Studies
in
Quantitative Linguistics
12

Radek Čech
Gabriel Altmann

Problems in Quantitative Linguistics 3

RAM - Verlag

Problems in Quantitative Linguistics 3

by

Radek Čech
Gabriel Altmann

Dedicated to Reinhard Köhler on the occasion of his 60th birthday

2011
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

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.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.

ISBN: 978-3-842303-08-8

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

Preface

The third volume of *Problems* has the same structure and the same aim as the first two volumes. The problems can be used for writing essays for final examinations, dissertations or, in European context, "habilitations". The authors would be very pleased if solutions to some problems would be published in some appropriate journals. They also hope that each problem presented here will be developed further both in empirical and theoretical directions. In many cases only suggestions and hints are given but the authors hope that the readers will go their own ways and systematize the given problem, i.e. subsume it under a more general problem, link it with other phenomena and, at last, propose a theoretical derivation with linguistic substantiation. However, even empirical testing using other texts/languages and empirical generalizations would be of great value.

Each problem automatically evokes other problems which are either collateral or hierarchic, i.e. concern problems at the same linguistic level or at a different one. As a matter of fact, local problems can be solved in isolation adequately or ad hoc, but in theoretical research the only criterion of a useful solution is systematization. Hence we recommend the reader not to stop at the solution of a problem for a single text or text sort or author or language but to extend the analysis at least by means of empirical generalization to different objects. Since the readers are linguists, the extension of the scope of study to other languages may not be a serious problem.

In the present volume the problems are classified into six groups: phonology and script, grammar, semantics, synergetics, text analysis and a group of mixed problems. If there is a hint to a similar problem in the previous volumes, it is recommended to solve first the simpler task and generalize the problem step by step.

The mathematics necessary for solutions is simple and can either be found in text-books of statistics, or, if not, then the reader finds instructions directly in the "Procedure" accompanying each problem. It is recommended to read at least some of the works given in the References. Many times we quote old works in order to show that the problem itself is nothing new but obtains a different look in its new quantitative costume.

Acknowledgments
We are obliged to Eric Wheeler who read the book and helped us to improve our style. Radek Čech was supported by the Czech Science Foundation, grant no. 405/08/P157.

Radek Čech
Gabriel Altmann

Contents

1. Phonology and script

1.1. Phoneme distribution: structural gaps

Hypothesis

The greater the number of phonemes in an inventory (*P*), the smaller is the proportion of realized phoneme combinations, i.e. the greater the proportion of structural gaps (*G*)**.** Test the hypothesis.

Procedure

There is no language in which all phoneme combinations would be exploited (cf. Goldrick 2004; Kawasaki, Ohala 1981; Kawasaki-Fukimori 1992). In every language there are structural gaps, i.e. missing phoneme combinations e.g. /bp/ or /pb/ within a morpheme or syllable in English. It can be expected that the greater the phoneme inventory, the more gaps can be (proportionally) expected, since language strives for "easy" combinations (clusters) and sufficient redundancy and does not need to exploit all possibilities. The omitting of certain combinations can be compensated by means of functional equivalents like tone, stress, word prolongation, etc.

State all phoneme combinations (*R*) in many languages, i.e. collect them from the existing literature or analyse extensive corpora. No frequencies are necessary. Then for a class *i* of languages having the same phoneme inventory size (P_i) compute the mean proportion of gaps. The size of a gap in an individual language can be computed as

$$G = 1 - \frac{R}{P^2} = 1 - \frac{number\ of\ \ phoneme\ combinations}{square\ of\ \ phoneme\ inventory\ size}$$

and the mean is the sum of *G* of all *k* languages having the same inventory size *P* divided by *k*. One needs about 200 languages to obtain reliable results. If some *P* class is not sufficiently represented, then several classes should be pooled and also the mean *P* should be stated.

First obtain the empirical values *of* <*G*,*P*>, then derive the function $\bar{G} = f(\bar{P})$ from some assumptions and substantiate the derivation linguistically.

In order to facilitate the collecting of data, we reprint the cases prepared by E. Kelih (2009) who proposed also some other relationships between the number of vowels, consonants and phoneme combinations and presented all references.

Table 1
Inventory sizes (*P*) and realized phoneme combinations (*R*) in 75 data (from E. Kelih 2009)

Language	**P**	**R**		**Language**	**P**	**R**
Rotokas	11	85		Pāli	33	335
Hawaiian	13	104		Sirionó	33	580
Maori	15	125		Old Church Slavic	34	540
Huichol	20	172		Sarakatšan	35	483
Kaiwa (Guarani)	20	185		Hungarian	37	922
Kurija	21	187		Czech	37	826
Hittite	21	233		Ardchama-gadchi	32	292
Kikongo	22	222		Amuesh	32	477
Sierra Nahuat	23	339		Aromanian	38	627
Spanish (Costa Rica)	23	290		Arabic (Cyprus)	38	736
New Islandic	23	325		Ukrainian (2)	38	826
Guarani	24	244		Vedic	39	740
Luba	24	241		Russian (2)	39	910
Lomongo	25	303		Russian (1)	39	908
Ayacucho-Quechua	25	379		Polish-Dialect (1)	40	610
Basque (Maya)	26	277		Polish-Dialect (2)	40	696
Slovenian	26	519		Polish (2)	40	883
Mvera	26	363		Polish (1)	41	920
Cuicateco	26	402		Sanskrit (1)	41	869
Greek B	26	378		Sanskrit (2)	41	868
Duala	27	319		Sanskrit (3)	41	860
Attic	27	550		Lituanian-Dialect 1	42	584
Totonaco	27	562		Lituanian-Dialect 2	42	666
Old Japanese	28	294		Sanskrit (4)	43	853
Lingala	29	266		Arabic (Egypt)	43	1267
Indonesian (1)	29	459		Ukrainian (1)	43	867
Indonesian (2)	29	473		Kashmiri	43	840
Modern Japanese	30	328		Belorussian-Dialect 1	44	742
Songe	30	420		Belorussian-Dialect 2	44	817
Toyolabal (Mayan)	30	447		Ganda	45	668
Kikuju	31	503		Slovak	46	1066
Serbocroatian (1)	31	690		Bambara	48	649

Serbocroatian (2)	31	542	Khmer (1)	49	1025
Maharasti	32	300	Khmer (2)	49	870
Mahadhi	32	305	Malyalam	51	763
Šauraseni	32	318	Punjabi	51	1001
Macedonian	32	618	Lituanian	53	719
American English	32	661			

References

Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language 51, 586-603.*

Kawasaki, H., Ohala, J.J. (1981). Acoustic basis for universal constraints on phoneme combinations. *Journal of the Acoustical Society of America 70.*

Kawasaki-Fukimori, H. (1992). An acoustical basis for universal phonotactic constraints. *Language and Speech 35, 73-86.*

Kelih, E. (2009). Phonemverbindungen und Inventarumfang: Empirische Evidenz und Modellentwicklung. *Glottotheory 2(1), 60-74.*

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

Kleinlogel, A., Lehfeldt, W. (1972). Zur Problematik einer syntagmatisch-phonologischen Sprachklassifikation. In: Jäger, S. (ed.), *Linguistik und Statistik: 51-64.* Braunschweig: Vieweg.

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

1.2. Stress placement in isosyllabic poems

Problem

In isosyllabic poetry each verse has the same number of syllables, however, the stress does not need to be distributed deterministically, i.e. on the parallel syllables in all verses. Perform Fourier analysis (Howell 2001; Stein, Weiss 1971) of the stress placement.

Procedure

Take an isosyllabic poem and compute the proportion of stresses in each position of the verses. You obtain an irregular oscillating sequence.

Express the sequence mathematically using either Fourier series or difference equations of higher order. The order should not surpass the half number of syllables in the verse.

Perform the analysis for a whole poetical "school" or a period and perform the following characterisations.

(1) Join the sequence of proportions in individual positions with straight lines and compute the length of the arc. The longer the arc length, the more deterministic is the stress placement, the stronger is the placement tendency. Explain why.

(2) In order to be able to compare poems with different verse lengths, compute for every length the maximum arc length. Then divide the observed arc length by this maximum in order to obtain relative (comparable) measures. Normalize this measure to obtain an indicator in the interval <0,1>. The simple division by the maximum does not yield an indicator in <0,1> because an arc of zero length does not exist here.

(3) Consider the proportions of stresses in individual positions and compare them with the discrete uniform distribution. Use the difference as an indicator of the strength of the stress-placing tendency.

(4) Study the historical development of the strength of stress-placing tendency in the isosyllabic poetry of a language. Is there an increasing or decreasing tendency? Set up a test for comparing your indicator in different poems.

(5) Compare stress-placing in isosyllabic poetries of two different languages. If there are differences, is it caused by language, author, or individual texts? Use your indicator.

(6) Consider the course of stress placement in two similar poems (i.e. having identical syllable numbers in the line) and perform a simple homogeneity test. Then collect all poems with the same isosyllabicity in one language and compare them as a whole with a similar group in another language. Use the information discrimination test.

References

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

Howell, K.B. (2001). *Principles of Fourier Analysis*. Boca Raton – London – New York – Washington, D.C.: Chapman & Hall/CRC Press.

Stein, E.M., Weiss, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton: Princeton University Press.

1.3. Vowel duration

Hypotheses

(1) Test the hypothesis that the duration of vowels in a text measured in milliseconds (x) abides by the function $y = ax^b exp(cx + d/x)$.
(2) Test the hypothesis that the duration of an individual vowel depends on word length (measured in terms of phoneme numbers) and the dependence abides by a power function (Best 2008).

Procedure

Take a text a let it read loud by a test person. Use a program which enables you to measure the vowel length. Then ascribe the vowels of given length to respective intervals. Consider X the mid of the intervals and Y the number of vowels whose duration falls in the given interval. Fit function (1) to the data and state whether it is adequate. The text must be long enough. If you want to transform (1) in a distribution (with a as normalizing constant), remember that each interval should contain at least 5 observations in order to be reliable.

(a) Compare the results originating from the same text but from different test persons. (b) Fit the function to different texts spoken by the same person. If you obtain differences in parameters, study them and draw conclusions. If the above function does not hold, change the hypothesis and substantiate it linguistically (cf. Geršić, Altmann 1988; Santen 1992; Bergem 1993; Schiavetti et al. 2004).

Use the same spoken texts but this time study the duration of the (phonemically) same vowel in words of different length. Show that the length of an individual vowel depends on the length of the words in which is occurs. The length of words should be measured in terms of syllable numbers. That is, state the duration e.g. of an /a:/ in words of length 1 syllable, 2 syllables etc. Does the dependence follows a power function ($y = ax^{-b}$)? If not, formulate a new hypothesis and substantiate it linguistically.

References

Bergem, D.R.v. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication 12 (1), 1-23.*

Bergsveinsson, S. (1941). *Grundfragen der isländischen Satzphonetik*. Copenhagen: Munksgaard.

Best, K.-H. (2008). Gesetzmäßigkeiten der Lautdauer. *Glottotheory 1(1), 1-9.*

Geršić, S., Altmann, G. (1988). Ein Modell für die Variabilität der Vokaldauer. In: Schulz, K.P. (ed.), *Glottometrika 9, 49-58.* Bochum: Brockmeyer.

Laziczius, J. v. (1939). Zur Lautquantität. *Archiv für vergleichende Phonetik 3, 245-250.*

Santen, J.P.H.van (1992). Contextual effects on vowel duration. *Speech Communication 11(6), 513-546.*

Sievers, E. (1876). *Grundzüge der Lautphysiologie zur Einführung in das Studium der indogermanischen Sprachen.* Lepzig: Breitkopf & Härtel.

Schiavetti, N., Metz, D.E., Whitehead, R.L., Brown, S., Borges, J., Rivera, S., Schultz, C. (2004). Acoustic and perceptual characteristics of vowels produced during simultaneous communication. *Journal of Communication Disorders 37(3), 275-294.*

1.4. Homophones

Problem

Homophones destroy redundancy and increase the decoding effort of the hearer. Some languages reduce their rise by suprasegmental means, by increasing the number of phonemes in the inventory, by prolonging the words, by admitting more phoneme combinations, etc. (1) Set up a control cycle with all factors exerting influence on homophone/homonym formation and specify the requirements involved (cf. Köhler 2005; Miozzo, Caramazza 2005; Caramazza et al. 2001; Jescheniak et al. 2003).

(2) Ogura and Wang (2006) studied the development of homophones in Japanese and English and showed two regularities:

(a) There is a tendency to brake the increase of the number of homophones, i.e. there are more homophones containing 2 items than 3, 4,… items. Test the hypothesis using Ogura and Wang´s data in Japanese and English using Tables 1 and 2 below.

(b) Creating of homophones depends also on word length, however, this tendency need not be monotonously decreasing. Using the tables of Ogura-Wang propose a bivariate distribution expressing this dependence. You may begin with univariate fitting some distribution to the rows of the tables and then constructing a bivariate distribution.

Procedure

(1) Elaborate a list of possible factors and substantiate their influence linguistically. Then draw a figure analogous to those in Köhler (2005) with homonyms/ homophones in the mid and arrows from the influencing factors. Place a plus or minus on the arrow according to the kind of influence. In the second step perform analyses in several languages and for each relationship derive a simple formula. Test the formula using your data.

(2) Tendency (a) can be studied using the last columns of Table 1 and Table 2.

(b) Fit a distribution to individual rows or to the row containing the totals. If the result is satisfactory, create a bivariate distribution and fit it to the inner columns and rows of the tables. Substantiate the distribution linguistically.

Perform all procedures using other languages and obtaining data from dictionaries that can be found on the Internet.

Table 1
Number of homophones in English (Ogura, Wang 2006)

#Homophone words	Word length (in terms of syllable numbers)							Total
	1	2	3	4	5	6	7	
2	3068	3888	1792	588	132	8	6	9482
3	900	477	123	60	3	0	0	1563
4	460	104	16	0	0	0	0	580
5	175	40	0	0	0	0	0	215
6	96	0	0	0	0	0	0	96
7	28	0	0	0	0	0	0	28
8	16	0	0	0	0	0	0	16
Total	4743	4509	1931	648	135	8	6	11980

Table 2
Number of homophones in Japanese (Ogura, Wang 2006)

#Homophone words.	Word length (in terms of mora numbers)						Total
	1	2	3	4	5	6	
2	26	446	1404	2268	108	12	4264
3	27	231	687	885	3	0	1833
4	32	132	420	520	0	0	1104
5	5	125	165	225	0	0	520
6	12	108	132	132	0	0	384
7	21	14	84	126	0	0	245
8	0	56	48	56	0	0	160
9	9	45	45	27	0	0	126
10	0	50	30	0	0	0	80
11	0	0	33	0	0	0	33
12	0	12	12	12	0	0	36
13	13	13	0	0	0	0	26
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0	0	16	0	0	0	16
Total	145	1232	2976	4251	111	12	8827

References

Antilla, R. (1989). *Historical and comparative linguistics.* Amsterdam: John Benjamins.

Caramazza, A., Costa, A., Miozzo, M., Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition 27(6), 1430-1450.*

Jescheniak, J.D., Meyer, A.S., Levelt, W.J.M. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa, et al. (2001) and New Experimental Data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29 (3)432-438.*

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

Miozzo, M., Caramazza, A. (2005). The representation of homophones: Evidence from the distractor-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition 31(6), 1360-1371.*

Ogura, M., Wang, W.S-Y. (2006). Ambiguity and language evolution of homophones and syllable number of words. *Studia Anglica Posnaniensia 42, 3-30.*

http://www.ling.ed.ac.uk/evolang/2004/ABSTRACTS/POSTERS/ogura-wang.txt (retrieved June 21, 2010)

Weber, S. (2005). Zusammenhänge (Interrelations). In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 214-226.* Berlin: de Gruyter.

Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types. *Glottometrika 14, 190-196.*

1.5. Phonetic difference and frequency

Problems

1. State whether the phonetic difference between the consonants in bi-phonemic clusters is normally distributed. If not, find the distribution using linguistic arguments. The role of frequency is played by the number of different clusters having the given phonetic difference. (X = phonetic difference, n_x = number of cluster with difference x)

2. State whether the phonetic difference between the members of a cluster and its frequency are correlated. Find the dependence function and substantiate it lin-

guistically. According to Saporta (1955) this distribution is normal. The frequencies are, as a matter of fact, weights of phonetic differences.

Procedure

If possible, use only already published results. Thereby you can save much work and especially avoid decisions about how to measure phonetic/phonemic similarity. The literature about this theme is overwhelming, you can choose several methods.

First collect lists of consonant clusters from individual languages. Either one lets a program analyze a corpus in its phonemic form or one takes ready-made results from the works concerning phoneme distribution (for a short list of sources cf. Altmann, Lehfeldt 1980; Lehfeldt 2005). Then compute the phonetic/phonemic difference between the members of the clusters. Use phonemic descriptions of the given language and some kind of similarity measure. There is a great number of such measures; strive for using the same method for all your data. Simple methods can be found in the references (cf. a Campo et al. 1989; Afendras et al. 1973; Altmann 1969; Austin 1957; Avram 1865/66, 1967, 1972; Baddeley 1966; Bailey, Hahn 2001, 2005; Black 1970; Frisch 1966; Frisch, Pierrehumbert, Broe 2004; Geršić 1971; Grimes, Agard 1959; Heike 1961; Imperl, Kacic, Horvat, Zgank 2003; Kloster-Jensen 1966; Kučera 1964; Kučera, Monroe 1968; Ladefoged 1970; Luce, Pisoni, Goldinger 1990; Marslen-Wilson, Moss, van Harlen 1996; Meyer-Eppler 1969; Miller, Nicely 1955; Naumann 1976; Perebyjnis 1970; Peterson, Harary 1961; Pierrehumbert 1993; Saporta 1955; Tolstaja 1968; Vinogradov 1966; Wang, Bilger 1973; Wickelgren 1965, 1966; Winitz, Bellerose 1963; Yu, Kim, Oh 1995). Of course, many other methods can be found.in the literature.

Determine the difference for each cluster and set up a difference scale. If the scale is discrete, pool some neighbouring difference classes containing few clusters; if the scale is continuous, set up intervals. For each difference state the number of clusters and, if you analyzed texts or a corpus, all frequencies. Determine the distribution or at least a curve which holds for all languages you analyzed. If this is impossible, search for boundary conditions which force you to modify the model. If you obtain different models, strive for incorporating them in a general model.

Do not start from the conjecture that any of the distributions must be normal, reckon with skewness and explain why it must be so.

If you used several similarity measures, compare their performance, derive their sampling properties (expected value and variance) and set up an asymptotic test for comparisons.

In addition, collect the complete literature concerning sound/phoneme differences and publish it in form of a historical survey showing the advantages and disadvantages of individual methods. Do not consider “advantage” equival-

ent to “truth” but choose other criteria (“easiness”, “completeness”, “ambiguity”, etc.).

References

a Campo, F.W., Geršić, S., Naumann, C.L., Altmann, G. (1989). Subjektive Lautähnlichkeit 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.*

Altmann, G. (1969). Differences between phonemes. *Phonetica 19, 118-132.*

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

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

Avram, A. (1965/66). La classification des phonèmes selon leur degré de parenté. *Acta Linguistica Hafnensia 9, 173-178.*

Avram, A. (1967). Sur la distance entre phonemes. *Cahiers de linguistique théorique et appliqué 4, 17-21.*

Avram, A. (1972). Sur la distance entre les traits phonologiques distinctifs. *Cahiers de linguistique théorique et appliqué 9, 171-175.*

Baddeley, A.D. (1966). Short term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology 18, 362–365.*

Bailey, T.M., Hahn, U. (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44, 568–591.*

Bailey, T.M., Hahn, U. (2005). Phoneme similarity and confusability**.** *Journal of Memory and Language 52(3), 339-362.*

Black, J.W. (1970). Interphonemic distance. *Actes du X^e Congrès Internationale des Linguistes III: 265-278.* Bucharest.

Frisch, S.A. (1996). *Similarity and frequency in phonology.* PhD thesis, Dept. of Linguistics, Northwestern University, Evanston Illinois. http://roa.rutgers.edu.

Frisch, S.A., Pierrehumbert, J., Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory 22, 179–228.*

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.*

Heike, G. (1961). Das phonologische System des Deutschen als binäres Distinktionssystem. *Phonetica 6, 162-176.*

Imperl, B., Kacic, Z., Horvat, B., Zgank, A. (2003). Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones**.** *Speech Communication 39(3-4), 353-366***.**

Kloster-Jensen, M. (1966). L´idée de resemblance phonétique. *Cahiers de linguistique théorique et appliqué 3, 91-97.*

Kučera, H. (1964). Statistical determination of isotopy. In: *Proceedings of the Ninth International Congress of Linguists: 713-721.* The Hague: Mouton.

Kučera, H., Monroe, G.K. (1968). *A comparative quantitative phonology of Russian, Czech and German.* New York: Elsevier.

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

Lehfeldt, W. (2005). Phonemdistribution. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 181-190*. Berlin: de Gryuter.

Luce, P.A., Pisoni, D.B., Goldinger, S.B. (1990). Similarity neighbourhoods of spoken words. In: Altmann, G.T.M. (ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives: 122–147*. Cambridge, MA: MIT Press.

Marslen-Wilson, W., Moss, H.E., Harlen, S.v. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22, 1376–1392.*

Mesgarani, N., David, S.V., Fritz, J.B., Shihab A., Shamma, Sh.A. (2010). Phoneme representation and classification in primary auditory cortex. http://www.engr.washington.edu/epp/iwaenc2008/proceedings/contents/papers/Phoneme_ASA.pdf (retrieved 17.7.2010)

Meyer-Eppler, W. (1969*). Grundlagen und Anwendungen der Informationstheorie.* Berlin: Springer.

Miller, G.A., Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27, 338–352.*

Naumann, C.L. (1976). *Grundzüge der Sprachkartographie und ihrer Automatisierung.* Hildesheim: Olms.

Perebyjnis, V.S. (1970). *Kil´kisni ta jakisni charakteristiki sistemi fonem v sučasnoï ukraïnskoï literaturnoï movi*. Kiev: Naukova dumka.

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

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. *Proceedings of the North East Linguistic Society (NELS) 23, 367–381.*

Saporta, S. (1955). Frequency of consonant clusters. *Language 31, 25-30.*

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science 237, 1317–1323.*

Storkel, H.L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language 29(2), 251-274*

Tolstaja, S.M. (1968). Fonologičeskoe rastojanie i sočetaemosť soglasnych v slavjanskich jazykach. *Voprosy jazykoznanija 3, 66-81.*

Tversky, A. (1977). Features of similarity. *Psychological Review, 84, 327–352.*

Vinogradov, V.A. (1966). Nekotorye voprosy teorii fonologičeskich oppozicij i nejtralizacii. In: *Problemy ilngvističeskogo analiza: 3-25*. Moskva: Nauka.

Wang, M.D., Bilger, R.C. (1973). Consonant confusions in noise: a study of perceptual features. *Journal of the Acoustical Society of America 54, 1248–1266.*

Wicklegren, W.A. (1965). Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America 38, 583–588.*

Wicklegren, W.A. (1966). Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America, 39, 388–398.*

Winitz, H., Bellerose, B. (1963). Phoneme-sound generalization as a function of phoneme similarity and verbal unit of test and training stimuli. *Journal of Speech and Hearing Research 6, 379-392.*

Yu, H., Kim, S.-J., Oh, Y.-H. (1995). Estimating fuzzy phoneme similarity relations for continuous speech recognition. *Proceedings of the 8th international conference on industrial and engineering applications of artificial intelligence and expert systems: 665-671.* Melbourne, Australia**.**

1.6. Script motifs

Problem

Every script consists of strokes which may differ by length, position, direction, thickness, curvature, etc. and can be differently joined with other ones. An identical set of singular or joined strokes can be called “script motif”. A motif is always the greatest repeated configuration of strokes (cf. also Problem 1.7: *Formal problems of rongorongo*)

(1) Quantify the above mentioned properties.

(2) Find the distribution of motifs in signs, i.e. $f(x)$ = number of motifs present in x signs, and derive its theoretical distribution.

(3) Analyze a simple script and set up an indicator of economy (motif repetitions).

(4) A script exploits the motifs in different ways; define an indicator of entropy and redundancy and interpret them.

Procedure

Take a simple script, e.g. Ogham, some runes or even Arial. Define the set of categories present (length, position, direction,…). Not all categories need to be distinctive. The simplest way is to partition the sign in individual strokes and

joined strokes and search for identities in other signs. For example the Arial A and L have the following partitioning

A	/	\	-	/-	/\	-\
L	|	_–				

Here one can distinguish length, position and direction of strokes, and the motifs are simple, double or threefold. Choose the most complex combination of strokes that is repeated in other letters and consider it as motif.

Here, the quantification of the properties of straight lines is simple: we have two positions beginning from top to bottom: top and mid; three directions: mid down to left, mid down to right (or seen from below given in degrees) and horizontal, and two lengths: full and half. Needless to say, these properties must be defined on the basis of the complete alphabet (script) and the individual properties may be quantified using ordinal numbers or degrees.

Take the individual strokes or motifs and count in how many signs they occur. Set up the frequency distribution and use some of its properties to define the economy of script, its transparency and redundancy. Count the proportion of motifs (= repeated structures consisting of more than one stroke) as the ratio of repeated motifs to all given multi-stroke structures.

Compare two different scripts after having set up appropriate tests. Take a present day form of a script and express its evolution from the ancient form. Which property changed? Express the evolution of properties by functions and describe the way of evolution. Begin to conceive one aspect of the theory of formal evolution of scripts.

Take an ideographic script, choose one of the forms of identical signs and find the motifs in it. Use, for example, modern Chinese script, Rongorongo, Egyptian hieroglyphs or the last version of Assyrian cuneiform script.

Using scripts that have not been written in standard form, take an idealized form. This procedure may injure the results but there is no other possibility to perform this analysis.

References

Bohn, H. (2002). Untersuchungen zur Chinesischen Sprache und Schrift. In: Köhler, R. (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 127-177.* [http:/ubt.opus.hbz-nrw.de/ volltexte/2004/279]

Köhler, R. (2008). Quantitative analysis of writing systems: an introduction. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties of characters and writing systems: 3-9.* Berlin-New York: Mouton de Gruyter.

Melka, T. (2009). Linearity, calligraphy and syntax in the Rongorongo script. *Glottotheory 2/2, 70-96* (and the references therein).

1.7. Formal problems of rongorongo

Problems

Rongorongo is the logographic script of the Eastern Island in Polynesia. The literature about deciphering, cultural setting, history, text explication, corpus, etc. is enormous (cf. Melka 2009a, b). However, the form of the script has never been studied. Solve the following problems:

(1) Make a list of the greatest repeated graphemic parts of the ideograms and state the participation of the given parts in ideograms. (cf. also the Problem 1.6: *Script motifs*)
(2) Compute the complexity of individual ideograms by usual methods.
(3) Compute the distinctivity of individual ideograms
(4) Compute the distribution of ideograms in the corpus.
(5) Compute the polytexty of ideograms.

Procedures

For the solution of the first problem compare every ideogram with the remaining ones and state whether they contain common parts. After having a list of repeated parts, set up the distribution of their occurrences, i.e. state the number of different parts, $f(x)$, occurring in exactly x ideograms. If you obtain a monotonously decreasing distribution, substantiate its form linguistically or psychologically.

To solve the second problem, use some of the methods for measuring script complexity (e.g. Altmann 2004). If you succeeded to solve the first problem, you can combine the method with composing the ideograms from parts.

The third problems means to find the formal differences between each ideogram and each other and for the given ideogram take the mean of its differences (Antić, Altmann 2009). The individual distinctivities are not equal, they form a kind of distribution. Find this distribution and substantiate it linguistically.

For the solution of this problem take all *rongorongo* inscriptions available and state simply the frequency of individual ideograms. Set up the frequency distribution, find a theoretical distribution and substantiate it linguistically.

Solving the fourth problem register for each ideogram the number of texts in which it occurs and at last set up the distribution of their polytexty. Consider the meaning of individual ideograms and interpret the polytexty linguistically or culturally.

References

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-73.*
Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 9, 46-53.*
Melka, T.S. (2009a). The corpus problem in the Rongorongo studies. *Glottotheory 1/2, 111-136.*

Melka, T.S. (2009b). Linearity, calligraphy and syntax in the Rongorongo script. *Glottotheory 2/2, 70-96*.

1.8. Symbol frequencies

Problem

This is a continuation of the problem "Letter frequency" from *Problems 1: 6*. In this problem, symbol is either a sound, a phoneme, a letter, or a grapheme. Take ready counts or – for the sake of security – evaluate symbol frequencies in several languages and set up their rank-frequency distribution. Distinguish vocabulary and text frequency. Omit blanks, punctuation marks and numbers. Elaborate the theoretical probability distribution or at least a theoretical function, find boundary conditions and, if possible, a beginning of a background theory.

Do not trust symbol frequencies published on the Internet. If you count letters or graphemes, do not distinguish between capitals and normal signs.

Procedure

Even within a language family you will find peculiarities which must be captured by your theory.

Since symbol inventories are very restricted, you must use a distribution with finite support or truncate your candidate on the right hand side. The former case would be more adequate. The observed distribution need not be convex everywhere, hence you must find a function which may take a concave form at the beginning and change to convexity after a turning point.

There may be gaps in the distribution, e.g. the first three frequencies are very high, then there is a gap, and afterwards there is a very regular convex continuation. If you use a mixed or a modified distribution for this purpose, you must find the boundary conditions.

There may be a distribution in which one of the frequencies "plays a clown" and deviates strongly from the smooth course of the function. You may use a modified distribution ascribing the problematic case its own probability and modifying all the rest, but in this case you must find some conditions.

Derive the main distribution from some axioms or conjectures and systematize the deviations by stating the boundary conditions. Start from the studies of Martindale, Gusein-Zade, McKenzie, Borodovsky (1996) and Strauss, Altmann, Best (2008).

If you have sufficient data, compare the development/change of distributions in one language.

Since simple-symbol frequency depends on the occurrence of larger symbols (e.g. words), study the simple-symbol distributions in different text sorts.

References

Altmann, G. (1993). Phoneme counts: Marginal remarks on the Pääkkönen article. *Glttometrika 14, 54-68.*

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

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie.* Bochum: Brockmeyer.

Andreev, N.D. (ed.) (1965). *Statistiko-kombinatornoe modelirovanie jazykov.* Moskva-Leningrad: Nauka.

Andreev, N.D. (ed.) (1967). *Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykoznaniii.* Leningrad: Nauka.

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology 46, 81-86.*

Bauer, F.L. (2000). *Entzifferte Geheimnisse.* 3., überarbeitete und erweiterte Auflage. Berlin-Heidelberg: Springer.

Bektaev, K.B. (1973). Alfavitno-častotnyj slovar´ slogov kazachskogo jazyka. In: *Statistika kazachskogo teksta 3: 566-611.* Alma-Ata: Nauka.

Belevitch, V. (1956). Théorie de l´information et statistique linguistique. *Bulletin de la Classe des Sciences Académie Royale de Belgique 419-436.*

Belonogov, G.G., Frolov, G.D. (1963). Empiričeskie dannye o raspredelenii bukv v russkoj pis´mennoj reči. *Problemy kibernetiki, Vyp. 9, 287-305.*

Berger, K.W. (1967). A study of printed Pilipino usage. *Phonetica 17, 31-37.*

Bergmann, H. (1986). Einige Ergebnisse der Phonemstatistik. *Abhandlungen der Heidelberger Akademie der Wissenschaften, Philosophisch-historische Klasse 1986, 5-19.*

Best, K.-H. (2003). *Quantitative Linguistik: Eine Annäherung.* 2., überarb. u. erw. Auflage. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2005a). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11, 9-31.*

Best, K.-H. (2005b). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21-32.*

Best, K.-H. (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovij visnik Černivec´kogo universitetu vypusk 231, Germans´ka filologija, 119-127.*

Beutelspacher, A. (1994). *Kryptologie.* 4., abermals leicht verbesserte Auflage. Braunschweig-Wiesbaden: Vieweg.

Bhagvat, S.V. (1961). *Phonemic frequencies in Marathi and their relation to devising a speed-script*. Poona: Deccan College.

Boldrini, M. (1948). *Le statistiche letterarie e i fonemi elementari nella poesia*. Milano.

Bosák, J. (1965). Frequency of phonemes and letters in Slovak and numerical expression of some phonemic relations. *Jazykovedný časopis 14, 120-130.*

Bourne, C.P., Ford, D.F. (1961). A study of the statistics of letters in English words. *Information and Control 4, 48-61.*

Bourdon, B. (1892). *L´expression des émotions et des tendences dans le langage*. Paris: Alcan.

Card, L.E., Eckler, R.A. (1975). A survey of letter frequencies. *Word waxs. The Journal of Recreational Linguistics 5, 81-85.*

Čistjakov, V.F. (1972). Častotnosti glasnych i soglasnych v 50 jazykach raznogo grammatičeskogo stroja. *Lingua Posnaniensis 16, 45-48.*

Denes, P.B. (1963). On the statistics of spoken English. *J. of the Acoustical Society of America 30, 892-904.*

Denes, P.B. (1964). On the statistics of spoken English. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 17, 51-72*

Dewey, G. (1923). *Relative frequencies of English speech sounds*. Cambridge, MA: Harvard University Press.

Dietze, J. (1982). Grapheme und Graphemkombinatorik der russischen Fachsprache. Eine Phonostatistische Untersuchung. *Glottometrika 4, 80-94.*

Doležel, L. (1963). Předběžný odhad entropie a redundance psané čestiny. *Slovo a slovesnost 24, 165-175.*

Džubanov, A.Ch. (1979). K voprosu o grafemnoj statistike kazachskogo teksta. In: *Voprosy kazachskoj fonetiki i fonologii 79-86.*

Estoup, J.B. (1916). *Gammes sténographiques. Méthode et exercises pour l´acquisition de la vitesse.* Paris: Institut sténographique.

