective after the noun (and $q = 1\text{-}p = 0.5$ of having the reverse order). In this case, the number x of languages in a cross-linguistic sample of size n having the predicted word order by chance is distributed according to the binomial distribution with parameters $p = 0.5$ and n.

Propose a statistical test and give a method to determine exactly a frequency borderline beyond which this hypothesis should be rejected.

Reference

Greenberg, J.H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (ed.), *Universals of Language: 73-113.* Cambridge, Massachusetts, and London, England: MIT Press.

6.7. Word order (2)

Problem

Test Greenberg's hypothesis from 'Word order (1)'.

Procedure

Re-analyse cross-linguistic (typological) samples. Determine the probabilities and perform significance tests. Do the data support Greenberg's hypothesis in its exact version?

Reference

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (ed.), *Universals of Language: 73-113.* Cambridge, Massachusetts, and London, England: MIT Press.

6.8. Phoneme sequences

Hypothesis

“In languages with both dissolvable and non-dissolvable medial clusters, the

former will be significantly more frequent than the latter. (A dissolvable cluster is defined as a sequence whose first part occurs in final position and whose second part occurs in initial position)" (Saporta 1966: 67). Test the hypothesis.

Procedure

Corroborate the Saporta hypothesis on at least one language. First obtain all medial consonant clusters in the language. Differentiate two cases: (a) occurrence in the dictionary, (b) occurrence in texts, where there may be more clusters caused by affixation or inflection. Then test the hypothesis distinguishing two alternatives: (1) Is the *number* of dissolvable clusters greater than that of non-dissolvable ones and (2) is their *frequency* greater?

As a matter of fact, there are four hypotheses each of which must be treated separately. For each of them use a statistical test; do not decide intuitively

According to Saporta (1966), the cause of this phenomenon is the general principle of economy: "the presence of the complex pattern implies the presence of the more simple one." Discuss this argument and find other phenomena being in line with this assumption.

Reference

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J.H. (ed.), *Universals of language* (2nd ed.): *61-72.* Cambridge, Mass: The M.I.T. Press.

6.9. Saporta's consonant sequences

Hypothesis

"The presence of C_1C_2- makes $–C_2C_1$ as likely as or more likely than C_1C_2-" (Saporta 1966: 68). Here C = consonant and the clusters indicate word-initial and word-final positions.

Procedure

First find some linguistic background for this hypothesis. Then choose a language with many consonant clusters, e.g. a Slavic language. Show that the hypothesis need not be accepted. Modify it, define boundary conditions, other features of the given language, etc. In other words, make the hypothesis reasonable.

Reference

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J.H. (ed.), *Universals of language* (2nd ed.): *61-72.* Cambridge, Mass: The M.I.T. Press.

6.10. Word frequency and analytism

Hypothesis

In texts of strongly analytic languages, the graph of the Zipfian power function $f(r) = cr^{-a}$ sits *above* the hapax legomena of the word-form rank-frequency sequence, in strongly synthetic ones *below* them (Popescu, Mačutek, Altmann 2009, 104 ff.). Test the hypothesis.

Procedure

Determine the rank-frequency distribution of word-forms and of texts from any language except those studied in the reference, compute and fit the above Zipfian function to the absolute frequencies. It is easier to show the effect on data from an extremely synthetic or analytic language.

The measure of the degree of analytism/synthetism can be calculated by means of the following indicator (Popescu, Mačutek, Altmann 2009, 106)

$$B = \frac{c}{(V - HL/2)^a}$$

where a and c are the parameters of the Zipfian function, which have to be estimated from the data, V is the vocabulary size of the text (= number of different word forms) and HL is the number of hapax legomena. Estimate the degree of analytism on data from several texts, computing their indicators B and finally computing the mean of these B's. Locate your language in Table 6.4.

Table 6.4
Mean analytism indicator *B* of 20 languages
(Popescu, Mačutek, Altmann 2009, 109)

	Language	**Mean *B***		**Language**	**Mean B**
1	Hungarian	0.2012	11	Marathi	12.302
2	Czech	0.7223	12	Italian	12.787
3	Latin	0.7982	13	Lakota	12.853
4	Romanian	0.8931	14	Tagalog	13.913
5	German	0.9372	15	English	14.514
6	Slovenian	0.9418	16	Marquesan	18.108
7	Kannada	10.378	17	Rarotongan	19.779
8	Russian	10.453	18	Samoan	21.465
9	Bulgarian	10.495	19	Maori	21.861
10	Indonesian	11.438	20	Hawaiian	50.815

References

Popescu, I.-I, Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies.* Lüdenscheid: RAM.

7. Synergetics

7.1. Frequency and polytextuality

Hypothesis

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

This hypothesis is a special case of a more general one: the more frequent a word is, the more contexts it occurs in, i.e. cotextuality depends on frequency (Köhler 1986).

Test the hypothesis.

Procedure

There are three possibilities to test this hypothesis on data from any language:

(1) Consult a frequency dictionary and determine the 20 most frequent verbs (for the more specific hypothesis). Then count the number of phrases, idioms etc in a monolingual dictionary which are given for these verbs.

(2) Find the 20 most frequent verbs in a text corpus, and determine the number of their different arguments.

(3) Set up a frequency word list of the texts in a corpus; then take the 100 most frequent ones and find the number of their different environments (co-text types). Consider verb "types", not "tokens" and investigate (a) the predecessors, (b) the successors, (c) both.

On this data, test the hypothesis that co-textuality (*CT*) is a specific function of frequency (*F*), viz.

$$CT = aF^b,$$

where *a* and *b* are parameters. In Gieseking (2002) and Köhler (2002) this relationship was reversed. Show that it holds in both directions. Obtain the parameters for different word classes and try to characterize them using the respective parameters.

See also the problems "Word length and polytextuality" (p. 84) and "Collocations" (p. 29) in Problems Vol. 1.

References

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 387 – 433:*. http://ubtopus.hbz-nrw.de/volltexte/2004/279 (March 24, 2008)

Köhler, R. (1985*). Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: de Gruyter.

Thompson, S.A., Hopper, P.J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (2001), *Frequency and the emergence of linguistic structure: 27-60.* Amsterdam/ Philadelphia: Benjamins

7.2. Polysemy and polytextuality

Hypothesis

Polytextuality increases with increasing polysemy according to a power function. Test the hypothesis.

Procedure

Polytextuality can be operationalized in various ways depending on the units under study. The most straightforward operationalization for words is the number of different texts in a corpus in which the given word occurs at least once. For morphemes or morphs, the number of words in a dictionary containing the given morph(eme) forms a reasonable measure. In general, the number of co-texts or contexts of a unit may constitute a measure of polytextuality.

Polysemy of words can be approximated by the number of senses given in a monolingual dictionary – even if the differentiation between homonymy and polysemy as presented in a dictionary may be as doubtful as the sub-categorization of meanings. There are also electronic versions of dictionaries which can be used. In the case of morph(eme)s, polyfunctionality can be determined either on the basis of exclusively semantic criteria or including grammatical features.

Collect the polysemy (or polyfunctionality) and polytextuality values for the units you study in one of the above-mentioned ways. Sort these value pairs according to polysemy, i.e. form groups of value pairs with identical polysemy. Calculate the mean polytextuality of the values in the individual groups. You obtain as many means as there are different polysemy values. The pairs <polysemy, mean polytextuality> constitute the data an appropriate function can be fitted to.

Köhler (1985, 2005) derives the function

$$y = Ax^b$$

from a differential equation as an element of the synergetic control cycle. Fit this function to your data. Alternatively, the extended version may be fitted with a better result (coefficient of determination), if an additional operator is introduced into the control cycle (cf. Köhler 2006):

$$y = Ax^b e^{cx} .$$

References

Köhler, R (1985*). Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: de Gruyter.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern - Eine Erweiterung des synergetisch-linguistischen Modells. In Rapp, R., Sedlmeier, P, Zunker-Rapp, G *(*eds.) *Perspectives on Cognition – A Festschrift for Manfred Wettler*. Lengerich: Pabst Science Publishers: 327-338.

7.3. Morph length and phoneme inventory

Hypothesis

"The mean length of morphs will be inversely related to the number of phonemes in the inventory." (Saporta 1966:70). Test the hypothesis.

Procedure

Collect data on mean morph length and phoneme inventories from several languages. If there are no frequency dictionaries of morphs, take at least one longer text from minimally five languages. In an ideal case, one should take the same text, e.g. a translation of a short text. First find a simple function expressing this relationship. If the fit (the coefficient of determination) is not satisfactory, add stepwise further "compensatory" factors as proposed by R. Jakobson and W. Gedney in a note to Saporta´s article, i.e. attempt to capture the relationship by adding further variables: the number of phoneme combinations used in the language, the number of homonyms, and presence of tone and/or stress. Construct a control cycle in which the dependent variable is mean morph length.

