Problems
in
Quantitative Linguistics
5

by

Gabriel Altmann

2015
RAM-Verlag

Studies in quantitative linguistics

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.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language*. 2014, VIII + 123 pp.
17. G. Altmann, R. Čech, J. Mčutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014. VI + 231 pp.

18. M. Kubát, V. Matlach., R. Čech,, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106
19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015, III+146 pp.

ISBN: 978-3-942303-31-6

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

Preface

The present volume is a continuation of the series dedicated to all linguists who want to solve linguistic problems in a non-classical way. Elementary knowledge of statistics is a necessary condition, however, even a collection of data in the prescribed way could be helpful for solving some problems. The comparisons, tests, finding a function or distribution can be made by a statistician but the linguistic background knowledge must be furnished by the linguist.

The volume is appropriate especially for those who try to enter the field of quantitative linguistics and seek the door leading to elementary problems.

The present volume contains 90 problems. To each problem some references are recommended but the reader can solve them in his own way. Unfortunately, qualitative linguistics contains many concepts and classifications rooted in opinions and leading to different descriptions. In the present volume the reader is forced to perform tests which corroborate or reject the primary concept formation and force him to create new data based on different definitions, concepts, criteria etc. The basic requirement is the testing of everything one says.

It is recommended to publish the results in a quantitative linguistics journal. In any case, all numbers should be presented in order to give other linguists the possibility of testing other hypotheses or to subsume the accepted results in a deeper theory.

Gabriel Altmann

Contents

1. General Problems 1

1.1. The problem of the problem 1
1.2. Hierarchies in language 2
1.3. Diversification of a language family 4
1.4. Formal diversification 6
1.5. Ways into the depth 8
1.6. Length levels 10
1.7. Hypotheses 12
1.8. Distance and similarity 13
1.9. Irregularity 15
1.10. Problem continuation 16

2. Text 18

2.1. Frequency motifs 18
2.2. Runs 19
2.3. Sequences in text 21
2.4. Frequency sequences 25
2.5. The world view of language 26
2.6. Climax types 27
2.7 Continuous modeling of sentence length 29
2.8. Thematic concentration 1 30
2.9. Thematic concentration 2 31
2.10. Thematic concentration 3 32
2.11. Thematic concentration 4 33
2.12. Denotative-connotative concentration 34
2.13. Text compactness 35
2.14. Conceptual inertia of texts 1 36
2.15. Conceptual inertia of texts 2 37
2.16. Conceptual inertia of texts 3 38
2.17 Adjective-verb ratio and text indicators 39
2.18. Unified word length modeling 41
2.19. Poetic and rhetoric figures 1 42
2.20. Poetic and rhetoric figures 2 44
2.21. The world view of a writer 45
2.22. Adjectives in texts 47
2.23. Adjectival motifs 49
2.24. Stylistic centrality 50
2.25. Text sorts 51

2.26. Nominativity vs. predicativity 1 52
2.27. Nominativity vs. predicativity 2 53
2.28. Predication 54
2.29. Sentence length 55
2.30. Sequence of valencies 57

3. Grammar 60

3.1. Adnominal modifiers: Symmetry 60
3.2. Morphological complexity 1 61
3.3. Morphological complexity 2 62
3.4. Morphological changes and frequency 63
3.5. Word classes 64
3.6. Parts-of-speech distribution 66
3.7. Parts-of-speech homogeneity 68
3.8. Clause centrality 69
3.9. Clause types 71
3.10. Clause lentgth 72
3.11. Topic – comment 74
3.12. Study of adverbials 1 77
3.13. Study of adverbials 2 80
3.14. Study of adverbials 3 81
3.15. Grammatical categories 82
3.16. Reduplication 83
3.17. Sequences of syntactic constituents 85
3.18. Noun phrase 87
3.19. Syntactic tags as NP components 88
3.20. Word class specification 89

4. Stage play 92

4.1. Stage play: 1. Sentence length 92
4.2. Stage play: 2. Speech acts 93
4.3. Stage play: 3. Sequences of illocutive speech acts 95
4.4. Stage play: 4. Transition matrix of speech acts 97
4.5. Stage play: 5. Polysemy in the speech of persons 98
4.6. Stage play: 6. Distant reaction 99
4.7. Stage play: 7. Aggregation of speech acts 100
4.8. Stage play: 8. Act compactness 101
4.9. Stage play: 9. Verb activity 102

5. Phonemics and script 104

5.1. Phonological complexity 104
5.2. Script motifs 106
5.3. Phonic similarity of words in proverbs 107
5.4. A variant of the Skinner hypothesis 109
5.5. Syllable complexity 110
5.6. Euphony 111
5.7. Distribution of syllable types 112
5.8. Phonetic symbolism 114

6. Semantics 117

6.1. Abstractness of nouns 117
6.2. Abstractness in the text 118
6.3. Polysemy 119
6.4. Measurement of verb activity 121
6.5. Word class specification 123
6.6. Metaphor 124

7. Other problems 127

7.1. Morphological motifs 127
7.2. Sentence motifs 128
7.3. Borrowings 129
7.4. Syllabic word length 130
7.5. Clause length 132
7.6. Word length and number of compounds 133
7.7. Word frequency and number of compounds 134

Author register 136

Subject register 143

1. General problems

1.1. The problem of the problem

Problem

Consider some disciplines of linguistics (especially quantitative linguistics, computer linguistics, corpus linguistics, grammar, text linguistics) and state what kinds of problems they have and how they try to solve them. Describe this way from the philosophy-of-science point of view.

Procedure

Take the last two issues of a special journal devoted to these domains. Read the articles and state what kind of problems they solve, what are their methods and aims. Classify the problems and judge the state of the discipline according to the following criteria:

(a) Are they purely descriptive/classificatory or written merely in form of instructions (e.g. for the computer)?

(b) Do they perform some kind of quantification and measurement?

(c) Do they mention an explicit hypothesis?

(d) Do they try to set up mathematical models?

(e) Do they test the models statistically?

(f) Do they strive for establishing laws?

(g) Do they set up a theory as a system of derived and corroborated hypotheses?

(h) Do they strive for explanations whose beginnings start with problem (d)?

(i) Do the authors think deterministically or admit also probability?

Evaluate the epistemological role of the individual levels giving them scores and apply the scores to the articles read. You can compare linguistics also with other scientific domains in order to estimate its scientific status.

If you analyzed some modern articles and stated the theoretical level of the discipline (based on the given issue of the journal), show how it could be advanced in order to obtain the status of an empirical science. You may apply your investigation also to individual linguistic journals in their historical development.

Show how the measurement of individual properties mentioned in the article could be performed. Set up (but do not test) hypotheses and conjecture how they may be linked with other ones.

If necessary, propose a (testable) mathematical model for the given problem.

References

Bird, A. (1998). *Philosophy of Science*. London: Routledge.

Boyd, R., Gasper, P., Trout, J.D. (eds.) (1991). *The Philosophy of Science*. Cambridge, Mass.: The MIT Press

Bunge, M. (1967). *Scientific research I. The search for system*. Berlin: Springer.

Bunge, M. (1979). *Treatise on Basic Philosophy Vol. 4. A World of Systems*. Dordrecht: Reidel.

Bunge, M. (1983). *Treatise on Basic Philosophy Vol. 5. Epistemology & Methodology I: Exploring the World.* Dordrecht: Reidel.

Bunge, M. (1983). *Treatise on Basic Philosophy Vol 6. Understanding the World.* Dordrecht: Reidel.

Curd, M., Cover, J. A. (eds.) (1998). *Philosophy of Science. The Central Issues.* New York: W.W. Norton & Co.

Fraassen, B.C. van (1980). *The Scientific Image*. Oxford: Oxford University Press.

Fraassen, B.C. van (2008). *Scientific Representation: Paradoxes of Perspective*, Oxford: Oxford University Press.

Godfrey-Smith, P. (2003). *Theory and reality: an introduction to the philosophy of science.* Chicago, London: The University of Chicago Press.

Klemke, E., et al. (eds.) (1998). *Introductory Readings in the Philosophy of Science.* Amherst, New York: Prometheus Books.

Ladyman, J. (2002). *Understanding Philosophy of Science*. London: Routledge.

Lange, M. (ed.) (2007). *Philosophy of Science. An Anthology*. Malden, Mass.: Blackwell Publishing Co.

Rosenberg, A. (2000). *Philosophy of Science: A Contemporary Introduction* London: Routledge.

Salmon, M., Earman, J., Glymour, C., Lenno, J.G., Machamer, P., McGuire, J.E., Norton, J.D., Salmon, W.C., Schaffner, K.F. (1992). *Introduction to the Philosophy of Science*. Upper Saddle River, New Jersey: Prentice-Hall.

Salmon, W.C. (1971). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.

1.2. Hierarchies in language

Problem

In Wikipedia (http://en.wikipedia.org/wiki/Hierarchy; - 08.11.2013) one finds the following definition:

“A **hierarchy** is an arrangement of items (objects, names, values, categories, etc.) in which the items are represented as being "above," "below," or "at the same level as" one another. Abstractly, a hierarchy can be modeled mathematically as a rooted tree: the root of the tree forms the top level, and the children of a given vertex are at the same level, below their common parent."

Show at least five domains of language in which one can easily state the existence of hierarchies. Describe them, and if you find some regularities express them mathematically.

Procedure

Consider some domains of linguistics, e.g. dialectology, textology, syntax, morphology, lexicology and find the hierarchies. Some examples are:

Dialectology: official language – dialect – sociolect – idiolect
Textology: hreb – sentence – clause – word – morpheme
Syntax: sentence – clause – phrase – word
Morphology: word form – morpheme – morpheme polysemy
Lexicology: lexical chains and nets arising from hypernymy
Semantics: ordering according to abstractness/concreteness or generality/specificity
Material domain: Menzerath's law in all material domains.

Consider the existing literature, describe the individual levels in the hierarchy and find some hypotheses. If there is a hierarchy, then the higher level exerts influence on at least the next lower level. Find this dependence and express it quantitatively.

Find indicators for height, width, complexity etc. of hierarchical nets. Derive them from some general hypotheses and test them on data.

Generalize the results in such a way that you show the common features of the hierarchies, i.e. the analogy between hypotheses and the commonality of their mathematical form. Strive for a theory of linguistic hierarchy. If possible, show the boundary conditions for some domains.

Show the place of individual linguistic "schools" in treating the hierarchies.

Find analogies to other phenomena, e.g. in biology, physics or sociology.

Prepare a possibly complete list of references to the individual forms of hierarchy in linguistics and publish at least these lists. The domain of hierarchies is not an "official" domain of linguistics but it is a step towards theory.

Now set up your own hierarchy. Take for example a class of words and ascribe to each member of the class some property (qualitative or quantitative). Then to each member having the same value of the first property ascribe a second property in order to obtain a third level. Continue in this way as long as possible. Compute the properties of the tree, of the paths, of the net. Then take another class and perform the same procedure, etc. At last, compare the trees, paths, nets, find their common features and derive a hypothesis. Test the hypothesis using your data and find a general feature of linguistic hierarchies. Do not forget boundary conditions!

References

Ahl, V., Allen, T.F.H. (1996). *Hierarchy Theory*. New York: Columbia University Press.

Allen,T.F. (2006). *A Summary of the Principles of Hierarchy Theory.* at:www.isss.org/hierachy

Allen.T.F.H., Hoekstra, T. (1992). *Toward a unified ecology*. New York: Columbia University Press.

Allen, T.F.H., Starr, T.B. (1982). *Hierarchy: Perspectives for Ecological Complexity*. Chicago & London: The University of Chicago Press.

Gaume, B., Venant, F., Victorri, B. (2006). Hierarchy in lexical organisation of natural languages. In: Pumain, D. (2006): *121-142***.**

Memar, P. (2009). *Hierarchie in der Baukunst. Architekturtheoretische Be trachtungen in Ost und West.* Mainz am Rhein: von Zabern.

Mesarović, M.D. et al., (1970). *Theory of Hierarchical, Multilevel Systems*. New York: Academic Press.

O'Neill, R.V., DeAngelis, D., Waide, J., Allen, T. F. H. (1986). *A hierarchical concept of ecosystems*. Princeton: University Press.

Pattee, H.H. (ed.) (1973). *Hierarchy Theory: The Challenge of Complex Systems.* New York: George Brazillier Publisher.

Pumain, D. (2006). *Hierarchy in Natural and Social Sciences*. New York: Springer-Verlag.

Simon, H.A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society (Philadelphia) 106 (6), 467–482*

Weiss, P.A. (1971). *Hierarchically Organized Systems in Theory and Practice*. New York: Hafner Publishing Company.

Whyte, L.L. et al. (eds,) (1969). *Hierarchical Structures*. New York: American Elsevier Publishing Company.

1.3. Diversification of a language family

Problem

Every language family diversifies. For comparison one mostly uses the similarity in the lexicon that is, merely a surface phenomenon. Perform different comparisons, show the family as a graph with weighted edges and draw consequences.

Procedure

Every property of language can be quantified and measured. Take, say, 20-50 sentences of the same text from each member of the language family. One can always find texts of this kind.

Consider the following properties:

(1) For each sentence in every language state the number of lemmas and define the difference as the mean difference in all sentences.

(2) Decompose the text into morphemes and state for each sentence (in two languages) the number of etymologically identical morphemes. Construct an indicator of similarity.

(3) For each sentence separately state the difference in the occurrence of grammatical categories. Set up an average measure of similarity.

(4) Study the difference in the word order comparing merely word-forms. Set up an indicator of difference and perform comparisons taking averages for any pair of languages.

(5) Study the differences in the use of parts-of-speech sentence by sentence.

Each comparison of two languages results in a vector of differences between identical sentences. Use the vector for computing the similarity/ divergence.

For each similarity/difference indicator derive its sampling properties and define a statistical test for establishing the significance of the divergence.

Set up a battery of hypotheses concerning the diversification of a language family in general, then those concerning only the family you analyzed.

Take into account different other properties and compare the texts sentence by sentence. Use e.g. the corrected indicators introduced by Greenberg. State which properties are more stable than other ones. Hypothesize why.

At last, venture the comparison of the same text in two non cognate languages. Some of the above properties can easily be applied. Do not use different texts and do not establish premature typological statements. Care for statistically correct comparisons. Do not compare religious texts.

References

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

Bisang, W. (2001). *Aspects of typology and universals*. Berlin: Akademie Verlag.

Comrie, B. (1989). *Language universals and linguistic typology.* Chicago: Chicago University Press.

Croft, W. (2002). *Typology and universals.* Cambridge: Cambridge UP.

Cysouw, M. (2005). *Quantitative methods in typology.* In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics: An International Handbook*: 554-578. Berlin: Mouton de Gruyter.

Eckert, G. (1986). *Sprachtypus und Geschichte. Untersuchungen zum typologischen Wandel des Französischen.* Tübingen: Narr

Finck, F.N. (1910). *Die Haupttypen des Sprachbaus*. Teubner: Leipzig.

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

Greenberg, J.H. (ed.) (1966). *Universals of language*. Cambridge, Mass.: The M.I.T. Press, Haarmann, H. (1976). *Grundzüge der Sprachtypologie*. Stuttgart: Kohlhammer.

Haspelmath, M.H., König, E., Oesterreicher, W., Raible, W. (eds.) (2001). *Language typology and language universals: An international handbook*. 2 vols. Berlin: de Gruyter.

Hinrichs, U. (ed.) (2009). *Die europäischen Sprachen auf dem Wege zum analytischen Sprachtyp*. Wiesbaden: Harrassowitz.

Ineichen, G. (1991^2). *Allgemeine Sprachtypologie*. Darmstadt: Wissenschaftliche Buchgesellschaft.

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

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

Lang, E., Zifonun, G. (eds.) (1995). *Deutsch – typologisch*. Berlin-New York: de Gruyter.

Lehmann, W.P. (1978). *Syntactic Typology: Studies in the Phenomenology of Language*. Austin: University of Texas Press.

Lewy, E. (1942). *Der Bau der europäischen Sprachen*. Tübingen: Narr (1964).

Nichols, J. (1992). *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Ramat, P. (1987). *Linguistic typology*. Berlin-New York-Amsterdam: Mouton de Gruyter.

Roelcke, Th. (1997). *Sprachtypologie des Deutschen*. Berlin-New York: de Gruyter.

Roelcke, Th. (2011). *Typologische Variation im Deutschen. Grundlagen – Modelle – Tendenzen*. Berlin: Schmidt Verlag.

Seiler, H. (ed.) (1978). *Language universals*. Tübingen: Narr.

Shopen, T. (ed.) (1985). *Language typology and syntactic description. 3 vols*. Cambridge: Cambridge Univeristy Press.

1.4. Formal diversification

Problem

Words (or better, stems) may diversify in different directions: there are phonemic variants like assimilations, change of a phoneme in a different morphological construction; morphological variations like inflections or intro-flections; derive-ations by means of affixes, and composition with various other stems. Restrict the investigation to one type of diversification, state the respective numbers of forms for each stem and construct individual distributions for: 1. Number of forms of individual stems, 2. Separately, the number of possibilities to build verbs, nouns, adjectives, etc. 3. If the stems belong to some part of speech, state the distributions within individual POS. You may adhere to the classical Latin classification of parts of speech. 4. Find a theoretical model for each distribution you obtain and substantiate it linguistically. 5. Comparing two good dictionaries study the development of a language.

Procedure

Take a dictionary and first find all phonological variants of the stems. It is sufficient to consider in the dictionary only stems beginning with the same letter. If you analyze your mother tongue, the procedure is simpler because you need not perform a mechanical search for each stem. Then consider x = the number of phonetic variants, $f(x)$ = the number of stems having x variants. Set up a distribution and find a model.

For each stem you identified, state the number of parts of speech in which it may appear, e.g. the German word *Tag* (day) may be transformed in an adjective/ adverb (*täglich*), verb (*vertagen*), pure adverb (*tagsüber*), noun (*Vortag*) and can be found in a number of compounds that can be found in some dictionary (*Feiertag,* names of days of the week, *Parteitag,* …). Let x = number of POS in which it may penetrate by some morphological procedure or the number of all forms your found, $f(x)$ = number of stems having x realizations. Set up the distribution, find a model and substantiate it.

The linguistic substantiation can be realized by finding another property of words and its relation to some of the parameters of the given distribution or to its mean etc. For example "morphological complexity of the language" expressed quantitatively.

A comparison of languages may be performed also by comparing the resulting distributions. Compute some properties of the distribution, express them by indicators and at least order the languages.

The evolution of the given language – seen from this point of view – can be studied using the changes in the given distributions. To this end two dictionaries published in different years or "the same" dictionary in some of the later editions may be employed.

References

Hanulíková, A., Davidson, D.J. (2009). Inflexional entropy in Slovak. In: Levická, J., Garabík, R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar: 145-151.* Brno: Tribun.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-775.* Berlin: de Gruyter.

Kostić, A., Mirković, J. (2002). Processing of inflected nouns and levels of cognitive sensitivity. *Psihologija 35, 287-297.*

Krajewski, G. Lieven, E.V.M., Theakston, A.L. (2012). Productivity of a Polish child's inflexional noun morphology: a naturalistic study. *Morphology 22, 9-34.*

Mačutek, J., Čech, R. (2012). Frequency and declensional morphology of Czech nouns. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: 59-68.* Belgrade: Academic Mind.

Milin, P., Filipović Durdević, D., Moscoso del Prado Martin, F. (2009). The simultaneous effects of inflexional paradigms and classes of lexical recognition: Evidence from Serbian. *Journal of Memory and Language 60, 50-64.*

1.5. Ways into the depth

Problem

Show that in linguistics the way into the depth is analogous to that in physics. But in linguistics we are always engaged with concepts, while physicists have more compact entities. If you come at a relative bottom, begin to theorize.

Procedure

In order to illustrate this procedure, consider the distribution of parts of speech in a language. Usually, one obtains 9-11 classes – as prescribed by the Latin grammar – but there are also systems with 100 classes.

Now consider only one of the classes, e.g. the adverbs. Again, one finds about 10 classes (place, time, mode, aim,…) – that is, concepts which allow us to perform a classification. If one orders the adverbs found in a text (or in a corpus) in the prescribed classes, one can search for the distribution of adverbial classes. The simplest ordering is according to the rank-frequency of classes but one can devise a number of other ordering criteria from the semantic point of view.

Now, omitting all but one class, one can study the logic of this unique class. If one considers e.g. adverbs of location, then location itself can be ordered. Some languages have special means for performing spatial orientation. How is the space oriented? Do the numbers obtained mirror this ordering? Is it possible to find a three-dimensional order? Does Man stay in the center? How is it with other adverbial classes?

The next step is, again, the reduction of the given class and considering merely one of the adverbs. It occurs in different environments (= polytexty) and displays a polysemy which can be found also in translations of the pertinent sentences into various other languages. Again, find the rank-frequency distribution of the individual meanings of the given adverb (= meaning diversification).

Now take that meaning of the adverb which is represented by the most occurrences, i.e. the first in the ranking scale of polytexty. Each occurrence may be realized in different contexts. Are all contexts identical or do some of them occur more or less frequently? Classify the contexts. Then set up the rank-frequency distribution of the polytexty of individual occurrences of the same meaning. Do it separately also for the other meanings. Find for all their rank-frequency distributions and show whether it is the same model or whether something changes when one goes to higher ranks?

Up to now, we passed 5 stages, i.e. we made 5 steps into the depth. Is it possible that the same frequency regime rules at all stages? If so, what does change in the model? Necessarily, the parameters obtain different values, but perhaps some parameters must be added, some may be omitted. Can one set up, say, a differential equation in which the parameters can be interpreted as representatives of Köhlerian requirements?

The way is in no case finished. We have a rank-frequency of polytexties and take, say, the most frequent class. What kinds of texts do we have? May we distinguish special classes of them? If so, then the given frequency can be represented as the frequency of text sorts in which the adverb (having the given meaning) occurred.

But now we made a step to text sort classification and reached a quite different domain. At each step in the hierarchy there are different steps possible according to our interest. Text sorts have properties which may be linked (or not) with those scrutinized by us. If we take – at whatever step – another frequency class, we may obtain a different result.

At last, there will be a net of links which will never be ready. This circumstance is caused not only by the extreme complexity of language but also by the fact that there are few linguists interested in this ladder into the precipice of our concept formation. The individual links must be derived and tested on many texts and languages in order to obtain language laws.

Needless to say, rank-frequency is only one of the ways that can be gone. The procedure can be performed also without rank-frequencies but one must have at least one property that can be traced down into the depth. Whatever way one takes, one will run against a boundary at which deductive work must necessarily begin.

The infiniteness of this enterprise is evident.

References

Ahl, V., Allen, T.F.H. (1996). *Hierarchy theory, a vision, vocabulary and epistemology*. Columbia University Press.

Allen, T.F. (2013). A summary of the principles of hierarchy theory. http://www.isss.org/hierarchy.htm (14.11.2014).

Altmann, G. (2006). Fundamentals of quantitative linguistics. In: Genzor, J., Bucková, M. (eds.), *Favete linguis: 15-27*. Bratislava: Slovak Academic Press.

Altmann, G., Dömötor, Z., Riška, A. (1968). Darstellung des Raumes im System der slowakischen Präpositionen. *Jazykovedný časopis 19, 25-40.*

Altmann, G., Dömötor, Z., Riška, A. (1968). The partition of space in Nimboran. *Anthropological Linguistics 8, 1-10.*

Altmann, G. (2014). Supra-sentence levels. *Glottotheory 5(1), 25-40.*

Coloma, G. (2014). Towards a synergetic statistical model of language phonology. *Journal of Quantitative Linguistics 21(2), 100-122.*

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

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

Salthe**, S.** (1985). *Evolving hierarchical systems: their structure and representation***.** New York: Columbia University Press.

1.6. Length levels

Problem

Investigate the problem of language levels based on length of pertinent entities in texts. Perform the study for every text separately. Compare texts, text sorts, languages.

Procedure

First define all language entities that have some measurable material length. The best known entities are syllable, mora, morpheme, word, rhythmic unit, phrase, clause, sentence, verse, speech act. You can define also classes, e.g. nouns, verbs, simple sentences, and even the size of classes in a classification, i.e. instead of length you study the cardinal numbers of special sets.

Take a single text and set up the distribution of specific length, that is, state the frequencies of entities having length 1,2,3,… State the length always in terms of immediate constituents, do no omit a level, i.e. do not compute e.g. the length of words in terms of phoneme numbers but either in syllables or in morphemes! Then using software fit the Zipf-Alekseev function $y = cx^{a + b \ln x}$ to the data, i.e. *frequency* = *f*(*length*). Having done this for all levels, state the value of the parameter *a* at individual levels. Does it change regularly when you pass from one level to the next? The lowest level is the phonic one, the highest can be considered e.g. that of speech acts or even the meaning set of the given entities. Hrebs can stay over all levels.

Now perform the same operation for several texts. Take the same level in all texts and study the relationship between the parameters *a* and b, i.e. $b = f(a)$. Can you state some regularity?

Now take the same entities in all texts and using the resulting formula compute the average of the parameter *a* (same level!) in all texts. Does average *a* change regularly with the change of level (from phonetic to semantic)? Express this change by a formula and substantiate it linguistically.

If possible, perform the same investigation in another language and compare the results.

If you want to make the next theoretical step, consider any other property of the given unit, that is a special level, and search for its link to the parameter *a* you obtained for length distribution. Take inspiration from language synergetics.

Remark. The relation length-frequency is an integer part of language synergetics. According to Zipf, it is rather length that adapts to frequency; here we go the opposite way because with higher units like sentences the Zipfian way is not adequate.

Define new types of entities either theoretically or by classification. For example, study separately the length of individual parts of speech – either using a dictionary or using a text. Do not mix texts, perform each count separately. State whether they differ. If so, order the classes according to parameter *a* and interpret the order linguistically. For some entities, e.g. types of speech acts, there is still no ordering. Study the length of individual classes and set up an order.

Compare languages, text sorts and perform a numerical classification.

In order to make the problem more practical, here some tasks: State

(1) the number of morphs in terms of phoneme numbers;
(2) the number of syllables in terms of phoneme numbers;
(3) the number of words in terms of syllable numbers;
(4) the number of words in terms of morpheme numbers (count also phonemically not realized morphemes: zero morphemes);
(5) the number of phrases in terms of word numbers;
(6) the number of compounds in terms of stem numbers;
(7) the number of clauses in terms of phrase numbers;
(8) the number of sentences in terms of clause numbers;
(9) the number of speech act chains in terms of speech acts;
(10) the number of rhythmic units in terms of syllable numbers;
(11) the number of verses in terms of syllable numbers;
(12) the number of verses in terms of word numbers;
(13) consider stepwise all units and set up Köhlerian motifs, i.e. non-decreasing sequences of lengths as they occur in text;
(14) compute for each data the above Zipf-Alekseev function and study the parameter *a*.
(15) Define new units and study their length. Extend the study of hrebs and show which entities can constitute hrebs. See also the problem *Hierarchies in language* in this volume

References

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

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

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., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin-Heidelberg: Springer.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Lupea, M., Tatar, D., Altmann, G. (2015). *Quantitative analysis of poetic texts.* Berlin/Boston: de Gruyter Mouton.

Software: NLREG, TableCurves

1.7. Hypotheses

Problem

Study the properties of linguistic hypotheses considering all aspects known from the philosophy of science.

Procedure

Take 10 well known hypotheses from qualitative linguistics (structuralism, generative linguistics, historical linguistics, semantics, dialectology, etc.) and study their properties. Bunge (1967, Vol, 3: 222-291) explains the following aspects:

(1) Formulation
(2) Range
(3) Inferential power
(4) Order
(5) Precision
(6) Predicates
(7) Inception
(8) Ostensiveness
(9) Depth
(10) Ground
(11) Level of conjecture
(12) Testability
(13) Logical strength
(14) Function

Then take some hypotheses from quantitative linguistics, show their status scrutinizing the above points and show the differences.