Fähnrich, M., Meinold, G. (1973). Phonemstatistischer Vergleich zwischen Georgisch, Awarisch und Tschesarenisch. *Wissenschaftliche Zeitschrift 22, 109-117.*

Fairbanks, G.H. (1957). Frequency and phonemics. *Indian Linguistics 17, 105-113.*

Fant, C.G.M. (1958). Some notes on the relative occurrence of letters, phonemes, and words in Swedish. In: *Proceedings of the 8th International Congress of Linguistics, Oslo 1958: 815.*

Ferguson, C.A., Chowdhury, M. (1960). The phonemes of Bengali. *Language 36, 22-59.*

Findra, J. (1968). Frekvencia foném v ústnych prejavoch. *Jazykovedný časopis 19, 84-95.*

Förstemann, E. (1846). Ueber die numerischen Lautverhältnisse im Deutschen. *Germania 7, 83-90.*

Förstemann, E. (1852). Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen. *Zeitschrift für vergleichende Sprachforschung 1, 163-179.*

Fowler, M. (1957). Herdan´s statistical parameter and the frequency of English phonemes. In: Pulgram, E. (ed.), *Studies presented to Joshua Whatmough on his sixties birthday: 47-52*. s´Gravenhage: Mouton.

French, N.R., Carter, C.W., Koenig, W. (1930). Words and sounds of telephone communications. *Bell System Technical Journal 9, 290-325.*

Fry, D.B. (1947). The frequency of occurrence of speech sounds in Southern English. *Archives Néerlandaises de Phonétique Experimentales 20.*

Gačečiladze, T.G., Eliašvili, A.I. (1958). Statistika bukv sovremennogo literaturnogo gruzinskogo jazyka. *Soobščenija Akademii nauk gruzinskoj SSR 20, 565-567.*

Gaines, H.F.(1956). *Cryptanalysis: A study of ciphers and their solution.* New York: Dover.

Gerber, S.E., Vertin, S. (1969). Comparative frequency counts of English phonemes. *Phonetica 19, 133-141.*

Good, I.J. (1969). Statistics of language. In: Meethoun, A.R., Hudson, R.A. (Eds.), *Encyclopedia of information, linguistics and control: 567-581*. Oxford: Pergamon.

Grigoriev, V.I. (1980). O dinamike raspredelenija bukv v tekste. In: *Aktual´nye voprosy strukturnoj i prikladnoj lingvistiki. Sbornik statej: 40-48.* Moskva.

Grigoriev, V.I. (1980). Frequency distribution of letters and their ranks in a running text. In: Viks, Ü. (ed.), *Symposium: Computational linguistics and related topics. Tallinn: Academy of Sciences 1980: 43-47.*

Grzybek, P., Kelih, E. (2005). Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph. In: Altmann, G., Levickij, V.; Perebejnis, V. (eds.), *Problemi kvantitativnoi lingvistiki. Problems of Quantitative Linguistics: 159-179*. Černivci: Ruta.

Grzybek, P., Kelih, E, (2005). Häufigkeiten von Buchstaben/Graphemen/ Phonemen. Konvergenzen des Rangierungsverhaltens. *Glottometrics 9, 62–73.*

Grzybek, P., Kelih, E. (2005): Towards a general model of grapheme frequencies in Slavic languages. In: Garabík, R. (ed.), *Computer Treatment of Slavic and East European Languages: 73-87.* Bratislava: Veda.

Grzybek, P., Kelih, E., Altmann, G. (2004). Graphemhäufigkeiten (Am Beispiel des .Russischen). Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25-54.*

Grzybek, P., Kelih, E., Altmann, G. (2005). Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das 'ë'. *Anzeiger für Slavische Philologie 33, 117–140.*

Grzybek, P., Kelih, E., Altmann, G. (2006): Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen. In: Kozmová, R. (ed.): *Sprache und Sprachen im mitteleuropäischen Raum. Vorträge der Internationalen Linguistik-Tage, Trnava 2005, 661–664.* Trnava: Univ. sv. Cyrila a Metoda, Filozofická Fakulta.

Grzybek, P., Kelih, E., Stadlober, E. (2006). Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik. *Anzeiger für Slavische Philologie 34, 41–74.*

Grzybek, P., Kelih, E., Stadlober, E. (2009). Slavic letter frequencies: a common discrete model and regular parameter behavior. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics*: *17-33*.. Lüdenscheid: RAM-Verlag

Guirao, M., Garciá Jurado, M.A. (1990). Frequency of occurrence of phonemes in American Spanish. *Revue québécoise de linguistique 19(2), 135-150.*

Gusein-Zade, S.M. (1988). O raspredelenii bukv russkogo jazyka po častote vstrečaemosti. *Problemy Peredači Informacii 24(4), 102-107.*

Hayden, R.E. (1950). The relative frequency of phonemes in general-American English. *Word 6, 217-223.*

Herdan, G. (1958). The relation between the functional burdening of phonemes and the frequency of occurrence. *Language and Speech 1, 8-13*.

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

Hultzén, L.S., Allen, J.H.D., Miron, M.S. (1964). *Tables of transitional frequencies of English phonemes.* Urbana, Ill.: University of Illinois Press.

Hussien, O.A. (2004). The Lerchianness plot. *Glottometrics 7, 50-64.*

Isengel´dina, A.A. (1973). Faktory, opredeljajuščie otnositel´nuju častotnost´ fonem. In: *Statistika kazachskogo teksta 3, 659-662.* Alma-Ata: Nauka.

Kelih, E. (2009). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle. *Glottometrics 18, 53–69.*

Kelih, E. (2010). Kumulierte Ranghäufigkeiten von slawischen Graphemen: Modell und Parameter-Interpretation. *Glottotheory. 3(1), 69–82.*

Kerkhoffs, A. (1883). *La cryptographie militaire*. Paris.

King, R.D. (1966). On preferred phonemicisation for statistical studies. Phoneme frequencies in German. *Phonetica 15, 22-31*.

Kosonovskij, A.I. (1968). Nekotorye predvaritel´nye dannye o častotnosti grafem i fonem sovremennogo literaturnogo jazyka Chindi. In: Beskrovnyj, V.M. et al. (eds.), *Jazyki Indii, Pakistana, Nepala i Cejlona: 167-180.* Moskva: Nauka.

Krámský, J. (1965). Some statistical observations on the role of the place of articulation in languages. *Philologica Pragensia 8, 245-250.*

Kubáček, L. (1994). Confidence limits for proportions of linguistics entities. *J. of Quantitative Linguistics 1, 56-61*.

Kučera, H. (1963). Mechanical phonemic transcription and phoneme count in Czech. *International Journal of Slavic Linguistics and Poetics 6, 36-50.*

Kučera, H. (1963). Entropy, redundancy and functional load in Russian and Czech. In: *American Contributions to the Fifth International Congress of Slavists, Vol. I, 191-219.* The Hague: Mouton.

Kučera, H., Monroe, G.K. (1968*). A comparative quantitative phonology of Russian, Czech, and German*. New York: American Elsevier.

Kullback, S. (1976*). Statistical methods in cryptanalysis*. Laguna Hills, CA: Agean Park Press.

Küpfmüller, K. (1954). Die Entropie der deutschen Sprache. *Fernmeldetechnische Zeitschrift 7*, 265-272.

Kuzina, V. (1977). Statistika bukv v tekstach raznych tipov sovremennogo latyšskogo jazyka. In: *Statistika un valodas funkcionālie stili: 97-106.* Riga: Zinatne.

Lua, K.T. (1990). Analysis of Chinese character stroke sequences. *Computer Processing of Chinese & Oriental Languages 4(4), 375-385.*

Lua, K.T. (1992). Linearization of Zipfian distribution for Chinese characters. *J. of Information Processing 15(1), 10-16.*

Ludvíková, M., Königová, M. (1967). Quantitative research of graphemes and phonemes in Czech. *Prague Bulletin of Mathematical Linguistics 7, 15-29.*

Macrea, D. (1941/43). Frecvența fonemelor în limba română. *Dacoromania 10, 39-49.*

Marinova, M., Marinov, A. (1964). Statističeski izsledovanija na fonemite v bъlgarskija knožoven ezik. *Blъgarski ezik 1964, 2-3.*

Martin, A. (2009). Why are phoneme frequency distributions skewed?
http://www.linguistics.ucla.edu/people/grads/amartin/presentations/Martin_LSA_2009_phoneme_frequencies.ppt (retrieved 4April 4, 2011)

Martindale, C., Gusein-Zade, S., McKenzie, D.P., Borodovsky, M.Y. (1996), Comparison of equations describing the ranked frequency distributions of phonemes and graphemes *Journal of Quantitative Linguistics. 3*, 106-112.

Mázlová, V. (1946). Jak se projevuje zvuková stránka češtiny v hláskových statistikách. *Naše řeč 30, 101-111; 146-151.*

Mdivani, R.R. (1968). Zamečanie k modeli obščego izčislenija fonem. *Voprosy jazykoznanija 1968(3), 124-125.*

Meier, H. (²1967). *Deutsche Sprachstatistik*. Hildesheim: Olms.

Messner, D. (1976). A statistical approach to Potuguese. In: Schmidt-Radefeldt, J. (Hrsg.), *Readings in Portuguese Linguistics: 425-446.* Leiden: North-Holland.

Moīnfar, M.D. (1973). *Phonologie quantitative du Persan.* Paris: Editions Jean-Favard.

Moreau, R. (1961). Au sujet de l´utilisation de la notion de fréquence en linguistique. *Cahiers de lexicologie 3, 140-159.*

Nagórko-Kufel, A. (1975). Z badań nad częstościami elementów tekstowych języka polskiego dla potrzeb pisma niewidomych. *Poradnik językowy 1, 7-13.*

Naranan, S., Balasubrahmanyan, V.K. (1993). Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *J. of Scientific and Industrial Rresearch 52, 728-738.*

Nemetz, T., Szilléry, A. (1979). Nyelvstatisztikai táblázatok. *Alkalmazot Matematikai Lapok 5, 69-87.*

Newman, E.B. (1951). The pattern of vowels and consonants in various languages. *American Journal of Psychology 64, 369-379.*

Nikonov, V.A. (1960). Konsonantnyj koefficient. *Lingua Posnaniensis 8, 228-235.*

Novak, L.A. (1971). Statistica delle lettere e delle combinazioni di lettere nella lingua rumena scritta. In: Tagliavini, C. (ed.), *Statistica linguistica: 291-322.* Bologna: Patron.

Novak, L.A. (1968). Statistika bukv i bukvosočetanij v rumynskom pis´mennom jazyke. In: Alekseev, P.M., Kalinin, V.M., Piotrovskij, R.G. (eds.), *Statistika reči: 228-230.* Leningrad: Nauka.

Ohlmann, N. (1958). Subject-word letter frequencies with applications to superimposed coding. In: *Proceedings of the International Conference of Scientific Information 2: 903-915.* Washington.

Pandit, P.B. (1965). *Phonemic and morphemic frequencies of the Gujarati language.* Poona: Deccan College.

Pääkkönen, M. (1993). Graphemes and context. *Glottometrika 14, 1-53*

Penkov, V. et al. (1962). Frequencies of letters in written Bulgarian. *Comptes rendus de ´Académie bulgare des Sciences, 15, No. 3.*

Perebejnos, V.I. (1965). Častota i sočetaemost´ fonem sovremennogo ukrainskogo jazyka. In: *Seminar – Avtomatizacija informacionnych rabot i voprosy prikladnoj lingvistiki: 25-30.* Kiev.

Perebyjnis, V.S. (1970). *Kil´kisni ta jakisni charakteristiki sistemi fonem sučasnoï ukraïnskoï literaturnoï movi.* Kiïv: Naukova dumka.

Pierce, J.E. (1957). A statistical study of consonants in New World languages (I) Introduction, (II) Data. *International Journal of American Linguistics 23, 36-45, 94-108.*

Piirainen, I.T. (1971). Grapheme als quantitative Größen. *Linguistische Berichte 13, 81-82.*

Proskurnin, N. (1933). Podsčety častoty liter i komplektovka šrifta. *In: Revoljucija i pis´mennost´. Sbornik I: 72-82.* Moskva-Leningrad.

Pukui, H.K., Elbert, S.H. (1957). *Hawaiian-English dictionary*. Honolulu: University of Hawaii Press.

Rachmanov, D.A.O. (1988). *Statistiko-distributivnyj analiz azerbajdžanskogo teksta. na urovne grafem i fonem*. Baku: Diss.

Ramakrishna, B.S., Nair, K.K., Chiplunkar, V.N., Atal, B.S., Ramachandran, V., Subramanian, R. (1962). *Some aspects of the relative efficiencies of Indian languages.* Bangalore.

Roberts, A.H. (1965). *A statistical linguistic analysis of American English.* The Hague: Mouton

Roceric-Alexandrescu, A. (1968). *Fono-statistica limbii române*. Bucureşti: Editura Academiei RSR.

Rocławski, B. (1975). Ze studiów fonostatystycznych nad kaszubszczyną. Rozklad częstości występowania fonemów. *Gdańskie Studia Językoznawcze Zakład Narodowy Im. Ossolińskich 1975, 107-130.*

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

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

Rūłe, V. (1951). Lauthäufigkeit in der lettischen Schriftsprache. In*: Slaviska instituts vid Lunds universitetet årsbok 1948/1949: 153-164*. Lund.

Savický, N.P. (1966). Ob ustojčivosti otnositel´nych častot lingvističeskich elementov. *Československá rusistika 11, 214-217.*

Segal, D.M. (1969). K statističeskoj charakteristike pol´skogo jazyka na fonologičeskom urovne. In: *Issledovanija po pol´skomu jazyku: 20-52*. Moskva: Nauka.

Segal, D.M. (1972). *Osnovy fonologičeskoj statistiki*. Moskva: Nauka.

Seiden, W**.** (1960). Chamorro phonemes. *Anthropological Linguistics 2, 6-35.*

Sigurd, B. (1968). Rank-frequency distributions for phonemes. *Phonetica 18, 1-15.*

Singhal, R., Toussaint, G.T. (1978). Probabilities of occurrence of characters, character-pairs, and character triplets in English text. *ALLC Bulletin 6, 245-253.*

Širokov, O.S. (1964). O sootnošenii fonologičeskoj sistemy i častotnosti fonem. *Voprosy jazykozanija 1964(1), 53-60.*

Siromoney, G. (1963). Entropy of Tamil prose. *Information and Control 6, 297-300.*

Solso R.L., King, J.F. (1976). Frequency and versatility of letters in the English language. *Behavior Reserch Methods and Instrumentation 8, 283-286.*

Steffen, M. (1957). Częstość występowania głosek polskich. *Biuletyn polskiego towarzystwa językoznawczego 16, 145-164.*

Stolze, F. (1891). Die Iterationsverhältnisse der Laute in der lateinischen Sprache für die Kurzschrift.. *Magazin für Stenographie. 1891, 47-48.*

Strauss, U., Altmann, G., Best, K.-H. (2008). Phoneme frequencies. http://lql.uni-trier.de/index.php/Phoneme_frequency *(retrieved April 4, 2011)*

Svacevičius, B.I. (1966). K voprosu o častote vstrečaemosti fonem v litovskoj pis´mennoj reči. *Materialy kollokviuma: 19-22*. Vilnius: Pedagogical Institute Press.

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

Tambovcev, J.A. (1982). Empiričeskoe raspredelenie častotnosi fonem v jazyke kazymskich kanty [v kazymskom dialekte chantyjskogo jazyka]. In: *Lingvostatistika i vyčislitel´naja lingvistika: 121-135.* Tartu.

Tambovcev, J.A. (1983). Phonostatistical study of Komi Zyryan vowels and consonants. *Finnisch-ugrische Forschungen 45, 164-167.*

Tambovcev, J.A. (1983). Empiričeskoe raspredelenie častotnosi fonem v oročskom jazyke. In: *Kvantitativnaja lingvistika i stilistika: 124-125.* Tartu.

Tambovcev, J.A. (1984). Empirical distribution of the phonemes in Orokh. Typological analysis. *Archiv orientální 52, 285-294.*

Tambovcev, J.A. (1988). Nekotorye fonostatističeskie charakteristiki jazyka barabinskich tatar. In: *Fonetika i grammatika jazykov Sibiri: 135-139.* Novosibirsk: IIFF.

Tambovcev, J.A. (1988). Phonostatistical characteristics of different dialects of Eskimo. In: *6th Inuit studies conference. Copenhagen, October 17-20, 19888: 11-17.*

Thorndike, E.L. (1948). The psychology of punctuation.. *American Journal of Psychology 61, 222-228.*

Tobias, J.V. (1959). Relative occurrence of phonemes in American English. *J. of the Acoustical Society of America 31, 631-633*

Trnka, B., Kanekiyo, T., Koizumi, T. (1968). *A phonological analysis of present-day standard English.* Alabama: University of Alabama Press.

Tuldava, J. (1988). Opyt kvantitativnogo analiza sistemy fonem estonskogo jazyka. *Acta et Commentationes Universitatis Tartuensis 838, 120-133.*

Tuldava, J. (1995). Quantitative analysis of the phonemic system of the Estonian language. In: Tuldava, J., *Methods in Quantitative Linguistics, Chapter 10, 161-187.* Trier: WVT.

Veenker, W. (1979). Zur phonologischen Statistik der komipermjakischen Sprache. *Finnisch-Ungarische Mitteilungen 3, 13-27.*

Veenker, W. (1979). Zur phonologischen Statistik der vogulischen Sprache. In: Gläser, Ch., Pusztay, J. (eds.), *Festschrift für Wolfgang Schlachter zum 70. Geburtstag: 305-346.* Wiesbaden: Harrassowitz.

Veenker, W. (1981). Problemy fonologičeskoj statistiki chantyjskogo jazyka. In: Ubrjtova, E.I., Kim Čer Len, Kuzmina, A.I., Ryžkina, O.A. (eds.), *Teoretičeskie voprosy fonetiki i grammatiki jazykov narodov: 84-96.* Novosibirsk.

Veenker, W. (1981). Zur phonologischen Statistik der mordvinischen Schriftsprachen. *Ural-altaische Jahrbücher 1, 33-72.*

Veenker, W. (1981). Zur phonologischen Statistik der votjakischen Sprache. In: Bereczki, G., Molnár, J. (eds.), *Lakó-Emlékkönyv – nyelvészeti tanulmányok 196-213*. Budapest.

Veenker, W. (1982). Konfrontierende Darstellung zur phonologischen Statistik der ungarischen und finnischen Schriftsprache. *Nyelvtudományi közlemények 84, 305-348a.*

Veenker, W. (1982). Zur phonologischen Statistik der syrjänischen Sprache. *Etudes Finno-Ougriennes 15, 435-445.*

Verglas, A. (1962). Remarques sur la relation entre rang et fréquence des lettres français. *Bulletin d´information du laboratoire d´analyse lexicographique 6, 29-40.*

Vértes, E. (1953). Statistische Untersuchungen über den phonetischen Aufbau der ungarischen Sprache. *Acta Linguistica Academiae Scientiarum Hungaricae 3, 125-158; 411-430.*

Weiss, M. (1961). Über die relative Häufigkeit der Phoneme des Schwedischen. *Statistical Methods in Linguistics 1, 41-55.*

Whitney, W.D. (1880). *On the comparative frequency of occurrence of the alphabetic elements in Sanskrit.* American Oriental Society Studies 10.

Wioland, F. (1972). Estimation de la „fréquence" des phonèmes en français parlé. *Travaux de l´Institut phonétique de Strasbourg 4, 177-204.*

Wioland, F. (1974). Contribution à l´établissement de constantes en relation avec la fréquence des phonèmes en français parlé. *Travaux de l´Institut phonétique de Strasbourg 6, 141-164.*

Zettersten, A. (1969). *A statistical study of the graphic system of present day American English*. Lund: Studentenlitteratur.

Zwirner, E., Zwirner, K. (1936). Die Häufigkeit von Buchstaben und Lautkombinationen. *Forschungen und Fortschritte 12, 23-24, 286-287.*

2. Grammar

2.1. Word-form diversification

Problem

Consider a strongly synthetic language, e.g. Latin, Hungarian or some Slavic language and for each lemma state the frequencies of all its forms. You may consider even compounds as pertinent word-forms of a lemma.

Show that the ranked distribution of individual forms has a regular form and find the theoretical probability distribution.

Procedure

Take 1000 lemmas from a large lemmatized corpus. Then state individually the frequencies of all their forms and for each lemma set up the ranked frequency distribution of forms.

(1) Find a theoretical distribution adequate for the majority of cases. For the rest of the cases generalize or specify the given distribution in different ways (adding a new parameter, generalizing, compounding, mixing, modifying, etc.) and show that the diversification of forms is not a haphazard process but develops in a certain well defined way.

(2) Show that with each lemma its shortest forms has the greatest frequency, i.e. the longer the form, the smaller is its frequency. You can measure the length of a form in terms of morphemes, syllables or even phonemes. If there are many length classes, find their distribution.

(3) Set up stepwise your own theory of form diversification and join it with other results in this domain (cf. esp. Altmann 2005).

References

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

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

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

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.). *Quantitative Linguistics. An International Handbook: 191-208*. Berlin: de Gruyter

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

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

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

Sanada, H., Altmann, G. (2009). Diversification of postpositions in Japanese. *Glottometrics 19, 70-79.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, 42-48.*

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

2.2. Case diversification in Finnish

Problem

Using the work of Pajunen and Palomäki (1982), Väyrynen, Noponen, Seppänen (2008) presented the proportions of cases in Finnish. Even if the absolute numbers are not published, one can reconstruct them partially using the sample size (20000).

(1) Show that the rank-frequencies follow a usual distribution used in different cases of diversification

(2) Compute the Popescu indicators *p* and *q* and compare them with those in other languages (Popescu et al. 2009)

See also *Problem 2.1*.

Procedure

(1) Use Väyrynen, Noponen and Seppänen's table given below and after ordering according to decreasing frequency compute the absolute frequencies.

Table 1
Cases in written Finnish
(Väayrynen, Noponen and Seppänen 2008)

Case	Percentage
Nominative	29.5
Genitive	20.3
Partitive	13.7
Inessive	7.1
Illative	6.3
Elative	4.4
Adessive	4.4

Accusative	3.1
Essive	2.6
Allative	2.3
Translative	2.2
Instructive	1.9
Abessive	0.2
Comitative	0.1
Ablative	1.0

As a matter of fact, the percentages can be multiplied by any appropriate number in order to obtain some positive integer values. First test the adequacy of the positive negative binomial distribution which was frequently used in Rothe (1991). If it is not adequate, find another distribution holding generally for case diversification. Use the data in Popescu et al. (2009). If you obtain a different distribution, substantiate it grammatically.

(2) Using the same data compute Popescu´s indicator

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

where $L_{max} = R - 1 + f(1) - f(R)$, R = maximal rank (r_{max}), $f(1)$ = greatest frequency, $f(R)$ smallest frequency, L is the arc length of the rank-frequency distribution defined as

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

and h is the h-point usually computed as

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

If there is a rank which is equal to its frequency, then $r = f_r$; if there is no such value, take (possibly neighbouring) values such that $f_1 > r$ and $f_2 < r + 1$. If $r_2 = r_1 + 1$, the formula can be simplified. If $f(R) > R$, one must transform the whole ranked sequence in $f^*(r) = f(r) - f(R) + 1$. Compare the resulting numbers with those for case diversification in other languages and with those obtained for other language phenomena. Generalize the results and substantiate them linguistically.

Process the cases in Hungarian or other strongly agglutinating language using several texts and compare the result with Finnish.

References

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

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 89-86.*

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

Pajunen, A., Palomäki, U. (1982). *Tilastotietoja suomen kielen muoto- ja lauseopillisista yksiköista*. Turku: Käsiskirjoitus.

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

Popescu, I.-I., Altmann, G., Köhler, R. (2009). Zipf´s law – a new view. *Quality and Quantity 44(4), 713-731*.

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.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, 42-48*.

Väyrynen, P., Noponen, K., Seppänen, T. (2008). Preliminaries to Finnish word prediction. *Glottotheory 1, 65-73.*

2.3. Inflection and derivation

Hypothesis

The more inflected words are there in a text, the more derived words it contains. Test the hypothesis.

Procedure

This is a classical typological hypothesis that should be tested on texts in several languages. It may hold in the given direction [$D = f(I)$] but not necessarily the other way round.

Define the measure of inflection in the usual “Greenberg-way” (Greenberg 1960) as

$$I = \frac{IW}{W},$$

where *IW* is the number of inflected words and those which can potentially be inflected, i.e. have a zero inflectional morpheme, and *W* is the number of all words in text. Sometimes even adverbs contain inflection (in Czech even the conditional particle *by*). Words which are at the same time derived and inflected belong to this class, too.

Define the measure of derivation as usually as

$$D = \frac{DW}{W}.$$

DW represents all words containing some kind of derivation. Even words of which only linguists know that they are derived, or words with affixoids (e.g. taken from Greek) or quasi-affixes (like the German *-wesen*, *-zeug*, etc.) may be ascribed to this class. The class *IW* and *DW* may have an intersection, of course.

After having analyzed several languages, show that $D = f(I)$. First show that there is a correlation between *I* and *D*, then derive the dependence from theoretical reflections.

Propose different measures of inflection and derivation. Show that inflection implies many other properties. Prepare a list of properties linked with inflection and process at least some of them.

Aim at a stepwise elaboration of a control cycle of many morphological properties and use texts for testing.

The fact that derivation is possible without the existence of inflection is not contained in the hypothesis. For this aspect set up a different hypothesis.

References

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

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

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

Krupa, V., Altmann, G. (1966). Relations among typological indices. *Linguistics 24, 29-37.*

2.4. Transitivity

Problem

Hopper and Thompson (1980) do not consider transitivity a punctual (i.e. object present or not present) but rather a continuous phenomenon expressing different

categories in different degrees. The categories are properties of verb semantics and grammar. The authors set up categories as presented in Table 2.4.1

Table 2.4.1
Categories of transitivity (Hopper, Thompson 1980)

Category	High	Low
A. Participants	2	1
B. Kinesis	action	non-action
C. Aspect	telic	atelic
D. Punctuality	punctual	non-punctual
E. Volitionality	volitional	non-volitional
F. Affirmation	affirmative	negative
G. Mode	realis	irrealis
H. Agency	A high in potency	A low in potency
I. Affectedness of O	O highly affected	O not affected
J. Individuation of O	O highly individuated	O not individuated

As can be seen, the categories are nominally classified (A to J) and the properties are dichotomous (except for the category J. Individuation).

(1) Find a superior category encompassing all categories (A to J) and ascribe the categories at least ordinal numbers. This is not an easy task.

(2) Ascribe degrees also to the properties – instead of binarity – i.e. quantify the properties on a higher scale; cf. the method proposed by Hopper and Thompson for the category *J. Individuation*, (1980, 256-257). Set up a vector of properties.

(3) Analyze two texts and compare their transitivity vectors. Perform separate studies on poetry, press texts, scientific texts and using the vectors establish an indicator and its interval for different text sorts. Do the transitivity indicators differ with different text sorts? Compare several languages.

(4) At a higher level of research find the probability distribution of the property degrees in texts. Construct a test for text comparisons or test the differences asymptotically using the normal distribution.

(5) Alternatively, consider the 10 categories as elements of a vector which can attain only two values, 0 and 1. Compare two texts by computing the cosine of the angle between them. Let the first text be represented by the vector **a**.= $\{x_1,x_2,\ldots,x_{10}\}$, the second by **b** = $\{y_1,y_2,\ldots,y_{10}\}$ then compute

$$\cos\theta = \frac{x_1y_1 + x_2y_2 + \ldots + x_{10}y_{10}}{\sqrt{\sum_{i=1}^{10} x_i^2}\sqrt{\sum_{i=1}^{10} y_i^2}} .$$

The radian of this angle which is obtainable by using the arccos function is the

measure of dissimilarity. If in the vectors one uses only ones and zeros, one can take in the denominator sums of plain numbers (which are equal to squares).

References

Čech, R. (2009). Testing of the Transitivity hypothesis: double object verbs and aspect in Czech. In: Dočekal, M., Ziková, M. (eds.), *Czech in formal grammar: 29-38*. München: Lincom.

Čech, R., Pajas, P. (2009). Pitfalls of the transitivity hypothesis: transitivity in conversation and written language in Czech. *Glottotheory 2/2, 41-49.*

Hopper, P.J., Thompson, S.A. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Naess, Å. (2007). *Prototypical Transitivity.* Amsterdam-Philadelphia: Benjamins.

Olsen, M.B., Macfarland, T. (1996). Where's Transitivity? Paper presented at the *Seventh Annual Formal Linguistic Society of Mid-America conference, May 17–19 1996.* The Ohio State University.

Otani, N. (2008). The Transitivity Hypothesis and Object Case-marking in Japanese: On the transitivity of oyogu (Swim) in Japanese. In: *Workshop on East Asian Linguistics, Santa Barbara, February 23, 2008.*

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text.* Lüdenscheid: RAM-Verlag.

Thompson, S.A., Hopper, P.J. (2001). Transitivity, clause structure, and argument structure: evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 27-60.* Amsterdam-Philadelphia: Benjamins.

Tsunoda, T. (1985). Remarks on Transitivity. *Journal of Linguistic 21, 385–396.*

2.5. Morph length

Problem

The distribution of morph length in texts has been investigated very scarcely (cf. Best 2000, 2001a,b,c; Krott 1996; Pustet, Altmann 2005). Segment short texts in different languages having different morphological regimes (agglutinating, inflectional, introflectional, isolating,…) and find, if necessary, different models of their length distribution. Measure the length of morphs in terms of phoneme or grapheme numbers. Skip the so called zero-morphs.

Procedure

First state very exactly the identity and the boundaries of morphs. If they are discontinuous, you must make your own decisions or set up unambiguous criteria. Then partition the text in morphs without rest. In many languages the phonetic partitioning will drastically differ from the graphemic one. Propose an empirical model to capture the length regularity (probability distribution or function).

Compare the phonetic analysis with the graphemic one and use the difference between them to compute the redundancy of the graphemic version. Propose a measure for the surplus of the graphemic version. Compare e.g. English and French and state the extent of retardation in the evolution of the graphemic form behind the phonetic one. Compare Old English with Modern English of your own dialect. Compare the diversification of Romance languages beginning with Latin. Compare Old Church Slavic with modern Slavic languages. Can you indicate some directions of evolution?

Use some typological indicators of inflection and agglutination and place them in relation to the parameters of morph length distributions (functions). Is there a relationship indicating some latent mechanism?

Check a strongly isolating or a monosyllabic language. What can you conclude?

Do not confuse morphs with morphemes. At the very beginning define exactly what are your entities and how do you measure the length.

References

Best, K.-H. (2000). Morphlängen in Fabeln von Pestalozzi. *Göttinger Beiträge zur Sprachwissenschaft 3, 19–30.*

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

Best, K.-H. (2001b). Zur Länge von Morphen in deutschen Texten. In: Best, K.-H. (ed.) (2001a), 1-14.

Best, K.-H. (2001c). Wie viele Morphe enthalten deutsche Wörter? Am Beispiel einiger Fabeln Pestalozzis. In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica '99. Zborník na počesť Kláry Buzässyovej: 258-270*. Bratislava: Veda.

Best, K.-H. (2005). Morphlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 255-260*. Berlin-New York: de Gruyter.

Krott, A. (1996). Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics 3, 29-37.*

Pustet, R., Altmann, G. (2005). Morpheme length distribution in Lakota. *Journal of Quantitative Linguistics 12(1), 53-64.*

2.6. Distribution of morphological features

Hypothesis

The frequencies of morphological features are distributed according to a diversification model (Altmann 2005). Test the hypothesis.

Procedure

Extract the morphological features such as case, number, gender, tense, person, mode, aspect, etc. (inflectional and derivative ones) from the corpus you are working with. Count the number of occurrences of each of the values for each feature. You may also study appropriate feature combinations such as case-number-gender, which has the advantage of having more different values and hence more degrees of freedom for the distributions.

If your corpus is large enough, you can perform this analysis separately for text sorts, authors, etc. and compare the distributions.

Test the hypothesis by confronting your resulting data with suitable theoretical probability distributions taken from the works presented in the references.

References

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

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 648-659.* Berlin: de Gruyter.

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

Best, K.-H. (2008). Diversifikation der Phonems /r/ im Deutschen. *Glottometrics 18, 26-31.*

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

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 79-86.*

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

Laufer, J., Nemcová, E. (2008). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-25.*

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

Sanada, H., Altmann, G. (2009). Diversification of postpositions in Japanese. *Glottometrics 19, 70-79.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, 2009, 42-48.*

2.7. Paradigmatic expansion of words

Problem

Many words are able to extend their word-class membership either directly or by means of derivation or composition. For example *even* may be adjective, verb or adverb directly, and become noun by suffixation (*evenness*). Show that in the language you study (having *K* word classes) this expansion is restricted, and the number of classes in which a word may enter is regularly distributed.

Procedure

Take an extensive monolingual dictionary and study only the words beginning with the letter *A* or with a special sign (in Chinese or Japanese). To each word write the number of parts-of-speech classes which it enters (use a dictionary or a morphologically annotated corpus). Do not forget to find also compounds in which it takes the second place, e.g. *ever* is an adverb but *evergreen* is noun or adjective and *whatever* is a relative pronoun.

Having analyzed at least 200 words set up the frequency distribution of the variable "number of classes in which a word may enter". Then find the distribution of this variable. It must have a given finite support or it must be truncated at the right hand side. Substantiate your approach linguistically and take into account some requirements presented by Köhler (2005).

Then study the following hypotheses:

(1) Compare the empirical frequency distributions of two languages. Substantiate the hypothesis that the greater the skewness of the distribution, the more analytic is the language.