References

Saporta, S. (1966). Phoneme distribution and language universals. In: Greenberg, J.H. (ed), *Universals of language: 61-72.* Cambridge, Mass.: The M.I.T. Press.

7.4. Frequency and polysemy

Hypothesis

There is a "… direct relationship between the number of different meanings of a word and its relative frequency of occurrences" (Zipf 1945a:144). Test the hypothesis.

Procedure

In order to test the hypothesis, draw a sample of at least 100 words from a frequency dictionary or electronic source such as WordNet. In the latter case and in some frequency dictionaries, frequencies are given for the individual senses of a word. You should calculate on the basis of total word frequencies (sum up the frequencies of the individual senses)

If the hypothesis holds, the dependence $S = f(F)$ should be rather simple. However, some researchers warn against quick generalizations and recommend taking into account different boundary conditions (Ullmann 1966). In Köhler´s (2005) control cycle, there is no direct connection between frequency and polysemy: the relation is indirect through the medium of length (cf. also Guiter 1974) and can be expressed by one of the formulas

$$y = Ax^b \text{ or } y = Ax^b e^{cx}.$$

Find a direct dependence using averages and smoothing. Insert Köhler´s requirements in the model. Do not restrict yourself to English. Examine several languages.

There is a possibility that this relationship can be used for stylistic analysis, too. Take texts of different types, count the individual word frequencies and for each word get its polysemy using a monolingual dictionary. Set up the relation $S = f(F)$ which is most probably a power function and scrutinize the parameters of the functions. Are there text-type differences, e.g. between scientific and poetic texts, or are there also stylistic differences, e.g. between two lyrical poems?

References

Carloni, F. (2000). Le relazioni statistiche tra frequenza e significato delle parole nella lingua italiana. *Italica 77(4), 523-534.*

Guiter, H. (1974). Les relations fréquence-longueur-sens des mots (language romanes et anglais). *Atti del XII Congresso internazionale di linguistica e filologia romanza 14(4), 373-381. Napoli 1970, 15-20.*

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

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: de Gruyter**.**

Manin, D.Yu. (2008). Zipf´s law and avoidance of excessive synonymy. *Cognitive Science 32(7), 1075-1098.*

Ullmann, S. (1966). Semantic universals. In: Greenberg, J.H. (ed.), *Universals of language: 217-262.* Cambridge, Mass.: The M.I.T. Press

Zipf, G.K. (1945a). The repetition of words, time-perspective and semantic balance. *The Journal of General Psyhology 32, 127-148.*

Zipf, G.K. (1945b). The meaning-frequency relationship of words. *The Journal of general Psychology 33, 251-256.*

Zipf, G.K. (1949). *Human behaviour and the principle of least effort*. Reading, Mass.: Addison-Wesley.

7.5. Diversification distribution

Hypothesis

If an entity diversifies, the frequency of individual elements abides by a regular distribution.

Procedure

You may choose among all linguistic entities which can diversify, i.e. formal (phonetic, graphic), morphological, semantic, syntactic, lexematic, dialectal, sociolectal etc elements. An illustrative example is the study of the meanings of the conjunction "and". Count the frequency of occurrence in textual material of each of its senses separately. Then set up the empirical rank-frequency distribution of the individual senses. Find (a) a function, (b) a distribution which can successfully model it.

In case (a) apply first the usual Zipfian approach $f_r = c/r^a$ (r = rank, f_r = frequency at rank r) or the Zipf-Mandelbrot approach $f_r = c/(r+b)^a$, then try the Popescu approach $f_r = 1 + a*exp(-r/b)$. Calculate the goodness-of-fit of the determination coefficient. In case (b) find an adequate distribution. Begin with Zipf and Zipf-Mandelbrot (as distributions), then add the Shenton-Skees-geometric distribution (cf. Shenton, Skees 1970; Wimmer, Altmann 1999; Mačutek 2008)

$$P_x = pq^{x-1}\left[1 + a\left(x - \frac{1}{p}\right)\right], \quad x = 1,2,3,...$$

where a and p are parameters, $0 < p < 1$, $q = 1\text{-}p$, $0 < a < 1/q\text{-}1$. Test the goodness-of-fit by means of a chi-square test. Find the "best" model and analyse in this way all conjunctions.

Describe the behaviour of this kind of diversification. Which model yielded the "best" fit? Are the parameters of the individual models interrelated?

Study the diversification of other phenomena (cf. Rothe 1991) and compare them with that of conjunctions.

Explain the existence of the observed regularity.

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: de Gruyter.

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 2000, 69-81*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length a meaning diversification in English. *Glottometrics 17, 2008, 82-89.*

Mačutek, J. (2008). On the distribution of graphemic representations. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties and characters of writing systems: 75-78.* Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 2008, 97-111*.

Popescu, I.-I., Altmann, G., Köhler, R. (2008). Zipf´s law – another view. *Quality and Quantity* (online)

Rothe, U. (ed.) (1991). *Diversification process in language: grammar*. Hagen: Rottmann.

Shenton, I.R., Skees, P. (1970). Some statistical aspects of amounts and duration of rainfall. In: Patil, G.P. (ed.), *Random counts in scientific work, Vol 3: 73-94*. University Park: The Pennsylvania State University Press.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

7.6. System boundaries and interactions

Problem

We cannot conceptualize any entity (system or not) without thinking of its boundaries. In linguistics, we are familiar with the problems connected to the segmentation of many linguistic units (e.g., sounds, syllables, or discontinuous units of any kind), i.e. the determination of their boundaries. Very little has been said so far about the boundaries of languages as systems. Elaborate on some boundaries. Consider language as a system having subsystems, units, properties.

Procedure

Consider the kinds of boundaries connected with the concept of language. There are boundaries which separate different languages from each other (what about dialects?), boundaries within a language (between subsystems such as syntax and morphology, syntax and lexicon etc., between inventories such as the inventory of morph(em)s and the one of lexemes etc, between (stylistic, sociolinguistic etc) registers, between the "languages" / idiolects of individuals (with respect to a cognitive concept of language) etc. Give more examples of parts of a language and study the boundaries between them.

References

Before looking into this subject more closely you should be familiar with the basic concepts and the way of thinking in modern systems theory as presented in e.g.

Altmann, G., Koch, W. (eds.) (1998). *Systems. New Paradigms for the Human Sciences.* Berlin: de Gruyter.

Bowler, T. D. (1981). *General Systems Thinking.* New York/Oxford: North Holland.

Bunge, M**.** (1979). *Treatise in Basic Philosophy, Vol. 4. Ontology II: A world of systems*. Dordrecht/Boston/London: Reidel..

7.7. Language and text

Problem

What do you think about the boundary and the interaction between language and text?

Procedure

Some linguists consider not only language as a system but also text. First, determine the difference between the two kinds of systems (with respect to their dynamics, function, and structure). Then, draw a diagram describing the boundries, interfaces, and interrelation(s) between the two systems. Finally, reflect your findings on the background of the idea that (1) language is the realm of potentiality whereas text (parole) the realm of reality, and (2) language is formed by its use in communication, and (3) language is nothing than a construct which is used to describe regularities of linguistic communication.

References

Bybee, Joan. (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J., Hopper, P. (2001). *Frequency and the emergence of linguistic structure: 1-26.* Amsterdam/Philadelphia: Benjamins

7.8. Frequency and age of words

Hypothesis

The older a word, the more frequent it is. Test the hypothesis.

Procedure

There are two problems connected with this hypothesis: (1) The age of words cannot be determined exactly. For our purposes it can be estimated from the year/ century of its first appearance in written documents. (2) Words can also die out; hence the hypothesis must be specified.

The simplest way to get the age is to look in a historical (etymological) dictionary. If only centuries are given, take the mid of the century. Take a random sample from the given dictionary or select only words of a special class. Then determine the frequencies of the given words by means of a corpus or a frequency dictionary. Show that there is at least a correlation between age and frequency. Examine other languages than English.

If your data corroborate the elementary hypothesis above, fit an empirical formula to the hypothesized dependence. Skip words denoting industrial products because these are young and frequent. Take always the averages of the words with the same age.

References

None

7.9. Word length and age

Hypothesis

The older a word the shorter it is. Test the hypothesis.

Procedure

If words prevail for a long time, they must be frequent enough. But in that case they become shorter. Study sets of words of different word classes separately. Obtain, if possible, their age in the same way as in the problem “Frequency and age of words”. Study which classes abide by the above hypothesis, i.e. specify

the hypothesis. Investigate a language with many non-monosyllabic words. Develop a proto-theory of shortening with age, i.e. find the relevant formulas.

References

None

7.10. Valency and polysemy

Hypothesis

The greater the polysemy of a word (verb, noun, or adjective) the greater is its valency.