At last, consider only one of the hypotheses and study its history. How did it begin and what is its state today? Examine especially the development from description through conjecture to law. Describe the history of a law.

Consider especially some hypotheses from synergetic linguistics (Köhler 2005) which have today the status of laws. Originally, they were formulated only qualitatively. Show their development and at every stage refer to one of the above mentioned 14 points. You can study also the development of each point separately.

References

Bunge, M. (1967). *Scientific Research I. The Search for System*. Berlin/Heidelberg/New York: Springer.

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

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

1.8. Distance and similarity

Problem

According to Skinner’s hypothesis (1939, 1941, 1957), parts of a text positioned in mutual vicinity are phonetically more similar than distant ones. This is given by the activation of special brain processes. The hypothesis has been positively tested many times not only in the domain of phonetics. Test the hypothesis that similarity decreases with increasing distance, applying it to grammatical phenomena.

Procedure

First define a grammatical phenomenon, e.g. parts of speech, types of sentences, types of clauses, types of phrases, degrees of predication, types of speech acts, length of sentences, dependence structure, word and morpheme complexity, valency of verbs, etc.

Then rewrite the text in terms of sentences, i.e. each sentence is a unit with the given properties or structure.

Define an indicator of similarity. If you compare numbers (e.g. degrees, lengths, etc.), you can use any of the known indicators. If you compare symbols, structures, sequences, sets, you must use different indicators. Compute the similarity between entities positioned in distance 1, 2, 3,… (in terms of sentence or verse numbers positioned between the repetitions) and take a mean similarity for each distance.

Consider the means and present them in form of a function. Find the respective function. You can begin inductively, trying to find an adequate representation using software. If you obtain a similar result in many cases, begin to model the phenomenon and substantiate it linguistically, neurologically, psycholinguistically, etc. Search for boundary conditions bringing about stylistic, text-sort, language level and other differences. In the first step, identify the boundary condition adding a parameter to your function.

Strive for deciphering this mechanism as thoroughly as possible. First, find a function expressing this relation (distance vs. similarity) on each level separately. Then study the form of the given function, e.g. the change of parameters according to the level and different units within the level (phonetics, grammar, semantics). Strive for finding a law. To this end you must perform many tests and place the discovered regularity in a control cycle (cf. e.g. Köhler 2005)

References

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

Altmann, G., Köhler, R. (2015). *Forms and degrees of repetitions in texts. Detection and analysis*. Berlin/Munich/Boston: de Gruyter Mouton.

Cha, Sung-Hyuk (2015). Comprehensive survey on distance/similarity measures between probability density functions.
http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.154.8446 (20.05.2015)

Deza E., Deza M.M. (2006). *Dictionary of distances*. Elsevier

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

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

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

Popescu, I.-I., Lupea, M., Tatar, D., Altmann, G. (2015). *Quantitative analysis of poetic texts.* Berlin/Boston: de Gruyter Mouton.

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

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

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

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

1.9. Irregularity

Problem

Irregularity of an entity may be measured absolutely or locally. “Absolutely” means taking into account all changes that are possible with an entity (in morphology, composition, sentence, etc.), “locally” means the number of changes which are actually present with an entity as used in the given text or conversation. Count the numbers of irregularities with each word separately, set up the distribution and find a model of the distribution. You may use some “basic form” of the word or of the stem and compare the topical form with it (cf. Corbett et al. 2001; the problem *Distances and similarity* in this volume). Corbett et al. (2001) studied in this way Russian.

Procedure

Take a text and measure the number of topical changes of each word and at the same time the number of its absolute (possible) changes. Construct the sequence of the given numbers. Write each sentence in a separate line.

(1) Study the complete sequence of the text and find some of its properties. You can use any type of indicator or a time series. Do it both for absolute and for local irregularities.

(2) Study the distribution of irregularities in the complete text and find a preliminary model. Do it both for absolute and for local irregularities.

(3) Do the same with the translation of the text in another language and compare the languages. Do it both for absolute and with local irregularities.

(4) Study the distances between equal irregularities, characterize them by an indicator and find their distribution. Do it both for absolute and local irregularities.

(5) Now consider the individual sentences. For each of them you have a vector of irregularities. Compare the vectors of two neighboring sentences (i.e. those whose distance is 1 step) computing any of the known similarity measures. Then compute the *mean* of the similarities for this first step. In the next step, compute the similarity of each pair of sentences in distance 2, i.e. separated by one sentence. Compute again the mean similarity in distance 2. Continue increasing the distance. At last, you obtain a series of mean similarities for distances 1,2,3,… State whether the hypothesis “the greater the distance the smaller the similarity” holds for this aspect. Find a mathematical expression of the curve. This is a special case of the well known Skinner hypothesis applied to higher than phonetic level.

(6) Compare the results with those in other languages (same text) and state whether the function found in the given text holds also for other texts of the same language.

References

Cha, Sung-Hyuk (2015). Comprehensive survey on distance/similarity measures between probability density functions. http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.154.8446 (20.05.2015)

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

Deza E., Deza M.M. (2006). *Dictionary of distances*. Elsevier

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

Schreuder, R., Baayen, H. (1997). How complex simplex words can be. *Journal of Memory and Language 37(1), 118-139.*

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

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

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

Zörnig, P. (1984). The distribution of distances between like elements in a sequence, part I. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6, 1-15;* part II. In: U. Rothe (ed.), *Glottometrika 7, 1-14*. Brockmeyer, Bochum.

Zörnig, P. (1987). A theory of distances between like elements in a sequence. In: I. Fickermann (ed.), *Glottometrika 8, 1-22*. Brockmeyer, Bochum.

1.10. Problem continuation

Problem

The present problem is very difficult but it can give you many perspectives. Take the omnibus volume: *Quantitative Linguistics. An International Handbook.* Berlin: de Gryuter (2005), read several individual articles concerning one special domain and for each of them show the research continuation. What could and should be made in order to develop the given problem?

Show what type of theory may/must be developed in order to make the problem itself more theoretical. Some articles care for the linguistic substantiation of the background, other ones merely describe and apply some model. In the first case, formalize the problem, subsume it under a theoretical background; if necessary and possible, set up the differential equation or formulate a stochastic process which gives rise to the given phenomenon.

In the second case, collect the literature concerning the problem, prepare a survey and substantiate the problem and the applied model linguistically. Extend the testing to several languages, search for boundary conditions, show the problem of data collecting, and insert the hypotheses you obtained in the Köhlerian control cycle.

Devote special attention to the requirements of speaker and hearer by which the given type of link between properties is created. Use systems theoretical graphs and strive for a synergetic substantiation.

References

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

Köhler, R. (2003). *Semiotik und Synergetik.* In: Posner, R., Sebeok, Th.A. (eds.), *Semiotics. A Handbook on the Sign-Theoretic Foundations of Nature and Culture: 2444-2452.* Berlin/New York: de Gruyter.

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

Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). *Quantitative linguistics. An international handbook.* Berlin: de Gryuter.

Roelcke, Th. (2005). Sprachliche Ökonomie: Kommunikative Effizienz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). *Quantitative linguistics. An international handbook: 775-791.* Berlin: de Gryuter.

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

2. Text

2.1. Frequency motifs

Problems

Study the properties of frequency motifs in texts. Characterize them using known or your own indicators and finally show the interrelations between pairs of properties.

Procedure

Take a text and state the frequencies of some entities, e.g. types of syllables, syllable lengths, word lengths, individual lexemes or word forms, grammatical categories, parts-of-speech, polysemy of the words, morphological complexity of the words, sentence length (measured in terms of clause numbers), etc. Then transcribe the text in terms of these frequencies, i.e. replace each entity by its frequency. You obtain a sequence of numbers, just as in the problems *Frequency sequences* and *Sequences in text.* Now construct frequency motifs: a motif is defined as a non-decreasing sequences of numbers. If one has, e.g. the sequence 1,2,8,3,9,2,1,1, one obtains the motifs <1,2,8>; <3,9>; <2>; <1,1>.

Study the following problems:

(1) State the frequency of individual motifs and set up their rank-frequency distribution. Find a model of this distribution and compare different texts, e.g. using the chi-square test for homogeneity..

(2) State the lengths of individual motifs represented by the number of elements in them, e.g. the motif <1,2,8> has length 3; set up the length disribution and find a model of this distribution. Compare several texts, compare text-sorts and languages.

(3) State the average length of the elements in each motif and set up a new sequence. E.g. the mean of the motif <1,2,8> is (1 + 2 + 8)/3 = 3.67. Apply the methods mentioned here to the new sequence.

(4) Compute for each motif its range, i.e. the difference of the first and the last element, e.g. in <1,2,8> the range is 8 – 1 = 7. Replace the motifs by their ranges and study the new numerical sequence.

(5) Set up the discrete distribution of ranges and find a model of this distribution. Compare texts.

(6) For all distributions you obtained up to now, compute their entropy and Repeat rate. State the values of these indicators for different texts and text-sorts. Make a table of Ord's criterion <I, S> and show that the points are positioned in a small two-dimensional space. Compute the ellipse enclosing them or find the straight line if possible.

(7) State the inventory of motifs and set up a two-dimensional contingency table in which the frequencies of transitions from individual motifs to all the others are captured. Study this table using all the methods you know. State the (in)dependence of transitions, the symmetry of transitions, the strength of the diagonal, and the conspicuosity of individual cells. Search for reasons leading to these results. Compare texts, text-sorts and languages.

(8) Perform the same operations also with averages of motifs and ranges of motifs and draw consequences. Use different linguistic entities and different properties. Which properties display some regularities of motifs?

(9) Consider only the length of motifs (of any entity and property) and study the distribution of their runs.

(10) Study the same motifs in translations of the given text into other languages and draw consequences from the differences. Does this aspect have some relations to other properties of the given languages?

References

See *Köhler-motif* in *Problems Vol 1, 2, 3.*

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

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

Popescu, I.-I., Zörnig, P., Grzybek, P., Naumann, S., Altmann, G. (2013). Some statistics for sequential text properties. *Glottometrics 26, 50-94.*

2.2. Runs

Problem

Runs are uninterrupted sequences of some identical entities. Hence they can be stated only in texts. Consider five properties of the word and set up hypotheses concerning their run-behavior in texts. Test the hypotheses using various texts and if possible construct a control cycle of the serial behavior of these properties.

Procedure

Begin with the following properties of words: length, frequency, polysemy, synonymy, morphological complexity. Taking the first property, proceed as follows: (a) measure the length of each word in terms of syllable numbers and replace each word in the text by its length. You obtain a sequence of numbers. (b) State the number of runs and test whether this number is significantly large or significantly small. (c) Test whether the longest run is significantly long or not. (d) Test whether the sequence of run lengths displays some regularity; if so, set up an a posteriori hypothesis. Either there is a hierarchic construction beginning with runs and continuing with some of their "higher" behaviors or not. Express the hypothesis mathematically.

As to frequency, replace each word of the text by its frequency taken from a corpus, from a frequency dictionary or directly from the given text. There is surely a distribution of word frequencies. Do not consider word forms but lemmas. Compute the average frequency and partition the words in two classes: smaller than the mean frequency (A) and larger than the mean frequency (B). You obtain a sequence of letters A and B. Perform all the tests concerning runs and set up a hypothesis concerning too many or too few runs. The hypothesis has something to do with the type of language, or with the text-sort to which the given text belongs, or with the style of the author, or with the epoch in which it has been written. If you analyze many texts, set up a "frequency-runs" classification of texts.

Polysemy is the number of meanings of the given word in the usual monolingual dictionary. Replace the words by their polysemy values and study the behavior of runs in the text. Again, you can consider the average polysemy and partition the words in two classes or you can scrutinize the raw sequence. Is the number of polysemy runs too large, and if so, why? Set up a hypothesis, derive it in analogy to the above ones and show whether runs of length, frequency and polysemy can be integrated in a control cycle analogous to that presented by Köhler (1986, 2005).

Synonymy can be taken from a dictionary of synonyms which exist for many languages. Replace the words by their numerical synonymies. Perform the same analysis of runs as with polysemy. Then set up the distribution of run lengths both for dichotomized classes (larger/smaller than the average) and for raw runs. Do the same for the polysemies and study the difference between the distributions. If there is no significant difference, set up a hypothesis concerning the relation between polysemy and synonymy runs. Use the analogy to the Köhlerian cycle.

Morphological complexity of a word can be measured in many different ways. Use the simplest one: complexity means the number of morphemes in the word. Do not count "zero"-morphemes but count introflection, e.g. German *Vater* vs. *Väter*. The first word contains 1 morpheme, the second 2. The word *Vätern* contains 3. Take into account the fact that some compounds may be written separately, e.g. *Ministry of Foreign Affairs*, and represent 1 word with 6 morphemes. You may make such decisions in any way but you must describe exactly how the morphemes have been defined – for the sake of the comparability of your count with other ones. Then replace the words in the text by their numerical morphological complexity and study the runs. Find the relationship between the runs of word length and those of complexity, i.e. set up the frequency distributions of both kinds of runs and compare them using e.g. a chi-square test.

Having finished your computations, show that possibly all these run kinds are somewhat linked. First make (theoretical) conjectures which may represent hypotheses, than derive a formula for the link, test it and begin to draw a graph. Continue making further conjectures, use other texts and construct step by step a "teorita". Then consider further properties of words.

Find all chapters in this book concerning motifs. Then consider the sequences you obtained and segment them into motifs. Perform all the operations mentioned in the given chapters and explain the difference between runs and motifs.

References

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag.

Gibbons, J.D**.** (1985). *Non-parametric statistical inference*. New York: Decker.

Godbole, A. P., Papastavridis, G. (Eds.) (1994). *Runs and Patterns in Probability: Selected Papers.* New York: Kluwer.

Köhler, R**.** (1986). *Zur linguistischen Synergetik. Status 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: de Gruyter.

Lin, Yi-Ling, Jayawardhana, A**.** (2001). Theory of Runs. February 22, 2001. http://faculty.pittstate.edu/~ananda/STATMETHODI/Theory-of-runs.pdf (11.1.2013)

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol. 1.* Lüdenscheid: RAM-Verlag.

2.3. Sequences in text

Problem

Any linguistic units whose inventory is not infinite (as e.g. that of clauses, sentences) are repeated in text. However, the repetitions may underlie different regularities, trends, rules, oscillations, runs, distances or they may be chaotic or random. Analyze the regularities (or irregularities) considering the text as a sequence of units.

Procedure

Use one of the following entities

1. Sound types: according to place or manner of articulation or both
2. Syllable types: V, VC, CV, CVC, VCC, VCV,….
3. Syllable lengths (in terms of phoneme numbers)
4. Morpheme types: proclitic, prefix, stem, infix, introflection, suffix, postclitic, reduplication
5. Morph length (in terms of phoneme numbers)
6. Word classes (a) parts-of-speech: Noun, Verb, Pronoun, Adverb, Adjective, Preposition, Postposition, Interjection, Conjunction, Article, Particle, Numeral or one can use a syntactic definition with dozens of classes. (b) Involving: stem, derived, reduplicated, compound, inflected, derived-inflected, compound-inflected, compound-derived, compound-derived-inflected, reduplicated-inflected, etc.
7. Word length (in terms of syllable numbers)
8. Clause types (main, relative, causal,…)
9. Clause length (in terms of word numbers)
10. Sentence types (according to different criteria)
11. Sentence lengths (in terms of clause numbers)
12. Hreb members (or references)
13. Types of speech acts
14. Equal frequencies of (also different) words, i.e. sequence of frequencies
15. Alliteration (both in prose and poetry)
16. Assonance (repetition of vowel sequences)
17. Verb valency (cf. the problem *Sequence of valencies*)
18. Degree of verb activity (scaling!)
19. Types of noun attributes
20. Grammatical categories
21. Individual markers of a category (e.g. individual cases; times; numbers; persons,…)
22. Polysemy (= number of meanings of the given word in the dictionary)

First describe and capture quantitatively at least one of the different phenomena by evaluating many texts, i.e. take a property and transcribe the texts in terms of the given entities. Evaluate the repetition in form of distributions, runs, distances, auto-correlations, motifs, etc. If they represent numbers (i.e. if you have scaled the entities in some way), use also Fourier series. Then begin to generalize. Set up the first hypotheses and test them. Approach a theory from different sides.

Finally, formulate a theory of repetition of linguistic entities. Elaborate on boundary conditions for language, text-sorts, etc. Proceed in the following way:

Whatever entity you use, search for answers to the following questions:

I. Are there some tendencies concerning special words, author, text sort, age, education, historical time of text creation, language, etc.?

II. Which entities display an evident Skinner effect?

III. If you consider merely the class of nouns, how can e.g. “nominal style” be expressed (measured)?

IV. What are the properties of the distribution of distances between identical entities (moments, Ord’s indicators, skewness, asymmetry, etc.)

V. Can some laws be conjectured?

VI. How does one set up a theory of sequential structure?

VII. Does the Weber-Fechner law intervene?

VIII. Can a concrete hypothesis be derived from an existing repetition theory?

IX. If a tendency is found, how can it be interpreted, linguistically substantiated and derived from the background theory?

X. Which of the entities display random distances (using Zörnig’s model or the Poisson process) and which are not “quite” random. If they are not random, make conjectures about the background mechanism.

References

Altmann, E.G., Pierrehumbert, J.B., Motter, A.E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distribution of words, *PloS ONE, 4(11): e7678.*

Alvarez-Lacalle, E., Dorow, B., Eckmann, J-P., Moses E. (2006). Hierarchical structures induce long-range dynamical correlations in written texts. *Proc Natl Acad Sci USA.103:7956–7961*. [PMC free article] [PubMed]

Barabási. A-L. (2005). The origin of burstiness and heavy tails in human dynamics. *Nature 435, 207–211*. [PubMed]

Bell, A., Brenier, J., Gregory, M., Girand, C., Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language 60, 92–111.*

Corral, A., Ferrer-i-Cancho, R., Díaz-Guilera, A. (2009). Universal complex structures in written language. *arXiv: 0901.2924v1* (19 Jan. 2009).

Goh, K-I., Barabási, A-L. (2008). Burstiness and memory in complex systems. *Europhys Letters 81,48002.*

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics 32, 1489–1522.*

Grosz, B., Joshi, A., Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguististics 21,203–226.*

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

Herrera, J.P., Pury, P.A. (2008). Statistical keyword detection in literary corpora. *Eur. Phys. J. B. 63, 135–146.*

Katz, S.M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering 2, 15–59.*

Köhler, R., Tuzzi, A. (2015). Linguistic modelling of sequential phenomena. In: Mikros, G., Mačutek, J. (eds.), *Sequences in Language and Text: 109-123*. Berlin/ Boston: de Gruyter Mouton.

Kunz, M., Rádl Z. (1998). Distribution of distances in information strings. *Journal of Chemical Information and Computer Sciences 38(3), 374-378*.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics 15, 1-13.*

Malmgren, R.D., Stoufferm, D.B., Motter, A.E., Amaral, L.A.N. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sc.i USA 105, 18153–18158*. [PMC free article] [PubMed]

Mikros, G., Mačutek, J. (eds.) (2015). *Sequences in Language and Text*. Berlin/ Boston: de Gruyter Mouton.

Montemurro, M.A., Zanette, D.H. (2002). Entropic analysis of the role of words in literary texts. *Advances in Complex Systems 5, 7–17.*

Politi, M., Scalas, E.. (2008). Fitting the empirical distribution of intertrade durations. *Physica A 387, 2025–2034.*

Redner, S. (2001). *A Guide to First-passage Processes*. Cambridge: Cambridge Univ. Press.

Sarkar, A., Garthwaite, G.H., de Roeck, A. (2005). A Bayesian mixture model for term re-occurrence and burstiness. *Proceedings of the 9th Conference on Computational Natural Language Learning 48-55.*

Serrano, M.A., Flammini, A., Menczer, F. (2009). Modeling statistical properties of written text. *PLoS ONE.4,e5372*. [PMC free article] [PubMed]

Vázquez, A. (2005). Exact results for the Barabási model of human dynamics. *Phys Rev Lett. 2005 Dec 9; 95(24):248701.*

Vázquez,,A., Oliveira, J.G., Dezsö, Z., Goh,,K-I., Kondor, I., Barabási, A-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E 73, 036127.*

Wimmer, G., Altmann, G. (1996). The theory of word length distribution. In: Schmidt, P. (ed.), *Glottometrika 15, 112-133*. Trier: Wissenschaftlicher Verlag.

Yannaros, Y. (1994). Weibull renewal processes. *Annals of the Institute of Statistical Mathematics 46, 641–648.*

Zörnig, P. (1984). The distribution of distances between like elements in a sequence, part I. *Glottometrika 6, 1-15;* part II. *Glottometrika 7, 1-14.*
(In: Quantitative Linguistics, Vol. 25 and 26, Brockmeyer, Bochum.)

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22* (= *Quantitative Linguistics*, Vol. 32). Brockmeyer, Bochum.

Zörnig, P. (2010). Statistical simulation and the distribution of distances between identical elements in a random sequence. *Computational Statistics & Data Analysis. 54, 2317-2327.*

2.4. Frequency sequences

Problem

Study the sequence of word frequencies in two forms: (a) as word forms, (b) as lemmas. Compute the distances between equal frequencies and propose a function expressing the relation between x = distance, f(x) number of distances of size x.

Procedure

First, read the problem *Sequences in text* and use the references quoted there. Now take a text and compute the word frequencies in it. You obtain two variants: lemmas and word forms. For each variant separately, replace the words (lemmas) by their frequencies to obtain a sequence of numbers representing the frequencies. A simple program allows you to compute the distances between equal frequencies. The distance is considered as the number of steps necessary for coming from a number to the same number. It is simply the number of steps between the two identical numbers (= 1 + intervening numbers). Take into account only the next identical number (not all).

Now set up the distribution of distances, x = distance, f(x) number of distances of size x. Propose a model for this result. Do not consider it a discrete distribution, otherwise you get problems with pooling because many distances are not represented at all. That is, you may derive the model from whatever well substantiated background but at last, consider it simply a function (i.e. without normalization) and ignore the classes with frequency 0.

If you do not like complex derivations, use the Zipf-Alekseev function which is well substantiated, fit it to the data and state your two results (lemmas and word forms). The Zipf-Alekseev function is defined as

$$f(x) = cx^{a + b \ln x}$$

Where x = distance classes. The parameter c depends merely on the frequency $f(1)$, i.e. it is some function of the text size, or simply the number of distances of length 1. The above formula is a modification of “Zipf’s law”.

Perform the computation for many texts and state whether parameter b is linked in some way with parameter a. Find the form of the link.

If both dependencies (i.e. the Zipf-Alekseev relation and the link between parameters) hold, study texts in other languages. State whether there are outliers in one of the two functions and find the “cause” of this phenomenon, e.g. the style of the author, the text-sort, the morphological type of the language, etc.

The continuation of this research direction can be performed as follows: Since parameter a is the fundamental one in this relationship, set up simply the rank-frequency distribution of word/lemma frequencies in the text and fit the

Zipfian function (power function) to the ranks (*r*). You obtain $f(r) = m/r^k$. Study the relation of the parameter *k* to the parameter *a* in the Zipf-Alekseev formula.

If you drew an elementary control cycle for the distances between equal frequencies, add to this cycle that of rank frequencies. Then step by step add further properties.

References

Cf. References to the problem *Sequences in texts*

Popescu, I.-I., Zörnig, P., Grzybek, P., Naumann, S., Altmann, G. (2013). Some statistics for sequential text properties. *Glottometrics 26, 50-94.*

2.5. The world view of language

Problem

No language reflects the world in the same way, even if translations from one language to another are always possible. The concepts and the words are our creations. Things, processes, properties, circumstances, relations may be expressed differently, there is no one-to-one correspondence, not even between very near languages. Study the differences in a restricted domain and express them quantitatively.

Procedure

(1) Use the first hundred words of a bilingual dictionary and find the number of translations for each word of the basic language into the other. If there is only one corresponding word, then x = 1, for two translation words x = 2, etc. Set up the distribution of correspondences. The result is a mixture of semantic diversifications in the first language, of different world view of both languages and of a different classification. Using this background, set up a hypothesis, translate it in the language of mathematics, solve it and fit the model to the distribution data obtained empirically. If the empirical distribution is not quite smooth, add further words to your data.

(2) Take a special lexical domain, e.g. spatial prepositions (*in, on, to, from, above, below, behind,…*) in both languages, write those of the first language in one column and those of the other in a second. Then looking in the dictionary join the (spatial) translations of each preposition in the first language with those in the second using an edge. You obtain a bipartite graph. Using the literature on graph theory express the properties of this graph quantitatively.

(3) Take a longer text and its translation in another language. Prepare a contingency table: in the first column (left) write the prepositions of the original language, in the first (top) line of the table the individual translations. In the

translation (top line) write not only the prepositions of the second language but all means that were used in the translation. You obtain a table of correspondences in which the numbers express the strength of correspondence. Evaluate (a) the semantic diversification of each preposition in the first language, (b) evaluate the whole table using appropriate methods. Propose an indicator of divergence between the spatial systems of the given languages.

(4) Study other restricted semantic systems in two languages both in the dictionary and in texts. Show that there are differences because, in text, the style of the translator is a further factor. Do not use poetic texts.

(5) Study other restricted systems only in one language and describe them quantitatively.

References

Altmann, G., Dömötör, Z., Riška, A. (1968). The partition of space in Nimboran. *Beiträge zur Linguistik und Informationsverarbeitung 12, 57-91*

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

Asratian, A.S., Denley, T.M.J., Häggkvist, R. (1998). *Bipartite Graphs and their Applications*, Cambridge: Cambridge University Press.

Bollobás, B**.** (1998). *Modern Graph Theory*. Berlin: Springer.

West, D.B. (2001). *Introduction to graph theory.* Upside Saddle River: Prentice Hall.

2.6. Climax types

Problem

Climax is understood as the increase of some property in sentence or verse or text from the beginning to the end. In *Problems Vol 4.: 2.3. The course of polysemy in sentence*, we considered the polysemy of words. Generalize the problem to different properties – separately for verses, sentences and texts – and if you find some tendency, capture it formally.

Procedure

First take a poem and study a given property of words in each position of a verse separately. Consider each verse length separately because the positions of components are relative to the length of the construct. Consider at least one of the following properties: word length, polysemy, morphological complexity, frequency (in the given text), the number of hrebs to which it belongs, degree of abstractness vs. concreteness, degree of specificity vs. generality, number of associations (taken from an association dictionary), the number of grammatical categories it expresses. You can omit some words but describe exactly what you do. If you state a tendency, express it using a simple function.

Do the same for sentences. Replace the words by the values of their properties in the whole text and study the course of the value in each equally long sentence.

Then consider a text and take into account each sentence. After defining the properties of sentences, perform the same operations as above and find the respective tendencies. Consider also the properties of clauses and classify the sentences according to the number of clauses. Study the position of nouns in each sentence of the same length.

Study the increase of intensity, force, etc. in the plot of a story. Here, quite new methods of scaling are necessary. You can introduce new vistas.

Generalize the results in two ways: (a) How do properties in general behave and (b) how do levels (word, clause, verse, sentence) behave? In order to solve the last problem, compare several languages. The trends may be positive, negative or not existent at all.

Construct a theory of climax step by step. To this end express everything in the form of mathematical models, even if, at the beginning, you must apply inductive methods.

Be aware of the fact that you seek the existence and forms of climax and not just any kind of sequence.

References

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

Altmann, G., Štukovský, R. (1965). The climax in Malay pantun. *Asian and African Studies 1, 13-20.*

Groot, A.W. de (1946). *Algemene versleer.* Den Haag.

Hřebíček, L. (1996). Word frequency and word location in a text. *Archiv Orientální 64(3), 339-347.*

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-Boston: Mouton de Gruyter.