(2) If you can accept hypothesis (1), set up an indicator of analytism/ synthetism using only the properties of the distribution. Compare your indicator with other ones characterizing synthetism. This problem is not simple.

(3) The longer is a word, the smaller is the number of classes which it can enter.

(4) The greater is the polysemy of a word, the more classes it can enter.

(5) The greater is the frequency of the word, the greater is the number of classes it can enter. Here, however, distinguish the main meaning and class of each word. The hypothesis is rather complex.

Compare the problems connected with word classes in *Problems Vol 2*: 45, 47, 110f.

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.

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics 2*. Lüdenscheid: RAM.

2.8. Morpheme dynamics

Problem

Morphemes do not have unique form and need not be monosemantic. The first fact holds especially in strongly synthetic languages where a morpheme may change in form in different neighbourhoods and build morphs (c.f. Czech ***pes*** [nom., sg.], ***ps****-a* [gen., sg.], *dog*). The second fact holds especially for languages with short morphemes. Since the majority of morphemes in the majority of languages is monosyllabic (having also non-syllabic morphs) the hypotheses that follow may be tested in any language.

Before beginning one should set up a list of all morphemes of a language. For Slavicists this has been done by E. A. Karpilovska (2002) for Ukrainian but for non-Slavicist it is sufficient to prepare a random sample of about 1000 root-morphemes from a dictionary. In addition, all variants of the root, all derivatives in which it occurs and all compounds in which it occurs should be collected. This is rather a programming task but can be performed also with pencil and paper.

The hypotheses that follow are merely a beginning of the research, one should develop further ones.

Hypotheses

1. Find the distribution of homonyms, i.e. state how many roots have 0 homonyms, 1 homonym, 2 homonyms,… Set up the frequency distribution and find the theoretical distribution.

2. Find the distribution of the variants of the root (allo-roots), i.e. find the frequencies of roots having 0,1,2,… allo-roots. The variation will be greater in strongly synthetic languages. Using this distribution set up an indicator expressing the degree of synthetism of a language.

3. Is there a relation between the number of allo-roots and the number of compounds in which the root occurs? That is, probably a root having many allo-

roots occurs in more compounds than a root having few allo-roots. Find the dependence and express it by means of a function.

4. Is there a relation between the number of allo-roots and the number of derivates of the roots? State whether the hypothesis "The more allo-roots, the greater the number of derivates" holds.

5. Find the distribution of the canonical root forms: V, C, CV, VC, CVC,… (V = vowel, C = consonant), present it in two-dimensional form and find the two-dimensional distribution.

6. Find the dependence between the number of compounds of a root and the length of the root. Measure root length in two different forms: (1) In terms of syllable numbers, (2) in terms of phoneme numbers.

7. Are there synonymous roots? If so, test the hypothesis "the shorter a root, the more synonyms it has" and find the appropriate form of the dependence.

8. State the distribution of root polysemy and find its mathematical counterpart.

9. Is there a relation between root polysemy and root length? I.e. test the hypothesis that "the longer a root, the smaller is its polysemy".

10. Is there a relation between the number of allo-roots of a root and its polysemy, i.e. test the hypothesis "the more allo-roots a root has, the greater is its polysemy."

11. Develop further hypotheses and connect all in a control cycle (cf. Köhler 1986, 2005).

Procedure

Prepare a root dictionary in electronic form. If possible, use as starting points WordNet or other Internet sources and write the individual roots in a table containing columns for all variables mentioned above. Before you begin to work, set up all operational definitions. Publish any results with the complete table of data.

References

Karpilovska, E.A. (2002). *Korenevyj gnizdovyj slovnyk ukraïns'koï movy. Gnizda sliv z veršynami omografičnymy korenjamy.* Kyïv: Ukraïns'ka encyklopedija imeni M.P. Bažana.

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.

2.9. Cohesion of compounds

Problem

Develop a quantification method for measuring the cohesion of compounds. If you want to save work, use and modify the existing one (cf. Fan, Altmann 2007a,b). Then study the distribution of cohesion in texts of a special sort. Set up a model of this distribution and study its parameters in the course of time.

Procedure

After having defined operationally the concept of "compound", obtain a set of compounds from texts of a given sort. The definition may refer to the written form of compounds. If a compound is not written as one word, there are several ways of expressing the given concept, i.e. there are different degrees of cohesion. For example in bilingual English technical dictionaries one finds the concept *bottler, bottle filler, bottling machine, bottle-filling machine, a machine used for filling bottles*, out of which the first is a derivative, the other ones are compounds with different cohesion degrees. Languages may have different cohesion formation, and different text sorts may prefer some of them.

Take a number of press texts, extract all compounds, ascribe them cohesion degrees and set up the frequency distribution of degrees. If you have a continuous scale, make some reasonable intervals. First fit mechanically a curve to the frequencies using available software. Care for a minimum of parameters. Then "theorify" the curve, i.e. derive it from linguistic assumptions (e.g. by means of a differential equation) or devise other properties of compounds and associate them with the cohesion degree. One may use also (quantified) typological properties of the given language. That is, embed your result in a broader theoretical framework.

Perform the study historically and observe the change of degree representation in the given language. Use the same text sort.

At last, compare the distribution of cohesion degrees in different text sorts. This can, of course, be performed also without cohesion measurement, simply by comparing the numbers of compounds in individual formal classes.

References

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

Fan, F., Altmann, G. (2007b). Measuring the cohesion of compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and quantitative lexicology: 190*-209. Černivcy: Ruta.

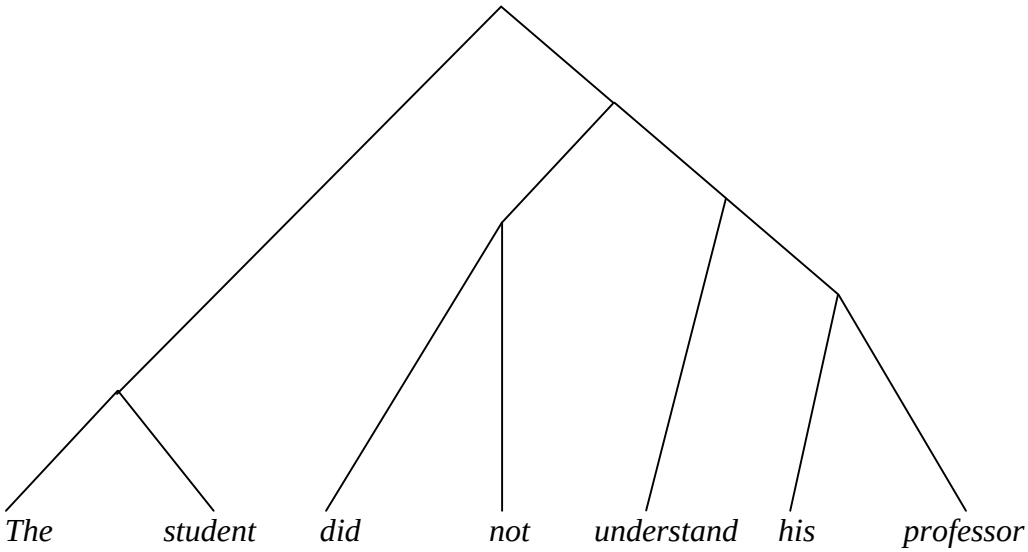
2.10. Symmetry of sentence structure

Problem

Present the syntactic structure of a sentence in form of a binary graph, i.e. a graph in which all vertices have outdegree 2 and the terminal vertices outdegree 0. Compute the (as)symmetry of each sentence and evaluate the properties of the text.

Procedure

Using an appropriate type of grammar display the syntactic structure of the sentence in a binary graph, e.g.



For each non-terminal vertex *i* compute the difference between the number of left and right vertices $|l_i - r_i|$, for example the top vertex has 2 vertices to the left and 5 vertices to the right, hence |2 – 5| = 3. Compute all differences and insert them in the formula

$$A = \frac{2\left(\sum_{i=1}^{n-1} |l_i - r_i| - NA(n)\right)}{(n-1)(n-2)}$$

where *n* is the number of terminal vertices (in the above graph there are 7 (= number of words), (*n*-1)(*n*-2)/2 is the maximal asymmetry and *NA*(*n*) is the necessary asymmetry of all binary graphs which do not have 2^k terminal vertices. In order to facilitate the computation we present the necessary asymmetries in

Table 2.10.1 up to sentence length n = 100. For sentences longer than 100 words use the formula presented by Sander and Altmann (1973).

Table 2.10.1
Necessary asymmetry of a binary graph with n terminal vertices

n	NA(n)	n	NA(n)	n	NA(n)	n	NA(n)
1	0	26	10	51	21	76	32
2	0	27	10	52	20	77	34
3	1	28	8	53	21	78	34
4	0	29	8	54	20	79	34
5	2	30	6	55	19	80	32
6	2	31	4	56	16	81	36
7	2	32	0	57	17	82	38
8	0	33	5	58	16	83	40
9	3	34	8	59	15	84	40
10	4	35	11	60	12	85	42
11	5	36	12	61	11	86	42
12	4	37	15	62	8	87	42
13	5	38	16	63	5	88	40
14	4	39	17	64	0	89	42
15	3	40	16	65	6	90	42
16	0	41	19	66	10	91	42
17	4	42	20	67	14	92	40
18	6	43	21	68	16	93	40
19	8	44	20	69	20	94	38
20	8	45	21	70	22	95	36
21	10	46	20	71	24	96	32
22	10	47	19	72	24	97	36
23	10	48	16	73	28	98	38
24	8	49	19	74	30	99	40
25	10	50	20	75	32	100	40

The symmetry *S* can be computed simply as the complement of *A*, i.e.

$S = 1 - A.$

Having analyzed a complete text, solve the following tasks:

1. Study the sequence of (as)symmetries and characterize it by autocorrelation, Hurst exponent, Lyapunov coefficient and fractal dimension.

2. Compare the results in Task 1 with those won from other texts and show that different genres have different characteristic sequences.
3. Compute the mean (as)symmetry for each text and set up a classification of texts/genres.
4. Find the distribution of (as)symmetries for each text separately. To this end count the frequencies in the intervals <0.0; 0.1>, (0.1; 0.2>, …, (0.9; 1.0). Find a theoretical distribution capturing the empirical data. Substantiate the distribution by deriving it from linguistic assumptions.
5. For each text compute Ord´s criterion and plot the values in a Cartesian coordinate system. Do the texts form a cloud or do different genres lie in different domains? Is it possible to draw straight lines for different genres?

References

Sander, H.-D., Altmann, G. (1973). Asymmetrie binärer Stammbäume. *Phonetica 28, 171-181.*

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

2.11. Transitivity as a text parameter

Problem

The quantification and the measurement of the degree of transitivity contained in a clause and the statistical processing of transitivity are non-solved problems.

Transitivity is considered to be a central phenomenon in the structure of human language and it is regarded as a language universal. Traditionally (cf. Tsunoda 2005) transitivity refers to the sentence property which is determined by the presence (or absence) of an object. Thus, clauses containing an object are assigned as transitive

(1) *John hit Paul*

while clauses with no object are assigned as intransitive

(2) *Mary smiles*.

From the semantic point of view, a transitive clause expresses an activity/action which goes from the subject to the object.

Despite the fact that the term transitivity is often used in a way which takes its content for granted, there is not any consensus about the character of transitivity among linguists (Naess 2007, Tsunoda 2005).

A traditional approach to transitivity assumed a binary character of this language property which corresponds to both structuralist and generativist linguistic tradition. However, some linguists have pointed out to inadequacy of 'binary' approach and they claimed that transitivity should be viewed rather as a continuum because it better reflects the nature of the human language. Further, transitivity started to be treated as prototype category (cf. Rosch 1978) which means that category of this kind admits of degrees of membership – it allows to characterize clauses as *more* or *less* transitive (e.g. Lakoff 1977; Hopper, Thompson 1980; Givón 1985; Kittilä 2002). Unfortunately, because of the lack of an empirical methodology also the "prototype" approach to transitivity has not led to consensus, not even about the prototypes of the category.

However, it is possible to take a prototype approach to transitivity as a starting point and define formal operational criteria which can be used for measurement of degree of transitivity. Needless to say, the proposed criteria (see below) do not reflect the "truth" of the character of transitivity, they only enable us to observe and test the transitivity properties empirically. Of course, the other criteria can be put forth.

Further, it is well known that the behaviour of some language categories is genre-dependent and it seems reasonable to expect that transitivity should be the category of this kind (cf. Thompson, Hopper 2001). So, the relationship between transitivity and genre can be studied.

Quantify transitivity. Then take 10 shorter texts in a selected language and measure the transitivity of all clauses.

Procedure

Let us define transitivity as any impact of an event or state expressed by a non-passive predicate on any entity expressed by dependent participant of predicate, except of a subject. This broad definition is in accordance with the full valency approach (see Problem 2.16 and Čech, Pajas, Mačutek 2010). As for the passive clauses, the situation is more complicated and, for the sake of simplicity, passive clauses are not considered here.

Based on the assumption that "the form of all speech-elements or speech-patterns is intimately associated with their behavior" (Zipf 1935, 19), the degree of transitivity of the clause is given by the form of dependent participants. Specifically, types of participants are determined as follows:

- direct (non-prepositional) noun (e.g., *Mary sees the* ***ball***);
  this form of participant is assumed to be prototypical, consequently, it is assigned maximum transitivity value, $t = 1$;
- prepositional noun (e.g., *John looks* ***for*** *the* ***book***);

the presence of the preposition generally indicates adverbial meanings, therefore this construction is assigned lower value, $t = 0.8$;

- infinitive verb (e.g., *He decided **to come***);
  this nominal form of the verb is a kind of situational participant, $t = 0.6$;
- subordinate clause (e.g., *He said **that he wants to get back***);
  the subordinate clause is a kind of situational participant, because its non-nominal form, it is assumed to express lower value than infinitive verb, $t = 0.4$;
- other participants (e.g., *She looks **beautiful***); the majority of cases are represented by adverbial participants; although they are not "real" participants in traditional sense, they are ruled by full valency mechanism of verb; $t = 0.2$.

Let us assume that the "transit" is the strongest in the clause which expresses only and only this transit and nothing more (e.g., adverbial) – in this case the attention of the hearer/reader is focused just on the transit (*John hit the ball*). The opposite side represents the clause with an absence of dependent verb participant (*Mary sleeps*). Based on this assumption, let us define the transitivity value of the clause (TC)

$$\mathrm{TC} = \frac{\sum_{i=1}^{n} t_i}{TC_{max}}$$

where t_i is the transitivity value of a particular participant i and TC_{max} is the maximal theoretical transitivity value of the clause which arises if all participants are expressed by non-prepositional noun (i.e., $t = 1$), hence

$$TC_{max} = n$$

where n is the number of participants in the clause. Hence TC is the mean transitivity of the clause and can be written as

$$TC = \frac{1}{n}\sum_{i=1}^{n} t_i .$$

Examples of computation:

John hit <u>the ball</u> ($t = 1$)

$TC = 1/1 = 1$

He put the <u>book</u> ($t_1 = 1$) *<u>on the shelf</u>* ($t_2 = 0.8$)

$TC = 1.8/2 = 0.9$

He said to me ($t_1 = 0.8$) *yesterday* ($t_2 = 0.2$) *that he is ill* ($t_3 = 0.4$)

$TC = 1.4/3 = 0.4666667$

The mean of all *TC*s can by used as a transitivity parameter of the entire text TC_{text}. Obviously, it is possible to use TC_{text} for text sort, language, authorship characterisation.

(1) Show the extent of transitivity in ten press texts. (2) Compute the mean *TC* for press texts and set up a 95% confidence interval for the mean. (3) Find the distribution of *TC*: first test for normality. (4) Compare press texts with fairy tales. (5) Compare the *TC* of an English text with the *TC* of its translations in a selected language.

References

Čech, R., Pajas, P., Mačutek, J. (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics 17, 291-302.*

Čech, R. (2011). Four reasons for a radical revision of the Transitivity Hypothesis. (submitted)

Givón, T. (1985). Ergative morphology and transitivity gradients in Newari. In Plank, F. (ed.), *Relational Typology: 89-108*. (Trends in Linguistic Studies and Monographs, 28). Berlin: Mouton.

Hopper, P., Thompson, S. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Kittilä, S. (2002). Remarks on the basic transitive sentence. *Language Sciences 24, 107-130*.

Lakoff, G. (1977). Linguistic Gestalts. *Papers from the 13th regional meeting of the Chicago Linguistic Society, 236-287.*

Naess, Å. (2007). *Prototypical Transitivity*. Amsterdam, Philadelphia: John Benjamins.

Rosch, E. (1978). Principles of categorization. In: Rosch, E., Lloyd, B.B. (eds.), *Cognition and categorization: 27-48*. Hillsdale NJ: Lawrence Erlbaum.

Thompson, S., Hopper, P. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure: 27-56*. Amsterdam, Philadelphia: John Benjamins.

Tsunoda, T. (2005). Transitivity. In: Brown, K. (ed.), *The Encyclopedia of Language and Linguistics: 4670-4677*. Oxford: Pergamon,

Zipf, G. K. (1935). *The psychobiology of language*. Boston: Houghton-Mifflin.

2.12. Transitivity and text deployment

Problem

Is there any relationship between transitivity and the deployment of text? Three hypotheses are possible: (i) the extent of transitivity increases, (ii) decreases, (iii) oscillates. Test the hypotheses.

Procedure

Follow the measurement of transitivity of clause as is presented in Problem *2.11. Transitivity as a text parameter* or devise your own measure.

(1) Compute the transitivity value (*TC*) for each clause of a short text (e.g., newspaper article, scientific article, short story, speech) and observe the sequence of *TC* in the given text.

(2) In the case of longer text (e.g., novel), compute the transitivity value (*TC*) for each clause of the chapter and (a) observe the development of *TC* in each chapter; (b) analogously to the TC_{text} in Problem *2.11. Transitivity as a text parameter* compute the mean of the chapter ($TC_{chapter}$) and observe the development of $TC_{chapter}$ in the entire novel.

The sequence of *TC*s will never be linear in longer passages. Since it lies in the interval <0,1>, it will have some sigmoid form or it will oscillate. The oscillation can be very regular: in that case use either Fourier analysis (Howell 2001; Stein, Weiss 1971) or a difference equation for capturing the regularity. Or it is very irregular: in that case use simply *the non-smoothness indicator* as shown in Popescu et al. (2010, 95ff) which has an easy testability.

In order to express the course of transitivity in other ways, use Hurst´s exponent, Lyapunov exponent or some other quantitative expression (cf. Hřebíček 2000).

Analyse a stage play. Does transitivity correlate with the dramatic course of the play? If so, substantiate this discovery linguistically and, after having analyzed several stage plays, set up further hypotheses

How do behave lyrical poems? Are they richer in transitivity then press texts?

Can genres be ordered according to the extent of transitivity?

References

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

Howell, K.B. (2001). *Principles of Fourier Analysis*. Boca Raton-London-New York-Washington, D.C.: Chapman & Hall/CRC.

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010)) *Vectors and codes of text.* Lüdenscheid: RAM-Verlag (= Studies in Quantitative Linguistics No 8)

Stein, E.M., Weiss, G. (1971). *Introduction to Fourier analysis on Euclidean spaces*. Princeton: Princeton University Press.

2.13. Transitivity and verb

Problem

Use the transitivity measurement, as defined in Problem *2.11. Transitivity as a text parameter*, for a verb characterisation. Compare the results with studies focused on transitivity prototypes (e.g. Naess 2007; Kittilä 2002, 2010; Rozas 2007; Kulikov et al. 2006; Brandt, García 2010)

Procedure

Use a syntactically annotated corpus, for example the *Alpino treebank* (Beek et al. 2001), the *CESS-ECE corpus* (Martí et al. 2007), the *Floresta synta(c)tica* (Alfonso et al. 2002), the *Italian Syntactic-Semantic Treebank* (Montemagni et al. 2003), the *Prague Dependency Treebank* (Hajič et al. 2006), the *Szeged treebank* (Csendes et al. 2005).

Compute the transitivity value of the clause (*TC*) and assign the observed value of the *TC* to the predicate (its lemma) of the given sentence. Analyze each clause in the corpus. Make a list of lemmas and compute the mean of *TC* for each lemma (TC_{lemma}).

Rank lemmas in descending order of TC_{lemma} and compare the results with the studies focused on transitivity prototypes (Naess 2007; Kittilä 2002).

Find a probability distribution of the rank order of verbs. Characterize texts according to the extent of *TC*.

Draw conclusions about the form of lemmas and their transitivity.

References

Afonso, S., Bick, E., Haber, R., Santos, D. (2002). *"Floresta sintá(c)tica": a treebank for Portuguese.* In: M.G. Rodríguez, C.P.S. Araujo (eds.), *Proceedings of LREC 2002, 1698-1703.*

Beek L. v.d., Bouma, G., Malouf, R., Noord, G.v. (2002). The Alpino Dependency Treebank. In: *Computational Linguistics in the Netherlands CLIN 2001, 8-22*. Rodopi, 2001.

Brandt, P., García, M.G. (2010). *Transitivity. Form, Meaning, Acquisition, and Processing*. Amsterdam-Philadelphia: John Benjamins.

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A. (2005). *The Szeged Treebank.* Heidelberg: Springer.

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Kittilä, S. (2002). Remarks on the basic transitive sentence. *Language Sciences 24, 107-130*.

Kittilä, S. (2010). Defining prototypical transitivity. *Lingua 118(12), 2012-2020*.

Kulikov, L., Malchukov, A., Swart, P. de, (2006). *Case, Valency and Transitivity*. Amsterdam-Philadelphia: John Benjamins.

Martí, M.A., Taulé, M., Marquez, L., Bertran, M. (2007). *CESS-ECE: A multilingual and multilevel annotated corpus*. Available for download from: http://www.lsi.upc.edu/ mbertran/cess-ece/publications.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Nana, N., Pianesi, F., Delmonte, R. (2003). Building the Italian Syntactic-Semantic Treebank. In: Abeillé, A. (ed.), *Treebanks: Building and Using Parsed Corpora: 189-210.* Dordrecht: Kluwer.

Naess, Å. (2007). *Prototypical Transitivity*. Amsterdam-Philadelphia: John Benjamins.

Rozas, V.V. (2007). A usage-based approach to prototypical transitivity. In Delbecque, N., Cornillie, B. (eds), *On Interpreting Construction Schemas*: *17-38.* Berlin: Mouton de Gruyter.

2.14. Transitivity and children development

Problem

"... we see a continuous developmental progression on which children gradually become more productive with novel verbs in the transitive SVO construction during their third and fourth years of life and beyond, evidencing a growing understanding of canonical English word order" (Tomasello 2003).

Use the transitivity measurement, as defined in Problem 2.11. *Transitivity as a text parameter*, for a characterisation of children development. Test the hypothesis concerning the age and transitivity value (*TC*). Analyze the problem separately for girls and boys, analogously to Popescu, Čech, Altmann (2011, chapter 9)

Procedure

Collect a sufficient set of children texts; for example, use web pages containing stories told or written by children (e.g,. for English: http://www.kids-space.org/ http://www.goodnightstories.com/stories.htm; for Czech:

http://sedmikraska.cz/dilna/cervotoc.php; http://zs.staravesno.indos.cz/dusan.htm)

Compute the transitivity value (*TC*) for each clause of the text and determine the TC_{text} (see Problem 2.11. *Transitivity as a text parameter*). Group particular TC_{text} in accordance with the age and compute the mean transitivity value for each year (TC_{year}). Observe the relationship between age and TC_{year}.

Interpret the results with regard to language development studies (e.g., Ingram 1971; Fisher 2000; Tomasello 2003; Tomasello; Bates 2002; Clark 2009).

References

Fisher, C. (2000). From form to meaning: a role for structural alignment in the acquisition language. In: Reese, H.W. (ed.), *Advances in Child Development and Behavior: 1-55*. London: Academic Press.

Clark, E.V. (2009). *First language acquisition*. Cambridge: Cambridge University Press (2nd edition)

Ingram, D. (1971). Transitivity in child language. *Language 47, 888-910*..

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The Lambda-structure of texts*. Lüdenscheid: RAM-Verlag (= Studies in Quantitative Linguistics No 10).

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition.* Cambridge-London: Harvard University Press.

Tomasello, M., Bates, E. (eds.) (2003). *Language development. The essential readings*. Malden: Blackwell Publishing Ltd.

2.15. Transitivity and aspect

Problem

"Aspect is systematically correlated with the degree of Transitivity of the verb: if the Aspect is perfective, the interpretation – other things being equal – has properties allowing the clause to be classified as more transitive; but if the Aspect is imperfective, the clause can be shown on independent grounds to be less transitive" (Hopper, Thompson 1980, 271).

"A correlation of aspect with transitivity is interesting in that it may function in both directions. On the one hand, languages with explicit morphological aspect marking may show differences in other properties of the clause co-varying with the aspect marking, so that clauses marked for imperfective aspect show a highly transitive case-marking pattern or other structural features associated with high transitivity, while clauses marked for imperfective aspect similarly show structural features associated with low transitivity" (Naess 2007, 118).

The relationship between transitivity and aspect seems to be well corroborated in linguistics, as is illustrated by a host of examples from many languages (cf. Hopper, Thompson 1980, 270-276). It has to be emphasized that all corroborations of this kind, to our knowledge, are based on *qualitative* analyses of language.

However, Čech, Pajas (2009) and Čech (2009) focused on a *quantitative* analysis of the relationship between aspect and transitivity (and ditransitivity) and revealed no significantly important correlation between both language properties. In these studies transitivity is defined in traditional sense: clauses containing an object are assigned as transitive, while clauses with no object are assigned as intransitive.

Modify the problem: use the transitivity measurement, as defined in Problem *2.11 Transitivity as a text parameter* and Problem *2.13. Transitivity and verb*, and test the hypothesis concerning the relationship between the transitivity value of verb (TC_{verb}) and aspect.

Procedure

(1) Make a list of pairs of verbs which have the same lexical meaning and which differ in respect to aspect, for example in

Czech

psát	–	*napsat*
[*write*, imperfective]		[*write*, perfective]

German

schreiben	–	*aufschreiben*

Hungarian

írni	–	*megírni*

Follow the procedure presented in Problem *2.13. Transitivity and verb* and compute the mean of *TC* for each verb lemma in the list (TC_{verb}). Test the differences between mean TCs_{verb} in each pair using the normal criterion

$$u = \frac{\overline{TC}_{verb1} - \overline{TC}_{verb2}}{\sqrt{Var(\overline{TC}_{verb1}) + Var(\overline{TC}_{verb2})}}$$

where *Var* are the empirical variances of means of individual verbs.

(2) Follow the procedure presented in Problem *2.13. Transitivity and verb* and rank lemmas in descending order of TC_{lemma}. Assign to each lemma aspect characterisation, if possible (bi-aspectual lemmas have to be excluded). Observe the relationship between TC_{lemma} and perfectiveness/imperfectiveness. Test the hypothesis: the higher TC_{lemma}, the more probably the verb is perfective.

References

Čech, R. (2009). Testing of the Transitivity Hypothesis: double object verbs and aspect in Czech. In: Dočekal, M., Ziková, M. (eds.), *Czech in Formal Grammar: 29-38.* München: Lincom.

Čech, R., Pajas, P. (2009). Pitfalls of the Transitivity Hypothesis: Transitivity in conversation and written language in Czech. *Glottotheory 2, 41-49.*

Hopper, P., Thompson, S. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Naess, Å. (2007). *Prototypical Transitivity*. Amsterdam-Philadelphia: Benjamins.

2.16. The distribution of full valency frames

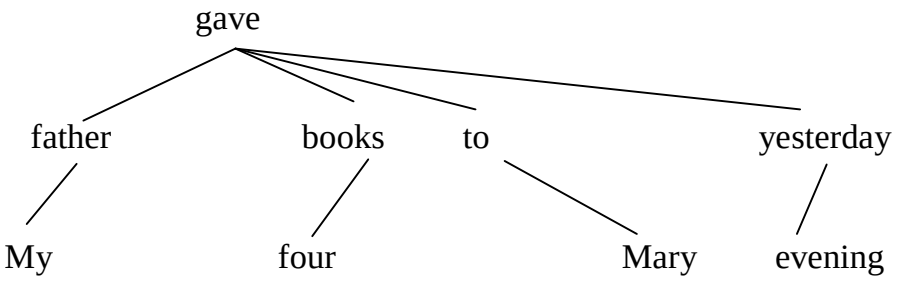
Problem

The "full valency" approach, proposed by Čech, Pajas and Mačutek (2010), is the reaction to fundamental deficiency of the traditional valency approach; specifically, the absence of clear and interpersonally acceptable criteria for distinguishing *complements* (obligatory arguments governed by the verb) and *adjuncts* (optional arguments) (cf. Rickheit, Sichelschmidt 2007; Herbst, Götz-Vottler 2007).

The term ''full valency'' means that all verb arguments which occur in observed language material are taken into account. A verb argument is an element of the sentence which is directly dependent on the predicative verb. For example, in the sentence

(1) *My father gave four books to Mary yesterday evening*

the words *father, books, to, yesterday* are assigned as arguments of the verb *gave* because they are direct dependents of the verb *give*.

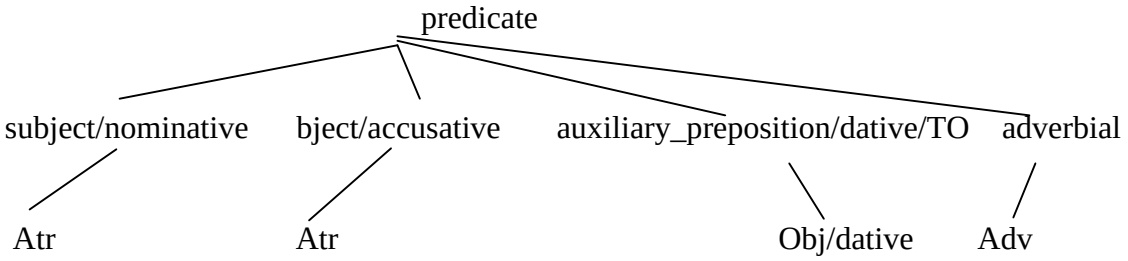


Since full valency is proposed to be a new linguistic category, it is necessary first to observe its advisability: it is well known that linguistic classification is judged as ‘‘good’’, ‘‘useful’’ or ‘‘theoretically prolific’’ if the taxa follow a ‘‘nice’’ rank-frequency distribution (Altmann 2005). So if full valency represents a ‘‘theoretically prolific’’ class, it should have a regular distribution.

Observe the frequency distribution of full valency frames in the language. Suggest a model.

Procedure

First, it is necessary to define a property of full valency frame for the testing of hypothesis. Čech, Pajas and Mačutek (2010) suggested *analytical functions* (e.g., subject, object), *morphological cases* (e.g., nominative, genitive), and *lemmas* (only in the case of prepositions) as properties which are used for an argument classification. These properties are assigned to each word in the sentence and, then, they are used as constituents of the full valency frame. As an illustration, if the sentence (1) adopts the annotation scheme as follows (it is based on Prague Dependency Treebank annotation, see Hajič et al. (2006))



Tthe full valency frame of the verb/lemma GIVE is

> GIVE: [subject/nominative; object/accusative; auxiliary_preposition/ dative/TO; adverbial]

Obviously, the properties which characterize arguments can be modified.

Next, one has to observe the number of full valency frames for each verb/ lemma in the language. Only formally unique full valency frames should be counted. This means that if the verb/lemma occurs in two or more identical full valency frames in the corpus, only one full valency frame is counted. In other words, only the number (not frequency) of different full valency frames is taken into account. Concretely, for the lemma COME presented in sentences (3) and (4),

(3) *Mary* *came* *early*
[subject/nominative] [predicate] [adverbial]

(4) *Tom* *comes* *late*
[subject/nominative] [predicate] [adverbial]

only one full valency frame is counted, viz.

COME: [subject/nominative; Adv].

Use the syntactically annotated corpus, for example the *Alpino treebank* (Beek et al. 2001), the *CESS-ECE corpus* (Martí et al. 2007), the *Floresta sinta(c)tica* (Alfonso et al. 2002), the *Italian Syntactic-Semantic Treebank* (Montemagni et al. 2003), the *Prague Dependency Treebank* (Hajič et al. 2006), the *Szeged treebank* (Csendes et al. 2005) and observe the number of unique full valency frames for each verb/lemma (not frequency). Count the number of verbs/ lemmas with 1,2,3,...,*n* full valency frames and rank it as follows:

Number of full valency frames	Number of lemmas/verbs with *x* full valency frames
x	*f(x)*
1	1200
2	890
3	610
…….	

Find the distribution empirically; use a statistical software which offers a large number of theoretical probability distributions. Then substantiate the distribution using linguistic arguments.

References

Alfonso, S., Bick, E., Haber, R., Santos, D. (2002). "Floresta sintá(c)tica": a treebank for Portuguese. In: M.G. Rodríguez, C.P.S. Araujo (eds.), *Proceedings of LREC 2002, 1698-1703.*

Altmann, G. (2005). Diversification processes. In: R. Köhler, G. Altmann, R.G. Piotrowski (Eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*: *646-659.* Berlin/New York: de Gruyter.

Beek v.d..L., Bouma, G., Malouf, R., Noord, G.v. (2002). The Alpino Dependency Treebank. In: *Computational Linguistics in the Netherlands CLIN 2001: 8-22.* Rodopi, 2001.

Čech, R., Pajas, P., Mačutek, J. (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics 17(4), 291-302.*

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A. (2005). *The Szeged Treebank.* Heidelberg: Springer.

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Linguistic Data Consortium, Philadelphia.

Herbst, T., Götz-Vottler, K. (eds.) (2007). *Valency. Theoretical, Descriptive and Cognitive Issues*. Berlin-New York: Mouton de Gruyter.