Procedure

Though valency can increase without increasing polysemy, it can be supposed that if polysemy increases, new valency instances arise. In order to test this hypothesis consult first a valency dictionary – which presents only a small subset of the lexicon of a language – and in addition a common monolingual dictionary in which the meanings can be found. Determine both values (valency and polysemy) for each selected word and find the dependence (separately for each part-of-speech). The dependence will not be linear. Then derive the dependence from assumptions or rely on synergetic reflections.

References

None

7.11. Complement to synergetic problems

Problem

In linguistic literature, a number of discourse-pragmatic functions of word order in sentences is discussed, among others *topic assignment, emphasis, conceptual closeness, foregrounding, certainty,* and *urgency*. Set up a synergetic model which shows word order as a multi-functional device and integrate some ideas concerning functional equivalences of word order with respect to the above-mentioned individual functions.

Procedure

Compile an overview on the functions of word order as discussed in the literature. Find functional equivalents of word order, which can be observed in

natural languages and consider their specific advantages and disadvantages with respect to the individual functions. Combine all these aspects into a synergetic model and derive at least one empirically testable hypothesis from it.

References

Behaghel, O. (1932). *Deutsche Syntax. Eine geschichtliche Darstellung*. Bd. IV., Heidelberg (Germanische Bibliothek. I. Sammlung germanischer Elemenar- und Handbücher. 1. Reihe: Grammatiken. 10. Bd.).

Croft, W. (2003). *Typology and universals*, 2nd edition. Cambridge: Cambridge University Press.

Givón, T. (1984). *Syntax: A functional-typological introduction*, Volume I. Amsterdam: John Benjamins.

Givón, T. (1990). *Syntax: A functional-typological introduction*, Volume II. Amsterdam: John Benjamins.

Givón, T. (1988). The pragmatics of word order: predictability, importance and attention. Studies in syntactic typology. In: Hammond, M., Moravcsik, E., Wirth, J. (eds.), *Studies in syntactic typology, 243-284.* Amsterdam: John Benjamins.

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (ed.): *Universals of Grammar. 2nd edition: 73-113.* Cambridge, Mass: MIT Press.

Haiman, J. (1983). Iconic and economic motivation. *Language 59.781-819.*

Haiman, J. (1985). *Natural syntax*. Cambridge: Cambridge University Press.

Hawkins, J.A. (1983). *Word order universals*. New York: Academic Press.

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

Jespersen, O. (1942). A Modern English Grammar on Historical Principles, IV. Munksgaard: Copenhagen.

Jespersen, O. (1949). *A Modern English Grammar on Historical Principles. Part 2* (*Syntax, vol. 1*). Copenhagen: Munksgaard; London: George Allen and Unwin.

Köhler, R. (1985). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: de Gruyter.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern – Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedlmeier, P., Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler: 327-338*. Lengerich: Pabst Science Publishers.

7.12. Phonotactics: Exploitation of linguistic material

Problem

Under-exploitation of material (in other words: of potential distinctions) is a well known fact in language. Example: only a small part of the possible combinations of phonemes is used by the languages to form units of higher levels, e.g. morphs. The same holds true – even more drastically – for the use of morph combinations in word formation etc.

The rate of under-exploitation seems to depend on the length of the units of the higher level, i.e. the longer the string of phonemes under consideration the fewer permutations are legal morphs.

Procedure

Set up a model which contains: (1) number of phonemes in the inventory of the given language, (2) morph length, (3) size of the morph inventory, (4) morph length (distribution), (5) morph similarity, (6) redundancy requirements, and (7) requirements of economy.

1. Define the concepts (1) – (7) in an appropriate way. Put some emphasis on the consideration of (5) *similarity* with respect to articulatory, auditive, and psycho- linguistic factors. Single out interrelations and influences between the system variables and set up a synergetic-linguistic model.
2. Derive individual testable hypotheses about distributions of properties of system variables and about functional dependences between system variables from the model. On this basis, determine the kind of data which can be used for empirical tests of the hypotheses and collect corresponding data. Perform the appropriate statistical tests.
3. Shift the model from the phonology/morphology level to the morphology/ lexicon level and consider which changes are needed.
4. Appropriate adjustments will allow for the application of the model on even higher levels such as syntax. Consider the problems to be solved.

Reference

Kelih, E. (2009). Phonemverbindungen und Inventarumfang: Empirische Evidenz und Modellentwicklung. *Glottotheory 1(2), 60-74.*

Köhler, R. (1985*). Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: de Gruyter.

Köhler, R. (2006). Frequenz, Kontextualität und Länge von Wörtern - Eine Erweiterung des synergetisch-linguistischen Modells. In: Rapp, R., Sedl-

meier, P, Zunker-Rapp, G. (eds.), *Perspectives on Cognition – A Festschrift for Manfred Wettler: 327-338.* Lengerich: Pabst Science Publishers.

7.13. Word length and polysemy in Chinese

Problem

In synergetic linguistics, the relationship between word length (*WL*) and polysemy (*P*) is usually given as $WL = aP^{-b}$. Evidently, it holds also vice versa as $P = c(WL)^{-d}$. However, there is a tiny problem: polysemy cannot be smaller than 1 (if we do not consider proper names as having no meaning) and a word contains at least one syllable. In some Slavic languages there are zero syllable prepositions but they are usually considered as proclitics. Hence, unity is the asymptote in both relations. Solve at least one of the following problems: (1) Correct the above formulas, (2) set up the differential equation and interpret it, (c) apply the new formula to Chinese data given below.

Procedure

Problem (1) is simple. Adding 1 to the above formulas warrants the convergence to 1 when the independent variable converges to infinity. Hence we obtain

(a) $WL = 1 + aP^{-b}$ and (b) $P = 1 + c(WL)^{-d}$.

Now construct the differential equation and interpret it. It is a non-homogeneous DE of the first order.

M.A. Breiter (1994) published the following data on Chinese (using weighted means of polysemy):

Length (*WL*)	Mean polysemy (*P*)
1	4.23
2	1.90
3	1.88
4	1.35

Though the series is short, fit (b) $P = 1 + c(WL)^{-d}$ to this data. If there are not enough classes for the goodness-of-fit test, estimate *c* and *d* from the first two classes.

Reference

Breiter, A.M. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1(3), 224-231.*

7.14. Length and frequency of affixes

Hypothesis

In strongly inflectional languages, the longer an affix, the smaller is its frequency. Test the hypothesis.

Procedure

This time the Zipfian relationship is tested in the reverse order.

For a first test you may apply the Spanish data published by Urrea (2000: 111-112). Test separately prefixes (which are mostly derivational) and suffixes (which are mostly inflectional). Determine the length of an affix (x) in terms of the number of phonemes. The frequency can be determined as the mean frequency (y) of all affixes of the same length. Propose a function as a model of the dependence. Substantiate the relationship.

Does the relationship hold also in strongly agglutinating languages? If not, why?

References

Urrea, A.M. (2000). Automatic discovery of affixes by means of a corpus: a catalog of Spanish affixes. *Journal of Quantitative Linguistics 7(2), 97-114.*

8. Philosophy of science and general problems

8.1. Degree of constituency

Hypotheses

"… the more often two elements occur in sequence the tighter will be their constituent structure" (Bybee, Scheibman 1999; Bybee, Hopper 2001: 14). Test the hypothesis.

Procedure

The first, most important problem is the design of a procedure by means of which it would be possible to measure the degree of tightness. No test of the hypothesis is possible without an operationalization of the concept of tightness and such a measure. Phonetic form and written pattern should be distinguished, i.e. two different criteria of tightness should be designed. As far as possible, scaling should be objective and phonetically applicable to all languages; as to written forms, scaling should be applicable to all languages using the same script.

On the basis of tightness values and relative frequencies of the words in a corpus, the dependency can be analysed. The weakest method is correlation analysis, a method that can show only linear relationship. The most scientifically prolific way is the theoretical derivation of a specific hypothesis about the form of the dependence and its mathematical formulation, which then can be tested against the data.

One of the possibilities is the segmentation of the texts into morphs and a frequency count; show that the more frequent a morph, the stronger is its crystallization, and the more often two morphs occur together, the greater is the degree of tightness.

As to the measure of tightness, take inspiration from Fan, Altmann (2007) but modify the computation appropriately.

Generalize the problem in the following way: give arguments for the conjecture that constituency is a continuous phenomenon, i.e. there is no clear boundary between phrase, compound and derivate, in spite of the fact that school grammars define it crisply admitting at the same time the existence of exceptions. One of the arguments may be supported by the fact that the above hypothesis speaks of "the more often" which is no crisp definition.

Analyze the crispness or fuzziness of any linguistic classes and study the role of frequency in class formation.

References

Boyland, J.T. (1996). *Morphosyntactic change in progress: a psycholinguistic approach.* Diss: Linguistic Department, University of California.

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P. (2001), *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 linguistics: 177-189.* Chernivcy: RUTA.

Krug, M.G. (2001). Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J., Hopper, P. (2001), *Frequency and the emergence of linguistic structure: 1-26.* Amsterdam/Philadelphia: Benjamins.