Průcha, J. (1967). On word-class distribution in Czech utterances. *Prague Studies in Mathematical Linguistics 2, 66-76.*

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

2.7. Continuous modeling of sentence length

Problem

If sentence length is measured in terms of the numbers of word, one cannot always find a discrete distribution applicable to data of this sort. Use published data, consider the individual lengths as averages of continuous intervals and find an adequate continuous function capturing the course of data.

Procedure

Consider, for example, the data presented by P. Grzybek (2013) concerning sentence lengths in the Russian novel Anna Karenina by L.N. Tolstoj. The distribution is bell-shaped. Grzybek fitted successfully a mixed negative binomial distribution having five parameters. Since one tries to avoid mixing – because it is not easily interpretable, especially if one does not know the boundary conditions which might cause it –, use the available software (e.g. *TableCurves*) and find an adequate continuous function with less than five parameters. Avoid polynomials. Then transform the function into a discrete distribution using the standard procedure proposed by Mačutek and Altmann (2007). Needless to say, you may skip the last step because modeling means merely finding a formalized and easily manipulative image.

Scrutinize further texts containing at least 100 sentences. First, use only texts of the same language; then extend your research to different languages. If you obtained the same result for all texts, set up the recurrence function of the discrete distribution and derive it from the unified theory (Wimmer, Altmann 2005). Interpret the parameters linguistically.

If you obtained several different functions (distributions), interpret them by linguistic boundary conditions, e.g. synthetism/analytism, style, etc. Strive for a theory.

Measuring sentence length in terms of word numbers means omitting the level of clauses. Clauses are the immediate constituents of sentence. Introduce a second independent variable represented by some function of clause length measured in terms of number of words, e.g. the product of clause lengths in the given sentence. You obtain a function, e.g. $y = f(x) + g(z)$, or $y = f(x)*g(z)$, etc. If you obtain plausible results, test other phenomena evaluated by omitting the immediate lower level.

References

Grzybek, P. (2013). Close and distant relatives of the sentence: Some results from the Russian. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: 44-58.* Belgrade: Academic Mind.

Mačutek, J., Altmann, G. (2007). Discrete and continuous modeling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

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: 781-807.* Berlin: de Gruyter.

2.8. Thematic concentration 1

Problem

In *Problems Vol. 1* (2008: 60f.) and *Vol. 3* (2011: 131 ff), thematic concentration of a text has been measured in relation to the h-point. Define it now as the frequency of autosemantic lemmas having the frequency $f_a > 1$ divided by the sum of frequencies of all lemmas whose frequency is greater than 1. That is, omit hapax legomena and consider only words occurring at least twice.

Procedure

Take a text and perform the usual word count. In strongly synthetic languages it is better to lemmatize the words, otherwise many forms will belong to the hapax legomena. Then set up the proportion of autosemantics occurring at least twice in the set of all words occurring at least twice, i.e.

$$TC = \frac{\sum_{r'=1}^{K} f_{a,r'}}{\sum_{r=1}^{K} f_r},$$

where *r* ist he rank, f_r is the frequency of a word (> 1) at rank *r; r'* is the rank of an autosemantic, $f_{ar'}$ is the frequency of an autosemantic (> 1), and *K* is the set of all words whose frequency is greater than 1.

Since this is a proportion whose expectation is 0.5, test the hypothesis (a) using the exact binomial test whether TC significantly differs from 0.5 and (b) using the asymptotic two-sided normal test for its deviation.

Perform the investigation on several tests in at least two text-sorts and show the difference. Perform the investigation on the same text-sort in two different languages. You may take also the translation of the same text, e.g. *Le petit prince*. Can you detect some differences?

Follow the development of a writer computing TC in his or her texts over the course of years. Is there some change in these texts?

References

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol 3.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Altmann, G. (2001). Thematic concentration in texts. In: Kelih, E., Levickij, V., Matskuliak, Y. (eds.), *Issues in Quantitative Linguistics Vol 2: 110-116.* 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.

Strauss, U., Fan, F., Altmann, G. (2008^2). *Problems in Quantitative Linguistics Vol 1.* Lüdenscheid: RAM-Verlag.

2.9. Thematic concentration 2

Problem

Quantify thematic concentration of a text by an indicator expressing the association of its sentences. First define exactly what an association is, then define an indicator and find its sampling properties.

Procedure

One of the possibilities is: (a) to number the sentences, (b) to define the association, e.g. two sentences are associated if they contain the same lemma, a synonym, a metaphor, or a reference; (c) to set up a matrix containing all associations or connections between sentences (the upper triangle is sufficient). Since the matrix represents a graph, you can use the connectivity of the graph as a measure of concentration. The simplest way is to take the ratio of the number of observed connections (edges or non empty cells of the matrix) to all possible connections $n(n\text{-}1)/2$ where n is the number of sentences in the text.

Characterize several texts in this way – take different authors, different text-sorts, different languages, different historical times in one language, etc. – and first classify the texts using a standard method. If you obtain “clear” classes, you have the first result. If not, perform tests for differences using the given ratio. Show that some texts significantly deviate from the value 0.5, i.e. they are significantly strongly or weakly connected.

Take any other property of texts and search for some kind of dependence between this new property and thematic concentration. If you find at least a significant correlation, fit a function to the dependence, derive it using a differential equation and substantiate the equation linguistically.

Add stepwise further properties and construct a control cycle (cf. Köhler 2005).

References

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol 3.* Lüdenscheid: RAM-Verlag.

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

Popescu, I.-I., Altmann, G. (2001). Thematic concentration in texts. In: Kelih, E., Levickij, V., Matskuliak, Y. (eds.), *Issues in Quantitative Linguistics Vol. 2: 110-116.* 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.

Strauss, U., Fan, F., Altmann, G. (2008^2). *Problems in Quantitative Linguistics Vol 1.* Lüdenscheid: RAM-Verlag.

2.10. Thematic concentration 3

Problem

Define thematic concentration as the mean squared deviation of the pre-h ranks of autosemantics from the h-point. Derive the sampling properties of this indicator, analyze several texts and compare them.

Procedure

First state the rank frequency distribution of the lemmas of a text. Then compute the h-point as indicated in the literature. Mark the ranks *r* of all autosemantics (A) smaller than *h* using an apostrophe, i.e. as *r'*. Compute the sum of squared deviations as $\sum_{r' \in A}(h-r')^2 f(r')$ and divide it by the sum of frequencies of all lemmas whose rank is smaller than *h*, say N_h. Now perform an asymptotic normal test for the difference between texts. The variable is here *r'*, while *h* and N_h are constants. You may use the mean rank of autosemantics in the pre-h domain.

Analyze several texts using lemmatizing software, state the value of this indicator in all texts and compare them. For the first, an ordering of texts (without testing) is sufficient.

Elaborate on characterizing text sorts. State whether there are other properties of texts linked with this indicator, for example entropy, repeat rate, etc. Interpret the results linguistically.

Set up other different indicators of thematic concentration and substantiate them linguistically.

Use short but complete texts, e.g. poems, press texts, but do not use text mixtures.

Performing the tests for similarity, state whether the given text sort, work of a writer, etc. are uniform, i.e. whether there are significant differences between the texts. Perform tests for each text compared with each other, set up a similarity matrix and use it to draw a graph of text similarity. The weights of edges are the results of the similarity tests. Then compute some properties of the graph, i.e. express the thematic concentration of the given set of texts by graph theoretical indicators.

References

Cf. *Thematic concentration* 1 and 2 in this volume.

2.11. Thematic concentration 4

Problem

Thematic concentration can be evaluated not only by taking into account the same words or lemmas but also the "same" meanings. Propose a method of evaluation.

Procedure

Consider a usual frequency list of word forms or lemmas. The lemma-list is always shorter, especially in strongly synthetic languages. Now join all entities expressing the same concept, for example "She is pretty. Her beauty is overwhelming." Consider "pretty" and "beauty" as the same concept. Or "quick", "quickly", "speed", "celerity" etc. may belong to the same set. Do not distinguish parts of speech but collect concepts. Insert in the same set also synonyms, metaphors, antonyms (which express merely the other extreme of the same concept).

Now, set up a new frequency distribution and study its properties. Use all previous indicators and show how the expression of thematic concentration changes beginning from word forms up to concepts. You may try various combinations, e.g. placing all pronouns in the same set, eliminating articles because they always belong to some noun, etc.

Strive for a well defined, linguistically well substantiated construction of conceptual sets. Try different variations. Finally, obtain some frequency distributions which can be evaluated in the usual way. Derive the appropriate distribution leaning against your linguistic substantiations – both qualitative and quantitative – and test your hypotheses comparing as many texts and languages as possible.

The writer uses words but he does not think "in words". Before he expresses something, he thinks in images. The image can be incorporated in different ways of expression. Your task is to capture the concentration of his mental images.

The problem does not concern only linguistics but involves a combination with psycholinguistics and literary science.

A very good object of analysis is a stage play in which one can distinguish not only acts but also individual persons and their “thematic restrictions”.

References

Cf. *Thematic concentration 1, 2, 3* in this volume.

2.12. Denotative-connotative concentration

Problem

Perform the weighting of elements of individual hrebs defined in some way. (1) Replace the entities of the text by their weights to obtain a time series. Evaluate the properties of the sequence. (2) Find a function capturing the frequency of the weights and define an indicator of denotative-connotative concentration.

Procedure

First take a short text and analyze it in hrebs. You may define them in any of the n ways. Then set up a scale for weighting the entities in the hrebs. Consider the fact that there are synonyms, antonyms, hypernyms, hyponyms, metaphors, agreement, government, references, associations (of different kind), connotations, suppletivism, etc. Some of them may obtain the same weight. Such a scale does not exist as yet.

Then construct a time series of the weights and evaluate the sequence, e.g. compute the mean, the variance, Ord’s criterion, auto-correlation, distances between equal weights, matrix of transition probabilities, etc. Compare the results with those obtained from other texts. Make the first steps toward the characterization of text sorts using your results.

If you defined the weights by cardinal numbers, set up the frequency distribution of the weights. You may use also a simple discrete or continuous sequence (without normalization). Compute the properties of the distribution and compare it with that of other texts. Compare some kind of prose with lyric poetry.

Study (a) the development of a text, e.g. strophe-wise or chapter-wise but even sentence-wise is possible; (b) if you analyzed several texts of an author, study his development; (c) study the development of a certain text sort, e.g. press texts, then compare it with the development of some other. Strive towards a textual development of a language.

At last, define an indicator of denotative-connotative concentration of the text which must follow from your computations. Do not mix it up with thematic concentration which has a different background. Derive the sampling properties of your indicator.

References

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

Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, T.G. (eds.*), Quantitative Linguistics. An International Handbook: 423-44*. Berlin: de Gruyter.

Ziegler, A., Altmann, **G.** (2002). *Denotative Textanalyse*. Wien: Praesens.

2.13. Text compactness

Problem

J. Mačutek and G. Wimmer (2014) defined text compactness as the relative number of sentence pairs associated with the same word. Generalize this approach (1) taking into account also the synonyms, (2) considering also references (pronouns, etc.), and (3) proposing an evaluation of the weight of associations.

Procedure

First compute the original indicator. Let *L* be the number of sentence pairs containing the same word. The number of all sentence pairs is $\binom{n}{2}$, or *n*(*n*-1)/2. Hence the relative measure is $LTC = 2L/[n(n - 1)]$. Compare *LTC* for several texts of two different text sorts. Compare all pairs of texts using the asymptotic normal test (cf. Mačutek, Wimmer 2014) and set up classes of texts.

Now, perform the same operation but extend the association of two sentences by taking into account also the synonyms of the given words and, if you want, also the antonyms, hypernyms, hyponyms, or metaphors. The overall text compactness will be, perhaps, greater than in the first case.

Continue taking into account also references of any kind. This is the extreme possibility to solve the problem without weighting, in a straightforward way.

Set up a text classification. Study the development of a writer or of children. Compare texts of the same text sort in two different languages.

The "highest" possibility is the weighting of associations. You must introduce a kind of scaling ascribing different degrees of association to identical words, synonyms, antonyms, hypernyms, hyponyms, metaphors, pronouns of different sort, referential associations, etc.

If you succeed creating such a weighted system, then study the number of sentence pairs associated by a certain weight. First find an empirical function capturing this relationship (degree vs. frequency). Use the parameters of the function for characterizing text sorts, authors, epochs, languages. Then derive the given function from a theoretical background. The background must be linguistically substantiated. Insert all this into a differential equation from which you can derive the given function. If the conditions inserted necessarily in the differential equation are different (e.g. the requirement of speakers and hearers-readers), derive the new function and fit it to the data. Do not remain on the empirical, inductive level but construct step by step an elementary theory.

Compare your results with those concerning *hrebs.*

References

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

Mačutek, J., Wimmer, G. (2014). A measure of lexical text compactness. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical approaches to text and language analysis: 132-139.* Lüdenscheid: RAM-Verlag.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

2.14. Conceptual inertia of texts 1

Problem

The subsequent sentences of a text are usually conceptually associated. This may be done not only by the repetition of the same word but also by its synonyms, metaphors, hypernyms, hyponyms, references, anaphoras, cataphoras, pronominal representations, etc. Subsequent associated sentences form chains called Belza-chains. A sentence may belong to several chains simultaneously. State the length and number of chains and the distribution of lengths.

Procedure

The problem cannot be solved by programming; unfortunately, it must be, performed “by hand”. Write each sentence in a separate line and search for concepts repeated in subsequent sequences. If there is a sentence not having a common concept with its predecessor and follower, it forms a chain of length 1. The proportion of chains of length 1 shows the conceptual interruptions but it can be interpreted in many different ways.

Find a model for the distribution of lengths; evaluate the proportion of length 1 and perform a comparison of texts in order to find some classes of texts, to study the evolution of a writer, a novel, or a stage play. Use the asymptotic normal test for the comparison of two proportions but you can use also the probability resulting from the binomial distribution.

If you have many texts in one language, analyze several texts in another one and compare them. Compare a strongly analytic and a strongly synthetic language. Can one see the difference from this point of view?

Consult the literature concerning text linguistics in which you find the different ways of conceptual association.

References

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

Chen, R., Altmann, G. (2015). Conceptual inertia in texts. *Glottometrics 30,73-85.*

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

2.15. Conceptual inertia of texts 2

Problem

In the previous problem you described the conceptual inertia in texts and performed a measurement of chain lengths. Take the same texts and measure the weight of conceptuality of individual components. Then find the distribution of weights of chains.

Procedure

The words or morphemes do not express the basic concept in an equal way. Direct naming is “stronger”, “more weighty” than for example. a pronominal representation or a reference or even an ellipsis. E.g. “father” is stronger that “he” or “who”. Ascribe weights to the respective words or morphemes related to the basic concept and measure the weights of individual chains adding all weights of the given concept. If there are two chains in a sequence, consider both separately. Thus a chain has not only a length but also a weight. Now set up the distribution of chain weights. Derive a function which captures this distribution and characterize the texts by the properties of this distribution.

Compare different texts: scientific, prosaic, poetic, press, didactic, etc. ones. Draw conclusions about text sorts.

Then compare languages distinguishing the individual texts sorts.

Then take a writer and analyze several of his works written in the same text sort. Study the development of the writer.

Pay special attention to stage plays and their development according to subsequent acts. Can one distinguish the features of the classical stage play theory?

References

See the previous problem 2.14.

2.16. Conceptual inertia in texts 3

Problem

What is the relation of conceptual inertia to *thematic concentration* and how can it be expressed? These two properties do not express the same because conceptual inertia takes into account also synonyms, metaphors, references, etc., while thematic concentration is a property of individual words (lemmas or word forms)

Procedure

Take a text and compute the vector of conceptual inertia leaning against the two previous problems. Find a characteristic feature, e.g. mean length of Belza-chains or some other properties of the fitting function. Then perform the usual word count and compute the thematic concentration (cf. e.g. *Problems Vol. 3: 133*; Popescu et al. 2009). You can apply any other well defined formula.

Perform these two operations for several texts (at least 10) and study the relation of the degree of conceptual inertia to that of thematic concentration. Express the relation by a function chosen inductively e.g. by a ready program. Then substantiate this relation linguistically. If it is not linear, lean against the unified theory (cf. Wimmer, Altmann 2005) and set up the differential equation. Does thematic concentration increase with increasing inertia or the other way round? Is the relation significant? Search as long as you find an acceptable function with as few parameters as possible. Do not use polynomials. Then take texts from a different text sort and perform the same operations. If the first function is adequate also here, take texts from another language.

Step by step, generalize the relationship and add also other properties of texts (cf. Popescu et al. 2009). Strive for constructing a control cycle similar to that proposed by Köhler (2005) or insert your cycle directly in Köhler's proposal if his control cycle contains some property scrutinized by you.

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.

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

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

2.17. Adjective-verb ratio and text indicators

Problem

In *Problems Vol. 3* (p. 124) the relation of Lambda to adjective-verb ratio has been studied. Show that the adjective-verb ratio, defined as $Q = V/(V+A)$ can be linked with other text indicators (V = number of verbs, A = number of adjectives in text).

Procedure

First compute Q for several texts (at least 10) of the same text sort. Then compute the frequencies of all words and the following indicators: Gini’s coefficient for rank-frequency of words, text compactness, Ord’s criterion, the probability of the given number of runs (of A and V), the arc of the rank-frequency distribution, entropy, Repeat rate of frequencies, and some indicator of the distribution of sentence length. Then taking one indicator after another, find their relation to the adjective-verb ratio. Derive the respective function(s) from the unified theory. If the derivation is not yet possible, set up a function inductively.

Then compare all indicators with each other and set up, step by step, a control cycle. If Q is not linked with some of them, substantiate the lack of relation.

In this way, construct a partial theory of texts.

Consider now some other pairs of parts of speech (e.g. V and N), define for them an analogous Q and continue searching for links to other properties.

References

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

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

Altmann, G., Köhler R. (2015). *Forms and degrees of repetition in texts. Detection and analysis*. Berlin/Munich/Boston: de Gruyter.

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. et al. (2009). *Word Frequency Studies*. Berlin: Mouton de Gruyter.

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts*. Lüdenscheid: RAM.

Popescu, I.-I., Čech, R., Altmann, G. (2013). Descriptivity in Slovak lyrics. *Glottotheory 4(1), 92-104.*

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of Word Frequencies*. Lüdenscheid: RAM-Verlag.

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: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An international Handbook: 368-387.* 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 Linguistics. An International Handbook: 791-807.* Berlin: de Gruyter.

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

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

2.18. Unified word length modeling

Problem

Word length frequencies have been modeled in form of discrete probability distributions. Solve the following problems:

(1) Collect all publications concerning word length and set up a unique system of probability distributions – if possible.

(2) Consider word length a continuous property and find a unique function capturing all data.

(3) If possible, publish the complete bibliography, at least beginning from 1995.

Procedure

First find the last publications which show many models and the respective references, e.g. Best (1997, 2001), Grzybek (2006) or Đuraš (2012). Set up the relations between the distributions and show – if possible – that all are special or limiting cases of a more general distribution. This is not a simple task because the most general distribution must not have too many parameters. Ignore modifications of distributions and if you succeed to solve the problem at least partially, derive the general distribution using linguistic arguments based on language synergetics. For the derivation use difference equations.

If you do not succeed, consider the possibility of treating word length as a continuous variable and find a function which sufficiently captures all data. If you find anomalous cases, e.g. a local minimum or a systematic deviation in the smooth course of the function, modify the function only for the given anomalous class. At last derive the function from a differential equation which should be substantiated linguistically.

If there is a law behind word length, then the result obtained must show also the development of some parameters in a certain language in the course of time. Study Latin and Romance languages, or Old Church Slavic and modern Slavic languages or another family whose older stage is known.

Present a unified theory.

See esp. the problem *1.6. Length levels* in this volume.

References

Best, K.-H. (ed.) (1997). *Glottometrika 16. The distribution of word and sentence length.* Trier: WVT.

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

Đuraš, G. (2012). *Generalized Poisson models for word length frequencies in texts of Slavic languages*. Graz: Diss.

Grzybek, P. (ed.) (2006). *Contributions to the science of text and language. Word length studies and related issues.* Dordrecht: Springer.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

2.19. Poetic and rhetoric figures 1

Problem

A special aspect of style can be scrutinized analyzing the presence of poetic and rhetoric figures in the text. One can find complete lists of figures on the Internet or in books on poetics. Set up hypotheses about the relation of text-sorts to special poetic figures, perform a measurement and test your hypotheses.

Procedure

First prepare abbreviations of the individual figures in order to be able to present a text as a sequence of figures. You may adhere to a special school but you must present a list of figures you used for the analysis. Then take a text and analyzing it sentence by sentence set up a sequence of (abbreviations) of poetic figures. This part of the problem is not easy because it must be made by hand and every sentence must be thoroughly analyzed.

Begin with short texts of the same text-sort, e.g. press texts or poems. After having set up the vector of figures for the given text, perform the following analyses:

(1) Set up the rank-frequency distribution of the individual figures. Conjecture a hypothesis concerning the distribution (or function) and substantiate it linguistically, e.g. there are figures which are grammatically necessary, other ones are specific to the text sort, still other ones represent personal style, etc. What is the form of the distribution? Compare various texts and state the differences. The hypotheses will be somewhat difficult because nobody cared up to now for their deeper roots.

(2) Set up the spectrum of the figures, i.e. a distribution in which the independent variable is the occurrence x and the dependent variable y is the number of classes occurring exactly x-times. This can easily be performed using the resulting vector. Again, compare the texts for similarity using a chi-square or a rank test.

(3) Compute the distances between identical figures in the vector, set up an indicator of the distances, derive its variance and compare again various texts.

(4) Show that the proportion of poetic and rhetoric figures is not equal in the text. Set up confidence intervals of, say, rhetoric figures in individual texts, and compare the proportions in different texts.

(5) Can the poetic or rhetoric figures be scaled? This is a complex problem, not easy to solve. First define some properties or dimensions in which the figures may be situated. The simplest way is to study the aim of the figure. Then ascribe the individual figure a degree of the given property. The degree is the independent variable and the frequency is the dependent one. You obtain a distribution or a function. Find the form of the function. Make the first steps inductively, i.e. use software which finds many appropriate functions. Then choose that function which seems to be adequate for at least the texts of the same text sort or author. Derive the function from a (linguistically substantiated) differential or difference equation. Apply the function to all texts you analyzed and order the texts according to some parameter of the function; characterize the texts by an indicator computed from the given function (e.g. mean, Ord’s criterion, asymmetry, excess, entropy, repeat rate, etc.) and compare (or at least order) the texts in order to see whether there are differences. If possible, avoid polynomials and select a function with a small number of parameters..

(6) Strive for a theory resting on hypotheses, links between properties, derived functions and tests.

References

Adams, S. (1997). *Poetic Designs: An Introduction to Meters, Verse Forms, and Figures of Speech*. Amazon.

Gibbs, R.W.Jr. (2008). *The Cambridge Handbook of Metaphor and Thought***.** Cambridge University Press.

Harjung, J.D. (2000). *Lexikon der Sprachkunst. Die rhetorischen Stilformen. Mit über 1000 Beispielen*. München: Beck.

Lausberg, H. (1990): *Handbuch der literarischen Rhetorik. Eine Grundlegung der Literaturwissenschaft.* 3. Auflage, mit einem Vorwort von Arnold Arens. Stuttgart: Steiner, Stuttgart.

Meyer, U. (2007). Stilistische Textmerkmale. In: Th. Anz (Hrsg.), *Handbuch Literaturwissenschaft.* Band 1: *Gegenstände und Grundbegriffe: 81-110.* Stuttgart: Metzler

Plett, H.F. (2001). *Einführung in die rhetorische Textanalyse.* 9., aktualisierte und erweiterte Auflage. Hamburg: Buske.

Schüttpelz, E. (1996). *Figuren der Rede. Zur Theorie der rhetorischen Figur* (= *Philologische Studien und Quellen.* Bd. 136). Berlin: Schmidt

http://www.poetryfoundation.org/learning/glossary-terms?category=techniques-and-figures-of-speech

http://outre-monde.com/2011/01/16/a-short-glossary-of-rhetorical-and-poetic-devices/

2.20. Poetic and rhetoric figures 2

Problem

The writer usually does not know the names or the forms of poetic or rhetoric figures – just as a cook does not know the molecular composition of his material, – (s)he merely knows the effect (s)he wants to achieve. But if (s)he strives for a special effect or expression, (s)he spontaneously applies the same type of figures. Classify the figures according to the effect they should evoke – you can set up qualitative classes or quantitative scales – and transfer the results of your concept formation into the analysis of some texts.

Procedure

The simplest way is to take texts whose aim is known and thereafter to study which kind of figures were used. But even if the aim is known, the effect to be evoked may be obtained by various poetic means. Hence, begin to work exploratively: Take a text, state in it all poetic and rhetorical figures, let some test persons read the text and tell you their impressions. Classify the impressions and ascribe them to the topical textual means. After having analyzed several texts (according to your choice) state the possible association between effect and the present figures.

In the second step, perform a scaling of possible effects. You must propose some scalable properties and ascribe a given degree (or an interval) to the figures occurring in the text.

The other way round, you can take a list of figures and ascribe to each of them the possible effects. Then analyze a text and state the distribution of numerical effects. You obtain a relation between the effect-property and the kind of figures representing it.

Having chosen this way (without test persons), you can classify texts and search for links between effects expressed by figures, and other properties of texts. As a matter of fact, you would try to incorporate poetic and rhetoric figures into an embryonal theory of texts. This is, of course, a very long way, but in any case you can strive for ascribing sets of figures to text-sorts.

If there is already an a priori classification of text sorts, you can study the kinds of figures occurring in the individual classes. At last, text sorts can be classified according to the figures occurring in them (significantly).

References

Adams, S. (1997). *Poetic Designs: An Introduction to Meters, Verse Forms, and Figures of Speech*. Amazon.

Gibbs, R., Steen, G. (eds.) (1999). *Metaphor in cognitive linguistics*. Amsterdam: Benjamins.

Gibbs, R., Colston, H. (eds.) (2007). *Irony in language and thought: A cognitive science reader*. New York: Erlbaum.

Gibbs, R.W.Jr. (2008). *The Cambridge Handbook of Metaphor and Thought.* Cambridge University Press.

Harjung, J.D. (2000). *Lexikon der Sprachkunst. Die rhetorischen Stilformen. Mit über 1000 Beispielen*. München: Beck.

Lausberg, H. (1990). *Handbuch der literarischen Rhetorik. Eine Grundlegung der Literaturwissenschaft.* 3. Auflage, mit einem Vorwort von Arnold Arens. Stuttgart: Steiner, Stuttgart.

Meyer, U. (2007). Stilistische Textmerkmale. In: Th. Anz (Hrsg.), *Handbuch Literaturwissenschaft.* Band 1: *Gegenstände und Grundbegriffe: 81-110.* Stuttgart: Metzler.

Plett, H.F. (ed.) (1996). *Die Aktualität der Rhetorik.* München: Fink

Plett, H.F. (2000). *Systematische Rhetorik.* München: Fink.

Plett, H.F. (2001). *Einführung in die rhetorische Textanalyse.* 9., aktualisierte und erweiterte Auflage. Hamburg: Buske.

Schüttpelz, E. (1996). *Figuren der Rede. Zur Theorie der rhetorischen Figur* (= *Philologische Studien und Quellen.* Bd. 136). Berlin: Schmidt

http://www.poetryfoundation.org/learning/glossary-terms?category=techniques-and-figures-of-speech

http://outre-monde.com/2011/01/16/a-short-glossary-of-rhetorical-and-poetic-devices/

2.21. The world view of a writer

Problem

The world view of a writer (restricted to the given text) can be observed and evaluated in various ways. In general, the central theme (word, term) is associated with other ones. Significant associations may be presented in form of a graph whose properties may be evaluated. There may be more than one graph representing the given text. The set of graphs represent the world view of the writer. It may be called also denotative concentration, etc. Find the set of associative graphs for a given text and evaluate the text.