Kittilä, S. (2002). Remarks on the basic transitive sentence. *Language Sciences 24, 107-130*.

Martí, M.A., Taulé, M., Marquez, L., Bertran, M. (2007). *CESS-ECE: A multilingual and multilevel annotated corpus*. Available for download from: http://www.lsi.upc.edu/□mbertran/cess-ece/publications

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Nana, N., Pianesi, F., Delmonte, R. (2003). Building the Italian Syntactic-Semantic Treebank. In: Abeillé, A. (ed.), *Treebanks: Building and Using Parsed Corpora: 189-210.* Dordrecht: Kluwer.

Rickheit, G., Sichelschmidt, L. (2007). Valency and cognition – a notion in transition. In: T. Herbst, K. Gotz-Vottler (eds.), *Valency. Theoretical, descriptive and cognitive issues*: *163-182*. Berlin-New York: de Gruyter.

2.17. Syntactic network analysis: hub/authority weight versus degrees

Problem

"Intuitively, the nodes having more links should be more important in the network. (…) However,… the number of connections can't indicate the real importance of nodes. A node which is pointed by a large number of unimportant nodes will be less authoritative than a node which is pointed by a few important nodes (Kleinberg 1998). Similarly, a node will be less "hub"-like, if it points to a large number of unimportant nodes, than that pointing to a few highly authoritative nodes. In other words, a node is only a good hub if it connects with many good authorities, and it is a good authority if it is linked with many good hubs. Hubs and authorities are in fact interdependent with each other, and cannot be identified independently."

"The positions of hubs and authorities are more intriguing, since they are determined by the global structure of the network, and not by the characteristics of individual nodes in isolation. The change of hubs and authorities may reflect the

change of the network structure, which in turn reflects the change of the language system in the speaker." (Ke, Yao 2008).

In the network analysis of development of child language Ke and Yao (2008) detected the dissociation between degree and hub/authority weight – some words with many links have low hub/authority weight and vice versa.

Generalize the problem. Analyze large language corpora (treebanks) and try to explain discrepancies between degree and hub/authority weight linguistically.

Procedure

Follow the procedure presented in Problem *4.15. Syntactic network analysis* and construct a directed network.

Compute out-degree, in-degree, hub-weight, and authority-weight of each word (or lemma) and rank the words (or lemmas) in descending order. Compare (a) out-degree and hub-weight ranks of particular words and (b) in-degree and authority-weight ranks of particular words. Observe words (or lemmas) which reveal discrepancies between ranks. Analyze whether the words with discrepancies could by grouped in accordance with some language properties (e.g., part of speech).

Better results should be got, if weighted ranks are used.

References

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

Kleinberg J. (1998) Authoritative sources in a hyperlinked environment. In *Proc. 9th ACMSIAM Symposium on Discrete Algorithms, 668-677*. Extended version in *Journal of the ACM* 46(1999). Appears also as *IBM Research Report RJ 10076*, May 1997.
[available at http://www.cs.cornell.edu/home/kleinber/auth.pdf]

2.18. Syntactic network analysis: hub/authority weight as a text parameter

Problem

"The change of hubs and authorities may reflect the change of the network structure, which in turn reflects the change of the language system in the speaker" (Ke, Yao 2008).

Analogically, does the change of hubs and authorities reflect the change of genre, authorship, or diachronic development of language? Set up hypotheses and test them.

Procedure

Follow the procedure presented in Problem *4.15. Syntactic network analysis* and construct a directed network. However, do not use entire corpora, but only particular texts – for example, (1) take texts of different genres written by the same author or (2) take texts of the same genre written by different authors; or (3) take texts of the same genre written in different decades or centuries and for each text create an individual network.

Compute hub-weight and authority-weight of each word (or lemma) in each network and observe the differences of hub/authority weights. Scrutinize the changes of hub/authority weights of particular words (if they appear). Focus on function words which should not reflect thematic differences of texts. Order the function words according to the measure of hub-weight/authority-weight and set up a hypothesis joining hub-weight/authority-weight with some other properties of the given words.

References

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford: Oxford University Press.

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

Mehler, A. (2007). Large text networks as an object of corpus linguistics studies. In: Lüdeling, A., Kytö, M. (eds.), *Corpus linguistics. An international handbook of the science of language and society: 328-382.* Berlin-New York: de Gruyter.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review 45(2), 167-256.*

Newman, M. (2010). *Networks: An Introduction.* Oxford: Oxford University Press.

2.19. Proper names as a high Transitivity feature

Problem

Hopper and Thompson (1980) claim that the more an object of the sentence is individuated, the higher transitivity is expressed by the sentence. Specifically, "[a]n action can by more effectively transferred to a patient which is individuated

than to one which is not" (Hopper, Thompson 1980, 253). Properties of nouns which indicate high individuation of object (left column) and low individuation (right column) are shown in Table 2.19.1.

Table 2.19.1. Properties of nouns which indicate high and low individuation of object (Hopper, Thompson 1980, 253)

INDIVIDUATED	NON-INDIVIDUATED
proper	common
human, animate	inanimate
concrete	abstract
singular	plural
count	mass
referential, definite	non-referential

Implement this statement into the approach presented in Problem no. *2.11 Transitivity as a text parameter* and test the following hypotheses:

- the higher the transitivity value of the clause, the more probably a proper name occurs in the clause;
- the higher the transitivity value of the text, the more proper names occur in the text.

Moreover, you can extend these hypotheses and observe the other properties of high individuation (cf. Table 2.19.1) which can be assigned to proper names, i.e. human&animacy, singular. Test the following hypotheses:

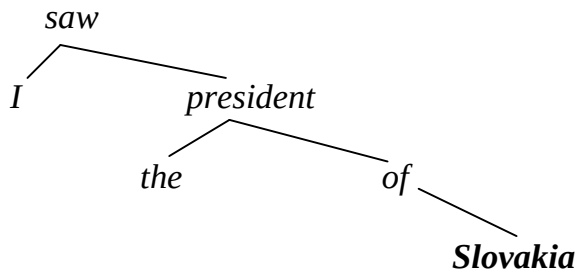
(3) if a proper name is animate (i.e. it denotes human or animal), it should occur in clauses with higher transitivity value more probably than a proper name which denotes inanimate entities;
(4) if a proper name has the singular form, it should occur in clauses with higher transitivity value more probably than a proper name in the plural form;
(5) the higher the transitivity value of the text, the more animate proper names occur in the text (in comparison with inanimate ones);
(6) the higher the transitivity value of the text, the more proper names in the singular form occur in the text (in comparison with inanimate ones).

Further, you can refine the problem by testing of following hypotheses:

(7) if a proper name is both animate (i.e. it denotes human or animal) and has the singular form, it should occur in clauses with higher transitivity value more probably than inanimate proper names in the plural form;
(8) if the proper name is both animate (i.e. it denotes human or animal) and has the singular form, it should occur in clauses with higher transitivity value more probably than animate proper names in the plural form;

(9) if the proper name is both animate (i.e. it denotes human or animal) and has the singular form, it should occur in clauses with higher transitivity value more probably than inanimate proper names in the singular form;
(10) the higher the transitivity value of the text, the more animate proper names in the singular form occur in the text in comparison to inanimate proper names in the plural form;
(11) the higher the transitivity value of the text, the more animate proper names in the singular form occur in the text in comparison to animate proper names in the plural form;
(12) the higher the transitivity value of the text, the more animate proper names in the singular form occur in the text in comparison to animate proper names in the plural form.

However, the hypotheses can be falsified because the proper name can occur not just as a direct dependent element of verb, see the graph of the sentence *I saw the president of* ***Slovakia***



Obviously, the proper name *Slovakia* has no impact on the transitivity value of the clause (*TC*). Moreover, the approach presented in Problem no. *2.11 Transitivity as a text parameter* neglected (consciously) the subject position in which proper names often occur. Hence, if necessary, modify the hypotheses.

Procedure

Follow the procedure presented in Problem no. *2.11. Transitivity as a text parameter* in this volume and compute the transitivity value (*TC*) of each clause in a text (or in a corpus) and the transitivity value of an entire text (TC_{text}).

Hypothesis no. 1 should be tested as follows: (a) group clauses with the same *TC* (or state some intervals); (b) in each group compute the mean *P* of a clause, i.e. divide the number of proper names in a group by the number of clauses in the same group; (c) rank groups in a decreasing order of *TC* and to each group assign the mean *P*; (d) if the hypothesis is adequate, the mean *P* should decrease in accordance to decreasing *TC* - test the relationship between *TC* and *P*.

Analogously, test hypothesis no. 2: (a) take many texts and for each text compute its transitivity value TC_{text} ; (b) group texts with the same TC_{text} (or state

some intervals); (c) in each group compute the mean *P* of the group, i.e. divide the number of proper names in a group by the number of texts in the same group; (d) rank groups in decreasing order of TC_{text} and to each group assign the mean *P*; (e) if the hypothesis is right, the mean *P* should decrease in accordance to decreasing TC_{text}. That means, test the relationship between TC_{text} and *P*.

Hypothesis no. 3 should be tested as follows: (a) from the text (or a corpus) collect only clauses containing at least one proper name; (b) to each proper name assign an information about its animacy/inanimacy, i.e. for a proper name denoting human or animal $A = 1$, for a proper name denoting inanimate entity $A = 0$; (c) group clauses with the same *TC* (or state some intervals); (d) in each group compute the mean *A* of a clause, i.e. divide the number of animate proper names in a group by the number of clauses in the same group; (e) if the hypothesis is right, the mean *A* should decrease in accordance to decreasing *TC* - test the relationship between *TC* and *A*.

The same procedure may be used for testing of hypothesis no. 4; just change point (b), i.e. replace the animacy by the grammatical number (for singular $S = 1$, for plural $S = 0$).

For the testing of hypothesis no. 5 use the following procedure: (a) take many texts and for each text compute its transitivity value TC_{text} ; (b) to each proper name assign an information about its animacy/inanimacy, e.g. for a proper name denoting human or animal $A = 1$, for a proper name denoting inanimate entity $A = 0$; (c) group clauses with the same TC_{text} (or state some intervals); (d) in each group compute the mean A_{text} of a text, i.e. divide the number of animate proper names in a group by the number of texts in the same group; (e) if the hypothesis is right, the mean A_{text} should decrease in accordance to decreasing TC_{text} - test the relationship between TC_{text} and A_{text}. Again, use the same procedure for testing of hypothesis no. 6; just change point (b), i.e. replace the animacy by the grammatical number (for singular $S = 1$, for plural $S = 0$).

The hypothesis no. 7 should be tested as follows: (a) from the text (or a corpus) select only clauses containing at least one proper name which is either animate and in singular form or inanimate and in plural form (animate proper names in plural form and inanimate ones in singular form have to be omitted); (b) to each proper name assign an information about its animacy&number, i.e. for a proper name denoting human or animal in singular form $AS = 1$, for a proper name denoting inanimate entity in plural form $AS = 0$; (c) group clauses with the same *TC* (or state some intervals); (d) in each group compute the mean *AS* of a clause, i.e. divide the number of animate singular proper names in a group by the number of clauses in the same group; (e) if the hypothesis is right, the mean *AS* should decrease with decreasing *TC* - test the relationship between *TC* and *AS*. Use the same procedure for testing of hypotheses no. 8 and 9; just change point (a), i.e. in the case of hypothesis no. 8 collect only clauses containing animate proper names and in the case of the hypothesis no. 9 collect only clauses containing proper names in singular form, and point (b), i.e. in the

case of the hypothesis no. 8, replace the animacy and number by a grammatical number (for singular *ASF* = 1, for plural *ASF* = 0), and in the case of the hypothesis no. 9 by animacy (for animate *SA* = 1, for inanimate *SA* = 0).

For the testing of hypotheses no. 10, 11, 12, use the procedure presented for testing hypothesis no. 5, just change the point (b) analogously to the testing of hypotheses no. 7, 8, 9.

You can use corpora with a specific annotation of proper names, e.g. *Prague Dependency Treebank 2.0* (Hajič et al. 2006). It annotates not only proper names themselves but offers a finer annotation, concretely

- given name;
- surname, family name;
- member of a particular nation, inhabitant of a particular territory;
- geographical name;
- company, organization, institution;
- product;
- other proper name: names of mines, stadiums, guerilla bases, etc.

References

Čech, R. (2010). Proper names as a High Transitivity feature: testing of the Transitivity Hypothesis. In: *Mnohotvárnost a specifičnost onomastiky*: 107-113. Ostrava: OU FF.

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Hopper, P., Thompson, S. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Thompson, S., Hopper, P. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure: 27-56*. Amsterdam-Philadelphia: John Benjamins.

Tsunoda, T. (2005). Transitivity. In: Brown, K. (ed.), *The Encyclopedia of Language and Linguistics: 4670-4677*. Oxford: Pergamon.

2.20. Proper names and aspect of verbs

Problem

This problem is an extension of Problems no. 2.15**.** *Transitivity and aspect* and 2.19. *Proper names as a high Transitivity feature*, and 2.4**.** *Transitivity*. According to Hopper and Thompson (1980), both the object of a sentence expressed by

a proper name and the perfectivity of predicative verb manifest high transitivity of a sentence, c.f. "[a]spect is systematically correlated with the degree of Transitivity of the verb: if the Aspect is perfective, the interpretation – other things being equal – has properties allowing the clause to be classified as more transitive; but if the Aspect is imperfective, the clause can be shown on independent grounds to be less transitive" (Hopper, Thompson 1980, 271), and "[a]n action can be more effectively transferred to a patient which is individuated than to one which is not" (Hopper, Thompson 1980, 253). Consequently, both properties should correlate in the sentence, according to the Transitivity Hypothesis.

Test the hypothesis in two different ways; (1) use the "classical" approach to transitivity; (2) implement the hypothesis into the approach presented in Problem no. 2.11. *Transivity as a text parameter*. Compare the results of both methods.

Procedure

1. "Classical" approach

Use a syntactically annotated corpus. Count separately the number of objects expressed by proper names, on the one hand, and by common nouns, on the other, dependent on perfective and imperfective verbs. Make a contingency table of the following form:

	perfective verb	imperfective verb
proper name object		
common noun object		

and test the results by the chi-square. Compare your results with Čech (2010) who used this procedure for analysis of Czech.

2. "Alternative" approach

Follow the procedure presented in Problem no. *2.11 Transitivity as a text parameter* in this volume and for each clause of a text (or a corpus) transitivity value of clause (*TC*). To each predicative verb of clause assign the information about the perfectivity, i.e. if a verb is perfective, $P = 1$, if not, $P = 0$. Further, to each noun directly dependent on predicative verb assign the information about its character, i.e. for proper names $PN = 1$, for common names $PN = 0$. Group clauses with the same *TC* (or state some intervals) and in each group compute the mean *P* of a clause, i.e. divide the number of perfective verbs in a group by the number of clauses in the same group, and compute the mean *PN* of a clause, i.e. divide the number of proper names in a group by the number of clauses in the same group. Rank groups in a decreasing order of *TC* and to each group assign the mean *P* and *PN*. If the hypothesis is right, both the mean *P* and *PN* should

decrease in accordance to decreasing *TC* - test the relationship between *TC* and *P*, *TC* and *PN*, and *PN* and *P*.

References

Čech, R. (2009). Testing of the Transitivity Hypothesis: double object verbs and aspect in Czech. In: Dočekal, M., Ziková, M. (eds.) *Czech in Formal Grammar: 29-38.* München: Lincom.

Čech, R. (2010). Proper names as a High Transitivity feature: testing of the Transitivity Hypothesis. In: *Mnohotvárnost a specifičnost onomastiky: 107-113.* Ostrava: OU FF.

Čech, R., Pajas, P. (2009). Pitfalls of the Transitivity Hypothesis: Transitivity in conversation and written language in Czech. *Glottotheory 2, 41-49.*

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Hopper, P., Thompson, S. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Naess, Å. (2007). *Prototypical Transitivity*. Amsterdam-Philadelphia: Benjamins.

Thompson, S., Hopper, P. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure: 27-56*. Amsterdam, Philadelphia: John Benjamins.

3. Semantics

3.1. Yesypenko´s linguistic world view

Problem

N. Yesypenko (2008) defined 47 conceptual classes within verbs, nouns, adjectives and adverbs and searched for their correlations in English adventure novels (Waugh, Swift, Twain). As a result, she displayed the associations in form of a graph. Find the properties of this graph.

The classes are as follows:

Verbs: I. Verbs of motion/removing. II. Engender verbs. III. Verbs of successful/unsuccessful action implementation. IV. Verbs of temperature phenomena. V. Verbs of communication. VI. Verbs of moral impact/effect. VII. Position verbs. VIII. Verbs of existence. IX. Modality verbs. X. Verbs of reference. XI. Verbs of emotional psychological impact. XII. Verbs of ownership/loss. XIII. Verbs of physiological state. XIV. Verbs of perception. XV. Verbs of subjective assessment. XVI. Verbs of emotional psychological state.

Nouns: I. Appearance/parts of the body. II. Proper names/nicknames. III. Establishments/groupings. IV. Diseases/defects. V. Abstract notions. VI. Food/ meals. VII. Weight/length/volume. VIII. Sound/fragrance/temperature/light. IX. Action/changes/movement. X. Time. XI. Speech. XII. Building/premises. XIII. Material/liquids. XIV. Vehicles. XV. Geographical notions.

Adjectives: I. Traits of character/emotions. II. Physical/natural condition. III. Intellectual capacity. IV. Temperature/sound. V. Shape/size. VI. Degree/ intensity. VII. Actions done to the subject. VIII.. Positive evaluation. IX. Material. X. Negative evaluation.

Adverbs: I. Repetition and frequency. II. Place and direction. III. Condition and consequence. IV. Manner. V. Degree and quantity. VI. Question adverbs.

Procedure

First, state the distribution of the vertex degrees of the graph. Then perform the same computations as Yesypenko for various other texts and treat each of them separately. Compare the degree distribution for each of them with Yesypenko´s result. Generalize the result.

Compute the mean distance in the graph. Compute the distribution of shortest ways. Study the connectivity of the graph, etc. (cf. West 2000; Caldarelli 2007).

Elaborate on the possibility of presenting the world view of a text taking into account individual words (lemmas), not word classes. Compute the graph of the resulting world view and compare it with those obtained from other texts.

Generalize your findings and establish some relations between the properties of the graph and other text properties. That is, incorporate Yesypenko's graph in a kind of a proto-theory. Combine it with Wilson´s (2002, 2006, 2008) approach (cf. Problem 6.6. *Psychoanalytic word categories*).

Use prosaic, poetic, journalistic, scientific, etc. texts

References

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford: Oxford University Press.

Langacker, R. (1990). *Concept, image and symbol: the cognitive basis of grammar*. Berlin: Mouton de Gruyter.

Schwarz, M. (1992). *Kognitive Semantiktheorie und neuropsychologische Realität.* Tübingen: Niemeyer.

West, D.B. (2001). *Introduction to graph theory*. Upper Saddle River, NJ: Prentice Hall. (2nd edition).

Wierzbicka, A. (1985). *Lexicography and conceptual analysis*. Ann Arbor: Karoma.

Wierzbicka, A. (1992). *Semantics, culture and cognition. Universal human concepts in culture specific configurations.* New York: Oxford University Press.

Wilson, A. (2002). The application of computer content analysis in sexology: a case study of primary process content in fictional fetishistic narratives. *Electronic Journal of Human Sexuality,* 5. [Retrieved October 31, 2006, from: http://www.ejhs.org/volume5/wilson.html].

Wilson, A. (2006). Development and application of a content analysis dictionary for body boundary research. *Literary and Linguistic Computing 21, 105-110.*

Wilson, A. (2008). The well-formedness of two psychoanalytic word categories in Portuguese texts. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 285-307*. Chernivtsi: ČNU.

Yesypenko, N. (2008). An integral qualitative-quantitative approach to the study of concept realization in text. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 308-327*. Chernivtsi: ČNU.

3.2. Polysemy and synonymy

Problem

Levickij and Wenhrynowytsch (2009, 82) presented data on polysemy and synonymy of German nouns using the Duden dictionary (1997). According to the

hypothesis of Ziegler and Altmann (2001) the mean number of synonyms is associated with the number of meanings by the relation $S = aP^b$ (S = number of synonyms, P = number of meanings, a, b = parameters). Test the hypothesis.

Procedure

Use the data in Table 3.2.1 presented by Levickij and Wenhrynowytsch or improve them by scrutinizing the cases "7 or more".

Table 3.2.1
Dependence of synonymy on polysemy of nouns in German
Levickij and Wenhrynowytsch (2009, 82)

Number of meanings (polysemy)	Number of nouns	Number of synonyms	Mean number of synonyms
1	33083	43145	1.30
2	2082	13002	6.24
3	432	3844	8.90
4	148	1305	8.82
5	47	630	13.40
6	21	231	11.00
≥ 7	21	291	13.86

Consider also other word classes and other languages and state whether the hypothesis is general enough. If you discover discrepancies, then search for special boundary conditions explaining them or generalize the hypothesis by adding a disturbance factor. Consult the preparatory chapter 3.10, p. 42 in *Problems Vol. 1* and chapter 3.3. in this volume.

Embed the hypothesis – if you can corroborate it – in Köhler's control cycle (1990, 2005).

References

Duden (1997). *Sinn- und sachverwandte Wörter. Synonymwörterbuch der deutschen Sprache.* Herausgegeben und bearbeitet von W. Müller. Mannheim-Leipzig-Wien-Zürich: Dudenverlag.

Köhler, R. (1990). Elemente der synergetischen Linguistik. *Glottometrika 12, 179-188.* 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.

Levickij, V.V., Wenhrynowytsch, A.A. (2009). Quantitative Charakteristika der substantivischen Synonymie im heutigen Deutsch. *Glottotheory 2(1), 75-85*.

Ziegler, A., Altmann, G. (2001). Beziehung zwischen Synonymie und Polysemie: In: Ondrejovič, S., Považaj, M. (eds.), *Lexicographica ´99: 218-225.* Bratislava: Veda.

3.3. The place of synonymy in self-regulation

Problem

Word synonymy is a semantic property associated with other word properties. Some of them have been mentioned in *Problems Vol 2* and in Problem No 3.2. *Polysemy and synonymy* in this volume and belong to the Köhlerian control cycle. Take the word properties mentioned in the procedure below and

(1) set up a hypothesis about the dependence of synonymy on the given individual property,

(2) perform computations of synonymy and the given other property either in texts, dictionaries, language history, dialect atlases, etc.,

(3) set up a preliminary intuitive hypothesis about the dependence,

(4) express the dependence mathematically as a function and test the goodness-of-fit,

(5) strive for a partial theory of synonymy and its place in the language self-regulation.

Procedure

Consider the following 28 word properties (further ones can easily be found):

1. Word length measured in terms of phoneme, letter, syllable, morpheme numbers (one of the counting alternatives is enough).

2. Word frequency in texts.

3. Polysemy as the number of meanings (senses) in a monolingual dictionary.

4. Polytextuality as the number of texts (of a corpus) or as the number of contexts (direct neighbours) in which the word occurs.

5. Morphological status of the word which can be simple, derived, composed or reduplicated.

6. The number of word classes to which the word belongs, directly or by conversion (without morphological change) (cf. Problem *2.7. Paradigmatic expansions of words* in this volume).

7. Productivity as the number of possible derivations, compounds and reduplications that are allowed for the given word. One can find them for some languages on the Internet.

8. The age of the word as the number of years or centuries from the first appearance of the word in the literature.

9. The provenience of the word counted in the number of languages through which a word arrived in the given language.

10. Verb valency counted according to the current or your own definition of valency.

11. Valency of nouns.

12. The number of grammatical categories of the word: conjugation, declination, time, mode, gradation or as the number of affixes that can be combined with the word.

13. Degree of emotionality vs. notionality of the word, e.g. the word "mother" is more emotional than a "paper-clip". In psycholinguistics one can find extensive dictionaries of this property.

14. Pollyanna, i.e. the place of the word on the good-bad scale (e.g. love against illness).

15. Meaning abstractness vs. concreteness, e.g. beauty vs. revolver.

16. Meaning specificity vs. generality, e.g. pen vs. instrument.

17. Degree of dogmatism, e.g. may vs. must, all vs. some, always vs. sometimes.

18. Number of associations of the word (= connotative potency) which can be taken from an association dictionary.

19. Number of possible functions in sentence, e.g. subject, predicate, occurrence in complement etc.

20. Diatopic variation, i.e. the number of sites in a dialect atlas in which the word can be found.

21. The number of dialectal variants (competitors) in a dialect atlas.

22. Discourse properties: does the word signalize an affiliation to a social group or not?

23. The degree of affiliation with the literary language.

24. Diversity: in how many word classes can the word enter by derivation, e.g. in German *Bild* (N), *bildhaft* (Adj and Adv), *bilden* (V).

25. Originality: is the word an original word, a calque or a borrowing?

26. Phraseology: in how many phraseological expressions can the word occur? One finds them on the Internet.

27. Verb transitivity in the sense of Hopper, Thompson (1980) and Thompson, Hopper (2001). This concept contains 10 categories each of which must be first quantified (cf. Problem *2.11. Transitivity*).

28. Full valency, see the Problems *2.16. The distribution of full valency frames* and *4.11. Full valency and synonymy* in this volume.

According to the given property the hypotheses will have different forms. Nevertheless, strive for a unified expression and if necessary, show that synonymy itself affects other properties. On the whole, strive for a theory of syn-

onymy and show that synonymy is a part of a dynamic system. If possible, link up your results with the Köhlerian self-regulation cycle.

References

Čech, R., Pajas, P. (2009). Pitfalls of the Transitivity hypothesis: transitivity in conversation and written language in Czech. *Glottotheory 2/2, 41-49.*

Hopper, P., Thompson, S. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Jiwei Ci (1987). Synonymy and polysemy. *Lingua 72(4), 315-331.*

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.

Murphy, M.L. (2006). Synonymy. In: Brown, K. (Ed.). *Encyclopedia of Language and Linguistics: 376-378.* Kidlington: Elsevier Science Ltd.

Thompson, S., Hopper, P. (2001). Transitivity, clause structure, and argument structure: evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 27-56.* Amsterdam-Philadelphia: Benjamins.

3.4. Distribution of semantic roles and frames (corpus)

Hypotheses

(1) Semantic roles and frame patterns are lawfully distributed.
(2) The frequency of a frame depends on the complexity of the frame.
(3) Both semantic role frequency and frame frequency depend on the text sort of the text under study in a specific way.

Test the hypotheses.

Procedure

Semantic roles (or thematic relations) are elements of some grammar models such as case grammar, sentence semantics, or functional grammar. They specify the semantic functions of parts of a sentence (or even of a discourse). Although there is no universally accepted list of semantic roles some typical ones occur in all proposals (agent, instrument, location, experiencer, ...).

Some languages (Korean, Hungarian, and others) express semantic roles on the surface by case marking (the Hungarian suffix *–ban/ben* indicates a location within an object, the suffix *–val/vel* the instrument, etc.) whereas other lan-

guages (English and many others) leave the interpretation of a grammatical function to context and knowledge of the hearer/reader (i.e., in English, the subject of a sentence can represent almost any semantic role).

A frame, in our context, is a cognitive representation of a stereotypical situation such as *commerce*. It consists of the stereotypical semantic roles which are connected to such a situation (e.g., the *commerce* frame is typically connected with the roles *seller*, *buyer*, *good*, *price*).

Find a corpus with semantic role annotation. Count the number of occurrences of each semantic role and of each frame in the corpus and in the individual texts. Determine the frequency distributions and find appropriate probability distributions. Show that Hypothesis (3) is compatible with the resulting data and specify the dependence.

References

Cook, W.A. (1989). *Case grammar theory*. Washington, DC: Georgetown University Press.

Fillmore, Ch. (1968). The case for case. In: Bach, E., Harms, R.T. (eds.), *Universals in Linguistic Theory*: 1-88. New York: Holt, Rinehart and Winston.

Fillmore, Ch. (1971). Types of lexical information. In: Steinberg, D. (ed.), *Semantics. An interdisciplinary reader in philosophy, linguistics and psychology: 370-392*. Cambridge: Cambridge University Press.

Minsky, M. (1975). *A framework for representing knowledge*. MIT-AI Laboratory Memo 306 [Reprinted in The Psychology of Computer Vision, P. Winston (Ed.), McGraw-Hill, 1975.]

Moreda, P., Navarro, B., Palomar, M. (2007). Corpus-based semantic role approach in information retrieval. *Data & Knowledge Engineering 61(3) , 467-483.*

http://framenet.icsi.berkeley.edu

3.5. Verb classes

Problem

B. Levin (1998) ordered the English verbs in 20 classes. Using this classification test the following hypotheses in any language:

(1) The class contains the more verbs the later the activity expressed by the class arose in our biological evolution.

(2) Different text sorts have different representation of these classes.

Procedure

For the first hypothesis take a dictionary of your language, obtain all verbs and ascribe them to the following classes used also by Levickij and Lučak (2005) (see also *Problems Vol 1. Verb profile*, p. 63).

1. Exchange verbs (*barter, buy, sell, exchange, pay, trade*).
2. Measure verbs (*bill, charge, cost, estimate, fine, measure, price, value, weight*).
3. Change of ownership verbs (*give, take, receive, borrow, lend, steal, return*).
4. Change of Position (*fall, drop, throw, slide, float, roll, fly, rotate, shift*)
5. Change of physical state (*melt, redden, soften, freeze, harden, dry, break*).
6. Circumstance verbs (*begin, start, stop, repeat, commence, continue, finish, halt, complete, quit, initiate, end, keep*).
7. Impact/Effect verbs (*cut, stab, crush, smash, pierce, bite, shoot, kill*).
8. Directed motion verbs (*enter, come, go, arrive, descend, ascend, raise, lower, exit, rise, depart, return, leave*).
9. Verbs of existence (*exist, live, dwell, loom, remain, reside, accumulate, aggregate, herd, gather, create, appear, disappear*).
10. Ingestion verbs (*chew, drink, eat, gobble, ingest, munch, sip, suck, swallow*).
11. Verb of mental process (*acquire, guess, know, learn, memorize, study, think*).
12. Load/Spray verbs (*scatter, spray, load, pile, pack*).
13. Manner of motion verbs (*bounce, dance, follow, hop, jog, jump, march, ride, sail, shuffle, stroll, track, walk, wander*).
14. Verbs of ownership (*belong, have, hold, keep, own, possess*).
15. Verbs of perception and communication (*ask, communicate, feel, hear, listen, look, notice, perceive, see, shout, smell, speak, talk, tell, watch*).
16. Position verbs (*remain, stay*).
17. Verbs of removing (*draw, eliminate, remove, empty, scrub, sweep, peel, shell*).
18. Orientation verbs (*aim, face, orient, point*).
19. Verbs of psychological state (*amuse, annoy, frighten, enjoy, fancy, hate, like*).
20. Verbs of sound emission (*bark, chatte, roar, yelp, rumble, strike, squeak, tick*).

Most probably you will not be content with this classification and will add some further classes or rename the old ones. To this end see also Levickij, Kiiko, Spolnicka (1996) who established 22 classes or Yesypenko (2009) who established 27 classes (cf. problem No. *3.1. Yesypenko's linguistic world view*). Classifications elaborated in more detail are shown in Ballmer, Brennenstuhl (1986).

In any case strive for a complete classification. If the classification is complete, set up the rank distribution of the number of elements in classes (cf. *Problems Vol. 1, Verb classification*: 25f). Consult an evolutionary biologist to learn the sequence of activities in our development. State whether there is a correlation between the size of a class and the developmental epoch.

To test the second hypothesis take simply tagged texts and order the verbs in them into the above mentioned classes. Test the difference between texts using asymptotic or nonparametric tests. What are the characteristic verb classes of individual text sorts? Do not compare verbs but only classes.

Note: do not care for the frequency of verbs, consider only the size of the classes. Check all verb classifications known to you and find the best theoretical distribution. Using it criticize some classifications and strive for linguistic substantiation of the "best" one. Characterize text sort by means of the parameters of the theoretical distribution.

References

Ballmer, T.T., Brennenstuhl, W. (1986). *Deutsche Verben*. Tübingen: Narr.

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

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

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

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

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

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

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

Silnickij, G.G. (1973). Semantičeskie tipi situacij i semantičeskie klassy glagolov. In: *Problemy strukturnoj lingvistiki 373-382*. Moskva: Nauka.

Silnicky, G. (1993). Correlation system of verbal features in English and German. In: Köhler, R., Rieger, B.B. (eds.) *Contributions to Quantitative Linguistics: 409-420.* Dordrecht: Kluwer.

Yesypenko, N. (2009). An integral qualitative-quantitative approach to the study of concept realization in the text. In: Kelih, E., Levickij V., Altmann, G. (eds.), *Methods of text analysis: 308-328.* Chernivtsi, ČNU.

3.6. Verb classes and valency

Problem

Do verbs of some individual semantic classes have greater valency/transitivity than those of other classes?

Procedure

First obtain verbs ascribed to classes as proposed in the problem *3.5. Verb classes*. Then write to each verb its valency/transitivity (a) using a ready valency dictionary or (b) applying a full valency approach, see the problem No. *2.16. The distribution of full valency frames.*

For each of 20 classes compute mean valency and state whether they differ significantly. Order the classes according to their mean valency and state whether mean valency correlates

(a) with the size of the class, i.e. test the hypothesis that the more verbs are in a class, the greater is its mean valency.

(b) In the problem *3.5. Verb classes* you correlated the class size with the developmental level; now state whether mean valency correlates with the developmental level.

References

None

3.7. Meaning specificity and compounding

Problem

Find the relationship between meaning specificity of a noun and its compounding propensity.