8.2. Exercises in the Philosophy of Science

> *„Unlike inborn patterns of behavior and unlike know-hows, scientific knowledge is entirely conceptual: it consists of systems of concepts interrelated in definite ways."*
> *(Mario Bunge 2007, p. 51)*

8.2.1.Concept

1. List some linguistic concepts and group them into (a) class concepts (nominal scale), (b) ordinal concepts (rank scale), and (c) metric concepts (interval scale or ratio scale). Choose one or two of these concepts and try to transform them into concepts of (a) a higher (b) lower scale.
2. Does a classification of linguistic items involve prior knowledge of the essential properties of objects or the intended classes? Does the result of a classification unveil the essential properties? Are there essential properties at all?
3. Differentiate the concepts of *frequency* and *commonness*; define the concepts exactly.
4. Which ones out of the following linguistic concepts refer to observable entities or properties?
   Word, part-of-speech, morph(eme), frequency, distribution, phrase type, verb valence, gender, length, co-occurrence, dependency, emphasis, iconicity, order parameter, requirement, inventory, production effort,

markedness, naturalness, projectivity, denotative meaning, connotation, text, language.

5. Which of these (or other) concepts are intervening concepts (non-observational concepts mediating between observational ones), which ones are hypothetical constructs?
6. Sharpen (to reduce vagueness of) an established linguistic concept (such as *lexical combinability*, cf. Levitskij (2005), *complexity* or *ornamentality* of characters or scripts).
7. Discuss the concept of fuzzyness in semantics. Can the concept of *fuzzy meaning* of a word be sharpened by replacing it with the concept of probability? If you conclude that this is possible only in certain cases characterise these cases. Discuss the consequences of such a replacement for the intension of the original concept.
8. Find an observational concept of fuzzy meaning (which would be a prerequisite for a measuring procedure for empirically determining the values of the membership function with respect to a fuzzy word meaning).
9. Discuss the differences between metaphors and imported concepts. How do you assess, in this respect, the cases of *mother node*, *parent language*, *word field*, *production effort*, *linguistic economy*, *entropy*?
10. List fundamental linguistic property concepts, i.e. concepts which are not constructed by means of other linguistic property concepts. Do the same with linguistic concepts in general.

References

Bunge, M. (2007). *Philosophy of science*. *Vol. 1: From problem to theory*. New Brunswick, London: Transaction Publishers.

Levitskij, V. (2005). Lexikalische Kombinierbarkeit. In: Köhler, R. Altmann, G., Piotrovskij, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook: 464-470.* Berlin, New York: de Gruyter.

8.2.2. Scientific Problems

11. Formulate a new scientific problem in the field of language and text studies. Check whether is meets all the formal and semantic conditions for a scientific problem (cf. Bunge 2005, chapter problems)).
12. How did you arrive at the new problem? By criticising known solutions, by allocation of known solutions to a new context, by generalising a solved problem, or by relating linguistic ideas to concepts in other disciplines?
13. Specify the logical structure of your problem, in particular, state which of the variables is the unknown one.

14. Do the same as in 11 with the following questions:
    (a) Is your first language a synthetic language?
    (b) What are the properties of word frequency?
    (c) Are there any languages without any verbs?
    (d) Which of the texts of your favourite author is the shortest one?
    (e) How are compounds formed in French/Russian?
    (f) Why do agglutinative languages display a flatter TTR function than analytic ones?
    (g) List some quantitative properties of sentences.
    (h) What is the mean length of Hungarian words?
    (i) How can syntactic ambiguity been measured?
15. Are the following problems well-formulated?
    (a) Is there any interrelation between word length and word valency?
    (b) Why is language a system?
    (c) How can a measure of aesthetic value be constructed?
    (d) Which of the world's languages has the largest lexicon?
    (e) Is it always possible to determine whether a symbol has a meaning?
    (f) Do languages enlarge their lexicons (alternatively: their grammars) without limit?
16. Classify the problems in 14 and 15 (and others) into (a) empirical, (b) conceptual, (c) methodological, and (d) valuational problems.
17. Identify the premises which are made with your problem statement from 10 or one of those from 13-15.
18. Consider the possibilities to generalise your problem, to transform it to another context, and to export it to another discipline.
19. How would a solution to your problem look like? Give an approximate description of the kind of answer that would solve the problem.
20. Identify some historical problems in linguistics using a textbook on the history of the discipline. Mention in which way, if any, these problems ended.
21. If linguistic properties are conceptual constructs, would you accept the opinion that language has a potentially infinite number of properties? What does their number depend on?
22. Is every property measurable?
23. If properties are conceptual constructs, how is it possible that they change? What changes?
24. Are there isolated properties in language?
25. If (22) holds, can a linguistic property attain an infinite value?

References

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

Bunge, M. (2007). *Philosophy of science*. *Vol. 1: From problem to theory*. New Brunswick, London: Transaction Publishers.
Polya, G. (1957). *How to solve it.* New York: Doubleday Anchor Books.

8.3. Rank-frequency, a general approach

Hypothesis

If a class of linguistic entities is "correctly" constituted and the elements of the class are ranked according to decreasing frequency, then the frequencies follow the function

$$f_r = 1 + a_1\exp(-r/b_1) + a_2\exp(-r/b_2)$$

where r = rank, f_r = frequency at rank r. Test the hypothesis.

Procedure

The hypothesis is a generalization of the Popescu-Altmann-Köhler approach (2009), originally restricted to word frequencies. Constitute any reasonable class of linguistic entities, e.g. all phonemes, syllables, colour names, pronouns, prepositions, conjunctions, clause types, types of compounds, word classes, etc. Get their frequency of occurrence in a long text, rank the frequencies and fit the above function to the data. If the number of ranks is small, the first component of the function is sufficient. If the number of ranks is large, sometimes a third exponential component must be added. Study the behaviour of the function when further components are added.

Clause types with American writers
(Data from J. Bojko 2005)

Clause type	Dreiser	Fitzgerald	Cronin	Steinbeck	Hemingway
Subject clause	6	2	4	23	32
Predicative clause	5	5	2	13	4
Object clause	647	306	246	173	208
Attributive clause	488	235	194	165	121
Time clause	211	193	153	159	114
Place clause	37	15	21	26	26
Causal clause	87	82	54	83	22
Manner clause	141	87	50	63	33
Result clause	8	13	5	16	6
Concessive clause	41	12	46	16	9
Purpose clause	12	3	2	10	3
Conditional clause	146	53	46	85	56

As an example use the data collected by J. Bojko (2005) concerning clause types in works of American writers as shown in the table above. Set up a rank-frequency sequence for each writer and fit the above function. Use only one exponential component. Which of the authors differs from the other ones (compare the b_1 parameters).

References

Boyko, J. (2005). Diferencijni parametri rečenija jak determinanta avtors´kogo stilju. In: Altmann, G., Levickij, V., Perebyjnis, V. (eds.), *Problems of Quantitative Linguistics: 292-305*. Černivcy: RUTA.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics* 1. Lüdenscheid: RAM (cf. Word frequency 3, p. 67)

Popescu, I.-I., Altmann, G., Köhler, R. (2009). Zipf´s law – another view. *Quality and Quantity: Online 9.5.2009.*

8.4. Universals, laws and theories

Problem

(a) Are 'language universals' laws? (b) Are 'grammar theories' theories? Try to answer the questions.

Procedure

(a) Study the concept of language universals such as those presented by Greenberg (1978). Do the statements of this kind meet the requirements for laws (as given in Bunge (1967))? If you conclude they are not laws - what else are they?

(b) Are 'grammar theories' such as HPSG, 'X-bar theory' etc. systems of universal laws, which explain the observed facts? If you conclude they are - make some of the law statements explicit and check whether they are really laws.

(c) The other way round: are linguistic laws language universals?

References

Bunge, M. (1967). *Scientific research I,II*. Berlin: Springer.

Cysouw, M. (2005). Quantitative methods in typology. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 554-578.* Berlin: de Gruyter.

Greenberg, J.H. (ed.) (1975). *Universals of Human Language, Vol 1: Method and Theory.* Stanford: Stanford University Press.

8.5. Observability

Problem

Determine the observability of linguistic entities.

Procedure

Set up a list of linguistic entities you are interested in (units, properties, systems, etc.) and examine their observability. You may differentiate between direct and indirect observation. Some philosophers, however, define observability exclusively with respect to the senses of the human organism, i.e. they do not consider observation by means of instruments as observation proper (cf. van Fraassen 1980).

1. Which linguistic entities are directly observable, which are not?
2. Which instruments do linguists use for indirect observation?
3. Is counting a method of direct observation? If not, what are the instruments used for counting? Is the conclusion you come to consistent with the view that sensual perceptions are direct observations (you use your cognitive apparatus also to recognize what you see or hear as a token of a known unit). Does the use of paper and pencil for counting make a difference?
4. Clarify the role of the langue-parole dichotomy with respect to the question of observability.
5. Consider the positions that realists and anti-realists would have to take when considering the langue-parole dichotomy.