Procedure

Take a longer text and consider each sentence a frame for associations. The definition of the sentence boundaries depends on your decision. The word may be represented also by its synonyms, metaphors, references, uses in other word classes, compounds, etc. Since you must use software, it would be better to replace all entities of this sort by the “main” word. Compute the coincidence of different words and set up a matrix (or a graph) containing the probability of the given or more extreme coincidence in sentences. At last, use the matrix and transform it in a graph containing only significant associations.

The graph represents the world view of the writer (for the given purpose).

References

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

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag.

Altmann, G., Köhler, R. (2015). *Forms and degrees of repetitions in texts. Detection and analysis.* Berlin/Munich/Boston: de Gruyter Mouton.

Berry-Roghe, G.I.M. (1973). The computation of collocations and their relevance in lexical studies. In: Aitken, A.J., Bailey, R.W., Hamilton-Smith, N. (eds.), *The computer and literary studies: 103-112*. Edinburgh: Edinburgh University Press.

Dannhauer, H.-M., Wickmann, D. (1972). Quantitative Bestimmung semantischer Umgebungsfeder in einer Menge von Einzeltexten. *Literaturwissenschaft und Linguistik 2, 29-43.*

Dolphin, C. (1977). Evaluation probabiliste des cooccurrences. In: David, J., Martin, R. (eds.), *Etudes de statistique linguistique: 21-34.* Paris: Klincksieck.

Geofroy, A., Lafon, F., Seidel, G., Tournier, M. (1973). Lexicometric analysis of co-occurrences. In: Aitken, A.J., Bailey, R.W., Hamilton-Smith, N. (eds.), *The computer and literary studies: 113-133.* Edinburgh: Edinburgh University Press.

Rieger, B. (1971). Wort- und Motivkreise als Konstituenten lyrischer Umgebungsfelder. Eine quantitative Analyse bestimmter Textelemente. *Zeitschrift für Literaturwissenschaft und Linguistik 4, 23-41.*

Rieger, B. (1974). Eine tolerante Lexikonstruktur. Zur Abbildung natürlichsprachlicher Bedeutung auf unscharfe Mengen in Toleranzräumen. *Zeitschrift für Literaturwissenschaft und Linguistik 16, 31-47.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts.* Lüdenscheid: RAM-Verlag.

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

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Vienna: Praesens.

2.22. Adjectives in text

Problem

Adjectives make a text more ornamental – even if they are not the only means to create ornamentality –, the expressions more exact, they are able to evoke images, emotions, etc. Describe the adjectival state of the text, a work that can show new ways in textology.

Procedure

First take "all" adjectives of the language and perform a classification. This can, of course, be done also with those taken from a frequency dictionary or from a corpus but there are also long lists in several languages on the Internet. There are various possibilities of classification, show several ones but choose only one of them.

If you have the classes, make the second step which is more difficult: determine a property which can be used for scaling the adjectives in each class. It may be a semantic criterion, a criterion expressing a certain attitude, a kind of gradation of a certain property (e.g. *nice, pretty, beautiful, …*), grammatical functions, etc. Then order the adjectives in each class according to this property. Some of them will stay at the lowest level, some of them at the highest, the other ones may form a "staircase" or stay at the same level. Some of them are not scalable because they serve identification.

In the third step, set up a scale from zero to one or from zero to ten, etc., and ascribe a degree to each group within the given class. That means, perform a kind of scaling. The second and the third step are the most problematic ones and whatever you do, they can be criticized, changed, reinterpreted, etc. But this is the usual way of science. You can restrict yourself only to one selected adjectival class.

In the fourth step, take a text, ignore all non-adjectives and replace the adjectives by their degrees. You obtain the "adjectival vector" of the text which can be further processed.

(1) First set up the distribution of degrees and compute some indicators, e.g. mean, variance, Ord's criterion, entropy, repeat rate, etc.

(2) Find a theoretical distribution or a function capturing the empirical distribution and interpret it qualitatively. For example, does the author strive for an extreme expression of a property (using e.g. superlatives), is he moderate, pejorative, etc.?

(3) State the rank-frequency distribution of the classes. Does the text prefer a certain class or are all classes represented uniformly. Does the representation of classes have a relation to the theme of the text? In texts on physics one will not find adjectives belonging to the "beauty" class but rather to that re-

presenting physical properties, etc. Can one ascribe the text to a special text-sort leaning against the occurrence of adjectives?

(4) Study the vector itself, i.e. consider it a time series. Is the oscillation random or can you find some regularities? If so, perform a Fourier analysis or fit an increasing or decreasing function. If necessary, perform the analysis applying moving averages. Find the breaks in the sequence.

Now, analyze another text and compare it with the first one. Where are the differences? Continue analyzing the same author in his/her development, study text sorts, compare translations of special works in different languages.

If you have some indicator of the given property, find its relation to other text properties, i.e. begin to construe a control cycle in which some of the properties of the adjectival vector are linked with other ones. Make the first steps towards a theory but do not simply collect data: state hypotheses and show their place in the control cycle. Study the adjectives themselves: is the given degree associated with another property, e.g. its phonological or morphological length, the length of the sentence (e.g. the longer the sentence, the higher the mean degree of adjectives occurring in it)? Is the degree of the adjective associated with the number of words derived from it? For different other properties and control cycles, see Köhler (2005).

References

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

Kullenberg, H. (2015). Functions of attributive adjectives in English. http://www.sol.lu.se/fileadmin/media/forskning/workingpapers/engelska/vol02/Helena.pdf (4.10.2105)

Schmale, G. (ed.) (2011). *Das Adjektiv im heutigen Deutsch: Syntax, Semantik, Pragmatik.* Tübingen: Stauffenburg.

Teyssier, J. (1968). Notes on the syntax of the adjective in Modern English. *Lingua 20, 225-249.*

Warren, B. (1984). *Classifying adjectives* (= Gothenburg studies in English No. 56). Göteborg: Acta Universitatis Gothoburgensis.

Warren, B. (1984). The Functions of modifiers of nouns. *Quaderni di Semantica V1, 111-123.*

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.

http://www.enchantedlearning.com/wordlist/adjectives.shtml

2.23. Adjectival motifs

Problem

Classify the adjectives using some of the possibilities (cf. e.g. Yesypenko 2008; http://de.wiktionary.org/wiki/Adjektiv) and symbolize the classes by abbreviations. Then state for a text the sequence of adjectives and set up R-motifs as proposed by Beliankou, Köhler, Naumann (2013). Evaluate the rank-frequency distribution and the length-frequency distribution of adjectival motifs.

Procedure

First set up the classes of adjectives. You can find a number of possibilities on the Internet. The classification depends on the aim of the linguist who determined them (there are e.g. for English, the classification according to function: attributive, predicative, postpositive, substantive). Each class should be symbolized, e.g. by a letter or number. Then take a text and write the symbolization in the order in which the adjectives occurred. You obtain a long sequence which should be segmented in R-motifs. The method can be found in Beliankou, Köhler, Naumann (2013). In practice, a new R-motif begins with the symbol that already occurred in the preceding motif but sometimes one must make decisions. Consider the length of motifs, i.e. the number of symbols in them, and set up the length distribution. Then set up the rank-frequency distribution, i.e. order the frequencies of individual motifs according to their occurrence; at last, set up the spectrum of occurrences (i.e. x = occurrence, y = number of different motifs that occurred x-times).

For each aspect propose either a distribution or a function and test it. Perform the analysis for various text sorts in order to show that different text sorts use different kinds of description.

Read all *Problems* in this volume concerning motifs.

References

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012, 33-43*. Belgrade.

Warren, B. (1984). *Classifying adjectives*. Gothenburg studies in English (No. 56). Göteborg: Acta Universitatis Gothoburgensis.

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.

2.24. Stylistic centrality

Problem

Consider stylistic centrality as a tendency of an author to use a special device in as many of his texts as possible. The device may be phonic (e.g. alliteration, assonance, type of rhyme, etc.), morphological (e.g. word complexity, special derivation type, compounds, etc.), syntactic (e.g. sentence length, clause length, type of sentence, poetic word order, etc.), relative to discourse (e.g. special types of speech acts, length of monologues, etc.), lexical (e.g. use of a special vocabulary, proportion of hapax legomena, rank-frequency distribution of word forms or lemmas, etc.), semantic (e.g. extent of polysemy of words, foreign words, old words, semantic structure of compounds, etc.). You can propose any property which is not constant but expressed in degrees. Compute an indicator expressing the given property, compare the texts for similarity and evaluate the similarity matrix.

Procedure

Take as many texts of a writer as possible or consider the individual chapters of a book as separate texts. Then compute an indicator already used in textology for each text. Propose a test – usually one applies the asymptotic normal test – for which you need the variance of the indicator. Then compare the texts with each other and set up a matrix of similarities.

Characterize the writer, say, by the mean of the similarities omitting the diagonal of the matrix which is always 0.

Then consider only those pairs of texts which display a u-value (normal variable) smaller than 1.96 (in absolute value). Omit the rest of the table. Then consider the number of these significant similarities and compute their proportion, i.e. their number divided by the number of cells in the matrix. Omitting the diagonal and taking into account the whole matrix there are $n(n - 1)$ cells. If you want to compare a writer with another one, use this proportion, derive its variance and use both for the comparison.

Another possibility is to consider the whole matrix (omitting the diagonal), considering only the significant similarities ($|u| < 1,96$) and for each row state their number. Divide the row sums by n-1 (number of compared texts). The vector of these values is characteristic for the writer, it is an indicator of his stylistic uniformity in the given domain. Now you may either set up a distribution or compare the vectors of two writers computing the arccos of the angle between the vectors.

The problem must be processed with the aid of a computer and with a team of collaborators. It is too extensive because it concerns many writers, even languages, but it is also a problem with a wide horizon.

Do not use text mixtures, consider each text as a separate unit. If you use a corpus, analyze each texts separately!

References

Altmann, G., Köhler, R. (2915). *Forms and degrees of repetition in texts. Detection and analyses*. Berlin/Munich/Boston: de Gruyter.

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin/New York: de Gruyter Mouton.

Popescu, I.-I. et al. (2009). *Word Frequency Studies*. Berlin/New York: Mouton de Gruyter.

Zörnig, P., Popescu, I.-I., Altmann, G. (2015). Statistical approach to measure stylistic centrality. *Glottometrics 32, 21-55.*

2.25. Text sorts

Problem

Text sorts have been defined in order to study the different structuring of texts and because we automatically classify the objects of reality in order to get better orientation. We may conjecture that if a text sort can be defined, it must differ in some sense from the other texts sorts. The most exact differentiation is that by menas of quantitative indicators. Study the properties of texts sorts and test the differences.

Procedure

Do not analyze texts but use those that were already analyzed by other scientists. Linguistic journals are full of analyses. Define two text sorts and find all sources concerning the given property. The number of properties is infinite, we can mention here only twenty of them: word length, sentence length, morphological complexity of words, sentence complexity, polysemy of the words in text, frequency distribution of words, word classes in text, verb-adjective ratio, entropy, repeat rate, Ord’s criterion, Gini’s coefficient, vocabulary richness, the association graph of the text, clause centrality, rhetoric a poetic figures, syllable types, syllable length, meaning abstractness/concreteness, meaning generality/ specificity, etc. This list can be extended according to the interest of the researcher or to the available data.

Chose one of the properties and find the complete literature containing data. Compute an indicator for all data; classify the texts according to the assumed text sorts. For each group of data belonging to the same text sort compute the mean of the chosen indicator. Consider the indicator a simple number. Then compute the variance of the indicator for each group and the variance of the mean (= variance of the indicator divided by the number of texts).

First order the texts according to the mean of the indicator in order to obtain a first image. This order shows you the behavior of texts and helps you to set preliminary limits to text sorts from the scrutinized point of view.

Then perform a t-text or a normal test for the difference of text sort pairs. Use the means and the variances of means. If you have several text sorts, you can present the result in form of a graph: the vertices are the text sorts, the edges are the similarities. Express the situation from the viewpoint of the given property in form of graph density, path lengths, etc. You may take into account not only the existence of an edge but also its strength (expressed by the similarity test)

If you obtained a satisfactory result, continue working with other indicators and elaborate, step by step, a theoretical background for text sort analysis.

Study the similarity of graphs constructed on the basis of two different indicators. At the beginning, you may use also factor analysis but later on use rather the theory of fuzzy sets.

References

Köhler, R. (1995). *Bibliography of Quantitative Linguistics*. Amsterdam/Philadelphia: John Benjamis.

All volumes of *Problems in Quantitative Linguistics*.

2.26. Nominativity vs. predicativity 1

Problem

Some texts prefer nominativity, i.e. registration of facts, other ones describe them using predicates of the first level, namely adjectives and verbs. Omitting auxiliary and modal verbs, test to what extent a text is nominative or predicative. Perform the procedure using individual texts of the same text sort. Then compare the text sorts.

Procedure

Take a text and count in it the number of nouns (N), adjectives (A) and verbs (A). Omit auxiliary verbs, copulas, modal verbs. Set up the vector (A,N,V). Then define a nominativity indicator, e.g. QN = N/(A+N+V). Since this is a proportion, derive its variance.

Perform the computation in several individual texts of a certain text sort. Order the texts according to increasing QN and search for an interpretation. Analyze poems of a certain author and order them according to the year of creation. Compute the above indicator a search for the development of the author in the given sense.

Take means of the indicator for a group of texts and compare them with text of a second group. Strive for a nominativity classification of text sorts. Perform the asymptotic normal test.

References

Altmann, G. (1978). Zur Anwendung der Quotienten in der Textanalyse. *Glotto metrika 1, 91-106.*

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag.

Altmann, G., Köhler R. (2015). *Forms and degrees of repetitions in texts. Detection and analysis.* Berlin/Munich/Boston: de Gruyter.

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

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol. 3.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Čech, R., Altmann, G. (2013). Descriptivity in Slovak lyrics. *Glottotheory 4(1), 92-104.*

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

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5(3), 269-280.*

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: Festschrift in honour of Luděk Hřebíček: 294-312.* Trier: Wissenschaftlicher Verlag.

2.27. Nominativity vs. predicativity 2

Problem

In the problem „2.26. Nominativity vs. predicativity 1“ you defined a vector and expressed the above properties by an indicator. Now consider the problems “6.4. Measurement of verb activity”, “6.1.Abstractness of nouns” and “2.22. Adjectives in text” and define a measure of nominativity vs. predicativity. Analyze texts and perform their classification.

Procedure

Before you begin to analyze texts, quantify the degree of a property, the degree of activity of verbs and the degree of abstractness/concreteness of nouns. Any trials will be preliminary but after having analyzed many texts, you obtain a wide horizon of possibilities. For decisions about the degree of N, A, V you can use also test persons and perform the Osgood scaling.

Now take a text and set up the sequence of A, N and V. Then to each of them write the degree of expressing the given property. As a matter of fact, you obtain now three vectors: one for adjectives and their property degree, one for verbs and their activity degree and one for nouns and their concreteness/ abstractness level. Then study the individual vectors (a) state whether they display some tendency from the beginning to the end of the text; if so, find the respective function; (b) study the correlation between the level of the noun and the degree of predicates (verb and adjective) belonging to it; find an appropriate function expressing these relations.

Define an indicator characterizing these properties of the text. Derive the variance of the indicator and an asymptotic normal test for comparing the means of two texts.

Search for other properties of the respective words, e.g. length, morphological complexity, polysemy, frequency, etc. and find their links to the above properties (degree, abstractness, activity). Strive for a synergetic control cycle in which there is place for these new properties.

References

See the problem “Nominativity vs. predicativity 1”

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

2.28. Predication

Problem

Here, predication is understood as a specification of the respective word. Thus nouns can be specified by verbs, adjectives, articles, conjunctions, …; verbs can be specified by adverbs, conjunctions, modal verbs, etc. Elaborate on the scaling of predication and analyze texts.

Procedure

Elaborating the scaling procedure and analyzing a text you can use any type of grammar – if it is suitable, but you can use also the categories from philosophy. The results will be different for each type of grammar – and *eo ipso* incomparable. Find a procedure which can be used by everybody without deeper knowledge of a grammar type.

You must make a lot of decisions. The principle is: the entity which specifies has a higher predication value than the specified entity. Begin with the topic (e.g. noun), state all parts of speech that can specify it and ascribe them value 1. Then consider those entities which specified the noun and search for those that specify the first level entities. Ascribe them value 2. Continue until you obtain the degrees of all. Then set up the list of words and their degree; if a word had several degrees, use always its highest degree.

The same can be achieved theoretically but the degrees may be different, e.g. some languages do not have articles, other ones write them together with the noun, some languages omit the copula, other ones write the preposition together with the word, etc. hence, an ad hoc solution is acceptable.

Then take a text, write the sequence of predication degrees separately for each sentence. Perform the following computations:

(1) Compute the mean of each sentence and set up the distribution of degrees.
(2) Find a theoretical function capturing this distribution. Consider it simply a (non-normalized) function.
(3) State the forces that may be present in forming a sentence (cf. Köhler 2005), and insert them in a theoretical model.
(4) You may begin to seek a function inductively and after having processed many texts you can begin to theorize.
(5) Find a link between the mean predication level and sentence length.
(6) Classify the sentences according to the grammar of the language; ascribe each sentence type the respective levels found in a text, find their means and order the sentence types according to the mean predication values.
(7) Using this order compare several texts by means of a non-parametric test.
(8) Compare also various text sorts. Since they differ in the types of predication, you may obtain a quite different view of text sorts.

References

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

2.29. Sentence length

Problem

Sentence length measured in terms of clause numbers has been studied in several investigations. One always tried to find a discrete probability distribution fitting well to the data. Change your mind and find a simple (non-normalized) continuous function capturing all available data. Interpret the differences of parameters as boundary conditions. Omit all works computing sentence length in terms of word numbers. Their number is enormous but they omit an intervening level, i.e. they do not measure the given entity (sentence) in terms of its immediate constituent. The more levels one omits, the more the results approximates a fractal.

Procedure

First, set up a bibliography of pertinent works (i.e. concerning measurement of sentence length in clause numbers) and publish it e.g. in *Glottotheory* or *Glottometrics*. Then exploit all available data. Begin to search inductively for an adequate continuous function. You may use *NLREG, TableCurves, Origin* or other software fitting mechanically functions to data. If you succeed in finding at least a family of functions, classify the texts, find the relation of the function to the text sort, to the language or to the development of a writer.

B. Niehaus (1997: 213) formulated ten questions which can be associated with sentence length research:

(1) Is sentence length characteristic for the style of an author or a text sort?
(2) Can sentence length be used as a criterion to solve problems of authorship?
(3) Are there sentence length changes of a speaker in the course of his life?
(4) Does sentence length develop from primitive forms up to complex scientific texts?
(5) Which factors are active in forming sentence length?
(6) Which mental processes are active in sentence generation?
(7) What is the link of sentence length to other properties of sentence and of other language entities, or, in other words, in which control cycles does it play a role?
(8) To what extent is sentence length a factor for the text difficulty?
(9) Are there mathematical models describing adequately the distribution of sentence lengths?
(10) Are sequences of sentence lengths chaotic, stochastic or deterministic ones?

The solution of at least one of these questions would open a way to theory building.

References

Altmann, G. (1988). Verteilungen der Satzlängen. In: Schulz, K.P. (ed.), *Glottometrika 13: 147-169.* Bochum: Brockmeyer.

Altmann, G., Köhler, R. (2015). *Forms and degrees of repetitions in texts. Detection and analysis.* Berlin/Munich/Boston: de Gruyter (pp. 57f.).

Brandwood, L. (1969). Plato’s seventh leter. Liège: Revue Lasla. Laboratoire d‘ analyse statistique des languges anciennes.

Buch, K.R. (1952). A note on sentence length as a random variable. In: Doležel, L., Bailey, R.W. (eds.), *Statistics and style: 76-79.* New York/London: Elsevier.

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In: Best, K.-H. (ed.), *Glottometrika 16. The distribution of word and sentence length: 213-275.* Trier: WVT.

Rottmann, O. (2001). Sentence length in Old Church Slavonic. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: Levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 251-255.* Trier: WVT.

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: Levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 266-282.* Trier: WVT.

2.30. Sequence of valencies

Problem

Valency of verbs has been discussed especially in Vol. 3 of *Problems*. Study the sequences of valencies in texts, their properties and compare both texts and text sorts. The definition and description of verb valencies can be found in many “qualitative”-linguistic studies. One may consult also dictionaries of verb valencies. In any case, one may be confronted with different definitions.

Procedure

Take a text and create a sequence containing the number of valencies of each verb, i.e. construct a numerical sequence. Define exactly your conception of valency and its evaluation. Having the sequence, study the following properties:

(1) Set up the empirical distribution of the valencies for each text separately.

(2) Compute some properties of the distribution, e.g. mean, variance, Ord’s criteria.

(3) Compare the homogeneity of distributions, e.g. by using the chi-square test, and state whether texts of the same text sort are “similar” or whether texts can be classified according to the degree of verb valency.

(4) Find a theoretical model of the distribution, i.e. derive it using linguistic arguments. A more comfortable way is to find inductively the distribution or a function using a software, then deriving the model a posteriori and substantiating it linguistically.

(5) Dissect the sequence of valencies in Köhler-Naumann's motifs and study their properties, i.e.
(6) Set up their length-distributions and characterize them.
(7) Compare the analyzed texts.
(8) Set up the rank-frequency distribution of motifs, characterize it and derive it theoretically.
(9) Compare the rank-frequency distributions of motifs in individual texts. If possible, classify text sorts.
(10) Perform the analysis using the same text in various languages.

Solve several problems enumerated in the problem *Sequences in text* in this volume.

References

Ágel, V. (2000). *Valenztheorie*. Tubingen: Narr.

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

Budai, L. (1997). *Morphosyntactic valency classes of English verbs*. Veszprém: Veszprémi Egyetemi Kiadó.

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol. 3*. Lüdenscheid: RAM-Verlag.

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

Dixon, R.M.W., Aikhenvald, A.Y. (eds.) (2000). *Changing valency. Case studies in transitivity.* Cambridge: Cambridge University Press, 2000.

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

Helbig, G. (ed.) (1971). *Beiträge zur Valenztheorie*. The Hague-Paris: Mouton.

Helbig, G. (1992). *Probleme der Valenz und Kasustheorie.* Tübingen: Niemeyer.

Helbig, G., Schenkel, W. (1991). *Wörterbuch zur Valenz und Distribution deutscher Verben.* Berlin: de Gruyter.

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

Ivanová, M. (2006). *Valencia statických slovies.* Prešov: Filozoficka fakulta Prešovskej university.

Karlík, P. (2000). Hypoteza modifikované valenčni teorie. *Slovo a slovesnost 61, 170 -189.*

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

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

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

Köhler, R. Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data analysis, machine learning and applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin-Beidelberg: Springer.

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.

Köhler, R. Tuzzi, A. (2015). Linguistic modelling of sequential phenomena. In: Mikros, G., Mačutek, J. (eds.), *Sequences in Language and Text: 109-123.* Berlin/Boston: de Gruyter Mouton.

Lopatková, M., Panevová, J. (2004). Valence vybraných skupin sloves (k některým slovesům dandi a recipiendi. In: *Čeština – univerzalia a specifika: 5, 348-356*. Praha: Nakladatelstvi Lidove noviny.

Nižníková, J., Sokolová, M. (1998). *Valenčný slovník slovenských slovies*. Prešov: Filozofická Fakulta Prešovskej Univerzity.

Panevová, J. (1999). Slovesná reciproční zájměna a slovesná valence. *Slovo a slovesnost 60, 269 –275.*

Panevová, J. (1998). Ještě k teorii valence. *Slovo sa slovesnost, 59, 1 – 13.*

Panevová, J. (2005). Sloveso: centrum věty; valence: centralni pojem syntaxe. In: *Aktuálne otázky súčasnej syntaxe: 73-77*. Bratislava: Veda.

Sanada, H. (2010). Distribution of motifs in Japanese texts. In: Grzybek, P., Kelih, E., Mačutek, J. (eds). *Text and Language. Structures – Functions – Interrelations.Quantitative Perspectives: 183-193*. Wien: Praesens.

Schumacher, H. (1986). *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben*. Berlin-New York: de Gruyter.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol 1.* Lüdenscheid: RAM-Verlag.

Svozilová, N., Prouzová, H., Jirsová, A. (1997). *Slovesa pro praxi. Valenčni slovnik nejčastějšich slovenských sloves.* Praha: Akademie věd České republiky,

Welke, K. (1988). *Einführung in die Valenz- und Kasustheorie*. Leipzig: Enzyklopädie.

3. Grammar

3.1. Adnominal modifiers: Symmetry

Problem

Adnominal modifiers may stay in front of the noun (= Left, L) or behind the noun (= Right, R). Some of them may be interrupted by the noun, i.e. embrace the noun. State the symmetry of the occurrence of adnominal modifiers and compare some text sorts.

Procedure

First, use the current literature to be able to identify adnominal modifiers. Now take several texts of the same text sort, e.g. written by the same author, and for each text compute the number of all adnominal modifiers (= n), the number of left modifiers (L) and the number of right modifiers (R). Omit all the other ones or take into account also embracing adnominals (E). In that case some computations will be different.

Then use the binomial test to state the extent of asymmetry of positions: if there are more than a half of L-modifiers, compute the probability (under the condition p = 0.5 and the given n) that the number of L-modifiers is as given or more, i.e. the sum of probabilities from L to n. If L is smaller than R, the procedure may be performed also for the smaller proportion (then from zero to L), the result is the same.

Another possibility is to perform the asymptotic normal test because the frequencies are usually large. The results are approximately equal.

Characterize a text-sort, the development of the writer, compare two text-sorts, e.g. poetry and science, compare two languages. Some languages prefer left modifiers, other ones right ones, still other ones use them in a balanced way. Order the languages (text sorts) according to the extent of asymmetry and link the results with other properties of language.

If you succeed to find several syntactic properties correlated with the adnominal symmetry, set up a syntactic control cycle and find the respective formulas.

References

Best, K.-H., Boschtan, A. (2010). Diversification of simple attributes in German. *Glottotheory 3(2), 5-9.*

Givón, T. (2001). *Syntax 1, 2.* Amsterdam-Philadelphia: Benjamins.

Gunkel, L., Zifonun, G. (2009). Classifying modifiers in common names. *Word Structure 2(2), 205-218.*

Halliday, M.A.K. (2004) *Introduction to functional grammar*, 3rd ed, London: Hodder Arnold.

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

Köhler, R., Altmann, G. (2014). *Problems in Quantitative Linguistics Vol. 4.* Lüdenscheid: RAM-Verlag.

Rijkhoff, J. (2004). *The Noun Phrase.* Oxford: Oxford University Press.

Zifonun, G., Hoffmann, L., Strecker, B. (1997). *Grammatik der deutschen Sprache*. Berlin: de Gruyter.

3.2. Morphological complexity 1

Problem

Collect the literature on morphological complexity, compare the individual qualitative descriptions, use the proposed indicators; derive for the indicators at least their variances and evaluate texts from different languages, text sorts and authors. Perform also a historical (evolutionary) comparison of your results.

Set up a hypothesis on the relation of morphological complexity of word to word length (distribution) and test it using the same data.

Procedure

Compare critically the individual definitions of complexity. If the indicator is a proportion, use the properties of the binomial distribution, otherwise derive the variance and compare the entities (languages, text-sorts, authors) using an asymptotic normal test.

Take into account the fact that a special morpheme can express at the same time different grammatical categories, e.g. number, case, gender, person, time, mode, aspect, etc. Further, consider the fact that some morphemes have only a phonetic value, some are interrupted, other ones can be detached from the stem, etc. Perform different scalings expressing all these peculiarities. Let your scaling open for further improvements.