Procedure

Take a random sample of 100 nouns from the greatest dictionary of your language. For each noun set up its definition chain, i.e. the sequence of hypernyms

up to the most general one. For example (simplified): *desk – furniture – object – entity*. Since *desk* is the fourth member of a definition chain, it obtains degree 4 of specificity (*x*). Omit circular definitions. Or you can extract data from WordNet.

Now, for each noun find the number of compounds of which it is part. Then for each specificity degree compute the mean number of compounds (*y*). Hence *y* represents the compounding propensity of the given specificity degree *x*. You obtain a sequence which is most probably not monotonous. Find empirically an appropriate function capturing this dependence.

Then test the adequateness of the Fan-Köhler-Altmann approach (2008) who proposed the beta-function on theoretical grounds:

$$y = Cx^a(M - x)^b$$

Compare the parameters you obtained for your language with those obtained for English. If possible, perform the analysis for several languages and study the change of parameters. Find the factor responsible for this change, i.e. continue developing the theory.

Reference

Fan, F., Köhler, R., Altmann, G. (2008). Compounding and meaning generality. Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe: 438-443*. Bratislava: Veda.

4. Synergetics

4.1. Arens' Law

Problem

Arens' Law states that the mean length of words is a power function of the mean length of sentences. Test the hypothesis using different sorts of texts. The resulting function should be monotonically increasing, i.e. the longer the sentence, the longer are its words (on the average). Obtain the respective power functions. Explain why this relationship exists without recourse to Menzerath's law.

Procedure

Up to now all computations were performed on measurements of sentence length in terms of the number of words (cf. Arens 1965; Altmann 1983; Grzybek 2010; Grzybek et al. 2007, 2008). However, word is not the immediate constituent of sentence and it is not very stable. It has an enormous variability and one needs very long texts in order to obtain reliable results. Therefore, perform two computations: first count sentence length in terms of words, then in terms of clauses, and compare the results.

(1) Take all sentences of length 1 (counted in terms of words) and compute the mean word length in them (in terms of the number of syllable). Continue with sentences of length 2, etc. At last, you obtain empirical data which should follow the relationship $\overline{WL} = aS^b$, i.e. mean word length is a power function of sentence length.

Take texts from different genres and languages and test the adequacy of the formula in different circumstances. How do the parameters change in different text sorts and languages?

(2) Measure the sentence length in terms of number of clauses and perform the same measurements of word lengths as above. Did you obtain different results? Using clauses you obtain more reliable length classes. Using words for sentence length measurement you must perhaps pool different classes.

Some authors consider the measurement of sentence length in terms of number op clauses as an expression of *sentence complexity* (cf. Levitsky, Melnyk 2011).

References

Altmann, G. (1983). H. Arens' "Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, M., Harweg, R., Lehfeldt, W., Wienold, G. (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: 31-39.* Tübingen: Narr.

Arens, H. (1965). *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute*. Düsseldorf: Schwann.

Fucks, W. (1955). Unterschied des Prosastils von Dichtern und Schriftstellern. Ein Beispiel mathematischer Analyse. *Sprachforum 1, 234-241*.

Grzybek, P. (2010). Text difficulty and the Arens-Altmann law. In: Grzybek, P., Kelih, E., Mačutek, J. (Eds.), *Text and Language. Structures • Functions • Interrelations. Quantitative Perspectives: 57-70*. Wien: Praesens.

Grzybek, P., Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length: an intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121*.

Grzybek, P., Stadlober, E. (2007). Do we have problems with Arens law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 205-217*. Berlin: de Gruyter.

Grzybek, P., Stadlober, E., Kelih, E. (2007). The relationship of word length and sentence length. The inter-textual perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis: 611-618. Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8-10, 2006*. Berlin-Heidelberg: Springer. [Studies in classification, data analysis and knowledge organization].

Levitsky, V.V., Melnyk, Y.P. (2011). Sentence length and sentence structure in English prose. *Glottometrics 21,* (in print).

4.2. Frequency and polytexty

Hypothesis

The more frequent a word is, the greater is its polytexty. Test the hypothesis in different ways.

Procedure

Polytexty is (a) the number of different texts in which an entity occurs, (b) the number of different environments of an entity, e.g. in a corpus.

Some aspects of "argument structure", valency, transitivity, etc. are special cases of this problem because they consider only the structural type of the environment.

(1) Select 100 verbs from a frequency dictionary. If possible, there should also be groups of verbs having the same frequency in order to form averages (because of greater reliability). For each verb state the number of different constructions in which it may occur. Two

constructions are different only on the basis of the a priori definitions of the properties of the environment which may be formal or semantic. For example, properties may be constant or ephemeral (*be tall – be drunken*), latent or manifest, physical or psychical, kinds of motion, kinds of grammatical categories, etc. Define exactly the types of environment, obtain data and compute the dependence which ought to have the power form or even be exponential. Substantiate the resulting formula linguistically.

(2) Select 100 verbs from a frequency dictionary of your mother language. Then, for each verb, write all prepositions with which it can be combined. Check your competence by sampling from a corpus. State whether the hypothesis holds and if so, find the form of the adequate function. Substantiate it linguistically.

References

Aarts, J., Aarts, F. (1995). Find and want: a corpus-based case study in verb complementation. In: Aarts, B., Meyer, C.F. (eds.), *The verb in contemporary English: 159-182.* Cambridge: Cambridge University Press.

Alsina, A. (1996). *The role of argument structure in grammar: evidence from Romance*. Stanford: CSLI Publications.

Durie, M. (1988). Preferred argument structure in an active language. *Lingua 74, 1-25.*

de Groot, C. (1989). *Predicate structure in a functional grammar of Hungarian.* Dordrecht: ICG Printing.

Hengeveld, K. (1992). *Non-verbal predication: theory, typology, diachrony*. Berlin: Mouton.

Hopper, P.J., Thompson, S.A. (1980). Transitivity in grammar and discourse. *Language 56, 251-299.*

Kärkkäinen, E. (1996). Preferred argument structure and subject role in American English conversational discourse. *Journal of Pragmatics 25(5), 675-701.*

Langacker, R. (1988). The nature of grammatical valence. In: Rudzyka-Ostym, B. (ed.), *Topics in cognitive linguistics: 91-125*. Amsterdam: Benjamins.

Noonan, M. (1985). Complementation. In: Shopen, T. (ed.), *Language typology and syntactic description, Vol II, 42-139.* Cambridge: Cambridge University Press.

Thompson, A.A., Hopper, Paul, J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 27-60.* Amsterdam-Philadelphia: Benjamins.

4.3. Frequency and noun generality

Hypothesis

The greater the frequency of a noun, the more general is its meaning. Test the hypothesis.

Procedure

First define the procedure of measuring the generality of a noun (cf. e.g. Sambor, Hammerl 1991; Köhler, Altmann 2009, 55).

Then select 100 nouns from a frequency dictionary or from a large general corpus. Measure their degree of generality (cf. Strauss, Fan, Altmann 2008, Chapter *Abstractness*; Köhler, Altmann 2009, Chapter 4.9) or, alternatively, their degree of abstractness which is a different property (cf. DeVito 1967; Flesch 1950; Gillie 1957; Kisro-Völker 1984). Prepare the empirical numerical relationship $G = f(Fr)$, taking average generalities for the nouns of the same frequency. Show that this is a simple proportionality relation yielding a power function. Substantiate the relationship linguistically.

References

DeVito, J.A. (1967). Levels of abstraction in spoken and written language. *Journal of Communication 17, 354-361.*

Flesch, R. (1950). Measuring the level of abstraction. *Journal of Applied Psychology 34, 384-390.*

Gillie, P.J. (1957). A simplified formula for measuring abstraction in writing. *Journal of Applied Psychology 41, 214-217.*

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. In: Boy, J., Köhler, R. (eds.), *Glottometrika 12, 139-151.* Bochum: Brockmeyer.

Köhler, R., Altmann, G. (2009). *Problems in quantitative linguistics Vol 2.* Lüdenscheid: RAM.

Paivio, A. (1966). Latency of verbal associations and imagery to noun stimuli as a function of abstractness and generality. *Canadian Journal of Psychology 20, 378-387.*

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

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in quantitative linguistics, Vol 1.* Lüdenscheid: RAM.

Wippich, W., Bredenkamp, J. (1977). Bestimmung der Bildhaftigkeit, Konkretheit und der Bedeutungshaltigkeit von 498 Verben und 400 Adjektiven. *Zs. für experimentelle and angewandte Psychologie 24, 671-680.*

4.4. Increase of word length

Problem

According to Kelih (2010) the mean word length increases from beginning to the end of a text. This is caused by the fact that the theme is explicated at the beginning with familiar words, later on the information must be more detailed and one adds more infrequent words (e.g. hapax legomena) which are automatically longer. Test the hypothesis using different sorts of text in different languages. [Cf. also Problem 4.5]

Procedure

First operationalize the measurement procedure, e.g. take steps of 20 sentences and measure the mean word-length in each of these groups. Or take steps of 100 words. Or skip all synsemantics (articles, prepositions, postpositions, conjunctions, pronouns) and take into account only autosemantics, deciding according to the given language.

Then compute the means of word-lengths and draw a figure of the points in Cartesian coordinates. If you do not obtain an increasing trend, change the grouping of words, e.g. take steps of 25 sentences or 150 words, etc. Possibly you obtain a sequence which decreases at the beginning and after a minimum it begins to increase.

Whatever the result, find a function capturing the given trend. At last, substantiate the function linguistically and derive it from some linguistic assumptions.

Use different texts, not only one. Short texts are not adequate for testing. Begin with a scientific text in which the above mentioned trend can be expected if the hypothesis is "correct". Do not consider poetic texts in which word lengths are mostly stable.

In any case, analyze a strongly synthetic and a strongly analytic language.

References

Kelih, E. (2010). Ist die Wortlänge abhängig von der Textlänge? Befunde aus russischen und bulgarischen Paralleltexten. (*in print*)

4.5. Text length vs. word length

Hypothesis

The longer a text, the longer the words may become, on average.

It may be a simple consequence of the fact that a topic/theme is described ever more precisely. New information comes in only if new words are introduced which specify the theme. But specification can be achieved especially by word prolongation (derivation, composition, reduplication), hence word length must increase as the text increases.

Test the hypothesis using texts of different sort in different languages. (Cf. also *Problem 4.4*)

Procedure

There are some contrary arguments against the hypothesis: (a) A text consists of sentences and sentences of clauses. Counting words one skips several levels between text and word, (b) languages may be monosyllabic, i.e. the condition that word length is a variable is not fulfilled. (c) Text sorts may differ and display different or even no tendency and (d) the morphological/syntactic structure of language may or may not support this phenomenon. Hence, all these boundary conditions must be taken into account.

This hypothesis automatically leads to another hypothesis, namely that average word-length increases from the beginning to the end of the text, presented in *Problem 4.4.* For testing, however, groups of sentences must be pooled (cf. Grzybek, Stadlober 2007), otherwise the data would oscillate extremely – even if the hypothesis could be corroborated.

Several further consequences have been stated in Grzybek, Kelih, Stadlober (2008) signaling that research in this direction has just begun.

References

Grzybek, P., Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length: an intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121.*

Grzybek, P., Stadlober, E. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 205-218*. Berlin: Mouton de Gruyter.

4.6. Word length and phoneme inventory

Problem

Based on very small samples, skipping all other factors known from language synergetics (Köhler 1986, 1987, 1993, 2005) and measuring word length in terms of phoneme numbers, Nettle (1995, 1998) tried to show that mean word length

depends on the size of phoneme inventory. The dependence can be expressed by the power law $L = aI^{-b}$ His results are presented in Table 4.6.1.

Table 4.6.1
Word length and phoneme inventory for 20 languages
(Nettle 1995, 1998)

Language	Phoneme inventory	Mean word length
Hawaiian	18	7.08
Nahuatl	23	8.69
Turkish	28	6.44
Italian	30	7.00
Fula	33	6.42
Georgian	34	7.74
Hausa	35	5.68
Tamasheq	36	5.26
Hindi	41	5.57
German	41	6.44
Songhai	42	4.96
Bambara	49	4.86
Ngizim	52	5.32
Mandarin	53	5.40
Edo	53	4.42
Igbo	58	4.62
Mende	71	4.70
Thai	76	3.65
Ewe	81	4.16
!Xũ	119	4.02
Vata	164	4.56
Vute	195	3.94

Show that the hypothesis does not hold in the form of the power law and correct it.

Procedure

First show that the power law is not adequate. Then obtain random samples from different languages (at least 1000 words) using standard dictionaries. Measure word length in terms of syllable numbers. State the size of the phoneme inventories according to the present state of the art. Compute the mean word length and set up a table of inventory vs. mean word length, as shown above. Find a function fitting the data with a determination coefficient at least $R = 0.90$. If the function is not adequate, add data, add a second independent variable and repeat the testing. Continue until you obtain positive results.

References

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

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

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.

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

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics 5(3), 240-245.*

4.7. Word length distributions

Problem

In Table 4.7.1, word length distributions in 152 Slovenian texts taken from Antić, Kelih, Grzybek (2005) are presented. The length is measured in terms of numbers of syllables taking into account also zero-syllable words (prepositions *k, s*) usual in Slavic languages.

(1) Find a common simple distribution for all data, or, if not possible, reduce the number of models or choose a general one.

(2) Plot the data in an Ord´s scheme (cf. *Problems Vol. 1*: 111f.) and interpret it.

(3) Study either the parameters of the distribution(s) or the location in Ord's scheme in connection with the year of origin of the given texts. If there is any, express it at least verbally.

Table 4.7.1
Word length distributions in 152 Slovenian texts
(Antić, Kelih, Grzybek 2005)

Text	0	1	2	3	4	5	6	7	8	9	Year
1	11	266	194	93	35	3					1907
2	8	478	325	130	33	3					1907
3	9	507	315	164	37	6					1907
4	6	376	250	131	32	1					1907
5	6	381	280	119	19	3	1				1907
6	8	434	237	151	50	10					1907

7	16	441	306	157	46	7					1907
8	26	672	449	270	52	4					1907
9	17	423	288	169	37	5					1907
10	13	560	336	181	39	5					1907
11	12	441	269	165	49	1					1907
12	12	566	339	213	71	2					1907
13	25	726	477	283	61	11					1907
14	14	466	265	156	48	7					1907
15	12	645	423	230	69	9					1907
16	15	573	361	185	58	10	1				1907
17	17	585	340	188	67	6					1907
18	7	136	94	46	16	4					1907
19	43	1126	944	500	197	22	3	1			1904
20	42	1099	872	527	203	29	2	1			1904
21	31	1397	1057	579	180	23	4				1904
22	40	1669	1104	581	174	18	2				1904
23	62	1961	1444	780	252	43	5				1904
24	63	1675	1223	592	180	25	3				1904
25	36	1326	895	573	218	37	5				1904
26	48	1472	1005	497	165	25	8				1904
27	24	1131	832	439	168	17	5				1904
28	23	581	477	255	96	15	1				1920
29	41	1993	1524	658	208	22	6				1920
30	8	452	313	130	59	14	2				1903
31	38	1386	886	474	143	15	2				1903
32	28	1460	918	389	101	6					1903
33	18	1424	924	406	99	19					1903
34	42	1540	1131	554	162	26	3				1903
35	11	474	353	214	51	10	1	1			1903
36	12	430	272	150	49	9					1903
37	15	508	315	210	46	7					1903
38	16	336	226	118	32	4					1903
39	13	434	288	177	62	8	2				1903
40	8	302	216	128	32	7	1				1903
41	24	1067	695	404	148	20	2	1			2001
42	15	692	462	289	102	17	1				2001
43	20	691	446	270	76	9	1				2001
44	15	643	399	278	115	21	2				2001
45	24	890	615	360	110	21	3				2001
46	10	400	271	182	53	10					2001
47	18	1021	744	433	159	29	2				1892
48	45	2102	1599	805	340	47	6				1892

49	37	1724	1282	784	283	45	2				1892
50	97	3101	2398	1327	473	68	13				1892
51	57	2117	1589	917	327	56	11	1			1892
52	60	2381	1770	974	344	50	8	1			1892
53	1	72	62	34	2						1888
54	119	66	33	7	3						1877
55	50	39	11	1							1901
56	27	38	16								1882
57	4	75	48	22	5						1880
58	21	14	9	4							1882
59	35	27	7								1882
60	3	58	42	18	3						1901
61	2	77	63	42	4						1882
62	26	5	6								1908
63	42	34	5								1902
64	26	25	10	1							1872
65	2	98	44	16	6						1879
66	29	29	10	1							1864
67	27	26	15								1908
68	106	63	21	3							1880
69	2	59	37	24	1						1882
70	1	29	33	6	2						1882
71	1	45	55	8	1						1879
72	1	104	84	35	2						1878
73		99	50	12	6						1879
74	14	278	226	125	9	2					1882
75	1	78	47	14	2						1882
76		65	49	15	2						1881
77	1	50	48	21							1882
78		65	49	14	1						1882
79		25	23	10	1						1872
80	1	104	91	45	6						1882
81	1	48	34	11	3						1882
82		30	33	4	3						1882
83	2	103	69	23	1						1882
84		107	63	11							1979
85	3	151	117	59	6						1878
86	4	123	89	32	3	1					1878
87		40	41	12	1						1870
88	1	57	53	24	2						1871
89		22	23	5							1879
90	1	61	49	25	2						1876

91	1	131	93	28	4						1879
92	1	81	58	34	3						1878
93	2	64	55	32	3						1806
94	1	55	84	19	7						1806
95		23	27	9	1						1795
96	1	53	62	10	1						1798
97		36	25	10	1						1790
98		10	6	6	1						1795
99	2	115	85	54	10	1					1811
100		38	31	17	1						1810
101		69	67	19	3						1888
102	2	187	145	70	7	2					1888
103		144	117	37	7	1					1888
104	10	267	167	145	96	32	5	2			2001
105	9	167	127	122	68	20	6				2001
106	34	699	484	443	210	75	15	5	1		2001
107	6	278	236	163	77	15	5	1			2001
108	4	142	82	99	38	19	6				2001
109	5	124	76	82	25	6	1				2001
110	5	170	113	120	54	27	5	1			2001
111	9	132	94	110	72	23	5	5			2001
112	9	220	155	134	60	12	2	1			2001
113	22	564	359	368	209	52	5	3			2001
114	15	280	201	174	87	38	5				2001
115	2	121	69	90	45	9	5	1	1		2001
116	6	259	185	147	58	27	4	1			2001
117	17	179	139	128	94	26	5	2			2001
118	7	87	81	95	36	7	5	1			2001
119	6	362	256	182	99	30	7				2001
120	5	326	269	216	134	24		7			2001
121	2	54	47	26	13	1					2001
122	3	187	108	85	62	10	8				2001
123	4	103	61	76	29	14	7	1			2001
124	3	178	112	77	46	23	1	1	0		2001
125	2	97	60	48	31	13	3	2			2001
126	16	254	174	191	127	19	9	3			2001
127	11	295	200	201	80	41	8	1			2001
128	5	74	80	43	17	3	2				2001
129	1	65	61	33	30	10	3				2001
130	11	150	118	86	49	17	1		1		2001
131	8	164	89	88	37	11	2	2	1		2001
132	6	227	137	149	62	27	2	2			2001

133	6	156	104	79	51	14	2				2001
134	9	170	103	68	43	8	2	3			2001
135	16	202	174	174	98	26	4	4			2001
136	9	141	121	105	57	10	2	3			2001
137	9	148	104	96	54	22	5	1			2001
138	3	66	50	45	24	4	2				2001
139	1	54	38	39	25	11	1	1			2001
140	4	71	38	43	45	18					2001
141	3	108	94	84	31	13	1	1	2		2001
142	1	54	30	32	19	2	1				2001
143	3	95	52	58	21	7	3				2001
144	4	86	49	50	22	5	2				2001
145	5	101	72	90	40	16	4	2			2001
146	9	307	200	189	90	35	4	2			2001
147	3	42	23	24	16	9					2001
148	3	107	73	61	39	19					2001
149	1	69	36	53	32	7	2			1	2001
150	2	73	49	52	16	9	2				2001
151	2	52	41	40	27	2					2001
152	3	46	33	49	20	6	3		2		2001

Procedure

Proceed inductively: use software for fitting different distributions to data (e.g. *Fitter*). For each text note all distributions fitting well and note the first three moments (mean, variance, third central moment). At last, choose the minimal number of distributions covering all data. The distributions should belong to the same family (= a more general distribution). Order the texts according to the given distribution and note its parameters. Compute Ord´s coordinates <I,S> for each text and plot them.

If you did not succeed in finding a family of distributions, skip simply the values of length 0. The interpretation of zero-syllabic prepositions in Slavic languages is not definitively solved: one can consider them as independent words or as clitics. Perform the fitting again and repeat all previous steps.

Are all texts in Ord´s scheme placed on a straight line? If not, which texts deviate? Compute the straight line. Does a parameter of the common distribution correlate with the year of text origin? If so, show the relation of the distribution (= some of its parameters) with the year. Does mean word length increase in time? Cf. also *Problems Vol. 1: 68*.

Some data are bimodal. Consult the problem with a specialist in Slovenian. Use a mixed, compound or generalized distribution if necessary. If nothing helps, set up a new distribution using a difference equation of second order.

References

Antić, G., Kelih E., Grzybek, P. (2005). Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language. Word length studies and related issues: 117-156.* Dordrecht: Kluwer.

4.8. Full valency and the frequency of verbs

Hypothesis

"The more frequent the verb, the more full valency frames [it has]" (Čech, Pajas, Mačutek 2010, 295).

The relationship between frequency and full valency is proposed analogically to the relationship between valency and frequency:

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

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

The idea is clear: a more frequent verb occurs in more contexts, so it seems reasonable to expect that it should have more full valency frames.

Test the hypothesis.

Procedure

Follow the procedure presented in Problem 2.16. *The distribution of full valency frames*. Group verbs with the same full valency frames x and compute the mean frequency of verbs (y) in each group. Observe the dependence between the number of full valency frames x and the mean frequency y. Suggest the form of dependence and test it. Compare it with the result presented in Čech, Pajas, Mačutek (2010).

References

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

Čech, R., Pajas, P., Mačutek, J. (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics 17(4), 291-302.*

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

4.9. Full valency and the length of verbs

Hypothesis

"The shorter the verb, the more full valency frames it has" (Čech, Pajas, Mačutek 2010, 295).

A relationship between the length of the verb and the number of full valency frames of the verb should be the consequence of the relationship between frequency and length. It has been shown that length is a function of frequency (Köhler 1986; Popescu et al. 2009); the hypothesis is therefore based on the following idea: the more frequent a verb, the shorter it is, and consequently the shorter a verb, the more full valency frames it has.

Test the hypothesis.

Procedure

Follow the procedure presented in Problem *2.16. The distribution of full valency frames*. Compute the mean length (*l*) of verbs/lemmas having *x* full valency frames in a corpus. Measure the length of verbs either in terms of syllable numbers or morpheme numbers. Observe the dependence between the number *x* of full valency frames (not their frequency) and the mean length (*l*) of verbs/lemmas which have *x* full valency frames.

Suggest the form of dependence and test it. Compare it with the result presented in Čech, Pajas, Mačutek (2010).

References

Čech, R., Pajas, P., Mačutek, J. (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics, 17(4), 291-302.*

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

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: de Gruyter.

4.10. Full valency and polysemy

Hypothesis

The more full valency frames the verb has, the greater the polysemy of the verb. Test the hypothesis.

Procedure

Follow the procedure presented in Problem *2.16. The distribution of full valency frames* and assign to each verb/lemma the number of its full valency frames *x* (not frequencies) in a corpus.

Determine the number of meanings (*m*) of each verb/lemma using e.g. WordNet (see http://www.globalwordnet.org/) or a monolingual dictionary. Describe explicitly the way you obtained the given number of meanings.

Prepare a table containing the number of full valency frames and the number of meanings of individual verbs and observe the dependence between these two language properties. Suggest the form of dependence and test it. Interpret the results in the framework of synergetic linguistics (Köhler 1986, 2005), i.e. substantiate it linguistically.

If the result is positive, find its place in the Köhlerian control cycle and prepare a flow diagram. Then derive the result from the general theory (Wimmer, Altmann 2005) and interpret it.

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 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: 791-805*. Berlin-New York: de Gruyter.

4.11. Full valency and synonymy

Hypothesis

The more full valency frames the verb has, the richer is the synonymy of the verb. Test the hypothesis.

Procedure

Follow the procedure presented in Problem *2.16. The distribution of full valency frames* and assign to each verb/lemma the number of its full valency frames *x* (not frequencies).

Determine the number of synonyms (*s*) of each verb/lemma; use WordNet (see http://www.globalwordnet.org/) or a dictionary of synonyms in the given language.

By analogy with Problem *4.10 Full valency and polysemy* observe the dependence between the number of full valency frames (*x*) and the number of synonyms (*s*) of individual verb. Suggest the form of dependence and test it. Interpret the results in the framework of synergetic linguistics (Köhler 1986, 2005).

If the result is positive, incorporate the relationship in Köhler´s control cycle and prepare a flow diagram. Then derive the result from the general theory (Wimmer, Altmann 2005) and interpret it in terms of speaker and hearer impact. Since the relation between polysemy and synonymy is known, full valency must be only a new element in the control cycle but should not disturb it. If you do not want to tae into account the complete Köhlerian cycle, set up your own reduced one in which only the relationships of valency are contained.

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 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: 791-805*. Berlin-New York: de Gruyter.

4.12. Full valency and compounding

Hypothesis

The more full valency frames the verb has, the more compounds it produces. Test the hypothesis.

Procedure

Follow the procedure presented in Problem *2.16. The distribution of full valency frames* and assign to each verb/lemma the number of its full valency frames *x* (not frequencies) in a corpus.

Take a random sample of at least 300 verbs/lemmas and for individual verb/lemma observe the number of compounds (*c*) which are formed by the given verb/lemma. Use a monolingual dictionary or corpus. Group verbs/lemmas with the identical number of full valency frames and for each group compute the mean number of compounds (*mc*). Find a function expressing the dependence of compositionality and full valency: $mc = f(x)$. If you succeed, give a justification for the given function. If possible, set up the differential equation which leads to the given function and substantiate it linguistically. Chart your result in form of a simple 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 (Dec. 21, 2010)

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.

4.13. Full valency and derivation

Hypothesis

The more full valency frames the verb has, the greater its productivity, i.e., the more derivatives are formed from it. Test the hypothesis.

Procedure

Follow the procedure presented in Problem *2.16. The distribution of full valency frames* and assign to each verb/lemma the number of its full valency frames *x* (not frequencies) in a corpus.

Take a random sample of at least 200 simple, i.e. non-derived, verbs and for each individual verb/lemma observe the number of its derivatives (*d*). Define precisely the character of derivation. Use a monolingual dictionary or corpus.

Compute the correlation between the number of full valency frames *x* and derivation values (*d*). If there is a correlative relationship, find a function expressing the dependence. Interpret the results in the framework of synergetic linguistics (Köhler 1986, 2002, 2005).

Show that one of the parameters depends on the extent of derivation building in language. Test this assumption using a strongly analytic and a strongly synthetic language.

Show that there is a strong dependence between the given parameter and the derivativeness in language.

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 (Dec. 21, 2010).

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.

4.14. Full valency and Menzerath's law

Hypothesis

The more arguments the verb has in its full valency frame (counted by the number of words), the shorter are the arguments (counted by the number of syllables

or morphemes). Test the hypothesis.

Procedure

Follow the procedure presented in Problem *2.16. The distribution of full valency frames* and compute the number of arguments *x* for each verb (token) in the text (or treebank). Group the verbs (tokens) with the same number of arguments *x* and count for each verb (token) in the group the mean length of its arguments (counted by the number of syllables or morphemes). Then compute the mean length of arguments for each group.

Check whether the dependence between the number of arguments in full valency frame and the length of arguments has the form of a power function. If not, find another well fitting function and substantiate it linguistically.

References

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika 2, 1–10.*

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

4.15. Syntactic network analysis

Problem

Although the network theory (e.g., Barabási, Albert 1999; Newman 2003) has many applications in language sciences (cf. *Bibliography on Linguistic and Cognitive Networks*), up to now the majority of language network analyses have been merely descriptive, focused on global network characterisation, and there was an absence of linguistic explanation of language-based networks (cf. Ferrer i Cancho 2010; Mehler 2007).

For syntax, despite there being numerous achievements, some fundamental problems remain unsolved: although statistical properties typical for complex networks can be observed in all syntactic networks, the impact of syntax itself on these properties is still unclear (Liu, Hu 2008; Liu, Zhao, Huang 2010).

Test the following hypotheses and try to substantiate some syntactic network properties linguistically. All network properties which are taken as a variable in the hypotheses (out-degree, in-degree, hub-weight, authority-weight, betweenness centrality) express some kind of "global importance" of the node in the network; for more information of these properties see for example Caldarelli (2007) and Newman (2010). The testing of these hypotheses should reveal which

of these properties are (if ever) associated with linguistically well established language properties (word length, synonymy, and polysemy) (cf. Köhler 2005).

Hypotheses

1. The higher the out-degree of the word (or lemma), the shorter the word.
2. The higher the out-degree of the word (or lemma), the more synonyms it has.
3. The higher the out-degree of the word (or lemma), the more meanings it has.
4. The higher the in-degree of the word (or lemma), the shorter the word.
5. The higher the in-degree of the word (or lemma), the more synonyms it has.
6. The higher the in-degree of the word (or lemma), the more polysemic the word is.
7. The greater the hub-weight of the word (or lemma), the shorter the word.
8. The greater the hub-weight of the word (or lemma), the more synonyms it has.
9. The greater the hub-weight of the word (or lemma), the more polysemic the word is.
10. The greater the authority-weight of the word (or lemma), the shorter the word is.
11. The greater the authority-weight of the word (or lemma), the more synonyms it has.
12. The greater the authority-weight of the word (or lemma), the more polysemic the word is.
13. The higher the betweenness centrality of the word (or lemma), the shorter the word.
14. The higher the betweenness centrality of the word (or lemma), the more synonyms it has.
15. The higher the betweenness centrality of the word (or lemma), the more polysemic the word is.

Procedure

Use a syntactically annotated corpus, for example the *Alpino treebank* (Beek et al. 2001), the *CESS-ECE corpus* (Martí et al. 2007), the *Floresta synta(c)tica* (Alfonso et al. 2002), the *Italian Syntactic-Semantic Treebank* (Montemagni et al. 2003), the *Prague Dependency Treebank* (Hajič et al. 2006), the *Szeged treebank* (Csendes et al. 2005) and construct a syntactic complex network. In constructing the network, follow the method described in Ferrer i Cancho et al. (2004) and Liu (2008). For creating the network and computing its properties it is possible to use the free software *Pajek 2.03* (available at http://pajek.imfm.si/doku.php?id=pajek).

Construct a directed network in which each node of the network represents a word (or lemma). Two nodes are linked in the network, if there is a syntactic relationship (most usually syntactic dependency) between words (or lemmas) in the corpus. The direction of the links depends on the adopted syntactic formalism (Hudson 2007).

Compute out-degree, in-degree, hub-weight, authority-weight, betweenness centrality of each word (or lemma). Count the length (*l*) of each word in terms of syllable or morpheme numbers, determine the number of synonyms (*s*), and the number of meanings (*m*) of each word (or lemma). Test the hypotheses.

References

Alfonso, S., Bick, E., Haber, R., Santos, D. (2002). *"Floresta sintá(c)tica": a treebank for Portuguese.* In: M.G. Rodríguez, C.P.S. Araujo (eds.), *Proceedings of LREC 2002, 1698-1703.*

Barabási, A.L., Albert, R. (1999). Emergence of Scaling in Random Networks. *Science 286.5439: 509-512.*

Beek van der, L., Bouma, G., Malouf, R., Noord, G. van (2002). The Alpino Dependency Treebank. In: *Computational Linguistics in the Netherlands CLIN 2001, Rodopi, 2001, 8–22.*

Bibliography on Linguistic and Cognitive Networks. http://www.lsi.upc.edu/~rferrericancho/linguistic_and_cognitive_networks.html.

Caldarelli, G.. (2007). *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford: Oxford University Press.

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A. (2005). The Szeged Treebank. In: Matoušek et al. (eds.) *Text, Speech and Dialog. 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings: 123-131*. Berlin-Heidelberg-New York: Springer.

Ferrer i Cancho, R., Solé, R.V., Köhler, R. (2004). Patterns in syntactic dependency networks, *Physical Review E 69, 051915.*

Ferrer i Cancho, R. (2010). Network theory. In: Colm Hogan, P. (ed.), *The Cambridge encyclopedia of the language sciences: 555-557.* Cambridge: Cambridge University Press.

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Linguistic Data Consortium, Philadelphia.

Hudson, R. (2007). *Language Networks. The New Word Grammar.* Oxford-New York: Oxford University Press.

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.

Liu, H. (2008). The complexity of Chinese syntactic dependency networks, *Physica A 387, 3048–3058.*

Liu, H., Hu, F. (2008). What role does syntax play in a language network? *Europhysics Letters 83, 18002.*

Liu, H., Zhao, Y., Huang, W. (2010). How do local syntactic structures influence global properties in language networks? *Glottometrics 20, 35-39.*

Martí, M.A., Taulé, M., Marquez, L., Bertran, M. (2007). *CESS-ECE: A multilingual and multilevel annotated corpus*. Available for download from: http://www.lsi.upc.edu/mbertran/cess-ece/publications

Mehler, A. (2007). Large text networks as an object of corpus linguistics studies. In Lüdeling, A., Kytö, M. (eds.) *Corpus linguistics. An international handbook of the science of language and society: 328-382.* Berlin/ New York: de Gruyter.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A. Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R. Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Nana, N., Pianesi, F., Delmonte, R. (2003). Building the Italian Syntactic-Semantic Treebank. In: Abeillé, A. (ed.), *Treebanks: Building and Using Parsed Corpora: 189-210.* Dordrecht: Kluwer.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review 45(2), 167-256.*

Newman, M. (2010). *Networks: An Introduction.* Oxford: Oxford University Press.