References

Van Fraassen, Bas C. (1980). *The scientific image*. Oxford-New York: Oxford University Press.

9. Different issues

9.1. Arc length and language evolution

Problem

The typological evolution of a language can be traced down using the arc length of the word-form rank-frequency distribution.

Procedure

Follow the procedure in the problem "6.1. Arc length and typology" but this time examine texts of a language from different centuries. Follow the change of the parameter *b* and determine whether the given language develops to a more synthetic or a more analytic language.

Show which of the Romance languages has the greatest tendency toward analytism. Check the disputed problem that German is developing toward analytism.

Study some Indonesian, Melanesian and Polynesian languages and show the geographical distribution of analytism in the family of Austronesian languages. This problem can be solved without command of theses languages as just word-forms have to be counted; texts in these languages can easily be found on the Internet.

Show the development of Slavic languages and the geographical distribution of synthetism.in Europe. Would you conclude that there are areal influences? Can you recommend this method as one of the methods of areal linguistics?

References

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18-44*.

9.2. Politeness

Problem

Consider some properties of "polite" words and expressions and compare them with "neutral" words.

Procedure

"Polite" words or expressions have some characteristic features which distinguish them from everyday (neutral) expressions. Collect polite words and exp-

ressions and compare them in whatever sense (prosodic, phonological, morphological, lexical, semantic, …) with their "normal" counterparts. Perform the comparison quantitatively, i.e. quantify the properties and their differences. Show that there are different degrees of politeness and express them on a quantitative scale. Consider also impoliteness and its properties.

Study some South-East Asian languages from this point of view. For Japanese, one can find some ready-made lists of graded polite expressions. But it is sufficient if you ask test persons to formulate a question in different politeness grades and let other persons perform an intuitive scaling.

References

Altmann, G., Riška, A. (1966). Towards a typology of courtesy in language. *Anthropological Linguistics 8, 1-10.*

Beeching, K. (2002) *Gender, politeness and pragmatic particles in French*. Amsterdam: John Benjamins Publishing Company.

Brown, P., Levinson, S. (1987) *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Ide, S. (1989). Formal forms and discernment: two neglected aspects of universals of linguistic politeness. *Multilingua 8(2/3), 223-248.*

Jemmy, H. (2007). *What is politeness? I've never heard of it before, can I put it in my mouth?* Wigan: Pieperback Books.

*Journal of Politeness Research. Language, Behaviour, Culture***.** Berlin-New York: Mouton de Gruyter.

Lakoff, R. (1975). *Language and woman's place*. New York: Harper & Row.

Matsumoto, Y. (1988). Reexamination of the universality of face: politeness phenomena in Japanese. *Journal of Pragmatics 12: 403-426*.

Mills, S. (2003). *Gender and politeness*. Cambridge: Cambridge University Press.

Stadler, S.A. (2007). *Multimodal (im)politeness. The verbal, prosodic and nonverbal realization of disagreement in German and New Zealand English.* Hamburg: Kovac.

Watts, R. J. (2003). *Politeness*. Cambridge: Cambridge University Press.

Watts, R.J., Ide, S., Ehlich, K. (eds.) (2006). *Politeness in language. Studies in its history, theory and practice.* Berlin-New York: Mouton de Gruyter.

9.3. Word class distribution in proverbs

Problem

The definition of the concept of proverb (such as "*All that glitters is not gold*") contains at least a formal and a pragmatic part, i.e. (1) a proverb consists always

of a full sentence (it represents a proposition) and (2) it is generally known and used.

Study the frequency distribution of the parts-of-speech occurring in proverbs. Compare the result to other kinds of linguistic material.

Procedure

Assign each word in a proverb collection a part-of-speech tag, count these POS tags and arrange the data in the form of a rank-frequency distribution. Determine which theoretical probability distributions fit the data (you may expect one of the "diversification distributions" or, e.g. Zipf´s truncated zeta distribution $P_x = C/x^a$, $x = 1,2,3,\ldots, x_{max}$). Substantiate your finding.

References

Grzybek, P. (2004): A quantitative approach to lexical structure of proverbs. *Journal of Quantitative Linguistics, 1112; 79–92.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 2008, 97-111.*

Rothe, U. (ed.) (1991). *Diversification process in language: grammar*. Hagen: Rottmann.

9.4. Köhler motifs in proverbs

Problem

Study length-, frequency-, polysemy-, polytextuality sequences (or "motifs") with respect to their rank-frequency, length etc. distributions. Do proverbs display more regular patterns than other linguistic material?

Procedure

Replace the words, morphs, syllables etc. in proverbs by the values of the above-mentioned variables. Form motifs/sequences according to the method described in Köhler (2006) and Köhler/Naumann (2008) and determine whether the distribution types confirm the findings in the literature. Do the parameters of the distributions significantly differ from those of other material?

References

Grzybek, P. (2004): A quantitative approach to lexical structure of proverbs. *Journal of Quantitative Linguistics, 111−2; 79–92.*

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: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin, Heidelberg: Springer.

9.5. Semantic roles in proverbs

Hypothesis

Semantic roles display a rigid distribution pattern in proverbs.

Procedure

(1) Assign the syntactic constituents in a proverb collection semantic roles (such as AGENT, OBJECT, INSTRUMENT) and determine their frequency distribution.
(2) Each proverb can be described by means of a role pattern (such as AGENT-OBJECT-INSTRUMENT). Count the frequencies of these patterns in your collection and determine its distribution.
(3) Try to interpret the findings on the background of a synergetic-linguistic consideration of the function of proverbs (Hint: consider proverbs as linguistically coded means for everyday explanation).

References

None

9.6. Number and length of proverbs

Problem

Do you expect a relationship between the number of proverbs in a language (in a collection) and their mean length? Justify your opinion.

Procedure

Set up the length distribution of proverbs in a collection and draw a conclusion. Compute the length in two ways: In terms of number of words and in that of clauses.

References

None

9.7. Sentence structures in proverbs

Hypothesis

There is a special, very skewed rank-frequency distribution of sentence structure types in proverbs. Test the hypothesis.

Procedure

Analyse the sentences in a proverb collection with respect to their structure in terms of clauses, e.g.

Too many cooks spoil the broth	C	main clause
As you make your bed so you must lie on it	m+C	subordinate clause of manner + main clause
Don‘t count the chickens before they are hatched	C+t	main clause + subordinate temporal clause

Set up the rank-frequency distribution of the clause patterns and determine the corresponding theoretical frequency distribution. Interpret the result on the background of the function of proverbs in communication.

References

None

9.8. The recognition of variants of phraseological elements

Problem

Phraseological elements (such as proverbs, phrases, sayings, ‘citations’) are often used in a (intentionally or unintentionally) defect or modified way. What kinds of similarities/dissimilarities between an actually used variant and the corresponding original form can be postulated? How can these similarities be measured? How similar must a variant be in order to be recognized?

Procedure

Collect material from newspapers and other sources (newspaper article headings are often variants of phrases, book or movie titles etc; proverb collections contain sometimes variants of proverbs; interpersonal communication is a rich source, too). Determine the kinds of (phonological, morphemic, lexical, syntactic, …) variations in the material. Define measures of similarity/dissimilarity, e.g. based on Levenshtein edit distance. Consider different methods to combine the individual kinds of similarity (e.g. a multidimensional similarity vector).

References

Gonzalo Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33(1), 31–88.*

Levenshtein, V.I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. In: *Doklady Akademii Nauk SSSR, 163(4) S. 845–848, 1965* (Russian). English translation in: *Soviet Physics Doklady, 10(8), 707–710, 1966.*

9.9. Synonymy and (im)politeness

Problem

The more impolite a word, the more synonyms it has. Test the hypothesis.

Procedure

Consider either a dictionary of slang or a language in which politeness levels are lexically or grammatically distinguished, e.g. Javanese or Japanese. Obtain a list of impolite words (forms) and their neutral and polite equivalents. Scale the impoliteness, collect the synonyms and after performing a count express the hypothesis formally.

References

None.

9.10. Death process in dialectology

Hypothesis

The more distant two sites (places) in the area of one language are, the greater is the dialectal differences between them.

Procedure

Design a simple stochastic death process showing the decrease of similarity (phonetic, lexical, etc) with increasing geographical distance between the dialects. Use the Poisson-process replacing time by distance. Integrate a parameter representing some boundary conditions (natural barriers, bad communication …) into the formulas. Test the hypothesis using the abundant data from dialectometry.

Alternative procedure: use diffusion theory from biology, sociology etc.

Apply different methods for measuring similarity. The Levenshtein distance is a somewhat raw measure and does not express very fine phonetic differences. Consult also older literature.