Then take texts and replace the individual words by their complexity values. Then do the same with word length measured in terms of syllable numbers. You obtain two vectors for each text. Express the distribution of complexity and that of word length by a probability distribution (or a simple function), then find the relationship (at least some kind of regression) between complexity and length comparing the two vectors.

If you do not find a link, check your definitions, scaling and data. If they are satisfactory, set up a new hypothesis.

References

Altmann, G. (2014). On morphological complexity in Indonesian. *Glottometrics 29, 59-69.*

Altmann, G., Roelcke, Th. (2015). Morphological complexity of the word. *Glottotheory 6(3), 93-111.*

Anderson, S.R. (2012). Dimensions of morphological complexity. In: M. Baerman, D. Brown & G.G. Corbett (eds.), *Understanding and measuring morphological complexity*. Dept. of Linguistics, Yale University http://cowgill.ling.yale.edu/sra/dimensions_revised.pdf

Bane, M. (2008). Quantifying and measuring morphological complexity. In: Ch.B. Chang, H.J. Haynie (eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics: 69-76*. Somerville, MA: Casca-dilla Proceedings Project. http://www.lingref.com/cpp/wccfl/26/paper1657.pdf

Juola, Patrick (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics 5(3), 206–13.*

Nichols, J. (2007). The distribution of complexity in the world's languages. *81st Annual Meeting of the Linguistic Society of America.*

Rice, K. (2012). Morphological complexity in Athabaskan languages: A focus on discontinuities. Presented at *SSILA workshop on Morphological Complexity in the Languages of the Americas, Annual Meeting of the Linguistic Society of America*, Portland, Oregon.

Shosted, R. (2006). Correlating complexity: A typological approach. *Linguistic Typology 10, 1–40.*

3.3. Morphological complexity 2

Problem

Compute the morphological complexity of individual words in a text and express the properties of the text by various methods.

Procedure

Use e.g. the measurement proposed by Roelcke, Altmann (2014). The complexity of each word is given by the sum of its complexity degrees. Set up the vector of the text.

Having a time series compute:

(1) The distribution of complexities in form of a continuous function.

(2) The distances between equal complexities. Set up the distribution of these distances and find a model for it. Compute the properties of this distribution and compare them with those obtained from other texts. Test whether there is a difference between text sorts or between languages. Since you obtain positive real numbers for complexity, you can find an adequate continuous function (without normalization) instead of a distribution.

(3) The number of runs of complexities in the text. If the number of runs is small, you have most probably to do with an analytic language. Use a property of the runs to characterize the degree of analytism/synthetism.

(4) Compute the Hurst exponent for the given sequence. Is the series volatile or persistent? What consequences can you draw for the given text/ language?

(5) Set up intervals of complexity, assort all words to individual intervals and study the forms of runs in text, e.g. their distribution. Take the same text translated into different languages and compare the number of runs using a statistical test. Is the number of runs linked with another property of the language? Find the links that must exist.

(6) Segment the text in complexity motifs. Set up the distribution of motif length and motif frequency.

References

See the problem *3.2. Morphological complexity 1*.
See the problems *7.1, 7.2, 7.3* in this volume
Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

3.4. Morphological changes and frequency

Hypothesis

Mačutek and Čech (2013) suppose that “the greater the magnitude of a morphophonetic change, the lower the frequency of word forms with the magnitude”. They apply it to Czech nouns and use inductively the formula $y = a(x + 1)^{b}e^{-cx}$, where y is the frequency, x is the number of changes in the word, and a, b, c are the parameters. Test the hypothesis in any language which is sufficiently synthetic and substantiate linguistically the above mentioned or a different formula.

Procedure

Take any language having declension or conjugation, then take a text and measure the extent of morpho-phonetic changes in all words of the given class (e.g. nouns, verbs,…). Do not evaluate the written but the phonetic form. Count the number of words displaying *x* changes and set up the distributions of individual parts-of-speech. Do not set up a common result for nouns and verbs etc., count them separately.

First apply the Mačutek/Čech formula and if it holds true in all cases, derive it from a differential equation interpreting the individual components as Köhlerian (2005) forces in terms of speaker, hearer and language norm. If the formula does not hold in all cases, search for boundary conditions leading to the different outcome, or modify the differential equation, derive a new formula and fit it to all data. Analyze the problem as long as you obtain satisfactory results in all cases.

Place your result in Köhler's control cycle, i.e. link "number of morphological changes" with other properties – not only frequency.

References

Bybee, J. (1985). *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: Benjamins

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.

Krug, M. (2003). Frequency as a determinant in grammatical variation and change. In: Rhodenburg, G., Mondorf, B. (eds.), *Determinants of Grammatical Variation in English: 7-67.* Berlin: de Gruyter.

Mačutek, J., Čech, R. (2013). Frequency and declensional morphology of Czech nouns. In: Obradović. I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: 59-68.* Belgrade: Academic Mind.

3.5. Word classes

Problem

Parts-of-speech or different word classes are usually determined on the basis of the traditional Latin grammar. But there are also syntactic criteria furnishing us about 100 classes. One can use also semantic criteria. But whatever our criteria, we do not obtain a classification that could be used generally, i.e. in all languages. We mostly forget that criteria are conventions and no real conditions.

Set up different classifications of words using different criteria, e.g. semantic, syntactic, morphological. Find criteria for the "most reasonable" classification. Start from the assumption that the "best" classification is that which links the results to other language properties, i.e. search for its synergetic substantiation.

Procedure

Take a monolingual dictionary and for each entry form a *set* of classes it may belong to. You can use also WORDNET. For example “keep” may have the defining set [transitive verb, intransitive verb, noun]; in German, all verbs in infinitive can be used as nouns, prepositions can be used as adverbs if they occur as separated verbal prefixes, many adjectives are at the same time adverbs, etc. Form as many different sets as necessary.

Do not forget that there is no “true” classification, each depends on our criteria. In order to strengthen the sense of your criteria, perform a count using a dictionary and state how many words belong to individual classes. Take the first 1000 words, later on analyze the complete dictionary. Then set up the rank-frequency distribution of the classes, i.e. order them. Find a probability distribution or a function capturing this ranked sequence. If you do not obtain acceptable results known from linguistics, redefine the classes and repeat the whole procedure until you obtain acceptable results. Realize that if in a language classes are isolable, they are in some relation to one another and this relation can be expressed quantitatively.

You can choose abbreviations for the classes and rewrite a text replacing the words by these abbreviations. In texts, (mostly) each word belongs to exactly one class but in the dictionary you may obtain complex classes, e.g. the German *schnell* belongs to the class *Adj-Adv.*

Now you can perform two operations: (1) Study the distribution of abbreviations (= classes) and different properties of this distribution; (2) study the sequence of abbreviations displaying distances, runs, autocorrelations, Markov chains, transition probabilities, co-occurrence tendencies, motifs, etc. That means, you can deepen the study of parts-of-speech (word classes) using some statistics.

If you obtain some functions, substantiate them by subsuming them in the unified theory (Wimmer, Altmann 2005).

Show some differences between languages: historical, genealogical, typological, areal. Make the first steps towards a theory. In practice, all models you use must be derived and substantiated linguistically, and there must be some links to other properties of language (or only words). Since word classes arose by diversification, study also this discipline.

References

Anward, J. (2000). A dynamic model of part-of-speech differentiation. In: Vogel, P.M., Comrie, B. (eds.), *Approaches to the Typology of Word Classes: 2-45*. Berlin: de Gruyter,

Bergenholtz, H., Mugdan, J. (1979). *Einführung in die Morphologie*. Stuttgart: Kohlhammer.

Bergenholtz, H., Schaeder, B. (1977). *Die Wortarten des Deutschen.* Stuttgart: Klett.

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

Bünting, K.-D., Bergenholtz, H. (1989). *Einführung in die Syntax.* Frankfurt: Athenäum.

Crystal, D. (1967). English word classes. *Lingua 17, 24-56.*

Grundzüge einer deutschen Grammatik (1981). Berlin: Akademie

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: Hřebíček, L. (ed.), *Glottometrics 11, 142-156*. Bochum: Brockmeyer.

Helbig, G. (1977). *Beiträge zur Klassifizierung der Wortarten.* Leipzig: Enzyklopädie.

Kaltz, B. (1983). *Zur Wortartenproblematik aus wissenschaftsgeschichtlicher Sicht*. Hamburg: Buske.

Köhler, R. (2012). *Quantitative Syntax Analysis.* Berlin-Boston: Mouton de Gruyter.

Ossner, J. (1989). Wortarten: Form- und Funktionsklassen. *Zeitschrift für Literaturwissenschaft und Linguistik 19, 94-117.*

Schweers, A., Zhu, J. (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (ed.), *Diversification Processes in Language: Grammar: 157-165.* Hagen: Rottmann.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative Analysis of Italian texts.* Lüdenscheid: RAM-Verlag.

Vincze, V. (2013). Domain differences and the distribution of parts of speech and dependency relations in Hungarian. *Journal of Quantitative Linguistics 20(4), 314-338.*

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

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: de Gruyter.

3.6. Parts-of-speech distribution

Problem

A. Ziegler (1998) analysed 21 Brazilian-Portuguese press texts and stated the rank-frequency distribution of word classes. In order to obtain good fits, he ordered the classes in each text according to frequency. Using an inductive approach he obtained two distributions, viz. the negative hypergeometric and the mixed Poisson. Restate the problem and find a linguistically founded background.

Procedure

First, define a fixed order of word classes. This can be made in such a way that one takes the mean rank of all words of the given class in all texts. The new ranks will not necessarily be discrete, they represent rather degrees.

Then using this order for each of the texts separately, ascribe them the frequencies found. You will not obtain in each case a monotonously decreasing sequence of frequencies. You have data displaying, perhaps, 21 different images.

Find a continuous function with a small number of parameters that can be fitted to all data. This can be done by means of a software, e.g. TableCurves. Choose a simple function adequate for all data, derive its differential equation and subsume it under the unified theory proposed by Wimmer and Altmann (2005). Take into account the systemic requirements of synergetic linguistics as proposed by R. Köhler (2005) in order to substantiate the individual parameters.

Search for distributions of parts-of-speech in the literature in order to obtain also results from other languages. Perform the same procedure as above, then test whether the word classes have similar mean ranks in all accessible languages. Continue to obtain results from many languages in which one can ascribe words to the 9 classical classes.

Strive for a theory. First give reasons for the given ranking in every language separately. Show the relationships with the grammar of the given language.

For extending and generalizing your research set up a different classification of words or use some known classification taken from the literature. Then analyse individual texts - do not use a corpus as a whole! - and repeat the whole procedure.

You can consider a more complex classification, e.g. first stating the part-of-speech of an entity, then its grammatical function. In this way each word can belong to different classes. The ranking yields then different results. Continue in constructing ever finer hierarchy. Can you predict the end of your hierarchization?

References

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

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

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5(3), 260-280.*

3.7. Parts-of-speech homogeneity

Problem

Set up a hypothesis concerning the equality of frequencies of individual parts-of-speech in texts in one language. Analyze several texts and test the hypothesis.

Comparing several languages make conjectures concerning the representation of word classes in languages and set up an elementary typology. Compute some properties of the distributions and link them with other properties.

Procedure

First use published data. Ziegler (2001) presented the frequencies of parts-of-speech in 20 Portuguese texts. Use his data. Prepare a table with 9 word classes and for each text ascribe the respective word class its rank in the distribution. You obtain a 9x20 contingency table.

Now test the columns for homogeneity using some non-parametric test.

Then test the homogeneity of all samples using directly the frequencies of individual word classes.

In these cases you obtain only one table for ranking and one table for frequencies.

Compare the results from Portuguese with other languages (cf. Best 1994, 1997, 2000; Hammerl 1990; Schweers, Zhu 1991; Ziegler 1998). Collect all published results and begin to generalize. State the role of individual word classes in the given language “type”. Characterize the distribution using a special indicator and find the link of this indicator to another property of the language, i.e. begin to construct a control cycle of the Köhlerian (2005) type in which different properties of word classes are taken into consideration.

References

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

Best, K.-H. (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. In: Best, K.-H. (ed.), *Glottometrika 16, 276-285.* Trier: WVT.

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

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

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

Schweers, A., Zhu, J. (1991). Wortartenklassifikation im Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (ed.), *Diversification Processes in Language: Grammar: 157-167.* Hagen: Rottmann.

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese texts. *Journal of Quantitative Linguistics 5(3), 269-280.*

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Linguistics Paradigm: Levels, Constituents, Construct. Festschrift in honour of Luděk Hřebíček: 266-282.* Trier: WVT.

3.8. Clause centrality

Problem

Compute clause centrality in texts of different text sorts and set up a hypothesis.

Procedure

Clause centrality can be measured in terms of the position of the finite verb in the clause. State the position of the verb and count the number of words in front of the verb (*f*) and behind the verb (*b*). You obtain a sequence

$$a_{f,}a_{f-1,}\dots,a_3,a_2,a_1V\ a_1,a_2,\dots,a_{b-1},a_b$$

where *V* is the finite verb and a_i are the individual words. Compute the centrality according to

$$C = 1 - \frac{|b-f|-\delta}{b+f},$$

where *b* and *f* are the greatest indices and

$$\delta = \begin{cases} 0 \text{ if } (b+f) \text{ is an even number} \\ 1 \text{ if } (b+f) \text{ is an odd number} \end{cases}.$$

and C = 1 when $b + f = 0$. *C* varies in the interval <0; 1>. The smaller is the difference between b and f, the more centralized is the clause. Compute C for all clauses in the text; then compute the mean centrality, $\bar{C}$. At last, set up the intervals <0; 0.10>, <0.11; 0.20>, <0.21; 0.30>,…, <0.91; 1.0> and state the number of clauses having the centrality in these intervals. Propose a model of the distribution of centrality. You can use a continuous function or transform the intervals in discrete values 0,1,2,…,9 and propose a discrete probability distribution or simply a discrete sequence. Test the function or sequence. If possible, use longer texts.

Now order the texts according to their average *C* or according to the parameter(s) of the proposed function/sequence. Show that different text-sorts have different averages. To this end compute the variance of *C* for each text and perform the asymptotic normal test for the difference of two averages, or compare the frequencies using e.g. the chi-square test.

In order to make a step towards theory, link the average degree of centrality of a text with other text properties which are already quantified. Perform this investigation inductively: compute the centrality and some other property for many texts and scrutinize the form of the "dependence". If you have indicators of other text properties at your disposal, set up an elementary control cycle and substantiate it linguistically.

At last compare texts belonging to the same text-sort in different languages in order to see whether text-sort or language is the influencing factor.

In some languages, the copula is not always expressed but latently present. In such cases decide where the boundaries of the clause are and where the copula should stay. Sometimes there are ellipses of verbs whose character must be decided ad hoc. The counting of words in front of and behind the verb is not simple: one must decide about the nature of compounds, clitics, numbers, detachable affixes, etc. that is, the qualitative problems must be solved before one begins to count. Decide whether the finite form is the auxiliary (or modal) verb or the main verb which may have an infinitive form. Or consider the complete verbal form as one verb (e.g. *I would like to go…*)

Be aware of the fact that decisions about grammar are no signs of truth but conventional criteria. They do not differ from the conditions in mathematical theorems: "Let be given …, then it holds that…". The "givenness" is the result of your analysis based on conventional criteria, but no feature of reality. If one sets up other definitions, other results may be expected.

Define and measure centrality of the clause in a different way and compare your results with those performed in the above way.

References

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

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

Householder, F.W.Jr. (1960). First thoughts on syntactic indices. *International Journal of American Linguistics 267, 195-197.*

Popescu, I.-I., Altmann, G. (2014). Clause centrality. *Glottometrics 28, 12-35.*

Pustet, R. (2005). *Copulas. Universals in the Categorization of the Lexicon.* New York: Oxford University Press.

Uhlířová, L. (2005). Word order variation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 598-606.* Berlin: de Gruyter.

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

3.9. Clause types

Problem

There are a number of classification possibilities for clause types because one may consider the number of their properties as never exhausted. For evaluation purposes one can exploit ready classifications and computations. In the article by Levickij, Pavlyčko, Semenyuk (2001) one finds a table (see below) containing the numbers of clause types in works of 4 German authors. (1) Test whether the frequencies with individual authors are internally uniform. (2) Test whether the four authors are homogeneous. (3) Set up the empirical rank-frequency distribution with individual authors and find an adequate distribution. (4) Characterize the concentration of the distributions.

Procedure

Consider the table displayed by the authors:

Table 1
Frequency of types in subordinated clauses

Types of clauses	Authors			
	Böll	Kant	Mann	Remarque
Subjective	16	12	6	16
Predictive	2	4	2	24
Objective	124	146	96	134
Attributional	194	158	282	166
Temporal	122	82	92	74
Local	42	6	18	8
Causal	18	30	12	28
Final	4	6	2	8
Comparative	20	42	20	28
Conditional	28	84	20	66
Modal	10	2	18	12
Concessional	6	8	16	8
Consecutive	14	20	16	28
	600	600	600	600

The data was collected from one work by Böll and 2 works by Kant, Mann and Remarque, in order to obtain the same number of clauses.

(1) Test the uniformity of a column using the simple chi-square homogeneity test. Most probably none of the four columns will be homogeneous. Take other texts and consider them individually. Compute the uniformity and express it with an indicator.

(2) Compare the individual texts of all authors for homogeneity. Then compare the results with other text sorts. Test again for homogeneity, considering all texts.

(3) For each text/author/text-sort propose a rank-frequency distribution and test it on the data. Start with empirical functions using software, later on derive the adequate function. How do the parameters differ? Are there differences between some indicators (properties of the distributions)?

(4) Dedicate a special study to the concentration of the texts to special clauses. For this purpose use any indicator that can be interpreted in this sense, e.g. mean, variance, excess, entropy, repeat rate, Ord's criteria, etc.

State whether stage plays, poems and press texts differ in this sense. If so, search for other properties of these text sorts and link them with some of the clause properties.

Continue constructing the control cycle connecting as many properties of clauses as possible. Strive for a theory.

References

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

Levickij, V.V., Pavlyčko, O.O., Semenyuk, T.G. (2001). Sentence length and sentence structure as statistical characteristics of style in prose. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 177-186.* Trier: WVT.

3.10. Clause length

Problem

Uhlířová (2001) studied clause length in terms of word numbers in Bulgarian and fitted to the empirical data the mixed negative binomial distribution. Since this distribution has 7 parameters, find a simpler (not normalized, continuous) function capturing the data.

Procedure

The data presented by Uhlířová (2001) are as follows:

Table 1
Clause lengths in three Bulgarian texts (Uhlířová 2001)

Length	Text 1	Text 2	Text 3
1	14	3	2
2	82	17	6
3	95	13	15
4	115	16	28
5	127	15	25
6	123	14	17
7	103	15	10
8	91	12	11
9	72	3	8
10	53	6	6
11	47	4	3
12	32	2	1
13	22	2	2
14	18	1	3
15	13	1	0
16	9	0	2
17	7	0	0
18	9	1	0
19	4	0	0
20	1	0	2

Find the appropriate function using software, i.e. find a model mechanically. You obtain several good results. Then derive the functions from differential equations and interpret their components. Keep the function whose interpretation is linguistically well substantiated. Lean against the unified theory (Wimmer, Altmann 2005)

If other languages or texts are at your disposal, (1) compare the present results and order the languages; (2) investigate the clause length in other text-sorts and construct, step by step, a typology. If possible, use the same text translated to languages you know and can analyze.

References

Popescu, I.-I., Best, K.-H., Altmann, G. (2015). *Unified modeling of length in language*. Lüdenscheid: RAM-Verlag.

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Lin guistics Paradigm: Levels, Constituents, Construct. Festschrift in honour of Luděk Hřebíček: 266-282.* Trier: WVT.

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: de Gruyter.

3.11. Topic – comment

Problem

Several linguistic schools studied topic and comment in the clause (or sentence), cf. the given references. The definitions differ slightly but in general, *topic* is that part of the clause/sentence about which a statement is made. *Comment* is the given statement. Problems arising with any definition of topic/comment must be solved conventionally, i.e. by an additional definition (or decision). A quite different approach is used by Beliankou, Köhler, Naumann (2013) based on rhetorical structure and resulting in tree-like structures.

For a given text solve the following problems:

(1) State the number of words in the topic and in the comment respectively in each clause/sentence. Then set up the distributions of topic and comment separately.

(2) State the position of the topic and count the number of words in front (F) of the topic and behind (B) it. Set up the empirical distribution of F and B separately and find a model.

(3) Compute the difference between the number of words in F and B separately for every clause/sentence and set up the distribution of these differences. This yields an indicator of symmetry or centrality.

(4) Study the sequence of these differences and capture its properties (e.g. roughness, distances between equal numbers, runs, etc.)

(5) Compute the centrality of the topic using some formulas (cf. e.g. Wimmer et al. 2003: 178; problem *Clause centrality*). Find the empirical mean and the variance of these numbers.

Procedure

Consider the criteria uttered by Gundel, Hedberg and Zacharski (2013):

“Topics need not be represented by noun phrases.”

“Topics need not be sentence initial.”

„Topics need not be continuous.”

“An utterance may contain no words associated with the topic.”

Take a text and identify the topic and the comment in each clause/sentence. Adhere to a certain grammar, otherwise you get problems. But even after accepting a unique definition of topic/comment, you will get problems with some sentences. This is a quite normal state of affairs. Omit units of this kind from your computations (but tell it). Then perform the first counting and computation using the given text. After having analyzed several ones, begin to propose models of distributions, to evaluate symmetry/centrality and to study the sequences of numbers you obtained. For solving problem (5) you can use the results from problem (3).

The research in this domain did not even begin, up to now researchers were concerned with definitions, identifications, and presenting of examples, hence the first step will be decisive for further research. Strive for a theory: interpret the results linguistically, derive the formulas and incorporate all in a system of links between them.

State and present exactly the criteria used for the identification of the topic because they are the elementary conditions for the validity of your results.

References

Arnold, J.E., Losongco, A., Wasow, T., Ginstrom, A. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76, 28-55.

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumenation motifs. In: Obradovič, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: 35-43.* Belgrade: Academic Mind.

Büring, D. (1997). *The Meaning of Topic and Focus — The 59th Street Bridge Accent.* London: Routledge.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In: Li, Ch.N. (ed.), *Subject and Topic: 25–55.* New York: Academic Press.

Daneš, F. (1970). One instance of the Prague school methodology: Functional analysis of utterance and text. In: Garvin, P. (ed.), *Method and theory in linguistics*. Paris, The Hague: Mouton.

Daneš, F. (1970). Zur linguistischen Analyse der Textstruktur. *Folia Linguistica 4, 72-78.*

Dik, S.C. (1978). *Functional grammar*. Amsterdam: North Holland.

Donselaar, W.van, Lentz, J. (1994). The function of sentence accents and given/new information in speech processing: Different strategies for normalhearing and hearing-impaired listeners? *Language and Speech 37(4). 375-391.*

Erteschik-Shir, N. (1997). *The dynamics of focus structure*. Oxford: Oxford U.P.

Féry, C., Ishihara, S. (2009). How focus and givenness shape prosody. In: Zimmermann, M., Féry, C. (eds), *Information Structure from Different Perspectives: 36-63.* Oxford: Oxford University Press.

Firbas, J. (1964). On defining the theme in functional sentence analysis. *Travaux Linguistique de Prague 1, 267-280.*

Frey, W. (2000). Über die syntaktische Position des Satztopiks im Deutschen. *ZAS Papers in Linguistics 20, 137–172.*

Gabelentz, G.v.d. (1891). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Leipzig: T.O. Weigel Nachfolger.

Givón, T. (ed.) (1983). *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: Arshdeep Singh.

Gundel, J. (1974). *The role of topic and comment in linguistic theory*. Ph.D. dissertation. University of Texas, Austin.

Gundel, J. (1985). 'Shared knowledge' and topicality. *Journal of Pragmatics 9. 83-107.*

Gundel, J.K. (1988). Universals of topic-comment structure. In: Hammond, M., Moravcsik, E., Wirth, J. (eds.), *Studies in Syntactic Typology: 209-239.* Amsterdam: Benjamins.

Gundel, J.K., Hedberg, N., Zacharski, R. (2013). Topic-Comment Structure, Syntactic Structure and Prosodic Tune**.** *http://www.sfu.ca/~hedberg/Helsinki_paper.pdf (10.10.2013).*

Hajičová, E., Partee, B.H., Sgall, P. (1998). *Topic–Focus Articulation, Tripartite Structures, and Semantic Content. Studies in Linguistics and Philosophy 71*. Dordrecht: Kluwer.

Halliday, M.A.K. (revised by C.M.I.M. Matthiessen) (2004). *An introduction to functional grammar*, 3rd ed., London: Hodder Arnold.

Hockett, Ch.F. (1958). *A Course in Modern Linguistics*. New York: The Macmillan Company.

Householder, F.W.Jr. (1960). First thoughts on syntactic indices. *International Journal of American Linguistics 267, 195-197.*

Jacobs, J. (2001). The dimensions of topic-comment. *Linguistics 39.641-681*.

Jäger, G. (1996). *Topics in Dynamic Semantics*. Ph.D. thesis, Humboldt University Berlin. CIS-Bericht 96-92, Centrum für Informations- und Sprachverarbeitung, Universität München.

Kadmon, N. (2001). *Pragmatics*. Blackwell Publishers.

Köhler R. (2012). *Quantitative syntax analysis*. Berlin/Boston: de Gruyter

Krifka, M. (2006). The origin of topic/comment structure, of predication, and of focusation in asymmetric bimanual coordination. *http://amor.cms.hu-berlin.de/~h2816i3x/Talks/BimanualCoordination.pdf (10-10-2013)*

Lambrecht, K. (1994). *Information structure and sentence form. Topic, focus, and the mental representation of discourse referents*. Cambridge: Cambridge University Press.

Li, Ch.N., Thompson, S.A. (1976). Subject and Topic: A New Typology of Languages. In: Li, Ch.N. (ed.), *Subject and Topic*: *457-490.* New York/San Francisco/London: Academic Press.

Mathesius, V. (1975). *A Functional Analysis of Present Day English on a General Linguistic Basis* (edited by Josef Vachek, translated by Libuše Dušková). The Hague – Paris: Mouton.

Molnár, V. (1991). *Das TOPIK im Deutschen und im Ungarischen*. Stockholm: Almqvist and Wiksell.

Most, R. B. Saltz, E. (1979). Information structure in sentences: New information. *Language and Speech 22, 89-95.*

Nooteboom, S.G., Kruyt, J.G. (1987). Accents, focus distribution and the perceived distribution of given and new information: An experiment. *The Journal of Acoustical Society of America 82(5).1512-1524.*

Payne, Th.E. (1997). *Describing morphosyntax: A guide for field linguists.* Cambridge: Cambridge University Press.

Portner, P., Yabushita. K. (1998). The semantics and pragmatics of topic phrases. *Linguistics and Philosophy 21. 117-157.*

Primus, B. (1993). Word order and information structure: a performance based account of topic positions and focus positions. In: J. Jacobs et al. (eds.), *Handbuch Syntax, vol. 1, 880–895*. Berlin and New York: de Gruyter.

Prince, E.F. (1986). On the syntactic marking of presupposed open propositions. *CLS 22, Parasession 208-222.*

Reinhart, T. (1982). *Pragmatics and linguistics: An analysis of sentence topics*. Bloomington, Indiana University Linguistics Club.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics 1, 75-116.*

Sgall, P. et al. (1987). *The meaning of the sentence and its semantic and pragmatic aspects*. Dordrecht: Reidel.

Vallduvi. E. (1990). *The information component*. Ph.D dissertation, University of Pennsylvania.

Vallduvi, E., Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics 34. 459-519.*

Ward, G.L. (1985). *The semantics and pragmatics of preposing*. Ph.D dissertation, University of Pennsylvania.