4.16. Typology

Problem

Construct a “typological“ control cycle of linguistic properties.

Procedure

Collect as many typological indicators as possible and show their mutual relations. Solve the relation by capturing it with a function and strive for a general theory, i.e. show that isolated properties do not exist; that at least some of them have the character of laws and that all relationships can be derived form a general theory.

Use some properties described in Altmann, Lehfeldt (1973) or Hoffmann (2005) such as effectivity of the system of distinctive features, aspects of vocalicity of the language, entropy of the phoneme system, properties of phoneme distributions, positional exploitation of phonemes, mean word length, synthetism/

analytism, sentence length, sentence depth, sentence centrality, sentence width and add other properties from the literature. Strive for a large list of quantified properties. Do not use dichotomous variables but quantify structures resulting from their combinations. Do not use universals – but if you use them, quantify them.

For every indicator derive its asymptotic variance and test some differences if necessary.

Do not classify languages. Strive for a theoretical construct.

References

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

Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology.* 2nd edition. Chicago: University Of Chicago Press.

Cysouw, M. (2005). Quantitative methods in typology. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 554-578.* Berlin-New York: de Gruyter.

Givón, T. (ed.) 1983. *Topic continuity in discourse: a quantitative cross-language study.* Amsterdam: Benjamins.

Greenberg, J.H. (1990). A quantitative approach to the morphological typology of language. In: Denning, K., Kemmer, S. (eds.), *On Language: selected writings of Joseph H. Greenberg: 3-25.* Stanford, California: Stanford University Pess.

Hoffmann, C. (2005). Morphologisch orientierte Typologie. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 578-598*. Berlin-New York: de Gruyter.

Justeson, J., Stephens, L.D. (1984). On the relationship between the numbers of vowels and consonants in phnological systems. *Linguistics 22, 531-545*.

Kasevič, V.B., Jachontov, E.S. (eds.) (1982). *Kvantitativnaja tipologija jazykov Azii i Afriki.* Leningrad: Izdateľstvo Leningradskogo Universiteta.

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

Krupa, V., Altmann, G. (1966). Relations between typological indices. *Linguistics 24, 29-37.*

Maslova, E. (2000a). A dynamic approach to the verification of distributional universals. *Linguistic Typology 4(3), 307-333.*

Maslova, E. (2000b). Stochastic models in typology: obstacle or prerequisite? *Linguistic Typology 4(3), 35-364*

Myhill, J. (1992). *Typological discourse analysis*. Oxford: Blackwell.

Nichols, J. (1992). *Linguistic diversity in space and time.* Chicago: Chicago University Press.

Sapir, E. (1921). *Language*. New York: Harcourt, Brace.

Silnickij, G.G. (1998). Correlational analysis of Indo-European morphological systems. *Journal of Quantitative Linguistics 5, 81-95.*

Skalička, V. (1979). *Typologische Studien.* Braunschweig: Vieweg.

Song, J.J. (ed.) (2011). *The Oxford Handbook of Linguistic Typology*. Oxford: Oxford University Press.

5. Text analysis

5.1. Text development

Problem

Texts are created in temporal succession. Even if one can perform corrections *a posteriori*, the deployment concerns both the content and the form. The deployment with regard to content is known e.g. from classical drama, detective stories, etc., but "material" deployment has not been studied frequently. However, there are a number of properties whose deployment can be characteristic for texts. Study the deployment of some properties.

Procedure

Take a text and partition it into its "natural" parts, e.g. chapters, strophes, paragraphs, sentences, which play the role of reference frames. For each part separately compute

(a) the sentence length distribution (in terms of clause numbers),
(b) the rank-frequency distribution of word forms,
(c) word length distribution,
(d) the entropy in (a), (b) and (c) (see Problems Vol. 1, 113),
(e) the repeat rate in (a), (b) and (c) (see Problems Vol 1, 113),
(f) Ord´s criterion in (a), (b) and (c) (see Problems Vol. 1, 111 f.),
(g) the angle of "writer´s view" in (b) (see Popescu, Mačutek, Altmann 2009, 24 ff.),
(h) the lambda-indicator (cf. Popescu, Čech, Altmann 2011a,b),
(i) the arc length (see Mačutek 2009),

and study the course of individual properties. Can you discern a tendency or a regular oscillation or did you obtain irregular oscillations?

Take those properties which exhibit some regularity and study them using different texts. Draw conclusions from the observations, substantiate them textually and capture the tendency at least with an empirical function. Then derive your function on the basis of your conclusions based on linguistic reasoning. If you use difference or differential equations, substantiate your procedure (e.g. the order of the equation) linguistically.

Do not restrict yourself to the above mentioned properties. Examine also other ones (e.g., different indicators of vocabulary richness, see Popescu et al. 2009; Popescu, Čech, Altmann 2011b).

Study the differences between special text sorts (poetry, scientific texts, press texts, etc.) and build the nucleus of a *dynamic stylometry*.

Perform the same examinations in some other languages and compare all your results. Do not omit to use statistical tests or, if necessary, develop ad hoc new ones in order to obtain objective results.

References

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

Mačutek, J. (2009). Arc length development and the highest word frequency. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 182-189*. Chernivtsi: ČNU.

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., Čech, R., Altmann, G. (2011a). *The lambda-structure of texts.* Lüdenscheid: RAM.

Popescu, I.-I., Čech, R., Altmann, G. (2011b). Vocabulary richness in Slovak poetry. (submitted)

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18-42*.

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

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM.

5.2 Cumulative text development

Problem

Some text properties change with text size.

(a) Which properties do?

(b) Find the form of change and analyze it as a kind of time series.

Procedure

Perform the same computing procedures as in the problem "5.1. Text development". However, this time do not compute them separately for each part of the text but after having evaluated the first part of the text add the second part and compute the characteristics from this whole, then add the third part, etc. This is not an addition of numbers from the above mentioned problem but addition of parts and computing anew.

Begin e.g. with hapax legomena, i.e. words occurring only once. Is the development of their number linear, convex increasing or concave increasing?

Where is the point at which no increase of hapax legomena can be observed any more?

Do text sorts differ in this aspect?

What is the dynamics of hapax legomena in a long epical poem whose parts are individual strophes?

What is the dynamics of synsemantics which are present in almost all sentences?

How do references develop in increasing text? For the definition of references refer to the recent literature.

Analyze a greater set of texts of the same sort and work out stepwise the dynamics of text development. Associate classes of entities with certain types of functions and substantiate your findings linguistically, psycho-linguistically and from the communication point of view.

Derive some of the developmental aspects from linguistic assumptions and test the goodness-of-fit of your functions. Begin to form the basis of a theory containing at least one (derived and corroborated) hypothesis. Continue in three directions:

(i) subsume your hypothesis under a more general theory,
(ii) derive some consequences from your hypothesis,
(iii) join your hypothesis with other ones showing a possible connection, i.e. begin to set up a control cycle.

References

See *5.1. Text development* in this volume and the references therein.

Köhler, R. (2006). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774.* Berlin-New York: de Gruyter.

Mačutek, J. (2009). Arc length development and the highest word frequency. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 182-189*. Chernivtsi: ČNU.

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

5.3. Experiments with arc length

Problem

J. Mačutek (2009) has shown that arc development of rank-frequency sequences of word-forms is significantly different in morphologically different languages

and pointed out that this property could be significant also with genres, authors, historical periods within one language. Test some of these hypotheses.

Procedure

Take a text and obtain the rank-frequency sequence of word-forms for the first 500 words and compute the arc length as

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

where L = arc length, r = rank, f_r = frequency at rank r, V = vocabulary. Then add the next 500 word-forms and compute L again. Continue until the end of the text. The individual L-values yield a curve or a straight line.

Then take a text in the same language from another author but in the same genre and perform the same procedure.

Compare the slopes of arc development of the two texts using the Sen-Adichie test described very clearly in Mačutek (2009).

Process many texts and set up classes of texts based on the slope of the cumulative arc length computation. If the regression lines are not parallel, order the regression lines in decreasing order and scrutinize the possible causes, i.e. find some boundary conditions (here, other properties of the deviating or extreme texts) under which some texts display a strong deviation.

State whether it is possible to distinguish (1) individual authors, (2) individual text sorts.

State whether texts of the same sort in one language display a kind of historical development. Compare the development – if there is any – in two different languages.

Compare the results obtained by examining the *same* text in different languages. A corpus based on an identical text in 11 Slavic languages has been set out by E. Kelih (2009, 2009a). Or use texts from the MULTEXT-East project (http://nl.ijs.si/ME/; Erjavec 2010).

Draw conclusions about the shaping of arc length in different styles (which must be defined independently).

Show the extremes of arc length both theoretically and empirically.

Interpret linguistically the behaviour of arc length.

References

Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (2010). *Proceedings of the Seventh conference on International Language*

Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), 2544-2547. (http://www.lrec-conf.org/proceedings/lrec2010/pdf/138_Paper.pdf)

Hollander, M. (1970). A distribution-free test for parallelism. *Journal of the American Statistical Association 65, 387-394.*

Hollander, W., Wolfe, D.A. (1999). *Nonparametric statistical methods.* New York: Wiley.

Kelih, E. (2009). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas´ stal´ (KZS)". In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 106-124*. Černivci: ČNU.

Kelih, E. (2009a). Preliminary analysis of a Slavic parallel corpus. In: Levická, J., Garabík, R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings: 175-183.* Bratislava: Tribun.

Mačutek, J. (2009). Arc length development and the highest word frequency. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 182-189*. Chernivtsi: ČNU.

MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. (http://nl.ijs.si/ME/)

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18-42*.

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

Potthoff, K.F. (1974). A non-parametric test whether two simple regression lines are parallel. *The Annals of Statistics 2, 295-310.*

5.4. The syntactic binary code of sentence

Problem

Compute the binary syntactic codes of all sentences in a text and study the resulting time series.

Procedure

The method has been shown only for German, Russian, and Czech up to now (cf. Altmann, Altmann 2008; Popescu et al. 2010), hence a more detailed description is in order.

(a) Depict the syntactic structure of a sentence in form of a tree following any background formalism (dependency, phrase structure etc.) and grammar, e.g. generative (Chomsky 1995), word grammar (Hudson 2006), Functional Gener-

ative Description (Sgall, Hajičová, Panevová 1986; Hajič et al. 2006). For example, one of the possibilities to depict the syntactic structure of the sentence "Wer reitet so spät durch Nacht und Wind" (the first line from Goethe´s "Erlkönig") can be as presented in Figure 5.4.1

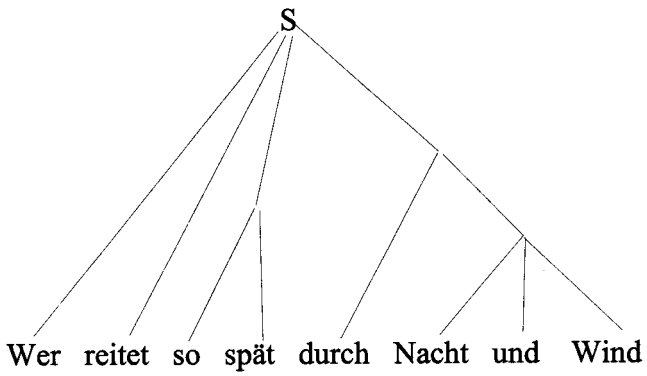


Figure 5.4.1. One of the possibilities

but other structures are possible.

(b) Now, numerate all vertices from top to bottom and from left to right in order to obtain Figure 5.4.2.

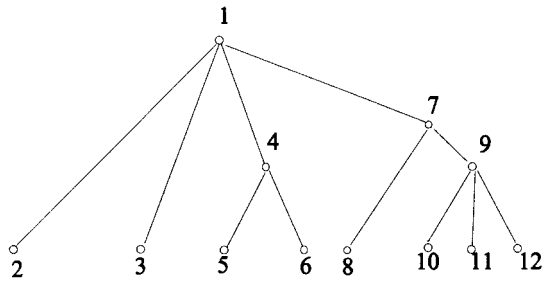


Figure 5.4.2. Sequence and adjacency of vertices

(c) Set up an adjacency matrix of vertices using only the part above diagonal, i.e. the upper triangular matrix. An existing adjacency obtains the value of 1, a non-existing one the value of 0, i.e.

$$(1) \qquad a_{ij} = \begin{cases} 0, & \text{the vertices } i \text{ and } j \text{ are not adjacent} \\ 1, & \text{the vertices } i \text{ and } j \text{ are adjacent (joined with an edge),} \end{cases}$$

For the given sentence analyzed in the given way one obtains the matrix in Table 5.4.1. The diagonal will be ignored.

Table 5.4.1
Adjacency matrix of the graph in Figure 5.4.2

v	1	2	3	4	5	6	7	8	9	10	11	12
1	-	1	1	1	0	0	1	0	0	0	0	0
2		-	0	0	0	0	0	0	0	0	0	0
3			-	0	0	0	0	0	0	0	0	0
4				-	1	1	0	0	0	0	0	0
5					-	0	0	0	0	0	0	0
6						-	0	0	0	0	0	0
7							-	1	1	0	0	0
8								-	0	0	0	0
9									-	1	1	1
10										-	0	0
11											-	0
12												-

The binary code (*BC*) is computed in form of a sum (c.f. Balakrishnan 1997):

(2) $BC = a_{12}2^0 + a_{13}2^1 + \ldots + a_{1n}2^{n-1} + a_{23}2^n + \ldots + a_{2n}2^{2n-3} + \ldots + a_{n-1,n}2^{k-1}$,

where a_{ij} are the weights given by formula (1), $k = n(n\text{-}1)/2$ and the summing begins with the cell (1,2). For the given matrix we obtain

$$BC = 1(2^0) + 1(2^1) + 1(2^2) + 1(2^5) + 1(2^{30}) + 1(2^{31}) + 1(2^{51}) + 1(2^{52}) + \\ +1(2^{60}) + 1(2^{61}) + 1(2^{62}) = 8{,}077{,}205{,}934{,}910{,}210{,}087.$$

In order to normalize this number, one divides it by the maximum which would be attained if all pairs of vertices would be adjacent, i.e.

(3) $$BC_{\max} = \sum_{i=0}^{\frac{n(n-1)}{2}-1} 2^i = 2^{\frac{n(n-1)}{2}} - 1.$$

For $n = 12$ we obtain $BC_{max} = 73{,}786{,}976{,}294{,}838{,}206{,}463$. Hence the relative binary code is given as

(4) $$BC_{rel} = \frac{BC}{BC_{\max}}.$$

In the example it is 8,077,205,934,910,210,087 / 73,786,976,294,838,206,463 = 0.1095.

(d) Compute the relative binary code for every sentence of a text in order to obtain a time series.

(e) Evaluate the properties of the time series obtained.

(f) Perform the above procedure for the same text in different languages. Use the properties of the computed time series to show the differences between languages. Use multilingual corpora (cf. Kelih 2009; Erjavec 2010; MULTEXT-East project, InterCorp).

(g) Use the method for showing genre differences.

(h) Use the method for showing uniformity or variation in works of one author.

(i) Use the method for showing change or stability in the history of a genre (e.g. press texts).

(j) Use other methods for obtaining a tree or dependence and repeat the whole procedure.

(k) If you analyze poetry, you may consider either lines or sentences.

References

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* Lüdenscheid: RAM.

Balakrishnan, V.K. (1997). *Graph theory*. New York: McGraw-Hill.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge: MIT Press.

Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D (eds.),. *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10): 2544-2547*, European Language Resources Association (ELRA). (http://www.lrec-conf.org/proceedings/ lrec2010/ pdf/138_Paper.pdf).

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Hudson, R. (2006). *Language Networks. The New Word Grammar*. Oxford– New York: Oxford University Press.

InterCorp. Project of Parallel Corpora. (http://www.korpus.cz/intercorp/)

MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. (http://nl.ijs.si/ME/)

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text.* Lüdenscheid: RAM.

Sgall, P., Hajičová, E., Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel.

5.5. The binary code of text

Problem

Read Problem *5.4 The binary code of sentence*. Then take a text and join the sentences having something in common (e.g. the same word, synonym, reference etc.). You obtain the same picture as if you joined the dependent words in a sentence. Set up the coincidence matrix and compute the binary code of the text. In poems you can also join the verses.

Procedure

First define the association between sentences in a clear way. If necessary, define other units as frame, e.g. verse, strophe, clause, or different punctuation marks as sentence boundaries (, ; : ? ! .). Enumerate the "sentences" currently and join all containing an "echo" or a reference to other sentences. Do not use oriented arcs. Prepare the text in this way, set up the upper triangle matrix of adjacencies and compute the binary code of the text. The procedure is presented in detail in Popescu et al. (2010, Chapter 8).

The code may be interpreted as a kind of text coherence. Solve the following problems:

(1) Show that simple forms (fairy tales, fables etc.) have greater coherence than complex forms (e.g. scientific texts).
(2) Show the place of different text types on the coherence scale.
(3) Compare the same text in two different languages and show whether they differ in this aspect. Use the variance of the binary code (cf. Popescu et al. 2010), set up an asymptotic test for the difference of two text codes and test whether they are different. Some languages may exploit references to a different extent.
(4) Study the works of one author and examine the development of coherences over the course of years. Is higher age correlated with higher text coherence?
(5) Study the text codes of individual acts of a stage play. Does it change or is it constant? In the second step separate the individual persons and study their isolated parts as wholes.
(6) Study the narrations of children in different ages. School essays are appropriate research objects.

Since the binary sentence or text codes are relative measures, the length of a text is not relevant. Of course, short texts tend to be more coherent because the theme is very concentrated. Therefore

(7) Study the effect of text length on the binary code and show that if it exists, after a certain length it becomes irrelevant.

References

Hřebíček, L. (1993). Text as a strategic process. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis 136-159*. Trier: Wissenschaftlicher Verlag.

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, R..B. (eds.), *Contributions to quantitative linguistics: 33-38*. Dordrecht: Kluwer.

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

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text*. Lüdenscheid: RAM-Verlag.

5.6. The development of writer´s view

Problem

The angle α representing "writer´s view" (cf. Popescu, Mačutek, Altmann 2009, 27ff) grows linearly with increasing text. Popescu, Čech, Altmann (2011) argue that the longer the text is, the more the writer loses his subconscious control over frequency of words and keeps only conscious control over the contents, the grammar, his aim, etc. As soon as parts of control disappear, the text develops its own dynamics and begins to abide by some laws which are not known to the writer but work steadily in the background. Test the above hypothesis.

Procedure

(1) Take a long text and compute first the rank-frequency sequence of word forms in the first 1000 words. Then state the quantities V = vocabulary size (= greatest rank), $f(1)$ = greatest frequency and h = the h-point as shown in the above mentioned reference.

(2) Compute the cosine α of the angle at the h-point.

(3) Add the next 1000 words to the text and repeat steps (1) and (2).

Register cos α at each step and observe its change. The formula for its computing is

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

(4) Study texts representing different text types in different languages. Show that in different languages the angle changes differently. Draw typological consequences.

(5) Take the same text in different languages and perform the whole procedure. Show that even here differences can be found.

If the change of the angle with increasing text length *N* is not linear, find an adequate function capturing all cases.

Substantiate the relationship linguistically.

References

Köhler, R., Altmann, G. (2009). *Problems in quantitative linguistics. Vol. 2*. Lüdenscheid: RAM (esp. Problems 5.10 and 5.11).

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 15, 45-52.*

Popescu, I.-I., Čech, R., Altmann, G. (2011). Some geometric properties of Slovak poetry. (submitted)

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

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts.* Lüdenscheid: RAM-Verlag.

5.7. Hřebíček´s hypothesis

Problem

According to a hypothesis of L. Hřebíček (1997, 2000) frequent entities (e.g. words) appear in text earlier than rare ones. Set up specific hypotheses concerning phonemes (letters, signs), syllables, morphs, morphemes, word-forms, lemmas, herbs; propose respective functions capturing this phenomenon and test them.

Procedure

Take a text and count the frequencies of units at a certain level (phonemes, syllables, morphs, morphemes, word-forms, lemmas, hrebs). Then replace the units by their frequencies and study the sequence of these numbers. Since the oscillation may turn out to be very pronounced, simplify the procedure by taking the mean of the frequencies in each sentence. If the hypothesis is correct, you should obtain some decreasing function (of frequency means in the sentence).

Alternatively, take only the greatest frequency in each sentence and test whether sentence order (x = order number of the sentence) displays a decreasing sequence of maximal frequencies (y).

Propose some other measurement methods but pay attention to the possibility of testing.

Analyze the situation on different levels in different languages Take into account both morphs and morphemes, lemmas and word forms but especially hrebs (cf. Ziegler, Altmann 2002) of different kinds. If you obtain positive results, begin to sketch a "teorita". Give linguistic substantiations, find boundary conditions and search for a family of empirical functions which could capture this phenomenon. If you find only counterexamples, supply arguments in support of the fact that the hypothesis cannot hold. Nevertheless, it may hold for some units but not for other ones.

The investigation is important for the subconscious strategy of communication, text control and text formation. Is there a relationship to the thema-rhema and topic-comment problem?

Do not use short texts (e.g. poems). The "higher" the unit you consider, the longer the examined text should be.

Do not use translated texts because here the text building is rigorously bound, it allows no spontaneity.

References

Hřebíček, L. (2007). *Lectures on text theory*. Prague: Oriental Institute.

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

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. (esp. Chapter 11).

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

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Presens.

5.8. Chaining and distances

Problem

The repetition of some identical entities in subsequent sentences gives rise to sentence chains. The chains bring about the impression of perseveration, echo, Skinner effect, inertia, etc.

(a) Set up a preliminary list of entities that can take part in sequential text structuring.

(b) Study some of them using a specific text sort and set up inductive hypotheses about their behaviour.
(c) Substantiate the hypotheses linguistically or psychologically and derive the hypotheses from a more general and abstract background, i.e. systematize this field of research.

Procedure

Partition a text completely in sentences giving an unambiguous list of phonetic or graphic signs that signal the end of sentence. If poems are analyzed, for some entities the line can be considered equivalent to sentence. Each sentence represents a point in the sequence. Then choose one of the possible entities that may be repeated in sentences. Some examples are: consonant clusters, syllables, morphs, words, references, speech acts, sentence types, dependency structures, descriptive expressions, active expressions, ornamentality, etc. This list will automatically increase with increasing research.

Replace the sentences by their order number and mark identically those that contain the respective repeated element. Now one can see uninterrupted chains of sentences (held together by the same element) having specific lengths. Set up a hypothesis about the distribution of lengths of chains. An exact description of the procedure has been presented in Popescu et al. (2010, Chapter 9).

Some of the entities seem to occur rather haphazardly but probably their mutual distances display a distribution. Count the distances and establish hypotheses about them. Make conjectures about the structuring of distances.

Begin with simple entities which can easily be found mechanically. For some entities one obtains more and longer chains (e.g. synsemantics); other ones will be more rare and shorter. Set up a preliminary hierarchy of entities. Strive for a theory and cooperate with a scientist engaged in neurology.

Is it possible to characterize styles or genres starting from the chaining structure of certain entities? Do languages differ in this respect?

Study the sequences of repetitions using different methods: Markov chains, fractal measures, dimensions, distance distributions, etc. Do not use very short texts though even in folklore quatrains one can find expressed repetitions. Epical poems display different kinds of phonetic chaining. Begin with phonetic entities and continue with syntactic, semantic and psycholinguistic ones. This discipline is not yet developed, nevertheless it may contribute both to linguistics, literary studies and psycholinguistics.

References

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

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

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

Lord, A.B. (1956). The role of sound patterns in Serbocroatian epic. In: Halle, M. (ed.), *For Roman Jakobson. Essays on the occasion of his sixtieth birthday: 301-305*. The Hague: Mouton.
Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text.* Lüdenscheid: RAM.
Skinner, B.F. (1939) The alliteration in Shakespeare's sonnets: A study in literary behavior. *The Psychological Record 3, 186-192.*
Skinner, B.F. (1941). A quantitative estimale of certain types of sound patterning in poetry. *The American Journal of Psychology 46, 64-79.*
Skinner, B.F. (1957). *Verbal behavior.* Acton: Copley Publishing Group.
Skorochoďko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten.* Bochum: Brockmeyer.
Zörnig, P. (1984a). The distribution of distances between like elements in a sequence I. *Glottometrika 6, 1-15.*
Zörnig, P. (1984b). The distribution of distances between like elements in a sequence II. *Glottometrika 7, 1-14.*
Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22.*

5.9. Sonnet 1

Problem

Take a collection of sonnets and compute their euphonic structure. Then solve some of the problems presented below.

Procedure

Use the method of computation of euphony presented in Strauss, Fan, Altmann (2008, 45 f.). Use only phonetic/phonemic transcriptions (i.e. no letters).

(1) Compute the overall euphony of the sonnet.
(2) Compute the course of euphony from the beginning to the end of the poem (line by line) and if possible capture it formally.
(3) Compute the euphonic structure of the verse beginnings (first sounds or sound combinations in the lines) in the sonnet.
(4) Set up a table of phoneme frequencies in all sonnets individually, compare them for homogeneity (using e.g. the chi-square test) and find the theoretical distribution they abide by.
(5) Mark the stressed vowels in each line and study the preference of stressing a certain vowel in the whole sonnet.

(6) Which phonemes contribute to euphony? Set up a rank-order of "euphonic" phonemes and generalize the result, if possible.
(7) Compare the results won from different authors in the same language.
(8) Compare the results won from sonnets in different languages.
(9) Study the evolution of euphony in sonnets in the given language from the beginning of this literary form till today. Take random samples from each century.
(10) Study the overall evolution of euphony in sonnets taking into account at least three languages.
(11) Since you have now the phonetic/phonemic transcription of sonnets, compute the phonetic similarity of lines. First define a similarity measure, then test the hypothesis that average phonetic similarity of lines decreases with increasing distance of lines. For distance 1 there are 13 cases, for distance 2 there are 12 cases, etc. Let x = distance, y = average similarity.
(12) Perform a subjective scaling of some feeling categories (e.g. sadness, melancholy, longing, gladness, etc.) in several sonnets and state whether the given feeling correlates with euphony. If necessary, use informants.

References

Altmann, G. (1966). The measurement of euphony. *Teorie verše 1, 259-261*. Brno: Universita J.E. Purkyně.

Sonnet Central. (http://www.sonnets.org/)

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM.

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

5.10. Sonnet 2

Problem

Take a collection of sonnets and compute some of their quantitative properties. Then solve some of the problems presented below.

Procedure

Study the following properties

1. Word-form frequencies. Set up the rank-frequency distribution for each sonnet separately and compute its properties in form of indicators. Fit

the Popescu function and study the number of strata in a sonnet (cf. Popescu, Altmann, Köhler 2010).

2. Belza-Skorochod´ko's chanining coefficient for sentences or verses. (cf. *Problems Vol. 2: 57*; Popescu et al. 2010).
3. For each sonnet compute the writer´s view, i.e. the angle formed by the straight lines joining the *h*-point with the greatest frequency and the greatest rank, and compute its possible correlation with the date of origin (cf. Popescu, Altmann 2007; Popescu, Čech, Altmann 2011a).
4. Compute the crowding of autosemantics in individual sonnets (cf. Popescu et al. 2009; *Problems Vol. 2: 59*).
5. Compute the vocabulary richness of individual sonnets. Use different indicators. (cf. Popescu, Čech, Altmann 2011b)
6. Compute the arc length of each rank-frequency distribution and the lambda-indicator from it (cf. Popescu, Čech, Altmann 2011c; Mačutek 2009).
7. Compute the thematic concentration of each sonnet using Popescu's formula associated with the *h*-point (cf. Popescu et al. 2009; *Problems Vol 1: 61*; Popescu, Altmann 2011). Combine this problem with that in 5.9 in this volume.
8. For each sonnet, construct the hreb-sets and state their rank-frequency distributions (cf. Ziegler, Altmann 2002; Ziegler 2005). Use the hreb for solving the previous problem.
9. Compute word lengths (measured in terms of syllable numbers) and word-frequencies in each sonnet separately. Then replace the words in the text by their lengths or frequencies respectively. You obtain a sequence of numbers. For these sequences compute the Hurst exponent and the Lyapunov coefficient. State Köhler's motifs for lengths and frequencies and set up their distributions (cf. Köhler 2006, 2008; Köhler, Naumann 2008; *Problems Vol. 2*).
10. For each sonnet state the rank distribution of word classes, fit the appropriate distribution and compare the sonnets.
11. Study the verbality and nominality of sonnets (cf. Ziegler, Best, Altmann 2002; Problem: *Verb classes*).
12. For each sonnet compute the significant word associations and draw an association graph. Compute some properties of this graph (cf. Ziegler, Altmann 2002; Ziegler 2005).
13. For each sonnet compute the golden section and compare them (cf. Popescu, Altmann 2007, 2009; Popescu et al. 2009; Tuzzi, Popescu, Altmann 2009; *Problems Vol. 2,* 66-76)
14. Transform the rank-frequency of words in each sonnet into a frequency spectrum. For each sonnet compute Ord´s criterion <I,S> and plot all sonnets in a Cartesian <I,S> coordinate system. Generalize the result (cf. Popescu et al. 2009).

15. For each sonnet compute the mean word length, partition all words into two classes: shorter than and longer than the mean. Then study the behaviour of runs (cf. Altmann, Altmann 2008; Popescu, Čech, Altmann 2011).

References

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitatven Textanalysen*. Lüdenscheid: RAM.

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

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

Köhler, R., Naumann, S. (2008). Quantitative text nalysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analsysis, Machine learning and applications. Proceedings of the Jahrestagung der Deutsche Gesellschaft für Klassifikation 2007 in Freiburg: 637-646*. Berlin-Heidelberg: Springer.

Mačutek, J. (2009). Arc length development and the highest word frequency. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 182-189*. Chernivcy: ČNU.

Popescu, I.-I., et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

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

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

Popescu, I.-I., Altmann, G. (2011). Thematic concentration of texts (submitted).

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf´s law – another view. *Quality and Quantity 44(4), 713-731*.

Popescu, I.-I., Čech, R., Altmann, G. (2011a). Some geometric properties of Slovak poetry. (submitted)

Popescu, I.-I., Čech, R., Altmann, G. (2011b). Vocabulary richness in Slovak poetry. (submitted)

Popescu, I.-I., Čech, R., Altmann, G. (2011c). *The lambda-structure of texts.* Lüdenscheid: RAM.

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text*. Lüdenscheid: RAM-Verlag.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). The golden section in texts. *ETC – Empirical Text and Culture Research 4, 30-41*.

Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 423-447*. Berlin-New York: de Gruyter.
Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.
Ziegler, A., Best, K.-H., Altmann, G. (2002). Nominalstil. *ETC – Empirical Text and Culture Research 2, 72-85.*
Internet: http://www.sonnets.org/

5.11. Text activity 1

Problem

Compare the "activity" of texts of different sorts, test their differences and set up at least an "activity rank-order" of text sorts.

Procedure

Take a text and count the number of adjectives (A) and active verbs (V). Consider a verb "active" if it expresses some activity (physical or mental). That is, verbs like *be, have, possess, must, repose*, etc. are not active. Define exactly which verbs and verb forms are considered "active".

Compute Busemann´s *verb-adjective ratio* in the form

$$(1) \qquad Q = \frac{V}{A+V}.$$

Since Q is a proportion, $Q > 0.5$ means increased activity, $Q < 0.5$ means increased descriptivity, and $Q = 0.5$ means active-descriptive equilibrium. However, it must be tested whether a tendency is significant. Since we have only two outcomes, for small $n = A+V$ the binomial distribution can be used; if n is large, one can use the normal approximation $z = (2Q\text{-}1)\sqrt{n}$.

Show the status of each text separately and order the texts according to their Q-value. Can you recognize some grouping of texts? E.g. are scientific texts more descriptive than press texts? For appropriate tests see Altmann (1978, 1988) or Wimmer et al. (2003).

Evaluate the individual acts of a stage play and show the course of activity/descriptivity. Are comedies structured differently than tragedies?

Does activity/descriptivity change in the subsequent chapters of a scientific book?

Compare the activity/descriptivity in a lyric and an epic poem. Is there a significant difference?

Compare several works (e.g. poems) of an individual author and state whether activity decreases with his increasing age (a consequence of Busemann's hypothesis).

Compare the *Q* indicator of texts told or written by children of different age. Is there a tendency for increasing activity with increasing age?

References

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

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

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag.

Antosch, F. (1953). Stildiagnostische Literaturuntersuchungen mit dem Aktionsquotienten. *Wiener Archiv für Psychologie, Psychiatrie und Neurologie 3, 65-73.*

Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio. In: Doležel, L. Bailey, R.W. (eds.*), Statistics and style: 57-65*. New York: Elsevier.

Bakker, F.J. (1965). Untersuchungen zur Entwicklung des Aktionsquotienten. *Archiv für die gesamte Psychologie 117, 78-101*.

Best, K.-H. (2006). Gesetzmäßigkeiten im Erstspracherwerb. *Glottometrics 12, 39-54.*

Boder, D.P. (1940). The adjective-verb quotient; a contribution to the psychology of language. *Psychological Revue 3, 309-343.*

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik.* Jena: Fischer.

Fischer, H. (1969). Entwicklung und Beurteilung des Stils. In: Kreuzer, H., Gunzenhäuser, R. (eds.), *Mathematik und Dichtung: 171-183*.. München: Nymphenburger Verlag.

Goldman-Eisler, F. (1954). A study of individual differences and of interaction in the behaviour of some aspects of language in interviews. *Journal of Mental Science 100, 177-197.*

Schlissmann, A. (1948/49). Sprach- und Stilanalyse mit einem vereinfachten Aktionsquotienten. *Wiener Zeitschrift für Philosophie, Psychologie und Pädagogik 2, 42-46.*

Tuldava, J. (2005). Stylistics, author identification. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 368-387.* Berlin-New York: de Gruyter.