References

Goebl, H. (1982) *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. (Denkschriften der Österreichischen Akademie der Wissenschaften, phil.-hist. Klasse, 157) Wien.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Thesis, Rijksuniversiteit Groningen

Nerbonne, J., Heeringa, W., Kleiweg, P. (1999). Edit distance and dialect proximity. In: Sankoff, D., Kruskal, J. (eds). *Time warps, string edits and macromolecules: The theory and practice of sequence comparison:* v-xv. Stanford: CSLI Press.

9.11. Length motifs

Problem

Determine the expected length of length motifs and compare it to the corresponding empirical mean length. Then do the same with the length of frequency and polysemy motifs.

Procedure

All kinds of motifs as defined in Köhler (2006) and also the F-motifs as earlier defined by Boroda (1982) for music have, in practice, limited lengths. As a motif is, by definition, a sequence of monotonously decreasing (or increasing) values we can regard the situation as a sequence of binary events: at each position in the sequence of the values under consideration, there is the probability p that the value is smaller than or equal to the preceding one (which would result in a length increase of the current motif by unity) and the probability $q = 1\text{-}p$ that the value at the given position is larger than the preceding one (which would end the current motif and establish the beginning of a new one). Therefore, the expected

length of a motif can be determined by means of the geometric distribution. The probability for a motif of length *x* is therefore

$$P(L = x) = qp^{x-1}, \quad x = 1,2,3,\ldots$$

Calculate *p* and *q* from your data by counting the relative number of transitions from all positions in the sequence of your values to a smaller or equal value ($= \hat{p}$) and use this number as estimation of the probability *p*. Test whether the geometric distribution fits to your data.

Consider also some problems concerning motifs in Problems Vol. 1: 50-52 and in this chapter.

References

Boroda, M.G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, J. K., Boroda, M.G., Nadarejšvili, I.Š. (eds.): *Sprache, Text, Kunst. Quantitative Analysen: 231-262.* Bochum: Brockmeyer.

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: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin, Heidelberg: Springer.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in quantitative linguistics, Vol 1.* Lüdenscheid: RAM.

9.12. Frequency and production effort (continuation)

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). Test the hypothesis on different cases.

Procedure

In Problems Vol 1, p. 75f., the hypothesis is discussed in general: its meaning, scope, formulation etc. Provided that you solved the initial problems take a linguistic phenomenon having different realizations and examine whether the hypothesis holds. Laufer (2009) examined the ways of expressing verbal aspect in German. Find other phenomena and make the hypothesis more concrete and more precise on the basis of authentic language use.

References

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

Laufer, J. (2009) (Personal communication).

9.13. Fourier analysis

Problem

Find 10 different sequences in texts which may display cyclic repetitions, "regular oscillation", some kind of wave-like movement. Capture them formally using Fourier analysis.

Procedure

Analyze either one text and take into account all well definable sequences, or consider only one type of sequence and study it in many texts. In any case, generalize the result to statements concerning the form of series on different language levels or to statements about the behaviour of the unique entity analyzed.

A practical introduction to Fourier analysis can be found in text-books on time series; a simple instruction can be found in Altmann (1988:197ff.). Comprehensive statistical software packages offer functions which do Fourier analysis automatically.

Some examples of cyclic repetitions are: (a) the number of dactyls in the sequence of verses, (b) the sequence of word or sentence lengths in text. (c) the positions of accents on words yielding a binary sequence 10010110…, (d) the sequence of distances between equal elements, etc.

Reduce the number of coefficients as far as possible. Substantiate your procedure and the results. To capture the oscillation, apply also difference equations and show whether the sequences of the same kind have the same order in all texts. Show what kind of (sequentially presented) properties are of a low order, what kinds of high order. Interpret the result and provide it with linguistic background.

References

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

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

Hřebíček, L., Altmann, G. (1996). The levels of order in language. *Glottometrika 15, 38-61.*

10. Pragmatics

10.1. Frequency distribution of speech acts

Problem

Speech acts can be considered linguistic units in a similar way as words, syllables, phrases, sentences, hrebs, etc. Hence they must abide by some regularities. Find some of them.

Procedure

Start with one of the available speech act classifications. Classifications and other descriptive means do not possess truth values; hence, there is no 'correct' one but only more or less suitable ones (with respect to a given purpose). Try different ones and test which of them is in the best agreement with quantitatively expressed regularities. Refer to the following classification (Bach, K: http://online.sfsu.edu/~kbach/spchacts.html: May 15, 2009):

> ***Constatives:*** *affirming, alleging, announcing, answering, attributing, claiming, classifying, concurring, confirming, conjecturing, denying, disagreeing, disclosing, disputing, identifying, informing, insisting, predicting, ranking, reporting, stating, stipulating*
>
> ***Directives:*** *advising, admonishing, asking, begging, dismissing, excusing, forbidding, instructing, ordering, permitting, requesting, requiring, suggesting, urging, warning*
>
> ***Commissives:*** *agreeing, guaranteeing, inviting, offering, promising, swearing, volunteering*
>
> ***Acknowledgments:*** *apologizing, condoling, congratulating, greeting, thanking, accepting (acknowledging and acknowledgment)*

Transcribe or annotate a drama and transcribe at least one act of a drama with respect to speech acts. Prepare two versions: (1) Distinguishing the dramatis personae, (2) taking the text as a whole.

Compute the frequencies of individual speech acts (if you prepare a file which contains just the sequence of speech act tags you can use a simple word counter). Version (2) can be attained from Version (1) by simple addition. Prepare the rank-order of frequencies for each person separately and also for the whole text (without distinguishing the persons).

1. Show that the rank-frequency distribution of speech acts follows one of the functions:

(a) $f(r) = ar^{-b}$

(b) $f(r) = 1 + ae^{-br}$

2. Show that the parameters *a* and *b* are different with individual persons. Is there a correlation between the parameters and some property of the person (e.g. dominance, servility, nervousness,…)? Of course, the properties of persons should be quantified, too, even if only on the ordinal scale, e.g. "dominance" between 0 – no dominance, and 10 – strong dominance.

3. Compute the mean, variance and the third central moment of the rank-frequency distribution of each person and plot them in an Ord scheme (cf. Problems Vol. 1, 111f.). How would you characterize the individual persons relative to their positions in the Ord scheme?

4. If you process a complete drama, distinguish both persons and acts. Trace down each person´s movement through the acts in the Ord scheme.

5. Compute the rank-frequency sequence and plot Ord´s functions in the Ord scheme for each act separately (do not separate the persons). Do you observe a movement from the beginning to the end of drama?

References

Alston, W.P. (2000*). Illocutionary acts and sentence meaning*. Ithaca: Cornell University Press.

Austin, J. L. (1962). *How to do things with words*, Cambridge, Mass.: Harvard University Press. (2005^2: Harvard University Press)

Bach, K., Harnish, R.M. (1979). *Linguistic communication and speech acts,* Cambridge, Mass.: MIT Press.

Cohen, A.D. (1996). Speech acts. In: McKay, S.L., Hornberger, N.H. (Eds.), *Sociolinguistics and language teaching: 383-420*. Cambridge: Cambridge University Press.

Doerge, F.C. (2006) *Illocutionary acts - Austin's account and what Searle made out of it*. http://tobias-lib.ub.uni-tuebingen.de/volltexte/2006/2273/

Grice, H.P. (1989). *Studies in the way of words*, Cambridge, Mass.: Harvard University Press.

Olshtain, E., Cohen, A.D. (1989). Speech act behavior across languages. In: Dechert, H.W. et al. (Eds.), *Transfer in production*: *53-67*. Norwood, NJ: Ablex.

Sander, Th. (2002). *Redesequenzen. Untersuchungen zur Grammatik von Diskursen und Texten*. Paderborn: mentis

Searle, J. (1969) *Speech acts: an essay in the philosophy of language.* Cambridge: Cambridge University Press.

Searle, J. (1975). Speech acts. In: Cole, P., Morgan, J.L. (eds.), *Syntax and semantics, 3: Speech acts*: *59–82*. New York: Academic Press. Reprinted in:

Davis, S. (ed.), *Pragmatics: A reader*: 265–277. Oxford: Oxford University Press. (1991)

Staffeldt, S. (2008): *Einführung in die Sprechakttheorie. Ein Leitfaden für den akademischen Unterricht.* Tübingen: Stauffenburg.

Tsohatzidis, S.L. (ed.) (1994). *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives.* London: Routledge.

Ulkan, M. (1993). *Zur Klassifikation von Sprechakten. Eine grundlagentheoretische Fallstudie*. Tübingen: Niemeyer.

10.2. Homogeneity, similarity and hierarchy of persons

Hypothesis

The performance of speech acts is different with each dramatis persona. Test the hypothesis.

Procedure

The hypothesis is reasonable. If each person of a play would perform equal speech acts, there would be no suspense, there would be no differentiation of roles. In order to test the hypothesis, use the following two methods:

(1) Use your classes of speech acts (cf. Problem 10.1) and their frequencies and perform different tests for homogeneity based on ranks. A number of such tests can be found in every text-book of non-parametric statistics. Compare each person with each other. Draw some consequences.