Wasow, T. (1997). Remarks on grammatical weight. *Language Variation and Change 9(1). 81-105.*

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

3.12. Study of adverbials 1

Problem

Čech and Uhlířová (2014) classified the adverbial expressions in the sentence in the following classes: Place, Time, Manner, Means, Aspect, Condition, Measure, Cause, Result, Origin, Purpose, Concession, Originator. Study their occurrence distances and positions in texts, develop some hypotheses and…..

Procedure

First take a text and create a sequence of adverbials using some abbreviations. Ignore the rest of the text, consider merely the sequence of adverbials.

In the next step, count the frequencies of individual adverbials. Set up the rank-frequency of classes and find a model of this distribution. The model can also be a usual (non-normalized) function or a series. Compare the rank-frequencies of various texts and state whether there are differences between text-sorts, authors, languages.

Now measure the length of the adverbial expressions in terms of word numbers. Thus “here” has length 1, “at home” length 2, “in the street” length 3, etc. Ascribe these numbers to individual classes. Now for each class you have a distribution of lengths. Find a probability distribution which is adequate for all classes. Compare texts, text sorts and languages. Again, you may apply a “good” function.

You have now the sequence of lengths of adverbials in the text. Study the properties of this sequence: Compute the Euclidean distances between the *subsequent* numbers, add them to obtain the arc and compute the mean arc for the given text. Find the empirical variance of the distances and compare texts, text-sorts, languages.

Then compute the distances between *identical* lengths using some variant of the Minkowski distance. Set up the distribution of distances and compare texts, text-sorts and languages.

Since lengths do not differ drastically, study the runs in the given sequence, express their distribution and perform comparisons.

Set up length motifs according to the Köhler-Naumann method and study separately the length, the mean and the range of *individual* motifs. You obtain three new sequences whose properties may be studied using the above methods.

Find other properties of adverbials, e.g. semantic ones, using different aspects, or their distance from the element on which they depend. Quantify these new properties according to some aspect and perform again all computations as given above.

Whatever property you take, find the transition frequencies between individual classes. Set up a two-dimensional contingency table and perform tests for independence, for the significance of individual cells, for the status of the diagonal and for symmetry. Use the table to test whether the transitions form a Markov chain of the first order.

Set up hypotheses expressing the status quo, substantiate them linguistically and connect all of them into a system of hypotheses in order to obtain a control cycle as used in systems theory (cf. Köhler 2005). Strive for a theory of adverbials for which a classification is merely the first step.

Study the position of the adverbials in sentence. They determine something hence they stay either in front of or behind the determined entity. For each class of adverbials state their frequencies in both positions, compare the proportions using a statistical test and make decisions about the preferred position of the adverbial class.

If possible, find an independent criterion which allows you to measure a property of classes of adverbials and order them not categorically but using the given quantity , e.g. prominence, weight, importance, historical priority, syntactic priority, special semantic features, etc. If you succeed to find such a quantity, study its relationship to the properties mentioned above.

References

Altmann, G. (1987). Tendenzielle Vokalharmonie. In: Fickermann, I. (ed.), *Glottometrika 8, 104-112.* Bochum: Brockmeyer.

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag.

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012.*

Čech, R., Uhlířová, L. (2014). Adverbials in Czech: Models for their frequency distribution (in print).

Diessel, H. (2001). The ordering distribution of main and adverbial clauses: A typological study. *Language 77, 343-365.*

Diessel, H. (2005). Competing motivations for the ordering of main and adverbial clauses. *Linguistics 43, 449-470.*

Diessel, H. (2012). Iconicity of sequence: A corpus based analysis of the positioning of temporal adverbial clauses in English, In: Janda, L.A. (ed.), *Cognitive Linguistics: The Quantitative Turn: 225-250.* Berlin-Boston: de Gruyter Mouton.

Ford, C.E. (1993). *Grammar in Interaction. Adverbial clauses in American English conversation.* Cambridge: Cambridge University Press.

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

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

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

Köhler, R., Naumann, S. (2009). *A contribution to quantitative studies on the sentence level*. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 34-57*. Lüdenscheid: RAM-Verlag.

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: 81-89*. Wien: Praesens.

Popescu, I.-I., Zörnig, P., Grzybek, P., Naumann, S., Altmann, G. (2013). Some statistics for sequential text properties. *Glottometrics 26, 50-95.*

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

Thompson, S.A., Longacre, R.E. (1985). Adverbial clauses. In: Shopen, T. (ed.), *Language Typology and syntactic description Vol II: 171-234.* Cambridge: Cambridge University Press.

Uhlířová, L. (1975). O frekvenci příslovečného určení v souvislém textu. *Naše řeč 58, 133–142.*

Verstraete, J.-C. (2004). Initial and final position of adverbial clauses in English: the constructional basis of the discoursive and syntactic differences. *Linguistics 42, 819-853.*

3.13. Study of adverbials 2

Problem

In order to set some hierarchies, study separately each class of the following adverbials: Place, Time, Manner, Means, Aspect, Condition, Measure, Cause, Result, Origin, Purpose, Concession, Originator. First define some quantitative properties, then state their distributions.

Procedure

Take a longer text and consider all adverbials of the given class. First study their properties mentioned in the previous problem (3.11. *Study of adverbials 1*).

(1) Then consider the occurrence of parts-of-speech in them. For example “in the street” contains a preposition, an article and a noun. For the adverbials of the chosen class prepare a frequency distribution of parts-of-speech concerning the given text. Do the adverbials of the given class tend to have the same form or is there a distribution?

(2) Compute the concentration of the tendency using the Repeat rate and the Entropy: The greater is the Repeat rate, the smaller is the variation of forms; the greater is the Entropy, the greater is the variation of forms.

(3) In the given class, quantify the complexity of the adverbial using your own definitions of complexity. It may be morphological, syntactic or semantic (also metaphoric), etc. Then state the distribution of complexities and compute its properties. For morphological complexity see Altmann, Roelcke (2015).

(4) Taking the mean of complexities, compare all groups of adverbials and rank them. Derive the distribution of complexities and substantiate this result linguistically.

(5) Compute Ord's criteria for all quantified properties separately,– it is possible even if one does not set up distributions – and enter the values <I,S> in a two-dimensional chart marking each adverbial class separately. You obtain points forming a straight line or a curve or an ellipse. Fit a simple function to these points and interpret it.

References

Cf. Problem "Study of adverbials 1".

Altmann, G., Roelcke, Th. (2015). Morphological complexity of the word. *Glottotheory 6(1), 95-111.*

3.14. Study of adverbials 3

Problem

Study the existence of runs of adverbials in texts. Test the alternative hypothesis that there are too many runs.

Procedure

Applying the classification of adverbials in: Place, Time, Manner, Means, Aspect, Condition, Measure, Cause, Result, Origin, Purpose, Concession, Originator, as performed by Čech and Uhlířová (2014) analyze texts in the following manner: Mark each class with a different letter (e.g. P = Place, T = Time,…) and set up a vector whose elements are the adverbials in text in the order of their appearance. For some languages/texts you may obtain a smaller set of classes, e.g. Yesypenko (2009) has for English seven classes (Time, Repetition and Frequency, Place and Direction, Condition and Consequence, Manner, Degree and Quantity, Question) . As a matter of fact, you have now a sequence of 10 (or fewer/more) different letters. Study the existence of runs applying usual methods.

In order to compare texts, prepare a vector of frequencies of individual letters/symbols. The elements of the vector should have the same order as given above. (1) For each pair of texts compute the distance of the vectors using the arccos-function. (2) State the mean distance among all texts of a given author and compare with it the mean of another author or text sort. (3) Compare the same text sorts in different languages. (4) Take the same text in different languages (e.g. The Little Prince by Exupéry) and compare the languages.

References

Problems: Study of adverbials 1 and 2 in this volume.

Čech, R., Uhlířová, L. (2014). Adverbials in Czech: Models for their frequency distribution (in print).

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

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*: *308-327*. Černivci: Černivec'kij nacional'nyj universitet.

3.15. Grammatical categories

Problem

Consider a grammatical category in a strongly synthetic language, e.g. case. First show that the frequency distribution of the cases has a stable background. Then show that the size of affixes depends on their frequency. At last, study the meaning diversification of individual affixes and find the theoretical distribution.

Procedure

Begin with grammatical cases. Take a longer text and state the occurrence of individual grammatical cases; then the morphemes expressing them (there may be several ones because of the interaction with gender, number, etc.) and the lengths of the morphemes. Show the relationship between frequency of the case and the number of morphemes (means) expressing it. Show the relationship between the frequency of the case and the average length of morphemes expressing it.

Then scrutinize the meaning of individual cases: what does express e.g. the genitive? State all meanings or functions and show that there is a regular rank-frequency distribution of frequencies of meanings/functions. Here, a grammar must be used, a unique text is not sufficient. Show that the more frequent a case, the more meanings/functions it has.

Study all parts-of-speech having the given category.

Study only inflectional languages. In strongly agglutinative languages some of the above relations do not hold true. Why?

Perform comparisons with existing analyses.

Study the behavior of all grammatical categories and of the morphemes expressing them.

References

Hanulíková, A., Davidson, D.J. (2009). Inflexional entropy in Slovak. In: Levická, J., Garabík, R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar research: 145-151*. Brno: Tribun.

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

Kostić, A., Mirković, J. (2002). Processing of inflected nouns and levels of cognitive sensitivity. *Psihologija 35, 287-297.*

Krajewski, G., Lieven, E.V.M., Theakston, A.L. (2012). Productivity of a Polish child`s inflexional noun morphology: a naturalistic study. *Morphology 22, 9-34.*

Mačutek, J., Čech, R. (2013). Frequency and declensional morphology of Czech nouns. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: 59-68.* Belgrade: Academic Mind.

Milin, P., Filipović Durdević, D., Moscoso del Prado Martin, F. (2009). The simultaneous effects of inflexional paradigms and classes of lexical recognition. Evidence from Serbian. *Journal of Memory and Language 60, 50-64.*

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

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

3.16. Reduplication

Problem

Reduplication exists in many (perhaps all) languages. Study its forms, frequencies and meanings. Set up distributions and show the links between their properties and other properties of language.

Procedure

Take a longer text in a language and prepare a list of all reduplications found. Set up classes of forms. Distinguish reduplication classes, e.g. full, partial, with variation, with assimilation, according to length, etc.

Then (1) set up the rank-frequency distribution of the classes obtained. Find a preliminary theoretical function or distribution expressing it. Compute some properties of the empirical distribution e.g. repeat rate, entropy, Ord's criterion, Gini's coefficient, steepness.

(2) The elements in each class express some general meaning, e.g. repetition, performer, instrument, some grammatical functions, extent, etc. For each form class set up the rank-frequency distribution of its semantic diversification. Compute again some properties of the empirical distribution for each class separately and compare the classes.

(3) Find the dependence, e.g. a class from (1) with great repeat rate has a small diversification stated in (2). Find the given relation and capture it by a formula. Substantiate the formula both theoretically and linguistically.

(4) State the parts of speech in the given language and study the occurrence of reduplication in the given part of speech class.

(5) State the syllabic length of reduplicated words and set up the respective distribution. Compute, again, some of its properties.

(6) Do lengths correlate with parts-of-speech classes? Do lengths correlate with the frequency? Do lengths correlate with the diversification?

(7) Some languages prefer special kinds of reduplication. Take some other languages and perform a comparative study.

(8) State whether the individual classes tend to contain some sound-symbolic phenomena. If so, show the extent of sound symbolism and relate it to the size of the class.

References

Botha, R.P. (1988). *Form and meaning in word formation: a study of Afrikaans reduplication*. Cambridge: Cambridge University Press.

Broselow, E., McCarthy, J.J. (1984). A theory of internal reduplication. *The linguistic review* 3, 25-88.

Bzdęga, A.Z. (1965). *Reduplizierte Wortbildung im Deutschen*. Poznan: Polska Akademia Nauk.

Fabricius, A.H. (1998). A comparative survey of reduplication in Australian languages. *LINCOM Studies in Australian Languages 3.* Munich: Lincom Europa.

Haeberlin, H. (1918). Types of reduplication in Salish dialects. *International Journal of American Linguistics 1, 154–174.*

Hurch, B., Mattes, V. (2009). Typology of reduplication: The Graz database. In: Martin Everaert, Simon Musgrave, Alexis Dimitriadis (eds.), *Use of data-*

bases in cross-linguistic studies: *301–328*. New York: Mouton de Gruyter.

Key, H. (1965). Some semantic functions of reduplication in various languages. *Anthropological Linguistics 7(3), 88-101.*

Moravscik, E. (1978). Reduplicative constructions. In: Greenberg, J. (ed.), *Universals of human language*. Vol. 3, *Word structure: 297–334.* Stanford, CA: Stanford Univ. Press.

Rubino, C. (2005). Reduplication: Form, function, and distribution. In: Hurch, B., Mattes. V. (eds.), *Studies on reduplication: 11–29.* New York: Mouton de Gruyter.

Schindler, W. (1991). Reduplizierende Wortbildung im Deutschen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 44, 597–613.

Scullen, M.E. (2001). New insights into French reduplication. In: Wiltshire, C., Camps, J. (eds.), *Romance Phonology and Variation: Selected papers from the 30th Linguistic Symposium on Romance Languages*. Philadelphia: Benjamins.

Skalička, V. (1936). Notes sur le redoublement. *Sborník Matice Slovenskej 14, 19-22.*

Stachowski, K. (ed.) (2014). *Standard Turkic C-type reduplications.* Kraków: Jagellonian University Press.

Thun, N. (1963). *Reduplicative words in English: A study of formations of the types tick-tock, hurly-burly, and shilly-shally*. Uppsala.

Van Huyssteen, G.B. (2004). Motivating the composition of Afrikaans reduplications: a cognitive grammar analysis. In: Radden, G., Panther, K-U. (eds.). *Studies in Linguistic Motivation: 269–292.* Berlin: Mouton de Gruyter.

Wiese, R. (1990). Über die Interaktion von Morphologie und Phonologie – Redup likation im Deutschen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 43, 603–624.

Zhu, J., Culp, Ch., Best, K.-H. (1995). Formen und Funktionen der Doppelungen im Chinesischen im Vergleich zum Deutschen. *Oriens Extremus 38(1/2), 183–208.*

3.17. Sequences of syntactic constituents

Problem

Parts of speech and syntactic constituents occur at different positions in the sentence. Köhler (2012: 84-92) counted the number of individual entities in positions 1 to 12 in the *Susanne corpus* and computed the entropy of the rank-frequency distributions. Use his data (p. 89-90) and compute

(1) the *relative* entropy in individual positions;

(2) find a theoretical function fitting the sequence of relative entropies;

(3) fit a theoretical function to the rank-frequencies and compare the parameters of the function in form of a sequence (1 to 12) or find a function capturing the change of a parameter in positions 1 to 12;

(4) show the changes of the rank-frequency distributions for 1 to 12 by computing Ord's criterion or the steepness of the distributions;

(5) compare the neighboring distributions applying a rank test; interpret the results.

Procedure

(1) Köhler (2012) computed the entropy for individual positions. Use his values and obtain the relative entropy simply as $H_{rel} = H/ld\ K$, where K is the inventory, i.e. the highest rank.

(2) First fit an empirical function to this sequence, then substantiate it linguistically. Set up the differential equation and interpret it.

(3) For computing the rank-frequency distribution/function use some well known models, e.g. Zipf's (zeta) distribution/function, Zipf-Alekseev, sinusoid function, etc. and interpret the results linguistically. Here you have to do with syntactic functions, not with words.

(4) The properties of distributions in individual position change. Computing some indicators you can discover some syntactic regularity, preferences of positions.

(5) The classes (Köhler 2012) are marked with letters. Mark each letter in each position with its respective rank for the positions 1 to 8. You obtain ranked samples that can be compared using some of the many rank tests. Use e.g. Mann-Whitney's U test for comparing only the neighboring positions. Having performed all tests conclude whether there is some trend in the positions; whether some of the positions has a special property, deviates significantly from its neighbors, etc..

Now, take a text consisting of maximally 200 sentences and analyze it "positionally" adhering to the method presented in Köhler (2012). Order the results and analyze them according to the above mentioned tasks. Then take a second text and compare the results of your analyses in both texts. Strive for a syntactic characterization of the given text sort. If you succeed, continue with other text sorts. At last, compare the results following from different text sorts and strive for a theoretical substantiation of your results. If necessary (or possible) include in your analysis also other properties of the text sort and make the first steps towards constructing a control cycle.

References

Köhler, R. (2012). *Quantitative syntax analysis*. Berlin-Boston: de Gruyter.

3.18. Noun phrase

Problem

This is a continuation of the problem in *Vol. 4: 19-20.*

Classify the noun phrases of a language as you are accustomed or use a ready classification. The literature is very rich. Study some selected properties of noun phrases in individual texts and develop models of their behavior. Use as many results from the non-grammatical research in quantitative linguistics as possible.

Procedure

Rewrite the text in terms of noun phrase types. Ignore everything else.

(1) As first, set up the rank-frequency distribution of types and find a model. You may use a discrete or a continuous function. Characterize the function by some indicators, e.g. mean, repeat rate, Ord's criterion, excess, Gini's coefficient, etc.

(2) Measure the length of individual NPs and set up the distribution of lengths; define length e.g. as the number of words in it. Characterize the distribution as above.

(3) Set up a vector of lengths (i.e. the sequence of lengths) and study its properties.

(4) Set up the matrix of transition frequencies from one type to another and evaluate it; you can use some kind of Markov chains; you can test individual transition cells for significance; which types do display a preference for neighborhood?

(5) Set up the distribution of distances between identical NPs and/or identical lengths and find a model for the distributions.

(6) If there are many runs, study their properties.

(7) Are there symmetric cells in the transition matrix? That is, for each pair of NP-types, test whether there are symmetric relations: are the cells, say, AxB and BxA symmetric (in frequency)? Perform the test for all pairs.

Comment each problem linguistically, use grammatical arguments and, if possible, compare the results with those obtained from other languages.

Compare texts of different text sorts, e.g. poetry with newspaper articles, and show the differences. Perform a classification.

Can your results be used for typological properties? How do behave strongly analytic and strongly synthetic languages? Do not care for the qualityative side of the NPs, consider only the quantitative results.

References

Cole, P., Morgan, J. (eds.) (1975). *Speech acts, syntax and semantics*. New York: Academic Press.

Fox, B. (1987). The noun phrase accessibility hierarchy revisited. *Language* 63(4).

Givón, T. (2001). *Syntax I, II.* Amsterdam-Philadelphia: Benjamins.

Gunkel, L., Zifonun, G. (2009). Classifying modifiers in common names. *Word Structure 2 (2), 205-218.*

Halliday, M.A.K. (2004). *Introduction to functional grammar,* 3rd ed, London: Hodder Arnold.

Köhler, R., Altmann, G. (2014). *Problems in Quantitative Linguistics Vol. 4.* Lüdenscheid: RAM-Verlag.

Rijkhoff, J. (2004). *The Noun Phrase.* Oxford: Oxford University Press.

Zifonun, G., Hoffmann, L., Strecker, B. (1997). *Grammatik der deutschen Sprache*. Berlin: de Gruyter.

Wang Hua (2012). Length and complexity of NPs in Written English. *Glottometrics 24, 79-88.*

3.19. Syntactic tags as NP components

Problem

Wang Hua (2012) described length and complexity of noun phrases in English and presented a thorough list of syntagmatic functions of NP components. Use her terminology and find (a) a possibility of scaling the components according to at least one property. (b) Set up the distribution of degrees and find a model. (c) Set up the vector of abbreviations of syntagmatic functions and evaluate the distances between equal NP components. (d) Compare the distributions obtained from two texts. (e) Compare different text sorts.

Procedure

Take a short text, e.g. a poem, and analyze it replacing the individual NP components by the abbreviations presented in Wang Hua (2012) or in other grammar you prefer.

(a) Scale the entities according to their “importance” in the NP. You can define the “importance” e.g. according to the level in the dependence tree. Other possibilities are not excluded. Replace the abbreviations by the degrees in order to obtain a vector of the text. Then count the individual degrees and set up the distribution. It may be discrete or continuous.

(b) Substantiate grammatically the presence of the given degrees, take into account the role of the speaker and hearer, of the forces and requirements of the text (cf. Köhler 2005) and construct a model, e.g. in form of a differential or difference equation. Solve it and fit the result to the distribution of degrees. If you do not want to work theoretically, characterize the distribution using some well known indicators (e.g. entropy, repeat rate, Ord's criterion, mean, h-point and many other). In any case you must obtain some values which are inter-textually comparable.

(c) Now use the vector of abbreviations and compute the relativized Euclidean distances between equal components. The distances give an image of text structuring. Then do the same for the degrees. First write the text in form of degrees and study their course. Can you observe some rhythm? Is there an autocorrelation? Then take the mean degree of each sentence separately and set up the vector of means. Do you observe some tendency? Is the distribution now "smoother"? Compute again the indicators of the distribution.

(d) Analyze several texts and compare their distributions and indicators. If possible, perform statistical tests; if not, then at least order the texts or – if you analyzed several texts – perform the usual classification using software.

(e) Perform your analysis systematically: (i) Take the works of the same author; (ii) analyze texts belonging to the same text sort and compare the different text sorts; (iii) compare the same text (translations) in different languages; (iv) compare different texts in different languages and observe the behavior of the individual entities (distributions, indicators).

At last, strive for a theory, i.e. link the results with other textological problems, construct a Köhlerian control cycle and after having analyzed several languages, propose a law.

References

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

Köhler, R. (2012). *Quantitative syntax analysis*. Berlin-Boston: de Gruyter.

Wang Hua (2012). Length and complexity of NPs in Written English. *Glottometrics 24, 79-88.*

3.20. Word class specification

Problem

If you solved at least one of the previous three problems concerning part-of-speech, continue analyzing the frequency of their specification. In the first steps consider only one language, use two different ways of specification and show which of them is more in agreement with the distribution.

Procedure

Take a text, restrict your analysis to only one word class, e.g. adjectives, verbs, or nouns. Classify the word class members according to some well known works (cf. Levin 1998; Ballmer, Brennenstuhl 1986; Jurčenko 1985; Silnickij 1966; 1973; Yesypenko 2009). Then compute the representation of individual classes in the text in form of frequencies. Show that a writer abides by some regularity which can be expressed by a distribution. Propose a distribution, derive it from theoretical consideration and substantiate it linguistically (stylistically).

Compare several texts of the same text sort and find a common distribution for all of them. In the first steps, you can apply also a simple (non-normalized) function. Later on, it can be transformed in a distribution.

Then consider another text sort and do the same. Can you apply the same distribution or not? If so, show the difference in some parameters. If no, propose a modification of the distribution based on some boundary conditions. Strive for a unified theory.

In the next step, take texts from another language, present the results and compare them with those of the first language.

Strive for a typology of writers, research in the development of a writer or of a text sort and for that of languages – if possible.

Since the resulting distribution has parameters, you can define some indicators and find their relation to other indicators, i.e. strive for finding links between word class specification and other properties of language.

References

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

Croft, W., Cruse, D.A. (2004). *Cognitive linguistics.* New York: Cambridge Univeristy Press.

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

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

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

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.

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: Kelih, E., Levickij V., Altmann, G. (eds.), *Methods of text analysis: 308-328.* Chernivtsi, ČNU.

4. Stage play

4.1. Stage play problems: 1. Sentence length

Problem

State the empirical distribution of sentence lengths in a complete stage play and find a theoretical distribution.

Procedure

Define a unit of measurement of sentence length. Begin with the number of clauses in the sentence. Count the sentence lengths for each person separately. Finally, add all the data and search for a probability distribution. Most probably you will obtain a very irregular distribution. First construct the distribution as a sum of weighted parts, e.g. using Fucks-Poisson, or some kind of Dacey distribution. If the number of parameters becomes too great, continue in two steps: (a) Find a simple distribution having maximally three parameters. If it does not work, take a general distribution containing an undefined function and replace the function with some simple functions. If you do not succeed, (b) partition the stage play in the speech of individual persons, i.e. consider your original data, and find for each person a distribution. The distributions may differ; the difference will not be drastic, maybe all persons may follow the same distribution but with different parameters. If not, strive for finding a family of distributions contained in the unified theory (Wimmer, Altmann 2005). Substantiate the distribution by the role the given person is playing. There may be persons uttering merely one-clause sentences and you obtain a deterministic distribution. Interpret the distribution by means of the roles the persons are playing.

If you had success, analyze further stage plays and generalize your results.

Analyze the sequence of sentence lengths in a stage play using standard methods some of which are presented in this volume. Study the autocorrelation, the differences between the lengths of the neighbours, the distances between identical lengths, the distances between the average length of the uninterrupted utterances of persons, etc. If there is some regularity, express it mathematically.

Study the evolution of sentence length historically. Take the oldest stage plays in your language and systematically analyze newer ones. What changed in the sentence length?

Study the same stage play translated into different languages, e.g. Shakespeare or Molière and compare the results.

References

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol 3.* Lüdenscheid: RAM (esp. pp 133 ff.).

Duraš, G. (2012). *Generalized Poisson Models for Word length Frequencies in Texts of Slavic Languages.* Graz, Diss.

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics Vol 2.* Lüdenscheid: RAM (esp. pp 126 ff.)

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: de Gruyter.

4.2. Stage play problems: 2. Speech acts

Problems

The study of speech acts in a stage play is a very fruitful problem because all the sentences of a conversation contain at least one of them. Study the following properties:

(1) For each person separately set up sets of different speech acts and state the frequencies of individual classes. You obtain two kinds of information: (a) The number of speech acts a person uttered; (b) their distribution.

(2) State whether the distributions are homogeneous or differ strongly.

(3) Consider a property that can be ascribed to each kind of speech act, e.g. dominance, urgency, etc. Prepare a scale for this property and ascribe its values to individual speech acts.

(4) Replace the respective speech acts of individual persons by these degrees and evaluate the set of each person. Set up a hierarchy of persons in relation to the given property.

(5) Study the development of the stage play in terms of the above degrees. Do they increase or decrease? If a change may be stated, perform the characterization of some stage plays in terms of this development. A classical tragedy is surely different form a modern comedy.

Procedure

Take a theoretical work in which “all” kinds of speech acts are described. Adhere to the given classification and state for each person separately the number of speech acts of different kinds. Consider the *proportion* of speech acts of each person separately and (a) perform a test for equality between persons. You may use the binomial distribution to obtain exact results, but if the stage play is long, you may use also the asymptotic normal test for the comparison of two proportions. Do not forget that in this case you must consider not only the variances but also the covariance of the proportions because the data come from the same sample. Set up the hierarchy of persons on the basis of the (significant) predominance of their speech act proportions. If you process several stage plays, show that the hierarchies may differ according to the character of the stage play, according to the author or according to the time of the first appearance of the stage play. Show the (possible) development from Greek dramas to modern ones.

If you have set up classes of speech acts, then compare the frequencies of classes of each person with those of each other using a homogeneity test. You can even rank the classes and perform a test based on ranks. In order to state the variety of roles, perform a double classification: that for speech act classes and that for persons; set up a contingency table and perform the chi-square test for this table, or devise an indicator expressing the state of this variety. If possible, derive also the sampling properties of this indicator – at least its variance – in order to be able to perform stage play comparisons.

Every property can be scaled because properties are our conceptual constructions. Characterize each person of a stage play by an indicator expressing the extent of this property. Derive the sampling properties of the indicator and classify the roles according to this indicator.

Perform this operation stepwise, i.e. for each act separately, and show the development of individual persons – if there is any. Then draw the scheme of this evolution and analyzing several stage plays begin to generalize. In the classical drama one recognized three stages but described them using words. Show that it is possible to do it with exact means. What kinds of speech acts are relevant and how is the change of the degree of the given property in the stage play? If it is linear, substantiate this fact. If it is not linear, what is the background of this development? What is its connection to the aim of the stage play or of the most important person?

Cf. especially Köhler, Altmann (2009: 118 ff.).