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

Zsilka, T. (1974). *Stilisztika és statisztika*. Budapest: Akadémiai Kiadó.

5.12. Text activity 2

Problem

Scale the "activity" of verbs, compute the distribution of the degrees and characterize the "activity" of texts using some indicator.

Procedure

Take a text and consider only its verbs. Order the verbs according the degree of their expression of activity. For example *be* could have degree 0 and *blow*, *run*, *jump* degree 10. There may be also negative degrees, e.g. with *suffer* or *die*. The task is psycholinguistic and should be performed with the help of questionnaires containing many verbs and the task to order them according to their degree of activity. One can also use the semantic classification of verbs presented e.g. in Yesypenko (2009) and ascribe preliminary degrees to individual classes.

Having obtained a kind of scaling, ascribe to each verb of a text its activity and compute some indicators, e.g. distribution of activity, mean activity, dispersion, the course of activity which may be interesting for stage plays or epic poetry.

Then consider individual problems discussed in Problem *5.11*: *Text activity 1*, namely

(i) Are comedies structured differently than tragedies with regard to their "activity"?

(ii) Does activity change in the subsequent chapters of a scientific book?

(iii) Compare the activity in a lyric and an epic poem. Is there a significant difference?

(iv) Compare several works (e.g. poems) of an individual author and state whether activity decreases with his increasing age.

(v) Compare the texts told or written by children of different age. Is there a tendency towards greater or smaller activity with increasing age?

Develop a conceptual framework of activity, set up some hypotheses, perform preliminary classifications and join activity with other properties of texts, express them mathematically i.e. begin to elaborate a theory

References

Croft, W., Cruse, D.A. (2004). *Cognitive linguistics*. New York: Cambridge UP.

Langacker, R. (1990). *Concept, image and symbol. The cognitive basis grammar.* Berlin: Mouton de Gruyter.

Wierzbicka, A. (1992). *Semantics, culture and cognition. Universal human concepts in culture-specific configurations.* New York: Oxford University Press.

Yesypenko, N. (2009). An integral qualitative-quantitative approach to the study of concept realization in the text. In: E. Kelih, V. Levickij, G. Altmann. (eds.), *Methods of text analysis: 308-327*. Černivci: ČNU.

5.13. Assonance

Problem

Set up an indicator of global vocalic assonance in isosyllabic verses and test the hypothesis that the mean assonance of verses that are close to one another is greater than in more distant ones. That is, show that with increasing distance of verses the mean assonance decreases.

Find the form of this decrease and test it using several poems in different languages.

Procedure

Take a poem consisting of isosyllabic verses, i.e. verses having the same number of vowels which are the nuclei of syllables. Or, alternatively, take a poem, find the shortest line (having *n* syllables) and consider only the vowels placed in other verses up to syllable *n*. The latter procedure is somewhat restricted but one can scrutinize at least the syllables up to *n*.

For each line set up the vector whose elements are the vowels in individual positions, e.g. in Goethe´s poem "Erlkönig" the first line is "Wer reitet so spät durch Nacht und Wind", hence the vector would be

V_1 = <e:,ai,e,o:,ä:,u,a,u,i>.

It is a question of definition whether one ignores the length or tone of the vowel. If we ignore the length in German, we obtain the first four lines in "Erlkönig" as

V_1 = <e,ai,e,o,ä,u,a,u,i>
V_2 = <e,i,e,a,e,i,ai,e,i>
V_3 = <e,a,e,a,e,o,i,e,a>
V_4 = <e,a,i,i,e,e,e,i,a>.

Since in each line there are 9 position and all vowels may occur in all positions, one can easily state that in the first two lines there are 3 equal vowels in the same position, i.e. $P(V_1,V_2) = 3/9$, further $P(V_2,V_3) = 5/9$, $P(V_3,V_4) = 4/9$. Taking the mean of the three neighbouring line pairs we obtain the mean vocalic parallelism for distance 1 as $P_1 = 4/9 = 0.4444$. The two-step distance follows from $P(V_1,V_3)$

= 2/9 and $P(V_2,V_4)$ = 2/9. The mean is P_2 = 2/9 = 0.2222. Of course, one should process the whole poem, take averages and study the decrease of assonance.

It is advisable to *leave out the vowels occurring in the rhyme position* because they automatically increase the assonance of the given distance. If in the above vectors we leave out the last vowels, we obtain for distance one: (2/8 + 5/8 + 3/8)/3 = 1.25/3 = 0.4167, i.e., a slightly smaller number, while P_2 = 0.2500, a slightly greater number.

The hypothesis is in agreement with Skinner´s conjecture of formal reinforcement. If not all vowels may occur at all positions, another methods must be developed. In the same way one can consider consonants, syllables or other units.

Use the method for characterizing the poetry of one author and analyze more languages.

References

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

Skinner, F. (1957). *Verbal behavior***.** Acton, Mass.: Copley. New York: Appleton-Century-Crofts, Inc.

5.14. Frequency and position of words in sentence

Hypothesis

This hypothesis is a consequence of another one (cf. *Problems Vol 1*: "Word length and position in sentence", 65). If long words have the tendency to occur later in the sentence, then short words stay in the anterior positions in the sentence. But since these words are shorter, they occur more frequently (cf. Köhler 2005). Hence *the position in the sentence is negatively correlated with the frequency of words.* Another conjecture is given by Fenk and Fenk-Oczlon (2005, 162f.) who start from "an increase of content words and a decrease of function words during a sentence." But "the token frequency of function words is higher than the token frequency of content words", consequently the position-in-sentence is negatively correlated with frequency. Test the hypothesis.

Procedure

Take a long text and using one of the many counting programs state the frequency of each word. In some languages lemmatization will be necessary, according to whether one counts lemmas or word-forms.

Then replace each word in the sentence by its frequency. Order the sentences in groups according to their length (measured in terms of word numbers).

For each position in the length group *n* compute the mean frequency in individual positions. State whether the hypothesis may hold. That is, compute some kind of regression of average position frequency on position. Begin with linear regression.

If there is a significant regression, then the longer the sentences the more insignificant it will be. That is, the regression coefficient will decrease. Fit a function to the decrease of the regression coefficient.

Perform this investigation on several texts of one language, then scrutinize also some other ones.

References

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 277, 517-556.*

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

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.

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

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

5.15. Word length and position in sentence

Problem

The hypothesis that word length increases in the different positions of the main clause has been tested in *Problems*, Vol. 1, p. 85. Scrutinize the evolution of mean word length in the individual positions of whole sentences and find a regularity – if there is any.

Procedure

Take a longer text and compute word length (measured in terms of syllable number) in each position of each sentence. Then group together sentences having the same length (= the same number of words). Now, for each group separately, compute the mean word-length in each position. For each sentence length you

obtain a sequence of numbers. Since the independent variable *x* is the position in sentence, compute the linear regression of mean length on position. You obtain as many regressions as there are length groups. Care for the following conditions: skip sentences of length 1 and 2; each length (3, 4,...) must be represented by at least 10 sentences, otherwise the results will not be reliable.

Compare all linear regressions and observe the behaviour of the regression coefficient. You may state that it decreases monotonically with increasing sentence length.

(1) Express this decrease by an appropriate function. Since the values of the regression coefficient also attain negative values, one must find a function fulfilling this condition, i.e. it cannot be a power or an exponential function because these functions converge to zero but do not have also negative values.

(2) Substantiate this phenomenon linguistically. Use the fact that short sentences mostly contain only the main clause but in longer sentences several clauses are present. Take into account the fact that clauses are joined with conjunctions (which are short), contain pronominal references, etc.

(3) Analyze several languages and compare your results with the existing ones.

(4) How do Menzerath's law and Arens' law interact with this problem?

(5) Is there any relationship between the hypothesis and the functional sentence perspective (Firbas 1992)?

References

Fan, F., Grzybek, P., Altmann, G. (2010). Word length in sentence. *Glottometrics 20, 70-109*.

Firbas, J. (1992). *Functional sentence perspective in written and spoken communication.* Cambridege: Cambridge University Press.

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

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

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

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

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

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

5.16. Development of rhyme

Problem

Study the development of different types of rhyme in a given language, set up hypotheses and test them.

Procedure

First state the different kinds of rhyme. Lists for different languages can be found e.g. in the Wikipedia. Choose only one class, e.g. masculine-feminine-dactylic (differing in the position of accent) or open-closed (i.e. ending with vowel or with consonant). Then take a collection of poems (e.g. from Project Gutenberg, http://www.gutenberg.org) and for different years of origin, count the proportion of individual classes of rhymes. If the rhyme structure is e.g. A B A B then count twice A and twice B, i.e. each rhyme-word must be counted separately because rhymes need not be perfect and even if the words are rhymed, the accent may be different. The simplest problem is to count the number of open rhyme-words and closed ones; it has been shown, for example, that in Slovak the proportion of open rhymes decreases from the beginnings until 1960. Each year or decade must be represented reliably, i.e. not only one poem but several ones should be analyzed.

Now, knowing the proportion of the given classes formulate a hypothesis about the development of the given rhyme sort, fit to the data an empirical function and venture/compute a prediction. Analyze some further texts from the 21st century and check whether your hypothesis can be accepted.

Since proportions or percentages lay in <0,1> or <0,100> respectively, the function you have chosen must converge to one of the extremes of these intervals, They cannot converge to infinity or fall under zero. They should be "realistic".

In this way, analyze different types of rhyme – if possible, also in different languages; set up "local" hypotheses concerning types of poetry or language and at last, formulate a general hypothesis about the development of any kind of rhyme.

To continue, analyze different poetic phenomena, e.g. the evolution of forms of the hexameter from classical to modern languages and amplify your general hypothesis. Show the general development of poetic forms.

References

Bayer, C.M.M. (1990). Zur Entwicklung des Reimes in lateinischen metrischen Inschriften vom Ende des 8. bis zur Mitte des 13. Jahrhunderts. In: Köns-

gen, E. (ed.), *Arbor amoena comis: 113-132.* Stuttgart: Steiner.
Greber, E. (1994). *Textile Texte: poetologische Metaphorik und Literaturtheorie.* Köln: Bohlau Verlag.
Grotjahn, R. (ed.) (1981). *Hexameter Studies*. Bochum: Brockmeyer.
Keipert, H. (1977). Zur Entwicklung des Reims in Lomonosovs Odendichtung. *Zeitschrift für slavische Philologie 39(2), 251-269.*
Štukovský, R., Altmann, G. (1965). Vývoj otvoreného rýmu v slovenskej poézii. *Litteraria 8, 156-161.*
Štukovský, R., Altmann, G. (1966). Die Entwicklung des slowakischen Reimes im XIX. und XX. Jahrhundert. In: *Teorie verše I, 258-261.* Brno: Univerzita J.E. Purkyně.

5.17. Vowel auto-affinity in poems

Problem

Do the vowels of a strophe in a poem display some kind of auto-affinity? Perform tests using different longer poems.

Procedure

Transcribe the vowels(!) of a strophe phonetically/phonemically in the given order. Then write the same sequence under the original sequence but one step to the right, e.g.

```
a i u i e o o a a i o
  a i u i e o o a a i o
```

and compare how many vowels of the first series are equal to their vertical counterparts. Here we find only 2 (/o/, /a/). Since we compared 10 places and 2 were equal, we obtain the result 2/10 in the first step. Shift the lower line a further step to the right. In this step we obtain 1 equal pair (/i/). There were 9 comparisons, hence we get 1/9. In the third step we obtain 0/8 etc. This procedure can be performed mechanically with a program.

Note for the individual strophes of the whole poem the proportions in individual steps. Compute the means in individual steps and the general mean (mean of means). Then perform the usual analysis of variance, i.e. test whether the dispersion between steps is significantly greater than the dispersion within the steps. The appropriate formulas may be found in any text-book of statistics.

If you find a significant difference, test which of the step-means differs significantly from the general mean. Use the t-test or the normal test.

Interpret the result. Probably you will find affinity only in the rhyme position (if the number of syllables in all verses is equal). In that case there is no special affinity.

If you find affinity in some other step, can one interpret it as a sign of spontaneity? According to Skinner (1939, 1941) spontaneity may lead to the rise of affinities.

Analyze several poems of the same writer and of different writers. Compare the results.

Analyze a writer's development from the viewpoint of vowel affinity.

Are there poetic systems with prescribed auto-affinity?

References

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

Wimmer, G. et al. (2003). *Úvod do analýzy textov* (65-67). Bratislava: Veda.

5.18. The lambda indicator and Ord´s criterion

Problem

The lambda indicator for word frequencies can be found in Popescu, Mačutek, Altmann (2009), Popescu, Čech, Altmann (2011) and Ord's criterion in Ord (1972) or in Popescu et al. (2009). Take a set of texts by one author, e.g. individual poems or individual chapters of a book, and compute for all of them the lambda indicator. Show that in Ord´s scheme they are situated in a short interval, most probably on a straight line.

Procedure

First compute for each text the frequencies of word forms. Then compute the arc length between the frequencies as shown in the references, and transform the arc length into the lambda indicator. Set up intervals of length 0.1 beginning with the lowest lambda up to 2.3, i.e. <0.9, 1.0), <1.0, 1.1), <1.1, 1.2),…, <2.2, 2.3>. State the number of lambdas in the individual intervals. Then take the interval means, i.e. 0.95, 1.05, 1.15,…,2.25 and using them as a random variable, compute the I and S values for each writer or work separately.

Plot the <I,S> values of each set of texts in a Cartesian system and draw conclusions, e.g. about the given language, about the given text sort, about the given writer. State the left or right asymmetry of lambdas using S.

Possibly there is a certain kind of development in a language. Compare texts of equal text sort in the history of the given language.

Take texts narrated by children (written or oral) and perform the same procedure. Where are the differences? Is there an evolution with increasing age?

Study the texts of a certain literary direction, e.g. dada, or differentiate lyric and epic poetry.

Find the relationship of lambda to other text properties.

For easier processing we present the key formulas:

Arc length

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

f_r = frequency at rank r, V = vocabulary = highest rank

Lambda indicator

$$\Lambda = \frac{L(\log_{10} N)}{N}$$

N = text length

Ord's criterion

$$I = \frac{s^2}{\bar{x}}, \qquad S = \frac{m_3}{s^2}$$

$\bar{x}$ = mean of ranked frequencies, s^2 = variance of ranked frequencies, m_3 = third central moment of the rank-frequency distribution.

References

Ord, J.K. (1972). *Families of frequency distributions.* London: Griffin.

Popescu, I.-I. et al. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

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

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

5.19. The lambda-indicator and Busemann's adjective-verb ratio

Problem

Find the relation between the lambda-indicator and Busemann´s adjective-verb ratio.

Procedure

First compute both indicators for several texts. The formulas can be found in problem No. 5.11. *Text activity 1* (Formula 1, *Busemann's verb-adjective ratio*) and in problem No. 5.18. *The lambda indicator and Ord's criterion* (cf. Popescu, Čech, Altmann 2011).

Since active verbs and adjectives are autosemantic and frequently hapax legomena, their presence exerts an influence - at least indirectly - on the arc length of the rank-frequency distribution. Study the values of these two indicators in different text sorts and if your results display a kind of correlation, set up a corresponding hypothesis.

Test the hypothesis, and if it is significant, extend it adding further text properties. Construct stepwise a control cycle.

Since Busemann's indicator needs words expressing activity and descriptivity, you may restrict the investigation to those verbs which denote real activity, e.g. *jump, go, sing, think,…* and omit verbs like *be, have, sleep*. Nevertheless, you may choose verbs for comparability with the results of other researchers. Or you may choose a quite different procedure and ascribe to verbs degrees of activity.

References

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

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

Antosch, F. (1953). Stildiagnostische Literaturuntersuchungen mit dem Aktionsquotienten. *Wiener Archiv für Psychologie, Psychiatrie und Neurologie 3, 65-73.*

Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio. In: Doležel, L. Bailey, R.W. (eds.*), Statistics and style: 57-65*. New York: Elsevier.

Bakker, F.J. (1965). Untersuchungen zur Entwicklung des Aktionsquotienten. *Archiv für die gesamte Psychologie 117, 78-101.*

Boder, D.P. (1940). The adjective-verb quotient; a contribution to the psychology of language. *Psychological Revue 3, 309-343.*

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik.* Jena: Fischer.

Fischer, H. (1969). Entwicklung und Beurteilung des Stils. In: Kreuzer, H., Gunzenhäuser, R (eds.), *Mathematik und Dichtung; 171-183*. München: Nymphenburger Verlag.

Goldman-Eisler, F. (1954). A study of individual differences and of interaction in the behaviour of some aspects of language in interviews. *Journal of Mental Science 100, 177-197.*

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

Schlissmann, A. (1948/49). Sprach- und Stilanalyse mit einem vereinfachten Aktionsquotienten. *Wiener Zeitschrift für Philosophie, Psychologie und Pädagogik 2, 42-62.*

Tuldava, J. (2005). *Stylistics, author identification.* In: Reinhard Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An international Handbook: 368-387*. Berlin-New York: de Gruyter.

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

Zsilka, T. (1974). *Stilisztika és statisztika*. Budapest: Akadémiai Kiadó.

5.20. Sentence length in discourse

Problem

Show that discourse participants differ in sentence length. Set up a hypothesis about the hierarchy in discourse.

Procedure

Take a stage play and compute for each person the length of all her/his sentences. If possible, measure sentence length in terms of clauses. If this causes many problems, measure the length in terms of the number of words.

For each person set up the frequency distribution of sentence lengths. First show that the distributions are not homogeneous. Perform e.g. a chi-square or a 2I test. Alternatively, show that the means of distributions are significantly different using a t- or u-test for difference of two averages.

Starting from different properties of distributions, show that some of them can be used as an indicator of dominance in discourse. At the beginning, state, for example, whether sentence length correlates with the status of the given person. That means, quantify also the status of participants using other features. First compute the correlation coefficient of *sentence length vs. dominance*

feature, then find a quantitative model of this relationship, e.g. a regression curve.

Fnally, prepare an overall frequency distribution of sentence length, i.e. add the frequencies of all persons. If you measure the sentence in terms of clause numbers, see if this frequency distribution abides by the negative binomial distribution. If you measured in terms of number of words, see if it abides by the hyper-Pascal distribution.

Fit these distributions also to the data of individual participants and use the parameters of the distributions for solving the problem of hierarchy of participants.

Continue the examination using other properties of texts of individual persons. Use some indicators presented in Chapter 5 such as arc length, the binary code of sentence, text activity, the lambda-indicator, thematic concentration, Gini's coefficient, etc. Link these indicators with sentence length of individual persons and show the overall "structuring" of persons in discourse.

References

Cf. Different sections of Chapter 5 in this volume.
Altmann, G. (1988). Verteilungen der Satzlängen. *Glottometrika 9, 147-170.*

5.21. Sentence length sequences in discourse

Problem

Is there a tendency to vary the length of subsequent sentences in a discourse or to hold them in equal length? Great differences in length would indicate the dominance of one or more participants, non-significant differences would signal rather a democratic participation. Solve the problem and set up an hypothesis.

Procedure

Take a stage play and compute the length of each sentence either in terms of clause or words. Write the sequence of lengths as an array. Then prepare a contingency table containing all transitions from a given length to the neighbouring length, e.g. the sequence 4,2,3,1,2,1,4,4,3,2,2,3,3,3,1,1,4,2,3,1 could be written as

		Successor			
		1	**2**	**3**	**4**
Predecessor	**1**	1	1	0	2
	2	1	1	3	0
	3	3	1	2	0
	4	0	2	1	1

i.e. there are e.g. 2 transitions from 1 to 4, one transition from 4 to 4, etc. In your data, perform the overall chi-square test for independence. Then test the tendency in individual cells in order to show stereotype sequences, and, finally, test the tendency on the diagonal. If the diagonal is significantly strong, the sequence displays a kind of monotonicity; if it is significantly weak, it displays a contrastive discourse.

Compute the order of the Markov chain of the given sequence.

For testing the tendency in one cell n_{ij} use the criterion

$$u = \frac{n_{ij} - \frac{n_{i.} n_{ij}}{n}}{\sqrt{\frac{n_{i.} n_{.j} (n - n_{i.})(n - n_{.j})}{n^2 (n-1)}}},$$

where n_{ij} = frequency in cell (i,j), $n_{i.}$ = sum of frequencies in row i, n_{ij} = sum of frequencies in column j, n = sum of all frequencies.

For testing the diagonal see Altmann (1987), Schulz, Altmann (1988).

References

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

5.22. Properties of illocutive graphs

Problem

Study the properties of a weighted graph of transitions between illocutive speech acts in a stage play. Set up some hypotheses.

Procedure

Classify the illocutive speech acts using the classification in *Problems Vol. 2, p. 118* (cf. Bach, K. http://online.sfsu.edu/-kbach/spchacts.html), take a stage play and set up a matrix containing the numbers of transitions from one type to another. Since a matrix can be displayed as a weighted graph, study as many properties of the graph as possible.

Show that the quantitative properties of dramas differ from those of comedies. This may be expressed both by the values of some properties and the relationships among them.

Take a special kind of stage play and study its historical development in one language.

Set up a bipartite graph consisting of two distinct sets: that of persons of a stage play and that of illocutive speech acts. Then join the persons with the acts and ascribe each edge a weight according to the extent of use of the act by a person. Evaluate such a graph and interpret it linguistically. A quite simple possibility is given by using contingency tables with <Person, Act> classification and testing for independence.

If you analyze a stage play, observe also the change of the graph. If you present the data in form of a contingency table, compare the tables of individual stage play parts and develop a hypothesis concerning the change of dependence in the course of stage play.

Does this change correlate with some other properties of the stage play? Set up hypotheses, test them, quantify everything and begin to theorize.

References

Alston, W.P. (2000). *Illocutionary acts and sentence meaning*. Ithaca: Cornell University Press.

Austin, J.L. (1975). *How to do things with words*. Oxford: Oxford University Press.

Brock, J. (1981). An introduction to Peirce's theory of speech acts. *Transactions of the Charles S. Peirce Society 17, 319-326.*

Burkhardt, A.S. (ed.) (1990). *Speech Acts, Meanings and Intentions. Critical Approaches to the Philosophy of John R. Searle*. Berlin-New York: de Gruyter.

Doerge, F.C. (2006). *Illocutionary acts - Austin's account and what Searle made out of it*. Diss. Tübingen 2006. Available at http://deposit.ddb.de/ cgi-bin/dokserv?idn=979505232&dok_var=d1&dok_ext=pdf&filename =979505232.pdf

Erler, B. (2010). *The speech act of forbidding and its realizations: A linguistic analysis*. Saarbrücken: VDM Verlag Dr. Müller.

MacDonald, N.B. (2001). Illocutionary stance in Hans Frei's 'The eclipse of biblical narrative': an exercise in conceptual redescription and normative analysis. In: Bartholomew et al. (eds.), *After Pentecost: Language and Biblical*

Interpretation: 312-328. Grand Rapids: Zondervan (USA)/Carlisle: Paternoster (UK).
Searle, J.R. (1969). *Speech acts*. Cambridge: Cambridge Univ. Press.
Searle, J.R. (1975). A taxonomy of illocutionary acts. In: Günderson, K. (ed.), *Language, Mind, and Knowledge Vol. 7.* Minneapolis.
Searle, J.R. (1975). Indirect 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. (1991). *Pragmatics: A Reader*: *265–277*. Oxford: Oxford University Press.
Siebel, M. (2002). What is an illocutionary point? In: Grewendorf, G., Meggle, G. (eds.), *Speech Acts, Mind, and Social Reality. Discussions with John R. Searle: 125–139.* Dordrecht: Kluwer.
West, D.B. (2001). *Introduction to graph theory*. Upper Saddle River: Prentice Hall.

5.23. The frequency sequence of words in text

Problem

This is a continuation of problem 5.19 from *Problems Vol. 2*.

Compute different properties of the sequence of frequencies of words/ word forms in a text. Compare individual texts, genres and authors, and study the historical development of every property.

Procedure

Take a text and compute the frequencies of individual words/word-forms in it. Then replace the units by their frequencies. You obtain a sequence of numbers representing a kind of time series.

1. Compute the autocorrelation of frequencies. Having the autocorrelation for all lags, say from 1 to 20, plot the graph of the autocorrelation sequence, set up a hypothesis and test it on other texts. If they differ, search for boundary conditions and substantiate your findings linguistically. (For a good example see Eom 2006, 103ff.)

2. Set up a table of transitions from each frequency to each other. Evaluate your sequence and test the contingency table (a) for independence using the chi-square or 2I test; (b) each cell separately for a possible preference; (c) the tendency on the diagonal, i.e. the preference for frequencies following the same frequencies.

3. State the order of the Markov chain in your sequence.

4. Compute Hurst's coefficient for the sequence of numbers and the Hausdorff dimension of the sequence.

5. Use the box-counting method for computing the fractal dimension.

6. Compute the distance between equal frequencies and state whether the distribution of distances is random, i.e., compare it with Zörnig's model (1984, 1987). If it is not random, what type of distribution does it have?

7. Characterize the frequency distribution obtained from the distances by different indicators, e.g. entropy, repeat rate, moments, skewness, excess, position in Ord's scheme, all of which can be found in the three volumes of *Problems.*

References

Eom, J. (2006). *Rhythmus im Akzent*. München: Sagner.

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

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics Vol 2*. Lüdenscheid: RAM.

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

Zörnig, P. (1984). The distribution of distances between like elements in a sequence, I. *Glottometrika 6, 1-15;* II. *Glottometrika 7, 1-14.*

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

5.24. Frequency motifs in text

Problem

This is a continuation of the problem 5.23. *The frequency sequence of words in text* from which one can take over the ready sequences or prepare them starting from the description of the problem.

A frequency motif (F-motif) is a non-deceasing sequence of numbers. If we have for example a sequence 1,2,8,3,9,2,1,1 we obtain the following motifs: 1-2-8; 3-9, 2; 1-1.

Having obtained the F-motifs, evaluate their properties and characterize the text.

Procedure

First set up the set of F-motifs occurring in the text, i.e. identify them. Then solve at least one of the following problems:

1. Set up the rank-frequency distribution of individual F-motifs and characterize it using different indicators (e.g. entropy, repeat rate, moments, skewness, excess, position in Ord´s scheme, Popescu indicators).

2. Set up the distribution of F-motif lengths (e.g. the motif 1-2-8 has length 3) and characterize it in the similar way as in point 1.

3. Compute the maximal number of possible F-motifs using the result in Mačutek (2009) and characterize the F-motif richness of the text.

4. For each F-motif compute the mean of the numbers in it (e.g for 1-2-8 it is (1+2+8)/3 = 3.67) and set up the distribution of mean values of motifs. Characterize it.

5. Compute the autocorrelation of F-motifs and substantiate it linguistically.

6. Compute the distances between identical F-motifs and set up the distribution of distances.

7. Using the results in task 4 use the theory of runs in order to characterize the sequence of F-motif means and to find some tendencies.

References

Köhler, R., Naumann, S. (2010). A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures • Functions • Interrelations. Quantitative Perspectives: 81-89*. Wien: Praesens.

Mačutek, J. (2009). Motif richness. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 51-60.* Lüdenscheid: RAM.

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

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

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of texts.* Lüdenscheid: RAM.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts.* Lüdenscheid: RAM.

5.25. Thematic concentration

Problem

Study the thematic concentration in all poems of an author. Develop a test for comparing the significance of the difference between the thematic concentration

of two texts (cf. *Problems, Vol 1: 61* and *Problem 5.26. Dialog: thematic concentration of participants* in this volume).

Procedure

For the time being, there are three ways of measuring thematic concentration – use the rank frequency distribution of:

(1) word forms (cf. Popescu et al. 2009, Chapter 6),

(2) lemmas which represent a better approximation, or

(3) morpheme hrebs expressing optimally the concentration (cf. Ziegler, Altmann 2002; cf. *Problem 5.27. Text compactness* in this volume).

The first method consists of computing the frequencies of word forms, ranking them and using the Popescu et al. (2009, 96) formula. One can obtain at least a ranking of texts. The second method is a better approximation because in highly synthetic languages the forms of a word may be placed at different ranks and need not display the given word as "thematic". Hence lemmas and their ranking, using the same formula as above yield a more adequate picture of concentration and eliminate the influence of morphology.

However, there may be many references in text pointing at a special word which are lost with either of the above methods. If the thematic word is used only once and the rest are references, then even the "lemma"-method yields distorted results. For this reason, partition the text in "morpheme/word/phrase hrebs", set up the rank-frequency distribution of hrebs and compute the thematic concentration using the Popescu formula.

First take a short text and analyze it in the above mentioned three ways in order to see the difference. Then analyze several texts and perform a ranking of texts according to their thematic concentration. Are there differences between text sorts? For example, in scientific texts one expects a stronger concentration than in lyric poetry.

At last, derive the asymptotic variance of the Popescu formula of thematic concentration and use it for asymptotic testing the differences.

Use other methods of measuring thematic concentration and compare both the computing effort and the results. Find other text properties which are linked with thematic concentration.

References

Popescu, I.-I. et al. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Wilson, A. (2009). Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories. *Glottotheory 2(2), 97-107.*

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

5.26. Dialog: thematic concentration of participants

Problem

Study the thematic concentration of individual persons in a stage play.

Procedure

Choose a stage play and set up the rank-frequency distribution of word-forms for each person separately. Define the thematic concentration using Popescu´s indicator

$$TC = 2\sum_{r'=1}^{T}\frac{(h-r')\,f(r')}{h(h-1)\,f(1)},$$

where

h = h-point
r' = rank of a thematic word in the pre-h domain
$f(r')$ = frequency of a thematic word at rank r´
$f(1)$ = frequency of the most frequent word
T = number of ranks occupied by thematic words in the pre-h domain.

Thematic words are no synsemantics but either names or autosemantics. Nevertheless, in lyric poetry some pronouns (*I, you*) or the pronominal appellatives in some languages must be considered thematic. The indicator measures the weighted proportions of such words up to the h-point. In stage plays this concentration is possibly greater because a stage play concentrates upon a special theme without lengthy descriptions.

Compare the persons in the stage play, set up a hierarchy and correlate thematic concentration with the function of the given role.

Compare different stage plays and develop a purely conceptual theory of relation of thematic concentration with other properties.

Find the rank-frequency distribution of words of each role/person separately and set up a model. Compare the parameters of the model with different persons and set up some hypotheses.

References

Popescu, I.-I., et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM.

5.27. Text compactness

Problem

Compute text compactness for all poems of an author and state the development of the author from this point of view.

Procedure

Analyze a poem in terms of word-hrebs as proposed in Ziegler and Altmann (2002) or in *Problem 5.25* in this volume, i.e. set up sets of words based on the referential structure of the text. Let the number of hrebs be n and the number of words/tokens in text (= text size) N. Define *text compactness* as

$$C = \frac{1-\frac{n}{N}}{1-\frac{1}{N}},$$

i.e. as the normalized indicator of hreb-forming. The smaller the number of hrebs, the more compact the text is. The indicator lies in the interval <0,1>. Without any test, order all poems according to decreasing compactness and scrutinize whether this order correlates with the age of the author (or the year of origin). Set up an hypothesis and test it using the works of other poets.

Using combinatorial argument compute the probability of n, i.e. the probability of forming n hrebs out of N words. This is analogous to distributing N balls in at least n urns. Then derive the variance exactly or asymptotically and set up a test for comparing two texts.

Using the normal approximation, test whether or not the text compactness of a collection of poems as a whole differs significantly from that of a collection of prose texts.

Scrutinize different sorts of texts, set up hypotheses about text compactness and link it with as many properties of texts as possible. Begin to outline a theory (even if it is incomplete) and derive formulas linking the selected properties.

References

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

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

5.28. Frequency indicator *A*

Problem

In Popescu et al. (2010), Chapter 2, the hypothesis has been set up that the indicator *A* is related to some other properties of texts. It is defined as

$$A = \frac{M}{\log_{10} N},$$

where *N* is text length (in terms of word token number) and *M* is the modulus computed as

$$M = \frac{1}{h}\left(f(1)^2 + V^2 \right)^{1/2},$$

where *h* is the *h*-point, *f*(1) is the greatest frequency (i.e. at rank 1) and *V* is the extent of vocabulary (types) or highest rank. Test this conjecture.

Procedure

Collect all known properties of texts and quantify them if necessary (cf. Popescu et al. 2009, 2011). Then analyze many texts; compute *M* and *A* and some other properties and scrutinize the mutual associations. Develop a control cycle of properties and substantiate the result linguistically, textologically, psychologically, etc. Strive for a restricted theory.

The authors (Popescu et al. 2010, 4-25) computed and presented *A* and M for many texts in 28 languages and stated that the indicator is related to the degree of synthetism/analytism of a language. However, the above problem concerns individual texts and other properties, e.g. mean sentence length, parameters of the word-length distribution, sentence complexity, the distribution of speech acts, share of dialogues and author's speech, indicators of genre, etc.

First define a text property and compute it for ten texts. Compute its correlation with *A*. If it is significant, take the next property and associate it both with *A* and the first property. Extend the number of properties and texts step by step and construct a control cycle.

Then express the dependencies formally by some functions and test their goodness-of-fit.

At last substantiate the control cycle and the individual associations linguistically.

References

Popescu, I.-I. et al. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

Popescu, I.-I., Mačutek, J., Kelih, E., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and Codes of Text.* Lüdencheid: RAM.

5.29. A vocabulary richness indicator

Problem

Show that the indicator of vocabulary richness defined as $R = HL/N$ does not adequately express vocabulary richness. (*HL* is the number of hapax legomena = words occurring in a text only once; *N* is the text length = number of tokens or types in text).