(2) Compare the frequencies of speech act types of all pairs of persons using the chi-square test for homogeneity. Did you obtain the same results as with rank tests? Could you set up distinct classes of persons (of course, only if there are at least ten) or are they chained?

(3) Using the significance level of each pair of persons attained by the chi-square test, set up the graph of the given act or of the whole drama, i.e. join each person with another one only if their performance of speech acts is not significantly different.

(4) Evaluate the properties of the obtained graph (Balakrishnan 1997; West 2001). Evaluate the properties of each person. Elaborate on the structure of the given stage play.

(5) Apply a similarity indicator and evaluate the similarity of dramatis personae on the basis of the frequencies of speech act types. Set up a weighted graph of the persons´ similarities and consider the similarity indicator the weight of an edge. Since all persons are adjacent, set up the hierarchy of persons (centrality) using the sum of their similarities to other persons.

References

Balakrishnan, V.K. (1997). *Graph theory.* New York: McGraw-Hill

West, D.B. (2001). *Introduction to graph theory*. Upper Saddle River, NJ: Prentice-Hall.

10.3. Distances between equal acts

Hypothesis

The distances between equal speech acts in a stage play are structured in some way. Find this structure.

Procedure

One of the ways to find structure is the study of (positional) distances between equal speech acts (i.e. speech acts of the same kind). The text should be transcribed (annotated) to (with) sequences of speech act symbols, the persons are irrelevant; at least one complete act of the drama must be examined in order to get reliable results. Since persons have their own attitudes, speech customs and communication strategies, it can be conjectured that Skinner´s hypothesis holds, namely that there is an increasing probability of the appearance of the same unit in short distance (cf. Problems Vol. 1, 56). Hence, there are more short distances than long ones. One can measure the distance in terms of the number of different speech acts between two equal ones plus 1 or as the number of "steps" from one speech act to the next identical one.

According to the hypothesis, the distances are not distributed uniformly (equiprobably) but they represent a monotonously decreasing sequence. Approximate this sequence by means of the Zipf-Alekseev function

$$y = ax^{-b-c\ln x}$$

where x is the distance (x = 1,2,3,…), y is the number of occurrences of that distance and a,b,c are parameters. If the above function is not adequate, find a more adequate one.

Perform this procedure for each text part separately and scrutinize the development of the parameters from the first part to the last one. Then add the distances and analyse the drama as a whole.

Find other kinds of structure in the sequence of distances.

References

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.*
Skinner, B.F. (1957). *Verbal behaviour*. Acton: Copley
Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22.*

10.4. Scaling of speech acts

Problem

Set up some kind of scaling for the kinds of speech acts.

Procedure

At some point in the advancement of science, qualitative classification alone does not allow to gain a deeper insight into the mechanisms of the domain. Here Galileo´s dictum: "… measure the measurable and try to render measurable what is not yet" is to be followed. Here, forming a scale means to map the speech acts into a specific dimension. There can be different dimensions: the status of the speaker, the attitude of the speaker to the hearer, the emotion expressed by the given kind of speech act, the weight or the intensity of the speech act, e.g. *asking, begging, requiring, urging, ordering* have different "urgency" or "weight" parameters, and at the same time they express a certain attitude.

However, before one begins to construct scales, one should set up some hypotheses for whose testing such a scaling is necessary or at least reasonable. The following examples may illustrate this: (a) The more dominant a person in a drama, the more "weighty" are its speech acts, or (b) the degree of emotionality (of speech acts of a person) is a function of protagonism, etc. Dominance and protagonism have to be measured, of course, independently.

In this way one would obtain more exact quantitative concepts. A classification would, as a matter of fact, not be necessary any more; there would be measurable properties and the analyzed text could be transcribed as a sequence of attribute-value pairs (or, abbreviated in case of a single analysed dimension, simply as numerical values).

Every researcher engaged in studying speech acts has an intuitive idea of the "weight" of a phrase. If one tries to transform this intuition in numbers, the scale may be set up. Take the analogy to politeness (Chapter 9.2) which can intuitively be estimated by every native speaker.

References

None

10.5. Distribution of scaled values of speech acts

Hypothesis

The distributions of speech act values are not homogeneous in individual acts of drama. Test the hypothesis.

Procedure

If the individual speech acts of the drama are transcribed as values on a scale as established in Problem 10.4, one obtains a sequence of numbers. With the frequency of occurrence of individual degrees one can obtain a frequency distribution of the given property. Now, since a drama has a certain dynamic behaviour beginning with a conflict, continuing with its increase up to a climax and then the fall to catharsis, each act will contain speech acts of different degrees. It can be conjectured that all distributions will abide by the same principle but their forms (parameters) will be so different that they must display significant heterogeneity. Heterogeneity may be tested using the chi-square test and the dynamics of the drama can be described for example as a sequence of averages of degrees.

Further hypotheses about the genesis of the distributions will be possible as soon as the first data have been produced.

References

None.

10.6. Weight motifs

Problem

The subsequent weight of intensity values of speech acts (cf. the previous problems) form sequences in analogy to Köhler's motifs. In fact, weight value sequences are just another form of these motifs. Hence, all the investigations and methods that have been shown in the corresponding literature can be performed on speech act property values as well.

Procedure

Choose one of the possible speech act frames, i.e. complete texts, individual acts or chapters, speech acts of single persons etc. Start with a sequence of values resulting from one of the previous problems and form chunks according to the definition of Köhler motifs (cf. the references): A motif begins with the beginning of the frame (i.e. the text or act …) or where the previous motif ends. The current motif ends when the next value is smaller than the current one. Hence, a value sequence such as

2-3-3-5-3-4-4-2-1-3-5

is chopped into the motifs

2-3-3-5, 3-4-4, 2, and 1-3-5.

These motifs are units which can be studied with respect to their frequencies (i.e. how often does the motif 3-4-4 occur within the frame), to their length (e.g., the motif 2-3-3-5 has length 4) etc. You can analyse the rank-frequency distribution of the motifs (expect the Waring or the Zipf-Mandelbrot distribution), the length distribution (hyper-Pascal or hyper-Poisson d.) etc.

Moreover, these studies can be performed also on higher scales. e.g. you can scrutinize the length of length-motifs, i.e. you consider the sequence of length values of the motifs of the first order as a new level of analysis and form on this level new motifs following the above-given definition. This procedure can be repeated until very few motifs are left on the last level.

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. (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe 416-421*. Bratislava: VEDA.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg*: *637-646.* Berlin, Heidelberg: Springer.

10.7. Drama as a time series of speech acts

Problem

Find the difference equation of the lowest order simulating the sequence of scaled speech acts in a drama.

Procedure

If the speech acts are scaled, then a text can be represented as a sequence of numbers. But since a stage play is an incessant clash of speech acts and the persons speak alternately, the degrees of speech acts may alternate, too. The kind of alternation (oscillation) produced depends on the character of stage play. In

any case, it is easy to get the lowest degree of the difference equation capturing this oscillation. Appropriate software yields the results mechanically.

Having analyzed several stage plays find a relationship between the kind of stage play and the order of the difference equation. Strive for an interpretation.

References

None

10.8. Some properties of speech act sequences

Problem

Find and compute some other properties of the sequence of speech act values.

Procedure

Consider each act of a stage play separately. Replace the speech acts by the degree on some scale. Having the sequence of numbers, compute Hurst´s exponent, Lyapunov´s coefficient and Minkowski´s sausage, cf. Problems Vol. 1, p. 49, 53, 54. Consider the changes of these quantities in the course of the stage play. Do they change or are they constant? Can you explain the behaviour of speech acts or the background of the stage play on the basis of these numbers? Are these coefficients in correlation with other properties of stage plays?

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). *Lectures on text theory. Prague*: Oriental Institute.

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.

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

10.9. Drama and comedy

Problem

Compare a drama and a comedy in all aspects of speech acts.

Procedure

Compare at least one drama and one comedy by means of all quantitative data obtained in Problems 10.1 to 10.8. Which aspects yield the greatest differences? Interpret the differences and the equalities and establish some rules for stage plays.

References

None.

10.10. The development of drama

Problem

Since Classical Greek dramas differ from modern ones there is necessarily a kind of continuous change. Capture it on the basis of speech acts.

Procedure

Analyze dramas of different epochs in one or several languages. First collect all speech acts, process them quantitatively, compute the characteristic features (cf. Problems 10.1 -10.9) and show the development, cultural differences etc.

References

None

10.11. Speech act hrebs

Problem

Is it possible to set up hrebs based on speech acts?

Procedure

Transcribe the text of a drama in form of speech acts. Then consider all sentences of a person containing the same speech act as belonging to the same hreb. A sentence can belong to several hrebs. State the inventory of hrebs and its size of each person in terms of sentence numbers contained in them. Show that the dramatis personae differ in this respect. For the definition of hreb see Problems Vol. 1, p. 47-48.