References

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.*

Bach, K., Harnish, R. M. (1979). *Linguistic communication and speech acts,* Cambridge, Mass.: MIT Press.

Burkhardt, A.S. (ed.) (1990), *Speech Acts, Meanings and Intentions. Critical Approaches to the Philosophy of John R. Searle*. Berlin-New York: de Gruyter.

Čech, R., Altmann, G. (2011). *Problems in Quatitative Linguistics Vol 3*. Lüdenscheid: RAM-Verlag.

Cohen, A.D. (1996). Speech acts. In: McKay, S.L., Hornberger, N.H. (Eds.), *Sociolinguistics and language teaching: 383-420*. Cambridge: Cambridge University Press.

Grice, H. P. (1989). *Studies in the way of words*, Cambridge, Mass.: Harvard University Press.

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

Olshtain, E. & Cohen, A. D. (1989). Speech act behavior across languages. In H. W. Dechert, H.W. et al. (Eds.**),** *Transfer in production***:** *53-67*. Norwood, NJ: Ablex.

Sander, Th. (2002). *Redesequenzen. Untersuchungen zur Grammatik von Diskursen und Texten*. Paderborn: mentis.

Searle, J.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.

Staffeldt, S. (2008): *Einführung in die Sprechakttheorie. Ein Leitfaden für den akademischen Unterricht.* Tübingen: Stauffenburg.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol 1*. Lüdenscheid: RAM-Verlag.

Tsohatzidis, S. L. (ed.) (1994). *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives.* London: Routledge.

Ulkan, M. (1993). *Zur Klassifikation von Sprechakten. Eine grundlagentheoretische Fallstudie*. Tübingen: Niemeyer.

4.3. Stage plays: 3. Sequences of illocutive speech acts

Problem

Analyze a stage play, replace the words by some abbreviations of illocutive speech acts and set up the sequence of illocutive acts. Then perform the analysis of the given sequence in different ways.

Procedure

Take a stage play and rewrite it in the form of illocutive speech acts. The kinds of speech acts must be taken from a list that can be found on the Internet. You obtain a sequence – you can use letters of numbers or abbreviations for speech acts – which can be further processed.

(1) Compute the *distances* between identical speech acts; the distance is here the number of steps necessary to get from its occurrence to its next occurrence, e.g. the distance between the two A in the sequence A,B,C,B,D,A is 5. Compute the distances of all speech acts and set up their distribution, i.e. there are f_x distances of length *x*. If a speech act does not occur any more (after its last occurrence it would be infinite), omit this infinite distance.

(2) Find a model of the distribution of distances in the form of a discrete probability distribution.

(3) Compute the mean and the standard deviation of the distribution.

(4) Set up the above sequence for all persons separately. Their speech acts depend on the role they play, hence the distributions must be different. If you must derive different models, substantiate them textologically.

(5) For each person, compute separately the mean and the variance of its distance and compare them, i.e. perform a statistical test for the difference between means.

(6) Analyze the complete stage play but set up separate sequences for each act. Characterize the acts using some indicators and scrutinize the development of the indicator from the beginning to the end. Does classical drama differ from, say, a comedy?

(7) If you have scaled the speech acts according to some criterion, replace the individual speech acts by their degrees and scrutinize this numerical sequence. Can you observe some regular movement? If so, use some kind of analysis (time series, wavelets, Fourier analysis, etc.) to capture it formally. Do modern stage plays differ from classical ones? (Cf. Problems Vol. 4: 100)

(8) Construct Köhlerian motifs on the basis of scaled speech acts. A motif is a set consisting of the sequence of non-decreasing numbers. Analyze the motifs as follows:

(9) State the distribution of their lengths measured in terms of their cardinal numbers (= the number of its elements). Compute the basic indicators (average and variance) for the complete stage play and for the individual roles. Is there some difference between the individual roles? Characterize and test it.

(10) Compute the ranges of individual motifs, i.e. the difference between the last and the first element of the motif. The ranges are always positive. If the first and the last elements are equal, the range is zero. For the ranges of each role compute the average and the standard deviation; perform a test for equality of averages. Find the distribution of ranges and substantiate it textologically (as drama).

(11) Study the runs of ranges. A run is a sequence of equal signs/numbers. Compute the number of runs in the text and compare it with the expectation. If there is a significant difference, draw textological conclusions.

References

See the references in all problems of Chapter 4.

4.4. Stage play: 4. Transition matrix of speech acts

Problem

Transcribe a stage play in the form of a sequence of speech acts. Set up the transition matrix and study its properties.

Procedure

Take a stage play and replace the given speech acts by the class to which each belongs. You obtain a sequence of letters or some abbreviations. Prepare a two-dimensional contingency table whose first row and first column obtain the names of speech acts. Now study the transitions between the speech acts and record their number. This contingency table can be used for different tests. If you perform some of them, substantiate linguistically the hypothesis, do not perform mechanical testing. Study especially the following problems:

(1) Test the table for independence, i.e. test the independence of the following speech act on the preceding one.

(2) Perform the test for significance for each individual cell separately. There is a possibility that some speech acts evoke other particular ones, i.e. there may be special bigrams of speech acts. Collect the significant cells and interpret this state of affairs in the sense of the stage play.

(3) Test the diagonal of the table as a whole, i.e. ask whether an act evokes the same kind of act. Interpret the result psychologically or dramaturgically.

(4) Analyze several stage plays of the same kind and show whether there is a development in the individual dependencies. Can you differentiate classical dramas from modern comedies on this basis?

(5) Study Markov chains and state the order of the given data. Perform all analyses for various stage plays and show the differences between them. Is there some development from classical stage plays to modern ones? If you analyze separately each act, show the development of the stage play.

(6) Study the symmetry of transitions for each pair of cells separately. Substantiate textologically why some of them are symmetric and other ones not.

Using the results, set up hypotheses about the dynamics of the given stage play and generalize your results to the development of the stage play act-wise and the development of stage plays of certain kind historically.

Transfer the results won in phonemics – as given in the references – to the speech act domain.

References

Altmann, G. (1987). Tendenzielle Vokalharmonie. In: Fickermann, I. (ed.), *Glottometrika 8, 104-112.* Bochum: Brockmeyer.

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag (esp. P.24 ff.).

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

4.5. Stage play: 5. Polysemy in the speech of persons

Problem

The speech of persons in a stage play differs in various aspects. State whether individual persons differ in the polysemy of words uttered by them.

Procedure

Consider merely one act of a stage play. For each word in the act state its polysemy using a monolingual dictionary. Common phrases like “Good day!” have only 1 meaning but you can use your own definitions.

Replace the words (phrases) by their polysemy degree and distinguishing the persons compute for each of them (a) the average polysemy, (b) the variance, (c) the frequency distributions of polysemies.

Order the persons according to “their” average polysemy. Is this ranking linked with the given roles?

Compare the means of individual persons using, say, a t-test. Set up classes of persons – if there are many, otherwise it would be superfluous.

Compare the distributions of polysemy of individual persons.

Compute the entropy of individual empirical distributions and interpret it as dramma. What does it mean if a person has too much semantic entropy?

Find a theoretical model for the distribution of polysemy which holds for all persons and substantiate it as drama.

After replacing the words by their polysemies you obtained a vector of numbers. It represents a time series. Set up hypotheses about this time series and compute some of its properties. Are there some regularities or is the series chaotic?

References

Jastrzembski J.E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology 13, 278–305.*

Levickij, V.V. (2005). Polysemie. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 458-464.* Berlin: de Gruyter.

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

Pethö, G. (2001). *What is Polysemy? A Survey of Current Research and Results.* In: Németh T. E., Bibok, K. (eds.), *Pragmatics and the Flexibility of Word Meaning: 175-224.* Amsterdam: Elsevier.

Ravin, Y., Leacock, C. (eds.) (2000). *Polysemy. Theretical and Computational Approaches*. Oxford: Oxford University Press.

4.6. Stage play: 6. Distant reaction

Problem

Compute the distances between reactions to given speech acts in terms of intervening speech acts and find a distribution model for the distances.

Procedure

First transcribe an act of a stage play in terms of speech acts. Mark those which are reactions to a former speech act. The distance between them is given by the number of intervening speech acts. Hence you obtain a distribution whose variable (distance) takes on values d = 0,1,2,… Characterize the act of the stage play using some properties of the resulting distribution. Set up unequivocal criteria for constructing the sequence: some speech acts may stay within a more extensive speech act. Take into account only the nearest reaction, not all.

A person can react also to his own speech acts, hence you can perform a further analysis in “own reactions” versus “foreign reactions”.

The distances between reactions display the weight of the given speech acts. The greater the distances, the more weighty is the given speech act. Now consider merely identical speech acts, e.g. “questions”. Find the distribution of distances of reactions to the given kind of speech act, i.e. different distributions. Taking into account the properties of the given distributions, order the speech acts according to their weight. Find an indicator of the weight and characterize the act of the speech play.

If you perform this operation for a whole stage play (act-wise), you obtain an image of reactions and a description of the development of the stage play, and if you compare several stage plays, you obtain the mechanism of stage play structure, and you can compare the evolution of dramas, etc.

References

Dąmbrowska, E., Rowland, C., Theakston, A. (2013). The acquisition of questions with long-distance dependencies. In: Janda, L.A. (ed.), *Cognitive*

Linguistics: The Quantitative Turn: 197-223. Berlin-Boston: de Gruyter Mouton.

Vielliers, J. de, Roeper, Th., Vainikka, A. (1990). The acquisition of long-distance rules. In: Frazier, L., Vielliers, J. de (eds.), *Language Processing and Language Acquisition: 257-297.* Dordrecht: Kluwer.

4.7. Stage play: 7. Aggregation of speech acts

Problem

One may conjecture that in a stage play, near-by sentences are more similar concerning their speech act content than more distant sentences. Define a similarity measure and test the hypothesis. The problem is a special case of the Skinner effect on long-term memory.

Procedure

Take a stage play and rewrite it in terms of speech acts (use abbreviations). Write each sentence in a separate line. Consider the line as a vector or as a set. Now define a similarity indicator between two lines using a very extensive literature. Try with different similarity indicators. Then compute the similarities between neighboring lines (distance 1) and compute the mean similarity for distance 1. Continue with computing the similarities between lines 1-3, 2-4, 3-5,…, i.e. lines in distance 2, and compute the mean similarity. Continue computing the mean similarities for distances up to 20. Set up a table of mean similarities for distances d = 1,2,…,20 and state whether they decrease. If so, propose a function capturing this decrease and interpret it. You will discover an aspect of the dynamics of conversation.

Now consider each act of the stage play separately and perform the same operations. Compute the curves of similarity decrease and show the development of parameters. Here, you may discover some aspects of the dynamics of a drama. You can compare stage-plays, authors, the development of drama writing, etc.

References

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

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.

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

Burnaby, T. (1970). On a method for character weighting a similarity coefficient, employing the concept of information. *Mathematical Geology 2(1):25-38,*

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol 2*. Lüdenscheid: RAM-Verlag.

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.*

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.

Wang, X., Baets, B.de, Kerre, E. (1995). A comparative study of similarity measures. *Fuzzy Sets and Systems 73(2), 259-268.*

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.*

http://reference.wolfram.com/mathematica/guide/DistanceAndSimilarityMeasures.html

4.8. Stage play: 8. Act compactness

Problem

Study the sequence resulting from the evaluation of text compactness (TC) of individual acts in a stage play. Can you recognize a similarity with the classical course of a stage play? If not, compute TC for different stage play genres (kinds of drama, comedy etc.), set up intuitive hypotheses, test them and derive them theoretically.

Procedure

You may use the definition of text compactness as proposed by Mačutek and Wimmer 2014 (cf. Problem: 2.13. *Text compactness*, in this volume). Begin with the simple definition but if you can evaluate it in some other way proposed in the problem mentioned, perform the measurement with each method and act separately. You obtain a sequence of values which can further be evaluated. Hence first take a stage play with many acts.

Ask the following questions: is the development of TC constant or linear in some direction (increasing, decreasing), or is it necessary to capture it by means of a non-linear function. Do individual stage plays differ? In any case, begin to construct a differential equation containing all conditions present in a stage play evolution adhering to the unified theory (Wimmer, Altmann 2005).

Compare stage plays of the same author and study his development, then compare them with those of other authors and find a formula for the given text sort (in the given language). If possible, do not use translations but original stage plays.

References

Mačutek, J., Wimmer, G. (2014). A measure of lexical text compactness. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical approaches to text and language analysis: 132-139.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Altmann, G. (2008). Autosemantic compactness of texts. In: Altmann, G., Zadorozhna, I., Matskulyak, Yu. (eds.), *Problems of General, Germanic and Slavic Linguistics: 472-480*. Chernivtsi: Books XXI.

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: 781-807.* Berlin: de Gruyter.

4.9. Stage play: 9. Verb activity

Problem

Measure the verb activity in an individual stage play and analyze the course of activity (a) in the stage play as a whole, (b) considering the means of individual acts, and (c) compare a drama with a comedy.

Procedure

First consider the problem 6.1.*Measurement of verb activity* in this volume. Having solved the problem of measurement, (a) compute the activity of individual verbs and set up a time series. You can also use moving averages. Propose a model for the given time series. How would you characterize the given course?

(b) Compute the mean activities of individual acts and consider them as a time series; can you observe some trend, e.g. increase of activity up to the climax of a drama, then a sudden decrease? Find a model for the course of the sequence, if there is any. This can be done, of course, only if the stage play consists of several acts. How do comedies behave? Is there a special course for different types of stage plays?

(c) Compare different types of stage plays and set up a hypothesis. Then take a prose text of a special text sort, compute the activity of individual verbs and compare the result with those obtained from stage plays. Can you set up a typology of texts/text sorts based on verb activity?

Can you observe some general features of verb activity in stage plays or in other texts? If so, try to construct models.

If you computed other properties of the given texts and expressed them by some indicators, search for the links between the verb activity and the other properties. Set up step by step a control cycle whose central point is verb activity.

References

Cf. all problems concerning stage plays in this volume and the problem 6.1. *Measurement of verb activity*.

5. Phonemics and script

5.1. Phonological complexity

Problem

Describe the many facets of phonological complexity of a language, set up hypotheses and test them.

Procedure

Begin with sounds. There is a certain degree of "pronouncing difficulty" of individual sounds. It may depend on the muscular effort (which is measurable) or on the number of organs taking part in the pronunciation. Scale the difficulty and evaluate some languages.

Continue with the measurement of the transition difficulty, either taking into account the muscular effort or the difference of distinctive features between the neighboring sounds. Then set up the (distributional) table of sound/phoneme pairs, write in the table the phonetic/phonemic/muscular differences and evaluate the table statistically. Set up the frequency distribution of differences, find an appropriate model and test it on other languages. Test the tendencies in individual cells or parts of the table. What are the boundaries of effort? What are the differences between individual Slavic or Roman languages? How does a language develop (e.g. from Latin to French)?

Now take the individual syllables and state the mean difference between its sounds. Show that the smaller the mean difference, the more syllables of the given type exist. Find the frequency distribution or at least a continuous function expressing this relationship.

Take the canonical forms of syllables (V, CV, VC, CVC, CCVC, CVCC, …) and show the mean complexity of individual phonemes, the mean complexity of individual types and derive the respective dependence.

Substantiate each of the dependencies linguistically and if you set up models, interpret the parameters linguistically.

Take a short text in different languages and set up a time series (a) of phoneme complexities, (b) of syllable complexities measured in the same way (cf. also the problem 5.5. *Syllable complexity*). Analyze the series. Take a Latin text and its translation into a modern Roman language. Compare the time series and state the extent of change.

References

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

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

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

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

Bose, A., Colangelo, A., Buchanan, L. (2011). Effect of phonetic complexity on word reading and repetition in deep dyslexia. *Journal of Neurolinguistics 24 (4), 435-444.*

Coupé, Ch., Marsico, E., Pellegrino, F. (2009). Structural complexity of phonological systems. In: Pellegrino, F., Marsico, E., Chitoran, I., Coupé, Ch. (eds.) (2009). *Approaches to phonological complexity: 141-170.* Berlin: Mouton de Gruyter.

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

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

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

Kane, M., Mauclair, J., Carson-Berndsen, J. (2011). Automatic identification of phonetic similarity based on underspecification. *Human Language Technology 6562, 47-58.*

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

Lee, Ch.-Ch. (2007). *Relationship between jaw opening and phonetic complexity: A cross-language study*. Diss. Canterbury. http://ir.canterbury.ac.nz/bitstream/10092/1877/1/Thesis_fulltext.pdf

Lehfeldt, W. (1978). Zur Messung der phonetischen Lautdifferenz. Eine begriffskritische Untersuchung. In: G. Altmann (ed.), *Glottometrika 1, 26-45.* Bochum: Brockmeyer.

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

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

Maddieson, I. (2009). Calculating phonological complexity. In: Pellegrino, F., Marsico, E., Chitoran, I., Coupé, Ch. (eds.) (2009), *Approaches to phonological complexity: 85-110.* Berlin: Mouton de Gruyter.

Mauclair, J., Aioanei, D., Carson-Berndsen, J. (2009). Exploiting phonetic and phonological similarities as a first step for robust speech recognition. In: *17th European Signal Processing Conference (EUSIPCO 2009)* http://citeseerx.ist. psu.edu/viewdoc/download?doi=10.1.1.184.6327&rep=rep1&type=pdf

Pellegrino, F., Marsico, E., Chitoran, I., Coupé, Ch. (eds.) (2009). *Approaches to phonological complexity*. Berlin: Mouton de Gruyter.

Peterson, G.H., 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.

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

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

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

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

Cf. also *Problems in QL, Vol 1.2.3*.

5.2. Script motifs

Problem

Set up the distribution of motifs in a script and show some of its properties.

Procedure

Take an alphabetic script and find all kinds of strokes that are used in several letters. A motif is a stroke or a combination of strokes repeated in at least two letters. You may distinguish straight lines according to their length (e.g. short/ long), position (left, mid, right and bottom, mid, top) and angle (acute, obtuse); bows “opened” in different directions; and combinations of strokes; filled dots of different form; but you can omit some characteristics. For example, in the Arial script, the line “\” occurs in A, K, M, N, V, X, Y if one does not distinguish length and positions. The combined motif “V” occurs in M, V, W, Y. State all motifs in a script and for each motif state its exploitation in the script. The exploitation yields a probability distribution which must be derived. At the same time, it has different properties which can be used for the characterization of the script, for example economy associated with the number of simple (= not combined) motifs; repeat rate which furnishes us an indicator of the exploitation; one can also use the excess of the distribution for the same purpose.

Each letter or sign can be presented as a set of motifs. Either one begins to analyze taking into account the simplest motives and then the more complex ones, or one first finds the most complex motifs and after they are exhausted, one uses the simpler ones. E.g. W consist of two “V” motifs if one begins from the most complex ones, and of two slant line motifs if one begins from the simplest ones. The motif number in a letter is an indicator of economy, not of complexity (which must be measured in a different way).

Analyze also a syllabic script and the simple signs of Chinese. Devise different indicators of complexity, economy and motif content. Set up the distribution of complexity defined in different ways.

References

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

Melka, T.S., Altmann, G. (2014). Script complexity: *ogham* and *rongorongo* cases (submitted).

5.3. Phonic similarity of words in proverbs

Problem

Study the extent of alliteration and assonance in proverbs. Alliteration is the repetition of equal sounds at the beginning of words; assonance is the repetition of the same (not necessarily uninterrupted) sound sequence (mostly two vowels) in the word which can be considered a parallelism if the words stay at some prominent position. Set up a test for deciding whether the given alliteration or assonance is random or significant taking into account a collection of proverbs.

Procedure

First, use some known relative sound frequencies in the given language. For each sound *i* you have now its probability estimation *p*(*i*). Then take a collection of proverbs and scrutinize each proverb separately. Let the given proverb contain *n* words. Consider merely the beginnings of words. If the same sound occurs initially in *r* words, then compute

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

yielding the probability that the sound *i* occurs at the beginning of *r* or more words. You can transform this probability into some indicator (cf. Wimmer et al.

2003: 67 ff.) or simply interpret it. Before you begin to count, you must solve the following problems: (1) Are zero-syllabic prepositions e.g. in Slavic languages independent words? The same holds for clitics. (2) Solve the problem of liaison in French, sandhi in Hindi, assimilation, reduced forms of a word, e.g. in Hungarian "s" instead of "és", etc. (3) Should one consider merely autosemantics or all words? (4) One must distinguish the phonetic and the written form,.You must decide for one of them or study both separately.

If there are several groups of alliterated words, e.g. there are k_1 words displaying alliteration, k_2 words displaying a different alliteration and $n - k_1 - k_2$ words without alliteration, then one must compute the sum of multinomial probabilities

$$P(X_1 \geq j_1, X_2 \geq j_2, X_3 = n - j_1 - j_2) =$$
$$= \sum_{i \geq k_1, j \geq k_2}^{n} \frac{n!}{j_1!\, j_2(n - j_1 - j_2)!} p_1^{j_1} p_2^{j_2} (1 - p_1 - p_2)^{n - j_1 - j_2}$$

and analogically for more than two alliterated groups. The computation is somewhat lengthy, especially if there are many alliteration groups.

Evaluate proverb collections in different languages and compare the results.

Consider assonance only in cases when two words contain equal vowels in the same order. If you want to use the same method, you must have the probabilities of vowel sequences in the given language. If there is no such survey, use the sequences in the given proverb collection. The result can be obtained mechanically if you have an electronic collection of proverbs.

Again, define an indicator of assonance using the results. Do not finish your work with computation of percentages; find some linguistic, literary or cultural substantiations. Search for causes, forces, iconicity, psychological effectiveness, etc.

References

Gries, S.Th. (2013). Phonological similarity of multi-word units. In: Janda, L.A. (ed.), *Cognitive Linguistics: The Quantitative Turn: 177-196.* Berlin-Boston: de Gruyter Mouton.

Knauer, K. (1969). Die Analyse von Feinstrukturen im sprachlichen Kunstwerk. In: Kreutzer, V., Gunzenhäuser, R. (1969), *Mathematik und Dichtung: 193-210*. München: Nymphenburger Verlag

Krappe, A.H. (1921). *Alliteration in the Chanson de Roland and in the Carmen de Prodicione Guenonis.* Iowa City.

Pszczolowska, L. (1968). Some problems of the structure of assonance in folk and art poetry. *Teorie verše 2, 83-88.* Brno: Universita J.K.E. Purkyně.

Shewan, A. (1925). Alliteration and assonance in Homer. *Classical Philology 20, 193-210.*

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 estimate of certain types of sound patterning in poetry. *The American Journal of Psychology 54, 64-79.*

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

5.4. A variant of the Skinner hypothesis

Hypothesis

According to Skinner, the appearance of an entity increases the probability of its appearing in a near-by neighborhood. Test the hypothesis using texts and apply it to the alliteration in multi-word units (cf. Gries 2013).

Procedure

St. Gries uses verbs followed by nouns as direct objects. If the hypothesis holds true, then it may hold for any language. Take texts of different text sorts in any language and study the alliteration in all pairs "Verb - Noun as direct object". Count the cases of alliteration and no alliteration. Using the relative frequencies of phonemes (letters) in the given language (or alternatively, the frequencies at the beginning of words), compute the probability that the first phonemes (letters) of the given verbs and nouns are equal. Compare the resulting probability with the relative frequency of alliterations in your texts and state its significance. You may use the binomial distribution or, asymptotically, some version of the normal test. State whether there is a special text sort in which this phenomenon is significantly frequent (e.g. poetry).

Some other forms of the Skinner hypothesis have already been tested. Comment on your result and generalize it applying it to other forms of alliteration, e.g. beginning of phrases, clauses, sentences, verses, chapters; study especially proverbs. Take at least one other language than English.

Extend the Skinner hypothesis to other forms of repetition of sounds. Omit rhymes.

References

Gries, St.Th. (2013). Phonological similarity in multi-word units. In: Janda, L.A. (ed.), *Cognitive Linguistics: The Quantitative Turn. The Essential Reader: 177-196.* Berlin-Boston: de Gruyter.

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

Skinner, B.F. (1957). *Verbal Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

5.5. Syllable complexity

Problem

Perform a complexity measurement for all syllables in a language and find the relation of complexity to frequency.

Procedure

First define exactly what syllabic complexity is and propose a way of unambiguously measuring it. This has been done in various ways in the literature and you may use one or several variants. Nevertheless, take into account the various definitions of complexity used also in other sciences.

Set up a table containing all the syllables of a language and their complexity. Now, consider an individual text, state all the syllable complexities, propose a (continuous) distribution function and compute some of its properties. Consider or devise some indicators which may characterize this aspect of the text. Then compare your results with other texts in order to state whether complexity is a constant feature of texts. If possible use also a corpus in order to state some kind of convergence.

You may consider all individual syllables or you can consider average complexities of equal types of syllable (e.g. the canonical forms CV, CVC,…).

Having found an appropriate indicator, derive at least its variance in order to be able to compare texts. In the next step, consider the distribution of complexities, derive the formula based on linguistic arguments and fit it to your data.

Show the variability of texts and state whether scientific texts differ significantly from poetic ones?

It is to be remarked that there are many definitions of syllable for a given language and different syllabification algorithms. Use one of them mechanically.

References

Adsett, C.R., Marchand, Y. (2010). Syllabic complexity: a computational evaluation of nine European languages. *Journal of Quantitative Linguistics 17(4),269-290.*

Changizi, M.A. (2001). Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. *Journal of Theoretical Biology 211, 277-295.*

Fenk, A., Fenk-Oczlon, G., Fenk, L. (2005). Syllable complexity as a function of word complexity. In: V. Solovyev, V. Polyakov (eds.) (2005) *Text Processing and Cognitive Technologies, No 11: 337-346. Moscow: MISA,*

Fenk-Oczlon, G., Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In: Miestano, M., Sinnemäki, K., Karlsson, F. (eds.), *Language Complexity: Typology, Contact, Change: 43-66.* Philadelphia: Benjamins.

Maddieson, I. (2007). Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. In: M-J. Solé, P. Beddor, M. Ohala (eds.), *Experimental Approaches to Phonology: 93-103.* Oxford University Press, Oxford and New York.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol. 1.* Lüdenscheid: RAM-Verlag.

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

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

5.6. Euphony

Problem

Propose a new kind of measurement of euphony in a verse. Then study the course of euphony in the poem. Define a test for the difference in euphony of two verses. Define a test for the difference of euphony of two poems. Order the poems of an author according to increasing euphony and state whether it correlates with the age of the poems (or the poet). Choose a special poetry, e.g. Latin hexameters and compare their euphony with that in German hexameters.

Procedure

Take inspiration using the first trials in *Problems Vol. 1, 44f.* Do not use a limit of probability; simply evaluate the probabilities and consider them as degrees of euphony. For characterizing a verse, take the mean of the probabilities; to characterize the poem take the mean of all verses.

Set up the asymptotic normal test for the comparison of euphonies of two texts. Perform a classification of texts of an author.

Can you observe some tendencies in creating euphony? Which sounds are used, and to what extent, to create euphony?

Order the poems of an author according to increasing euphony and state whether there is some tendency.

In all analyses adhere to the phonetic image, not to the written one, otherwise you obtain quite false images e.g. for French or English.

Develop the problem in a synergetic sense: is euphony linked with other properties of the verse or poem? To this end, you must quantify another property of verse. First, set up a list of possible properties using the literature and consider one after another in their relation to euphony. Strive towards a theory. Start from the principle that there are no isolated properties in language or text. Hence make the first step in constructing a control cycle similar to that developed by Köhler (1986, 2005).

References

Altmann, G. (1966). Binomial index of euphony for Indonesian poetry. *Asian and African Studies 2, 62-67.*

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: de Gruyter.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol 1.* Lüdenscheid: RAM-Verlag.

5.7. Distribution of syllable types

Problem

Obradović et al. (2010) presented the two-dimensional distribution of canonical syllable types in Serbian and fitted the two-dimensional negative binomial distribution to the data. Another distribution can be found in Zörnig, Altmann (1993) for Indonesian data. Find a unified model.