Procedure

Take a simple sentence, e.g. a strophe of a poem. Its $R = 1$ because most probably all words are different. Add the next sentence/strophe. If all words are still unique, $R = 1$, although, logically, the text is richer because it contains more hapaxes. Now, take a text consisting of 100000 words. Although it is certainly richer than a text consisting of one sentence, its *R* converges to zero.

Hence, for very short and very long texts the indicator is not adequate as a measure of richness. Do words occurring twice not add something to the richness? Give the above indicator an adequate linguistic interpretation. What does it express? Study its behaviour in texts with $N < 100$, then in texts with $1000 < N < 10000$.

Before you develop a new richness indicator and add it to the set of other ones (cf. e.g. Wimmer, Altmann 1999; Popescu, Mačutek, Altmann 2009), concentrate on the behaviour of the above index, derive its expectation and variance and propose a test for comparing two texts with different *N*.

Analyze very thoroughly what the concept of "vocabulary richness" could mean; what should it capture; which parts of the text or of the vocabulary contribute to its formation; incorporate the *h*-point in your considerations and develop a variant of Tuzzi-Popescu-Altmann´s approach (2010, 126 ff.).

Further, compare didactic and scientific texts – in which repetition of certain words is necessary – with press and fiction texts, i.e. show that your indicator can distinguish genres. Do not forget to set up an asymptotic test for the difference of two texts in terms of your new indicator.

References

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag (p. 99).

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts*. Lüdenscheid: RAM-Verlag.

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

5.30. Vocabulary richness and lambda-structure

Problem

Show that there is a relation between Popescu´s vocabulary richness indicator R_1 and the lambda-structure of texts. Find the form of this relation.

Procedure

Take 100 (complete) texts of different size and for each of them compute the rank-frequency distribution of word forms. This can be done e.g. using the Gutenberg-project and a word calculator on the Internet. For each text compute the *h*-point, i.e. that number for which $r = f(r)$ (i.e. rank = its frequency). If there is no such number take simply $h = r + 0.5$. A more exact computation is shown in the references but for this purpose it is sufficient.

Now compute $F(r)$ = sum of relative frequencies from 1 to *h*, i.e. the distribution function up to *h* for each text separately. Then for each text set up the Popescu-indicator of vocabulary richness

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

where *N* is the text size (number of words in text) (cf. Tuzzi, Popescu, Altmann 2010:127).

Now for each text compute the lambda indicator in the following way: First compute the arc length between neighbouring frequencies as

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

where V is the size of the vocabulary (number of word-form types, the highest rank) and f_i are the individual frequencies. Arc length expresses the frequency structuring of the text. Using L, compute for each text the lambda indicator as

$$\Lambda = \frac{L(\log_{10} N)}{N}. \quad (3)$$

This indicator stabilizes the arc length and makes it independent of text length.

Prepare a table of $<\Lambda, R_1>$ values and show that $R_1 = f(\Lambda)$. Find an appropriate function. Finally, fit the function $R_1 = 1 - a \exp(-b\Lambda)$ (cf. Popescu, Čech, Altmann 2011, 7ff.). Show that this function may have different parameters for different languages, text sorts or even authors. It shows the relationship between vocabulary richness and a specific formal feature of texts.

Perform the computation for texts of one language only. Then consider the outliers – if there are any – and check whether the rank-frequency distribution has been constructed "correctly", i.e. whether the definition of the word-form has been taken into account by the counting program. Hence it is better to prepare the texts by performing changes where it is necessary.

References

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts*. Lüdenscheid: RAM.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts.* Lüdenscheid: RAM.

5.31. Gini´s coefficient and vocabulary richness

Problem

Compute Gini´s coefficient for rank-frequency distributions in 100 texts of the same sort in the same language and correlate the values with other indicators of vocabulary richness.

Procedure

Take, for example, all poems of a poet, compute for each text the rank-frequency distribution of word forms (or lemmas) and using it, the coefficient of Gini. For computation use the formula

$$(1) \qquad G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{r=1}^{V} rf_r \right),$$

where V is the vocabulary size (= number of different items), N is text size, r is the rank (r = 1,2,…,V) and f_r is the frequency at rank r. Vocabulary richness can be expressed by the complement

$$(2) \qquad R = 1 - G.$$

Show that the best fitting is $R_t = aN^b$, i.e. R depends on text size N. Consider the difference between the computed curve and the empirical values

$$(3) \qquad D_R = R_t - R$$

as an expression of vocabulary richness. This can be negative, too.

Order the texts according to D_R and find a relation (a) to the age of the author, (b) to other richness indicators, (c) to other properties of texts.

Show that two typologically very different languages may have quite different R and R_t. Use the parameters a and b for typological purposes.

References

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

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

5.32. Vocabulary richness project

Problem

Perform a thorough investigation of the vocabulary richness problem.

Procedure

1. Collect the complete bibliography concerning this problem and publish it. Begin with the most recent publication and proceed backwards. A great number of publications can be found on the Internet. Do not omit works written in Slavic languages and especially in French.
2. Read the literature and classify the individual procedures or indicators according to (a) their mathematical approach, (b) their application to

special texts, (c) the aims of characterization (texts, speakers, genres, styles, languages).

3. Analyze the different approaches and set up the mathematical background: (a) if the approach is an indicator, derive at least its variance and normalize its interval. (b) If it is a curve (e.g. a type-token relation), substantiate it linguistically and derive it by means of a difference or differential equation or use a stochastic process to obtain it.
4. Show the commonalities of individual approaches, their differences, backgrounds. Analyze the development of this study historically. Where and when did this research begin and what is the present-day state of the art?
5. Set up hypotheses about vocabulary richness, its relationship to other properties of texts and construct a control cycle for vocabulary richness.
6. Test your results on different texts in different languages and strive for formulating a theory, i.e. a set of interrelated statements which can be declared to be laws. The statements must be derived from a set of (preliminary) axioms and tested in many languages. Use both modern and archaic texts but only in their original form; do not use translations.
7. On the basis of your theory, study the development of vocabulary richness in the language of children, in the development of a single author, in the development of written documents in a language beginning from primitive forms up to modern novels, technical texts, etc.
8. Collect and study the different definitions of the concept of "vocabulary richness" (do not take into account the mathematical expression). Connect the results with those in point 2 above.

References

Since the elaboration of the references is part of the project, we mention only some recent publications:

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM.

Martynenko, G. (2010). Measuring lexical richness and its harmony. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language: 125-132*. Wien: Praesens Verlag.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts.* Lüdenscheid: RAM.

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

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

5.33. A textological project

Problem

Select a writer and order his works according to the date of creation (or first publication). If (s)he has written texts in different genres, e.g. novels and poems, differentiate the text sorts and analyze them separately.

Solve as many textological problems contained in this and the preceding volumes as possible (Strauss, Fan, Altmann 2008; Köhler, Altmann 2009). Order the problems thematically, in order to obtain a compact picture of the writer. Study the development of the writer with regard to the analyzed property, i.e. correlate the change of a text property with the date of origin. Find the relation between individual properties and set up a control cycle. Omit very complex problems but strive for programming the computation of each property.

Then take different textological books (Orlov, Boroda, Nadarejšvili 1982; Altmann 1988; Baayen 2001; Altmann, Altmann 2008; Popescu et al. 2009, 2010; Popescu, Mačutek, Altmann 2009; Tuzzi, Popescu, Altmann 2010; Popescu, Čech, Altmann 2011; for a bibliography see Köhler 1995) and solve the problems presented in them for your texts – as far as possible. Strive for a compendium of quantitative textology. If you suspect a relationship to be valid generally and to be a candidate for a law, take texts of other authors, other genres, other languages and test the hypothesis. If it does not hold for other texts, find the boundary conditions responsible for its validity in your texts. If it holds, derive the relationship using preliminary linguistic axioms or assumptions.

References

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

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM.

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

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

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics 2.* Lüdenscheid: RAM.

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

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text.* Lüdenscheid: RAM

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

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative aspects of Italian texts*. Lüdenscheid: RAM.

5.34. Text diffusity

Problem

Compare the lemmatic diffusity of texts belonging to different text sorts and set up the rank order of text sorts with respect to this property.

Procedure

Take a lemmatized text and state the frequency of individual lemmas in it. Then add to each lemmas its position in text. Consider only lemmas occurring at least twice. For each lemma with $f(lemma) \geq 2$ compute the indicator of diffusity defined as

$$D(lemma) = \frac{Last\ \ position - First\ \ position}{f(lemma)}$$

The diffusity of non- repeated words is zero.

Compute the mean diffusity and consider the result as characteristic for the given text.

Perform the analysis using texts of the given writer, of the given sort, of different sorts, and in different languages.

Set up a rank-order of text sorts based on diffusity.

Interpret the indicator as repetition pattern, as an aspect of Skinners formal reinforcement, as an indicator of vocabulary richness, as an indicator of thematic concentration, etc. Compare it with other indicators and construct mathematical links with them.

Derive the variance of $D(lemma)$ considering the positions as fixed and $f(lemma)$ as the variable.

Compare the means of individual texts using the asymptotic normal criterion and show the extent of dissimilarity of texts of a given text sort.

References

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.
Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 423-447*. Berlin-New York: de Gruyter.

5.35. Aggregation in poetry

Problem

In *Problems Vol 1* (Köhler, Altmann, 2009, 56f.), "phonetic aggregation" in texts has been proposed as a problem. The task was to show that phonetic similarity of verses decreases with increasing distance. Taking means one obtained a monotonically decreasing function.

Define further phonetic, morphological, semantic, syntactic, lexicological, referential etc. entities that can contribute to the similarity of verses. Take a longer poem in your language and test the hypothesis that the given similarity decreases with increasing distance. Consider always averages.

Procedure

Phonetic entities can be phonemes/sounds, syllables, consonant clusters, sequences of sounds, etc.; morphological entities can be grammatical categories, word classes, derivatives, compounds, etc; semantic entities can be semantic classes of verbs, nouns, adjectives, etc.; syntactic entities can be types of sentences, a special word order, etc.; lexicological entities can be combinations of morphological and semantic features, choice of vocabulary, metaphors, etc.; referential entities are direct lexical repetitions of a lemma or synonyms or references in further verses, etc.

Before you begin to count, choose an adequate similarity indicator. Their number is enormous. A few are listed in the references. Every book on classification contains several.

Construct a set of respective entities in the first verse in form

$$V_1 = \{a_1,a_2,b,c.d,e,f_1,f_{2,},f_3,g,h,i,j,k,\ldots\}$$

that is, if a unit occurs several times, index it. They are considered different units. However, this depends on the method of computing similarity.

Then taking a similarity measure compare all neighbouring verse pairs and take the mean of comparisons for distance $d = 1$. Then compare verses in dis-

tance $d = 2$ and take averages, etc. Most probably you obtain a decreasing sequence of means.

Set up hypotheses concerning the following problems (cf. Altmann 1988):

(1) Is there a development of aggregation in the life of a poet?

(2) Which similarity measure is adequate for your computations?

(3) Consult a psychologist engaged in problems of perseveration and seek the mechanism responsible for aggregation.

(4) Which entities tend to form aggregations.

Take the complete work of a poet. If possible, perform this investigation not only in an Indo-European language.

References

http://reference.wolfram.com/mathematica/guide/DistanceAndSimilarityMeasures.html

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

Ashby, F.G., Perrin, N.A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review 95, 124-150.*

Biberman. Y. (1994). A context similarity measure. In*: ECML '94: Proceedings of the European Conference on Machine Learning, pages 49-63*. Springer.

Bock, H.H. (1974). *Automatische Klassifikation*. Göttingen: Vandenhoeck & Rupprecht.

Burnaby, T. (1970). On a method for character weighting a similarity coefficient, employing the concept of information. *Mathematical Geology 2(1), 25-38.*

Chandola, V., Boriah, S., Kumar, V. (2007). Similarity measures for categorical data: a comparative study. *Technical Report 07-022*, Department of Computer Science & Engineering, University of Minnesota.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science 267, 843–848.*

Goodall, D.W. (1966). A new similarity index based on probability. *Biometrics 22(4), 882-907.*

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics 2.* Lüdenscheid: RAM.

Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel, J. (2005). Similarity measures for tracking information flow. In: *Proceedings of CIKM '05, 517-524*.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika 27, 125-140.*

Spertus, E., Sahami, M., Buyukkokten, O. (2011). Evaluating similarity measures: A large scale study in the Orkut Social Network. In: *Proceedings of 11th International Conference on Knowledge Discovery in Data Mining*: 678–684. New York: ACM Press.

Terra, E., Clarke, C.L.A. (2003). Frequency Estimates for Statistical Word Similarity Measures. *Proceedings of HLT-NAACL 2003, 165-17*2.

Wang, X., Baets, B.de, Kerre, E. (1995). A comparative study of similarity measures. *Fuzzy Sets and Systems 73(2), 259-268.*

Wilson, D.R., Martinez, T.R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research 6, 1-34.*

Wimmer, G. et al. (2003). *Úvod do analýzy textov.* Bratislava: Veda.

Zwick, R., Carlstein, E., Budescu, D.V. (1987). Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning 1(2), 221-242.*

6. Various issues

6.1. Rank-frequency distribution of classifications

Hypothesis

In quantitative linguistics there is a very general assumption that if in a full set of linguistic phenomena a classification is performed, the rank-frequency distribution of classes follows an honest probability distribution (cf. Altmann 2005). The same phenomenon always abides by the same distribution, however, the parameters may be different for different authors, text sorts, languages. Test the hypothesis using the data presented by Bojko (2005).

Procedure

In works of writers writing in English, Bojko classified the clauses in 12 classes and found their frequencies as presented in Table 6.1.1 Reorder the frequencies with individual writers, if necessary, in decreasing order, assign them ranks and find an appropriate theoretical distribution or at least a simple function with maximally two parameters. If none can be found, then either (a) re-define the clause classes or (b) analyze each text individually. Bojko combined three texts from each author. Remember that the agreement with a law-like hypothesis is the only independent criterion of the goodness of your classification. If the results are still unsatisfactory, analyze other texts on the basis of your own clause classification.

Show that the parameters of the fitted functions are different and find the external conditions (author, text sort, language, historical development, epoch, etc.).

Plot the data in an Ord-scheme (Strauss et al. 2008, 121 f.) and see if they lie almost on a straight line.

Take a text of each of the authors mentioned and count the frequencies of parts-of-speech. Show that the Ord-values of the POS rank-frequency distributions lie in different "Ord" domains.

In order to generalize the results, take texts from your own language and perform clause-type analyses. Do your results agree with those of Bojko – except for parameters – or is your primary hypothesis about the form of the distribution quite different?

Test your data leaning against different clause definitions (by different grammars) and if you attain consistent results, decide if you are on the track of a "teorita" or your results are sufficient only for text sort differentiation, or, finally, both results follow from your procedures.

If possible, perform analogous analyses with preliminary classification for other different text entities (sentence types, phrase types, parts-of-speech, mor-

pheme types, canonical syllable types) and develop the basis of classification of linguistic entities.

Table 6.1.1
Frequencies of clause types in English texts
(three texts by each author, data from Bojko 2005)

Clause Type	Dreiser	Fitzgerald	Cronin	Steinbeck	Hemingway
Object Clauses	647	306	246	173	208
Attributive Clauses	488	235	194	165	121
Time Clauses	211	193	153	159	114
Conditional Clauses	146	53	46	85	56
Clauses of Manner	141	87	50	63	33
Reason Clauses	87	82	54	83	22
Concessive Clauses	41	12	46	16	9
Place Clauses	37	15	21	26	26
Purpose Clauses	12	3	2	10	3
Result Clauses	8	13	5	16	6
Subject Clauses	6	2	4	23	32
Predicative Clauses	5	5	2	13	4

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: 646–659.* Berlin-New York: de Gruyter..

Bojko, J. (2005). Diferencijni parametri rečenija jak determinanta avtors´kogo stilju [Differential parameters of clauses as determinants of personal style] In: Altmann, G., Levickij, V., Perebyinis, V. (eds.), *Problems of quantitative linguistics: 292-305.* Černivci: Ruta.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in quantitative linguistics 1*. Lüdenscheid: RAM.

6.2. Ranking and classification of verbs

Problem

Solving Problem 6.1 show that the hypothesis “ranking as a criterion of verb classification” holds, too.

Procedure

Take the data published by Yesypenko (2009) concerning the frequency and classification of verbs in three English texts. First, rank the frequencies in individual texts and fit to each of them a function used in ranking, e.g. the power function, Mandelbrot's function, Popescu's function, etc. Omit zeros. The data are presented in Table 6.2.1. The texts are: E. Waugh, "A Handful of Dust", J. Swift, "Gulliver's Travels", M. Twain, "The Adventures of Tom Sawyer."

Table 6.2.1
Frequencies of verbs in three English texts (Yesypenko 2009)

Verbs	E. Waugh	J. Swift	M. Twain
Verbs of motion/removing	148	148	202
Verbs of process, change, development	24	28	26
Verbs of beginning/end of action	30	18	36
Verbs of physical action	113	84	132
Engender verbs	31	48	30
Destroy verbs	15	22	22
Successful/Unsuccessful action implementation	2	0	6
Verbs of attempt	3	4	8
Verbs of sound emisson	3	0	6
Verbs of ligth phenomena	3	2	2
Verbs of temperature phenomena	0	2	0
Verbs of nature phenomena	4	0	4
Verbs of communication	131	30	122
Verbs of moral impact/effect	41	24	20
Verbs of social activity	23	22	30
Position verbs	16	30	44
Verbs of existence	271	174	184
Modality verbs	91	70	42
Verbs of human relations	17	18	20
Verbs of reference	23	50	30
Verbs of emotional psychological impact	9	6	26
Verbs of ownership/loss	58	104	58
Verbs of physiological state	7	32	12
Verbs of perception	50	52	78
Verbs of mental activity	66	66	68
Verbs of subjective assessment	18	8	8
Verbs of emotional psychological state	39	22	52

(1) Find the best function yielding the highest determination coefficient. Note the parameters of individual functions and using their standard errors which can be found in the output of the software set up a normal test and test the difference between the pertinent parameters of individual texts. Extend the investigation to further English texts.
(2) Test the classification of nouns, adjectives and adverbs using further Yesypenko's (2009) tables. Extend the investigation to further English texts.
(3) Perform the same analysis in a language other than English, state whether the hypothesis can be corroborated and if so, compare the parameters.
(4) Is it possible to perform stylistic analysis using the above classification and ranking methods? Analyze short texts, e.g. some poems and press texts, perform the ranking and the testing of parameters for difference and draw conclusions.
(5) Apply all your results to classification of verbs (nouns) in other languages.
(6) Make a preliminary statement on the generality of the hypothesis in Problem 6.1 in this volume.

References

Croft, W., Cruse, D.A. (2004). *Cognitive linguistics*. New York: Cambridge University Press.

Schwarz, M. (1992). *Kognitive Semantiktheorie und neuropsychologische Realität.* Tübingen: Niemeyer.

Wierzbicka, A. (1985). *Lexicography and conceptual analysis*. Ann Arbor: Karoma.

Yesypenko, N. (2009). An integral qualitative-quantitative approach to the study of concept realization in the text. In: E. Kelih, V. Levickij, G. Altmann. (eds.), *Methods of text analysis: 308-327*. Černivci: ČNU.

6.3. Sentence-length development in German

Problem

According to Wittek (2001), sentence-length becomes shorter in the development of German. The hypothesis has already been set up by different authors (cf. Lüger 1995) using different text sorts. Wittek uses geographic texts, fits to sentence length (measured in terms of number of clauses) the positive Poisson distribution and observing the behaviour of parameter *a* corroborates the hypothesis. Test the hypothesis for older and newer German texts.

Procedure

Wittek shows that the parameter *a* of the positive Poisson distribution

$$P_x = \frac{a^x}{x!(e^a - 1)}, \; x = 1, 2, 3, ...$$

decreases with time (on the average), leading to the conclusion deducible from the Poisson distribution that the number of short sentences increases. Wittek found the following numbers:

Time interval	Average values of the parameter *a*
1896-1905	2.07
1929-1933	1.67
1959-1960	1.26
1993-1994	1.04

Since there are great distances between the time intervals, it is necessary both to scrutinize texts in the intervals not analyzed up to now and examine texts before 1896 and after 1994.

Can we restrict the hypothesis to special sorts of texts or does it hold for German as a whole? If it holds only for some texts, substantiate this finding.

Since the parameter *a* is the mean of the distribution, one does not need to fit any distribution, it is sufficient to study the mean sentence length and its evolution.

If sentence shortening exists in German, does it exist in other languages, too? If so, what can be the cause of this phenomenon?

One can extend the analysis to different text sorts and study the velocity of shortening – if there is any.

Does the hypothesis hold, if one measures sentence length in terms of word numbers? However, for performing a procedure of this kind very long texts are necessary because the frequency of occurrence of individual lengths will not be reliable. At least, discuss this problem.

References

Lüger, H.H. (1995). *Pressesprache.* Tübingen: Niemeyer.

Wittek, M. (2001). Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 219-247.* Göttingen: Peust & Gutschmidt.

6.4. Verb development

Problem

Consider all verbs of a language. Partition this set into subsets whose elements are verbs designating a certain stage in the development of life on the earth. Study the hypothesis that the more developed a level, the more verbs it contains. That is, there will be a small number of verbs designating "being", more verbs designating "movement", etc. and the greatest number of verbs designating psychological states or acts.

Procedure

First obtain all verbs of a language from a great dictionary. Then set up a scale corresponding to the (ontogenetic or phylogenetic) development of actions in living organisms (e.g. *to eat* preceeds *to doubt*; *to be* preceeds *to seize*) . If necessary consult a biologist. Then ascribe each verb to one of the given classes. There are different verb classifications which can be very helpful for making decisions and scaling. At last, test the hypothesis that the activities of living organisms are the more differentiated the higher their developmental level.

References

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

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

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

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

Scheibman, J. (2002). *Point of view and grammar*. Amsterdam-Philadelphia: John Benjamins.

6.5. History

Some problems in quantitative linguistics have a long development and a rich history. Unfortunately, there are few works describing the evolution of the problem, the proposed solutions and the data on which tests were performed. De-

scribe the history of some of the problems mentioned below at least for one decade.

1. Word length study
2. Sentence length study
3. Other sentence properties (complexity, difficulty, depth)
4. Phoneme/grapheme/sound frequency
5. Canonical syllable/morph/word structure
6. Rank-frequency distributions (Zipf´s law)
7. Menzerath´s law
8. Vocabulary richness
9. Rhythm study
10. Glottochronology
11. Quantitative syntax
12. Comparison of texts
13. Phoneme distribution
14. Sequences of linguistic entities (e.g. time series, Markov chains, Fourier analysis, Köhler´s motifs,….)
15. Relations between language properties – synergetic linguistics
16. Quantitative typology
17. Applications of information theory
18. Chaos, fractals, dimensions
19. Frequency dictionaries
20. Quantitative semantics
21. Historical and comparative linguistics
22. Disputed authorship studies
23. History of QL in a selected state
24. Quantification of verb valency
25. The quantification of transitivity

Do not propose new solutions but emphasize the criticism of approaches found in the described works. Concentrate on methods, aims and explanations/interpretations of whatever kind pronounced by the authors included in your history. Describe the basic philosophy behind the works mentioned.

References

Glottometrics (Chapters: *History of Quantitative Linguistics*)

Grzybek, P. (2006). *Contributions to the science of text and language. Word length studies and related issues.* Dordrecht: Springer.

Kelih, E. (2007). *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft*. Hamburg: Kovač.

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

http://www.nslij-genetics.org/wli/zipf

6.6. Psychoanalytic word categories

Problem

In the Portuguese translation of the Bible, A. Wilson (2009) computed the frequency spectra of words originating from the primary process thought (sensation oriented, concrete) and separately those from the secondary process thought (logical, oriented to time, space and society) for all gospels and some other texts. He stated that all frequency spectra follow the right truncated Zipf (zeta) distribution. The result shows that this is another kind of stratification abiding by Zipf´s law. Show that adding the two above strata the frequency spectrum can be captured by Popescu´s (2009) stratification function consisting of two components, viz.

$$g(x) = 1 + a_1 \exp(-x/b_1) + a_2 \exp(-x/b_2)$$

where x is the frequency, $g(x)$ is the number of cases with frequency x and a_i, b_i are parameters.

Procedure

Using the above formula we consider the frequency spectrum a usual function (not probability function). Take Wilson's tables and simply add the individual frequencies $g(x)$ of the given text. Since the words of the primary and the secondary process are different, addition is allowed. Test the fitting of the above function using the determination coefficient (not chi-square).

Perform this procedure with all Wilson's tables, state whether the above hypothesis holds, then analyze other different texts in different languages. If possible, use Martindale's (1975) *Regressive Imagery Dictionary* and perform the analysis also for other word classes, e.g. "…the categories of body boundary definiteness developed in Wilson (2006) and the categories of anality and orality constructed by Vanheule, Desmet, Meganck (2008). Finally, the most challenging task would be to integrate these empirical findings into a synergetic theory of text production that incorporates cognitive elements such as consciousness states and concept-word mappings. Researchers such as Roy (2004) and Spivak (2004) have already made interesting contributions in this direction, but there is still much work that remains to be done in developing a comprehensive model" (Wilson 2009, 305).

Draw conclusions about the stratification of texts.

Replace Zipf´s function in separate processes by Popescu´s function using only one component, i.e. $g(x) = 1 + a_1 \exp(-x/b_1)$. Perform a comparison of parameters in individual texts. Is there a difference in parameters in the primary and the secondary processes?

References

Bucci, W. (1997). *Psychoanalysis and cognitive science: A multiple code theory.* New York: Guilford Press.

Elter-Nodvin, E. (2000). *Computerized content analysis: a comparison of the verbal productions of high hypnotizable, low hypnotizable and simulating subjects*. Ph.D. dissertation, University of Tennessee, Knoxville.

Hogenraad, R. (2005). The Regressive Imagery Dictionary: a test of five versions (English, French, German, Portuguese, and Swedish). *Paper presented at the International Congress on Aesthetics, Creativity, and Psychology of the Arts, Perm, Russia, June 2005.*

Kristeva, J. (1996/1997). Freudian models of language: A conversation. *Psychomedia: Journal of European Psychoanalysis, 3/4.* [Retrieved December 8, 2006 from [http://www.psychomedia.it/jep/number3-4/ kristeng. htm] .

Martindale, C. (1975). *Romantic progression: the psychology of literary history*. Washington, DC: Hemisphere.

Martindale, C. (1976). Primitive mentality and the relationship between art and society. *Scientific Aesthetics 1, 5-18*.

Martindale, C., Fischer, R. (1977). The effects of psilocybin on primary process content in language. *Confinia Psychiatrica 20, 195-202.*

Popescu, I.-I., Altmann, G., Köhler, R. (2009). Zipf´s law – another view. *Quality and Quantity 44(4), 713-731*.

Roy, P.K. (2004). Stochastic resonance as an emerging technique for neuron modulation and pharmacolinguistics: using nonlinear dynamics to analyze drug-induced language transition and EEG. *Journal of Quantitative Linguistics 11, 49-77.*

Spivak, D. (2004). Linguistics of altered states of consciousness: problems and prospects. *Journal of Quantitative Linguistics 11, 27-32.*

Vanheule, S., Desmet, M., Meganck, R. (2008). Anal and oral word use in relation to dependency and self-criticism. *Poster presentation at the Winter Meeting of the American Psychoanalytic Association, New York, 2008*.

Werner, H. (1948). *Comparative psychology of mental development.* New York: International Universities Press.

West, A. (1991). Primary process content in the King James Bible: the five stages of Christian mysticism. *Computers and the Humanities 25, 227-238.*

West, A., Martindale, C. (1988). Primary process content in paranoid schizophrenic speech. *Journal of Genetic Psychology 149, 547-553.*

West, A., Martindale, C., Hines, D., Roth, W. (1983). Marijuana-induced primary process content in the TAT. *Journal of Personality Assessment 47, 466-467.*

Wilson, A. (2002). The application of computer content analysis in sexology: a case study of primary process content in fictional fetishistic narratives.

Electronic Journal of Human Sexuality, 5. [Retrieved October 31, 2006, from: http://www.ejhs.org/volume5/wilson.html].

Wilson, A. (2006). Development and application of a content analysis dictionary for body boundary research. *Literary and Linguistic Computing 21, 105-110.*

Wilson, A. (2009). The well-formedness of two psychoanalytic word categories in Portuguese texts. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 285-307.* Černivci: ČNU.

6.7. Language of children

Problem

In a conversation of children (adults may take part), show that the distribution of the number of different speech acts (cf. *Problems Vol 2: 10.1. Frequency distribution of speech acts*) in uninterrupted utterances of individual children (i.e. length of utterances determined in terms of the number of speech acts) is distributed according to the positive (= zero truncated) negative binomial distribution, substantiate it linguistically and characterize the speech and status of individual children.

Procedure

Take a recorded conversation of children and for each child separately count the length of its uninterrupted participations in the conversation in terms of speech acts. That means X = number of speech acts in child's participation.

Set up the frequency distribution of this length for each child (and adult) separately.

Test whether each of the distributions follows the zero-truncated negative binomial distribution given as

$$P_x = \frac{\binom{r+x-1}{x} p^r q^x}{1-p^r}, \quad x = 1,2,3,...$$

where the parameters r and p must be estimated from the data, $q = 1 - p$. Show why this distribution is adequate.

Interpret the parameters and use them for ascribing a child its role (importance) in the conversation, its attitude or its intelligence, etc.

State how many times a special speech act occurred in the conversation, count the speech acts of all kinds and show that the distribution of speech act classes displays a certain regularity. Find the model of this regularity in form of a theoretical probability distribution, derive it from some suppositions and test it.

Use the parameters of the resulting distribution as characteristics of the conversation which may be cheerful, controversial, dramatic, sad, etc.

Compare the results of your analysis in several conversations. Use also (spoken) interviews and (written) stage plays.

References

Beles, R.F. (1950). *Interaction process analysis: a method for the study of small groups.* Cambridge, Mass.: Addison-Wesley.

Horvath, W.J. (1965). A mathematical model of participation in small group discussions. *Behavioral Science 10, 164-166.*

Kadane, J.B., Lewis, G.H. (1969). The distribution of participation in group discussion: an empirical and theoretical reapproach. *American Sociological Review 34, 710-723.*

Kadane, J.B., Lewis, G.H., Ramage, J.G. (1969). Horvath's theory of participation in group discussions. *Sociometry 32, 348-261.*

Nowakowska, M. (1978). Some formal aspects of dynamics of dialogues. *Glottometrika 1, 73-90.*

Nowakowska, M. (1978). A model of participation in group discussion. *Behavioral Science 23, 209-212.*

Rothe, U., Altmann, G., Wagner, K.R. (1992). Verteilung der Länge von Sprechakten in der Kindersprache. In: Wagner, K.R. (ed.), *Kindersprachstatistik: 47-56.* Essen: Die Blaue Eule.

Stephan, F. (1952). The relative rate of communication between members of small groups. *American Sociological Review 17, 482-486.*

Stephan, F., Mishler, E.G. (1952). A distribution of participation in small groups: an exponential approximation. *American Sociological Review 17, 598-608.*

Tsai, Y. (1977). Hierarchical structure of participation in natural groups. *Behavioral Science 22, 38-40.*

6.8. Vocalic language

Problem

The concept of "vocalic content" of a language has been defined by Altmann, Lehfeldt (1983) in 23 different ways. Process all these definitions/indicators linguistically and statistically.

Procedure

In the definitions, different categories (sample, unit, type, occurrence/position, frequency) are used. Interpret the definitions linguistically and set up a network linking them: perform analyses of several languages/texts/dictionaries using several (possibly all) indicators and scrutinize their correlations. If one of the indicators is not correlated with the other ones, find the cause of this phenomenon.

For each indicator (all of them are quantified) find at least (a) the interval in which it lies; (b) the asymptotic variance of the indicator. Eliminate or correct indicators whose interval is not satisfactory. Having the variance, set up an asymptotic normal test and compare the texts/languages/samples.

Correlate the "vocalism" indicators with other properties, e.g. morphological, syntactic, word length, etc. Extend the control cycle step by step and find the functions linking individual pairs of properties.

References

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

Avram, A. (1964). Sur la typologie phonologique quantitative. *Revue Roumaine de Linguistique 9, 131-141.*

Isačenko, A. (1939-40). Versuch einer Typologie der slavischen Sprachen. *Linguistica Slovaca 1, 64-76.*

Krámský, J. (1946-48). Fonologické využití samohláskových fonémat. *Linguistica Slovaca 4-6, 39-43.*

Krámský, J. (1972). On some problems of quantitative typology of languages on acoustic level. *Prague Studies in Mathematical Linguistics 3, 15-26.*

Nikonov, V.A. (1960). Konsonantnyj koeficient. *Lingua Posnaniensis 8, 228-235.*

Pierce, J.E. (1962). Possible electronic computation of typological indices for linguistic structures. *International Journal of American Linguistics 28, 215-226.*

Skalička, V. (1958). Typologie slovanských jazyků, zvláště ruštiny. *Československá rusistika 3, 773-84.*

Skalička, V. (1966). Konsonantenkombinationen und linguistische Typologie. *Travaux Linguistiques de Prague 1, 111-114.*

Torsuev, G.P. (1966). Raznovidnosti tipologii jazykov i pokazateli fonetičeskoj i fonologičeskoj tipologii. In: *Strukturno-tipologičeskoe opisanie sovremennych germanskich jazykov: 261-268*. Moskva: Nauka.

Trubetzkoj, N.S. (1939). *Grundzüge der Phonologie*. Prague.

Voegelin, C.F. (1957). Six statements for a phonemic inventory. *International Journal of American Linguistics 23,78-84.*

Yegerlehner, J. et al. (1957). Frequencies and inventories of phonemes from nine languages. *International Journal of American Linguistics 23, 78-84.*

Author index

Subject index