Set up the frequency distribution of hreb size and derive a probability distribution or at least find inductively a distribution capturing it.

Do dramas differ from comedies in this respect?

Is it possible to determine units consisting of speech act sequences? Are there characteristic sequences for some persons?

Develop the speech act "theory" in this direction.

References

Cf. Problem 10.1.

10.12. Towards a theory of speech acts

Problem

Is the science of speech acts a theory or only a classified collection of defined concepts? Give arguments for one of these views.

Procedure

If the science of speech acts is more than a well classified collection of concepts, a part of which is well defined, another part rather fuzzy, give arguments for its theoretical status. Consult the available literature and collect hypotheses concerning some relationships, dependencies, development, systemic status etc of speech acts. If necessary, support the argumentation psychologically or sociologically. Express the hypotheses formally, derive them from some preliminary axioms (conjectures) or propose at least a formula. However, in any case show that the hypothesis is testable and give at least a hint at the possibility of an objective test.

If you do not find such a possibility in the literature, create some hypotheses.

References

Cf. Problem 10.1

10.13. Length of dialogue contributions

Problem

Determine the distribution of the contribution lengths in dialogues/polylogues.

Procedure

Determine the lengths of the individual contributions of the dialogues in stage plays, movie scripts, and in spontaneous speech as recorded in spoken language corpora. Assign the contributions to their originators and calculate the frequency

distributions of the contribution lengths (in terms of the number of sentences) (a) for the totality of the text, (b) for the individual participants. You will have to pool the length values in intervals such as 1-5, 6-10, etc in order to provide sufficient data in all classes.

(a) Which probability distribution do you expect to fit the data? Conduct a fit and test your hypothesis. Substantiate the distribution by linguistic arguments.

(b) Do the distributions of the individual participants differ significantly (i) within a text, (ii) among the texts and text sorts?

(c) Are the distributions and their parameters related to (i) the number of contributions of the individual participants, (ii) the number of participants of the polylogue, (iii) the (social) status of the persons in the dialogue/polylogue?

References

None

10.14. Discourse frequency (1)

Problem

Determine the rank-frequency distribution of grammatical categories in discourse.

Procedure

Select one or more grammatical categories depending on the language(s) you are going to study: case, number, gender, person, tense, aspect, diathesis, definiteness, etc. Collect texts from a single text sort such as novel, report, newspaper text, interview, dialogue, etc or use an appropriate (sub-)corpus. Count the different grammatical features according to the categories you selected and set up a frequency table (distribution). Arrange the different forms in the order of their frequencies.

Can you fit a theoretical probability distribution to the resulting data?

Hint: If your texts are very long, study individual texts one by one. Otherwise two methodological problems may arise: (1) Inhomogeneity of the data and (2) too big data sets, which may cause the chi-square test to fail.

If you do not succeed in fitting a probability distribution, use a simple continuous function or a series, i.e. skip normalization.

If your data still resist modelling, find the outliers and explain their deviation from the general trend using linguistic arguments.

References

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13, 105-120.*

Myhill, J. (2005). Quantitative methods of discourse analysis. In: Köhler, R., Altmann, G., Piotrovskij, R.G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*: *471-798.* Berlin, New York: de Gruyter.

10.15. Discourse frequency (2)

Hypothesis

The block-wise frequency distribution of grammatical categories in texts abides by Frumkina's law. Test the hypothesis.

Procedure

In analogy to the block-wise distribution of function words (shown by Frumkina (1962) and others given in the references) and of syntactic constructions/functions (cf. Köhler 2001), determine the number of text blocks (try block sizes of 50, 100, 200 words), in which there are 0, 1, 2, ... occurrences of selected grammatical categories (e.g., plural, dual, genitive, future tense, ...). Fit the negative hypergeometric distribution ("Frumkina's Law") to the data. Observe dependences of the parameter values on block length, number of blocks, category type.

The negative hypergeometric distribution is defined as

$$P(X = x) = \frac{\binom{M + x - 1}{x}\binom{K - M + n - x - 1}{n - x}}{\binom{K + n - 1}{n}}, \quad x = 0,1,...,n$$

where *K*, *M* and *n* are parameters.

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, 146-167.*

Bektaev, K.B., Luk´janenkov, K.F. (1971). O zakonach raspredelenija edinic pis´mennoj reči. In: Piotrowski, R.H. (ed.), *Statistika reči i avtomatičeskij analiz teksta: 47-112.* Leningrad: Nauka.

Brainerd, B. (1972). Article use as an indirect indicator of style among English-language authors. In: Jäger, S. (ed.), *Linguistik und Statistik: 11-32*. Braunschweig: Vieweg.

Francis, I.S. (1966). An exposition of a statistical approach to Federalist dispute. In: Leed, J. (ed.), *The computer and literary style: 38-78*. Kent, Ohio: Kent State University Press.

Frumkina, R.M. (1962). O zakonach raspredelenija slov i klassov slov. In: Mološnaja, T.N. (ed.), *Strukturno-tipologičekie issledovanija; 124.33.* Moskva: Akademija Nauk SSSR.

Köhler, R. (2001). The distribution of some syntactic construction types in text blocks. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.). *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 136-148*. Trier: Wissenschaftlicher Verlag.

Maškina, L.E. (1968). *O statističeskich metodach issledovanija leksiko-gramatičeskoj distribucii.* Minsk: Diss.

Mosteller, F., Wallace, D.L. (1964*). Inference and disputed authorship: The Federalist*. Reading, Mass.: Addison-Wesley.

Paškovskij, V.E., Srebrjanskaja, I.I. (1971). Statističeskie ocenki pis´mennoj reči boľnych šizofreniej. In: *Inženernaja lingvistika*. Leningrad: Nauka.

Piotrowski, R.G. (1984). *Text, Computer, Mensch*. Bochum: Brockmeyer.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

10.16. Discourse frequency (3)

Hypothesis

The proportion of simplicia, derivatives, compounds, compound-derivatives etc in a text depends on text type and age of the text. Test the hypothesis.

Procedure

Collect texts from different text types and authors. Each text type and each author should be represented by a number. Determine the number of words of each word-formation type and calculate the proportions for the individual texts. Perform a statistical test to show whether the differences are significant and may be characteristic of text types and/or authors. Determine those word-formations which are most specific for a given text, i.e. whose proportions significantly differ from those in other texts or groups of texts. Set up a rank-order of word-formations according to their discriminative power. Interpret the results and substantiate it linguistically or psychologically.

References

Text-books of statistics.

10.17. Rhetorical structure (1)

Problem

Study several texts with respect to the categories of the "Rhetorical Structure Theory" (RST) and determine the frequency distribution of the categories.

Procedure

Tag several texts according to RST analysis. Count the numbers of occurrences of the different tags without respect to their position in the structure.

Can you fit a theoretical probability distribution to the data?

Which distribution or what kind of distribution do you expect? Alternatively: How could the resulting distributions be interpreted or even explained?

References

Mann, W.C., Thompson, S.A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text 8(3), 243-281.*

http://www.sfu.ca/rst/01intro/intro.html (Sept. 23, 2009)

10.18. Rhetorical structure (2)

Problem

Does the distribution of RST tags depend on the position of the tags in the structure?

Procedure

Tag several texts according to RST analysis. Count the numbers of occurrences of the different tags with respect to their position in the structure: Count the frequencies separately for

(a) the levels of embedding

(b) position in the text in terms of the number of tags from the text beginning

(c) position in the sub-structure in terms of the number of tags from the beginning of the sub-structure.

Can you fit a theoretical probability distribution to the data?

Which distribution or what kind of distribution do you expect? Alternatively: How could the resulting distributions be interpreted or even explained?

References

Mann, W.C., Thompson, S.A. (1988). Rhetorical structure sheory: toward a

functional theory of text organization. *Text 8(3), 243-281.*
http://www.sfu.ca/rst/01intro/intro.html (Sept. 23,.2009)

10.19. Rhetorical structure (3)

Problem

Define some properties of RST units.

Procedure

As any unit, RST units can be given numerous interesting properties. Define properties such as complexity in analogy to the properties of syntactic structures as defined in Köhler (1999).

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*
http://www.sfu.ca/rst/01intro/intro.html (Sept. 23, 2009)

10.20. Rhetorical structure (4)

Problem

Set up hypotheses about interrelations between properties of RST units and test them.

Procedure

Postulate interrelations (dependencies) between properties such as node level in the RST structure, position in the text or in a given sub-structure, frequency or complexity of RST units.

Test these hypotheses on the data obtained by solving the problem "Rhetorical structure (2)". The analogous study on the syntactic level in Köhler (1999) can be used as orientation.

References

Köhler, R. (1999). Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics 6(1), 46-57.*
http://www.sfu.ca/rst/01intro/intro.html (Sept. 23, 2009)

Author index

Subject index