Procedure

Canonical syllable types are V, CV, CVC,… One can take them from an ordinary dictionary. Collect at least samples from different languages, and set up the table in this form

	V	VC	VCC	VCCC
V				
CV				
CCV				

where, e.g. CV on the left margin and VC on the top mean syllables of the type CVC. The syllables must be identified and counted in phonemic form (not letters!). Now find a general model of a two-dimensional distribution which would capture the situation in all languages. Fit the distribution to all available data.

Derive the two-dimensional distribution by means of a difference equation relying on the unified theory (cf. Wimmer, Altmann 2005) and interpret the parameters. If possible show the differences between languages, set up at least an order of languages. To this end characterize the data by proposing some indicators; derive their variances and compare the languages you have at your disposal. The method has been shown by Kelih, Mačutek (2013) who obtained the function

$$f(x,y) = c*exp(ax + by + rxy)*(x+1)^k(y+1)^m$$

Do these indicators display some links to other properties, for example to the (size of the) phoneme inventory, phoneme distribution, word length, etc.? Strive for setting up a control cycle.

References

Kelih, E. (2012). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation.* München/Berlin/Washington: Sagner.

Kelih , E., Mačutek, J. (2013). Number of canonical syllable types: A continuous bivariate model. *Journal of Quantitative Linguistics 20(3), 241-251.*

Kempgen, S. (2003). Phonologische Silbentrennung im Russischen. In: S. Kempgen, U. Schweier & T. Berger (eds.), *Ruststika – Slavistika – Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag: 195-211.* München: Sagner.

Lehfeldt, W. (1971). Ein Algorithmus zur automatischen Silbentrennung. *Phonetica 24, 212-237.*

Obradović, I., Obuljen, A., Krstev, C. & Radulović, V. (2010). Distribution of canonical syllable types in Serbian. In: P. Grzybek, E. Kelih & J. Mačutek (eds.), *Text and Language. Structures – Functions – Interrelations – Quantitative Perspectives: 145-157.* Wien: Praesens.

Unuk, D. (2003). *Zlog v slovenskom jeziku*. Ljubljana: Rokus.

Wimmer, G., Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Language: 320-337.* Dordrecht: Springer.

Zörnig, P., Altmann, G. (1993). A model for the distribution of syllable types. In: Köhler, R., Rieger, B. (eds.), *Glottometrika 14: 190-196.* Trier: VWT.

5.8. Phonetic symbolism

Problem

The sounds of language still preserve their iconic character and may be associated with some conceptual properties. The number of studies concerning phonosemantics is enormous. The problem is to state the degree of a property that can be ascribed to a sound. Test the following hypotheses:

(1) Using any type of scaling, ascribe the sounds degrees of the given semantic property. Use as many properties as necessary.

(2) Test whether a low degree of a property is linked with the frequency of sounds.

(3) Test whether the iconicity you found is general, e.g. comparing your results with those of Levickij (2013).

(4) Analyze at least two poetic texts in your language and compare the results.

(5) Take all words containing a sound with the given degree and analyze its properties (e.g. length, morphological composition, part of speech, etc.). Show the relation of these properties to the iconicity.

Procedure

(1) For scaling the iconic sound properties, use any available procedure e.g. the semantic differential by Osgood, Suci, Tannenbaum (1957) asking many test persons, or measuring muscular effort, occurrence in words of special classes, etc.

(2) Perform a sound frequency count of the language using ready-made counts which are available for many languages on the Internet. Then state whether the frequency has something incommon with the individual property degrees. If so, derive a function expressing this relationship, using linguistic argumentation. The function must hold for each property; however, it may have different parameters. If you can manage the same for 2 or 3 languages and obtain positive results, you are on the way to finding a law.

(3) Then compare your results with those of Levickij (2013). You may apply a statistical test or simply compare the vectors. If you compute correlations, do not forget to take into account the degrees of freedom. If your results differ from those of Levickij, find the sounds causing this difference and search for an explication in typology, ethnology, life conditions of the speakers, the choice of your informants, etc. i.e. search for boundary conditions leading to the differences.

(4) Take at least two poetic texts written in your language. Compute the frequency of individual degrees of meaning and perform a test for homogeneity in a 2xk contingency table using any of the usual methods. If the texts differ, perform a description of possible causes or motifs of the difference. Insert these boundary

conditions in your formulas expressing the relation between meaning degree and frequency.

If you also analyzed non-poetic texts, strive for an iconic text-sort classification. This does not replace a theory but you can begin to analyze other properties of the concerned texts and search for further links.

If you have two text sorts, you can perform comparisons. For each text sort separately multiply the weight of an individual sound with its *relative* frequency. You obtain a set of numbers that can be compared in various ways. (a) Order the products in decreasing order to obtain a ranked sequence and find an empirical function capturing it. Then compare the two functions (poetry and prose) mechanically using software. Here the quality of the sound does not play any role. (b) Order the sounds in the same way in the two text sorts, ascribe to them their respective weights, i.e. set up vectors of weights, and compute the distance of the two vectors, e.g. expressed in radians.

Take further text sorts, perform the same procedures, compare them and begin to theorize. Derive the hypothesis concerning the relation of weight and frequency; then make an hypothesis concerning the influence of the text sort on the weights. Translate the hypotheses into the language of mathematics, i.e. set up elementary differential equations leading to the given relations. Interpret the parameters of the equation in terms of human senses, mobility, emotionality, potency, evaluation, speed, cruelty, size, etc. (cf. Gnatchuk 2015).

(5) Interpret your results using the great number of available publications; look at both agreements and disagreements; look for other properties and construct, step by step, an elementary theory. If possible, include your results in the Köhlerian control cycle.

References

Bergen, B. (2004). The psychological reality of phonaesthemes. *Language 80(2), 290-311.*

Elsen, I. (2014). Lautsymbolik – ein vernachlässigter Forschungsgegenstand der Sprachwissenschaft. *Glottotheory 5(2), 185-218.*

Gnatchuk, A. (2015). A full bibliography of works on sound symbolism. *Glottotheory (submitted).*

Gnatchuk, A. (2015). Phonosemantic features of English and German consonants. *Glottometrics 30, 1-18.*

Hinton, L., Nichols, J., Ohala, J.J. (eds), (1994). *Sound Symbolism.* Cambridge: Cambridge University Press.

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

Levickij, V.V. (2013). Phonetic symbolism in natural languages. *Glottotheory 4(1), 72-91.*

Maduka-Durunze, O.N. (2001). Phonosemantic hierarchies. In: Voeltz, F.K.E., Killian-Hatz, Ch. (eds.), *Ideophones: 193-203.* Amsterdam/Philadelphia: Benjamins.

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

Yorkston, E., Menon, G. (2004). A sound idea: Phonetic effects of brand names on consumer judgement. *Journal of Consumer Research 31, 43-51.*

6. Semantics

6.1. Abstractness of nouns

Hypotheses

Test two hypotheses conjectured by S. Schierholz (1991):

(1) The more frequently a noun occurs, the higher is its degree of abstractness.

(2) The higher is the polysemy degree of a noun, the higher is its degree of abstractness.

Procedure

Since frequency can be mechanically computed from texts and polysemy directly from a monolingual dictionary, the only problem is the quantification of abstractness. Abstractness should not be mixed up with generality though there are cases in which both properties coincide. The most difficult task is the construction of a scale for abstractness. It need not hold for all languages in the same manner, one must, perhaps, differentiate. Further, one should not believe that the achieved scaling procedure has something to do with “truth”; it will be a trial to order nouns in a special way. The scale may be constructed or corroborated just by testing the given hypotheses. List of abstract nouns can be found easily on the Internet.

Now take a text, state the frequencies of nouns (in lemmatized form) and state their polysemies from a dictionary. Then ascribe to each noun its abstractness degree. In the first step, state simply whether there is a correlation between abstractness and the other properties. If so, derive an hypothesis expressing this link and interpret the parameters linguistically. Test the hypothesis using various texts. Can you see a difference in parameters for texts of different text-sorts? If the derivation of a mathematical hypothesis cannot be performed as yet, proceed inductively, i.e. find a function using some program.

Continue in the following directions: (1) Incorporate the result in the Köhlerian (2005) control cycle, i.e. find its place in a well developed theory. (2) Study the abstractness of other parts of speech, e.g. verbs and adjectives. (3) Characterize individual texts and different text sorts by their average abstractness. (4) After having measured the degree of abstractness of individual nouns, find the distribution of abstractness in individual texts. (5) Set up a model of the resulting distribution and test it on your data.

Present all numbers in tabular form and show all results.

References

Beauregard, M., Chertkow, H., Bub, D., Murtha, S., Dixon, R., Evans, A. (1997). The neural substrate for concrete, abstract, and emotional word lexica: A positron emission tomography study. *Journal of Cognitive Neuroscience, 9, 441–461.*

Bresson, D., Kubczak, J. (eds.) (1998). *Abstrakte Nomina*. Tübingen: Narr.

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6, 139-151*. Bochum: Brockmeyer.

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

Künne, W. (2007). *Abstrakte Gegenstände. Semantik und Ontologie*. Frankfurt: Klostermann.

Mikk, J., Uibo, H., Elts, J. (2001). Word length as an indicator of semantic complexity. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 177-186.* Trier: WVT.

Plaut, D.C., Shallice, T. (1991). Effects of Word Abstractness in a Connectionist Model of Deep Dyslexia. *Proceedings of the 13th Annual Meeting of the Cognitive Science Society, Chicago, IL, August 1991, 73–78.*

Schierholz, S. (1991). *Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive.* Tübingen: M. Niemeyer

Skipper, L.M., Olson, I.R.(2014). Semantic memory: Distinct neural representations for abstractness and valence. *Brain & Language 130 (2014) 1–10.*

Warriner, A.B., Kuperman, V., Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45, 1191–1207.*

6.2. Abstractness in the text

Problem

Measure the abstractness of sentences, set up the sequence of abstractness degrees of sentences, and find a function describing its distribution.

Procedure

First substantiate your decisions to consider a noun, verb, adjective or adverb as abstract or concrete. (You may restrict your analysis to nouns.) This may be a dichotomic decision; you need not perform scaling. Then take a text and replace the sentences by the number of abstract words you have found in each. You obtain a vector consisting of a sequence of small numbers (0,1,2,…). Set up the distribution: x = sentence abstractness expressing the number of abstract words in it, y = number of sentences with abstractness x. Needless to say, this is merely the first step in measuring abstractness of texts.

Find a theoretical distribution of X and compare various texts. The most abstract will be, of course, mathematical texts. Compare individual poetic texts, compare authors, compare text sorts and compare the same text translated into several languages.

Can you set up an order? Find a link between text abstractness and some other properties of texts, e.g. sentence length, word length, mean frequency of words, etc. That is, strive for incorporating sentence abstractness into a Köhlerian control cycle.

References

Beauregard, M., Chertkow, H., Bub, D., Murtha, S., Dixon, R., Evans, A. (1997). The neural substrate for concrete, abstract, and emotional word lexica: A positron emission tomography study. *Journal of Cognitive Neuroscience 9, 441–461.*

Bresson, D., Kubczak, J. (eds.) (1998). *Abstrakte Nomina*. Tübingen: Narr.

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6, 139-151*. Bochum: Brockmeyer.

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

Künne, W. (2007). *Abstrakte Gegenstände. Semantik und Ontologie*. Frankfurt: Klostermann.

Plaut, D.C., Shallice, T. (1991). Effects of Word Abstractness in a Connectionist Model of Deep Dyslexia. *Proceedings of the 13th Annual Meeting of the Cognitive Science Society, Chicago, IL, August 1991, 73–78.*

Schierholz, S. (1991). *Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive.* Tübingen: M. Niemeyer

Skipper, L.M., Olson, I.R.(2014). Semantic memory: Distinct neural representations for abstractness and valence. *Brain & Language 130 (2014) 1–10.*

6.3. Polysemy

Problem

Poddubnyy and Polikarpov (2013) presented frequencies of polysemy data using three Russian and two English dictionaries. (1) Find a unique continuous function expressing the extent of word polysemy. (2) Compare Russian and English, or, if the circumstances are positive, add your own language.

Procedure

Use the data presented by the above mentioned authors and search for a continuous function with as few parameters as possible capturing all the data. Apply software. Omit all zeroes in the data, i.e. consider only non-zero occurrences. Or, if you want to derive a discrete distribution, pool the empirical data below the first zero frequency. For the resulting function of continuous data set up the differential equation and interpret the parameters using the *unified theory*. For the discrete distribution, set up a difference equation or obtain the distribution directly by transformation of the continuous function.

Compare all data. You need not perform a test for difference; use rather some other method, e.g. show the place of the dictionaries in a two-dimensional system in form of Ord’s indicators <I, S>.

Compare your results with other quantitative studies of polysemy and interpret your result.

Classify the words you analyzed into parts of speech, state the mean polysemy within the given class, order the classes according to mean polysemy, rank the classes and find a continuous function capturing the observed course.

References

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

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol 3.* Lüdenscheid: RAM-Verlag.

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

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

Mačutek, J., Altmann, G. (2007). Discrete and continuous modeling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Ord, J.K. (1972). *Families of frequency distributions.* London: Griffin.

Poddubnyy, V., Polikarpov, A. (2013). Stochastic dynamic model of evolution of language sign ensembles. In: Obradović, I.. Kelih, E., Köhler, R. (eds.). *Methods and Applications of Quantitative Linguistics. Selected Papers of the 8th International Conference on Quantitative Linguisttics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 69-83*. Belgrade: Academic Mind**.**

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics Vol 1.* Lüdenscheid: RAM-Verlag.

Wimmer, G., Altmann, G. (1999). Verteilung der Polysemie in Maori. In: Genzor, J., Ondrejovič, S. (eds.), *Pange Lingua: 17-25.* Bratislava: Veda.

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: de Gruyter.

6.4. Measurement of verb activity

Problem

Find a method for measuring the activity expressed by the given verb. Then analyze a text and express numerically its activity. If you define a new indicator, do not forget to derive its sampling properties.

Procedure

Consider different verb classifications. For example, Yesypenko (2009) uses the following classes:

Verbs
Verbs of motion/removing
Verbs of process, change, development
Verbs of beginning/end of action
Verbs of physical action
Engender verbs
Destroy verbs
Successful/Unsuccessful action implementation
Verbs of attempt
Verbs of sound emission
Verbs of light phenomena
Verbs of temperature phenomena
Verbs of nature phenomena
Verbs of communication
Verbs of moral impact/effect
Verbs of social activity
Position verbs
Verbs of existence
Modality verbs
Verbs of human relations
Verbs of reference
Verbs of emotional psychological impact
Verbs of ownership/loss
Verbs of physiological state
Verbs of perception
Verbs of mental activity
Verbs of subjective assessment
Verbs of emotional psychological state

Silnickij (1993) mentions 20 classes of verbs.

You may ascribe an activity indicator/degree to individual classes or you can distinguish activity even within a class. In any case, show several examples of individual classes.

Take a text and create a sequence of verb activities as they occur in the text. Express the extent of activity by some indicator, e.g. the mean activity. This is simple and can easily be used for comparisons with other texts.

Realize that any researcher could set up a different scale. You can use any criterion.

Apply the indicator to different text sorts in order to see the power of your indicator.

Consider the sequence of activities and characterize it considering it a time series.

Find a model for the distribution of activities. If you analyzed different text sorts, compare them also graphically computing for each sort the Ord criterion.

You can perform all operations also on sentences, i.e. you construct a scale for determining sentence activity taking inspiration from speech act theory.

Another possibility is to take only those verbs which occur in the given text.

Still another possibility is to take into account not only verbs but also adjectives, adverbs etc. expressing some activity and construct a combined activity indicator.

Whatever indicator you propose, do not forget to show the possibility of testing the differences between texts, that is, derive at least the variance of the indicator and propose the asymptotic normal test.

References

Cf. also Problems Vol. 3, 67 ff. and a number of explanations on the Internet.

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

Croft, W., Cruse, D.A. (2004). *Cognitive linguistics*. New York: Cambridge University Press.

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

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

Levickij, V.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.

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

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

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

Silnicky, V. (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.

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.

6.5. Word class specification

Problem

If you solved at least one of the previous three problems concerning parts-of-speech, continue analyzing the frequency of their specification. In the first steps, consider only one language; use two different ways of specification and show which of them is more in agreement with the distribution.

Procedure

Take a text, restrict your analysis to only one word class, e.g. adjectives, verbs, or nouns. Classify the word class members according to some well known works (cf. Levin 1998; Ballmer, Brennenstuhl 1986; Jurčenko 1985; Silnickij 1966; 1973; Yesypenko 2009). Then compute the representation of individual classes in the text in the toform of frequencies. Show that a writer abides by some regularity which can be expressed by a distribution. Propose a distribution, derive it from theoretical consideration and substantiate it linguistically (stylistically).

Compare several texts of the same text sort and find a common distribution for all of them. In the first steps, you can apply also a simple (non-normalized) function. Later on, it can be transformed in a distribution.

Then consider another text sort and do the same. Can you apply the same distribution or not? If so, show the difference in some parameters. If not, propose a modification of the distribution based on some boundary conditions. Strive for a unified theory.

In the next step, take texts from another language, present the results and compare them with those of the first language.

Strive for a typology of writers, examine the development of a writer or of a text sort and that of languages – if possible.

Since the resulting distribution has parameters, you can define some indicators and find their relation to other indicators, i.e. strive for finding links between word class specification and other properties of language.

References

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

Croft, W., Cruse, D.A. (2004). *Cognitive linguistics.* New York: Cambridge Univeristy Press.

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

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

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

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.

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: Kelih, E., Levickij V., Altmann, G. (eds.), *Methods of Text Analysis: 308-328.* Chernivtsi, ČNU.

6.6. Metaphor

Problem

Using the rich literature – available also on the Internet – set up a classification of metaphor types. Consider not only semantic problems but also the possibility of the numbers of words or word lengths, etc. in the metaphor. Take into consideration also the possibility of scaling the strength and the distance from the background meaning. Set up hypotheses concerning the behavior of metaphors and test them using texts.

Procedure

First take a text and write out all the metaphors. Then classify them. To each metaphor write in parentheses its simplest (non-metaphoric) meaning.

(1) Perform the count of classes, i.e. set up the frequency distribution of the classes.

(2) Find a distribution or function – at the beginning do it inductively, later on, begin to theorize – capturing the distribution.

(3) Set up the distribution of metaphor length and find at least an inductive model.

(4) Scale the metaphors according to their distance to the expressions or words they represent.

(5) Find the distribution of this property and an adequate model.

Consider other texts of the same text sort, e.g. press texts. Perform the same procedures and compare the texts. Find a commonality. It is either one of the distributions or some of its properties. Perform tests for similarity/difference.

Then take another text-sort and do the same. Step by step, develop a relationship between your indicators and the text sort. If there is some relationship, find its form as a function.

Finally, take another language and begin to perform the same investigation. If your approach was correct, you will find similar phenomena in the other language, too. Perform tests and insert your results in a theoretical framework.

References

Black, M. (ed.) (1962). *Models and Metaphors*, Ithaca, NY: Cornell University Press.

Blumenberg, H. (1997). *Paradigmen zu einer Metaphorologie.* Frankfurt/Main: Suhrkamp.

Gibbs, R.W. (ed.) (2008). *The Cambridge Handbook of Metaphor and Thought.* Cambridge: Cambridge Univ. Pr.

Goatly, A. (2011^2). *The Language of Metaphors*. London, New York: Routledge.

Goschler, J. (2012). *Metaphern*. Tübingen: Julius Groos Verlag.

Haverkamp, A. (ed.) (1996). *Theorie der Metapher.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Hintikka, J. (ed.) (1994). *Aspects of Metaphor.* Dordrecht: Kluwer.

Knowles, M., Moon, R. (2006). *Introducing Metaphor*. London, New York: Routledge.

Kohl, K. (2007). *Metapher*. Stuttgart/Weimar: Metzler.

Kövecses, Z. (2010^2). *Metaphor. A Practical Introduction*. Oxford: University Press.

Kurz, G. (1982). *Metapher, Allegorie, Symbol*. Göttingen: Vandenhoeck und Ruprecht

Lakoff , G., Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Levin, S.R. (1977). *The Semantics of Metaphor*. Baltimore: Johns Hopkins University Press.

Ortony, A. (ed.) (1993). *Metaphor and Thought.* Cambridge, U.K.: Cambridge University Press

Rolf, E. (2005). *Metaphertheorien. Typologie – Darstellung – Bibliographie.* Berlin–New York: de Gruyter,

Sacks, S. (ed.) (1979). *On Metaphor*. Chicago: University of Chicago Press.

Skirl, H., Schwarz-Friesel, M. (2007). *Metapher*. Heidelberg: Winter.

7. Other problems

7.1. Morphological motifs

Problem

Define morphological motifs, study their occurrences in texts, study their properties, set up hypotheses and test them in at least two languages.

Procedure

Define the types of morphemes, e.g. stem, affix (prefix, infix, suffix), internal change, reduplicational morpheme, clitic, suppletivism, and transcribe a text in terms of these classes in the form of abbreviations. You obtain a sequence of symbols. Segment the sequence in Köhlerian R-motifs, i.e. a new motif begins with a symbol which occurred in the immediately preceding motif. You obtain units which need not correspond with your knowledge of language. Now study the following properties of your grammatical motifs:

(1) *Frequency*. Set up the spectrum of frequencies (i.e. x = occurrence, y = number of motifs occurring x-times), then set up the rank-frequency distribution, i.e. order the motifs according to their frequency. Express the resulting distributions by a probability distribution or by a function. Apply them to different texts and compare them.

(2) Propose an indicator based on frequencies characterizing the *type* of language. Construct an indicator in such a way that you can compute its sampling properties (at least its mean and variance) in order to be able to order the languages or to perform asymptotic tests for differences. Compare also texts of the same text sort in a given language.

(3) *Length.* Study the length of the motifs. There will be motifs having length 1 up to the maximal number of different abbreviations. Compute the frequencies of individual lengths and propose a distribution or a function capturing it. Characterize texts by the mean length; compare texts, text sorts, and languages.

(4) *Link.* Study the link between frequency and length. For example, compute the mean length of motifs occurring once, twice, etc. Then state whether there is some relation between these two quantities. In the positive case, express the relation by an inductive formula. You can obtain it using software. In the next step, express your formula either by a difference or a differential equation. Search in any case for the linguistic substantiation of the parameters. It is to be noted that the link may be different in different languages. If it is so, find a boundary condition expressing it. It may be done by a re-interpretation of parameters or by adding a third variable which must be quantified and measured, too.

If you have N different types of grammatical entities, express the variability of your text as a ratio of different observed and possible motifs. Since the

R-motifs do not allow the repetition of the same entity, it is easy to compute the number of possible ones (length x = 1,2,…,N).

References

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012, 33-43.* Belgrade.

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

Köhler, R. (2015). Linguistic motifs. In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 89-108.* Berlin/Boston: de Gryuter.

7.2. Sentence motifs

Problem

Study the problem 7.1. *Morphological motifs* and perform the same task with sentences.

Procedure

First define the types of sentences (declarative, interrogative, imperative, simple, complex, combined, etc.), decide whether the sentence must end with a dot, exclamation mark, interrogation mark, colon, semicolon, etc. and prepare a list of abbreviations for all types. Then take a text and transcribe it using your abbreviations.

Find all motifs defined as sequences of not repeated symbols. In this way you obtain a sequence of sentence motifs. Now state the frequency of individual motif types and set up their spectrum; then prepare the rank-frequency distribution.

Then solve all problems displayed in 2.1. *Morphological motifs*.

Compare the distributions of morphological motifs with that of sentences. Is there a difference in the distributions/functions? Were you forced to apply another distribution analyzing the sentence level? Explain the difference.

Analyze especially stage plays and compare the individual acts. How do the distributions change from the first act to the last? Can you distinguish the classical parts of a drama? Choose an indicator of the motif distribution and study its degree through the acts.

Perform the same analysis using the classes of speech acts (cf. *Problems Vol 4. 94-101* and this volume) but using R-motifs. Study the same problem concentrating on the age of the speaker, e.g. children of different ages. Texts can be found in the respective literature. The speech act analysis segments the text differently. If you performed the morphological, the sentential and the speech act analyses of the same text(s), you have made three steps in the text hierarchy. Can you draw some consequences?

Perform the analysis also for the translation of the same work in some other language and compare the individual levels.

References

Hindelang, G. (2010). *Einführung in die Sprechakttheorie: Sprechakte, Äußerungsformen, Sprechaktsequenzen.* Berlin/New York: Mouton de Gruyter.

Köhler, R., Altmann, G. (2014). *Problems in Quantitative Linguistics Vol. 4.* Lüdenscheid: RAM-Verlag.

Tsohatzidis, S. (ed.) (1994). *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives.* London: Routledge.

Ulkan, M. (1993). *Zur Klassifikation von Sprechakten. Eine grundlagentheoretische Fallstudie.* Tübingen: Niemeyer.

Problems: *Morphological motifs* and *Stage play 2* and *3* in this volume.

7.3. Borrowings

Problem

Test whether the number of new borrowings from the source language to the target language follows the Piotrowski law or derive a new model. Study the semantics of borrowed words.

Procedure

First read Problem “6.15. Borrowing” in *Problems Vol. 4: 129f.* Study the literature listed there.

Then take a regularly appearing text in a language other than English, e.g. a yearly catalogue or a newspaper from 2000 to 2014. Consider only one issue per year and study the anglicisms. Make a list of “English” words for each year separately. Prepare a table of (a) all English words occurring in the given issue in each year, (b) only new English words, i.e. omit repetitions in all following years.

For (a) test the homogeneity of the borrowing, i.e. are the numbers of borrowings in each year “similar”? Perform the chi-square test for homogeneity. If there is no homogeneity (the critical value for 14 degrees of freedom at $\alpha = 0.05$ is 23.7), then state which year is extremely deviant. Comment on the given year in your words.

Continue studying this problem and state whether there is some increasing tendency.

For (b) you have only the new words. Prepare a cumulative table, i.e. add the number in 2001 to that in 2000, then add 200+2001+2002, etc. You obtain an increasing sequence. Fit the Piotrowski law to this sequence. If it does not capture the data sufficiently, then either modify the model or find a new process that may lead to the rise of the given sequence.

Compare your results with those concerning other target languages. Omit words of other origin that came into your language through English.

Scrutinize other dynamic processes in language and strive for a theory.

References

Best, K.-H. (2003). Anglizismen – quantitativ. *Göttinger Beiträge zur Sprachwissenschaft 8, 7-23.*

Best, K.-H., Kelih, E. (2014). *Entlehnungen und Fremdwörter: Quantitative Aspekte.* Lüdenscheid: RAM-Verlag.

Köhler, R. Altmann, G. (2014). *Problems in Quantitative Linguistics Vol. 4.* Lüdensdheid: RAM-Verlag.

Müller-Hasemann, W. (1983). Das Eindringen englischer Wörter ins Deutsche ab 1945. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Spraxchwandelforschung 143-160.* Göttingen: Herodot.

Stuhlpfarrer, M. (2010). Anglizismen im Russischen. *Glottotheory 3(1), 97-109.*

7.4. Syllabic word length

Problem

Popescu, Best, Altmann (2014) proposed a model for any kind of length of linguistic units in the form

$$y = cx^{a + b \ln x}.$$

Test the model fitting it to as many data as you have. It is merely a generalization of the power law.

Procedure

Use software (e.g. NLREG, TableCurves, Origin etc.) and fit it to your data. Observe the values of the parameters *a* and *b*. It is to be noted that x cannot be 0.

In Slavic languages there are many zero-syllabic prepositions. If one considers them as independent words, one must use a modified model. Since they are usually proclitics of the next word, they can simply be omitted.

Test the model, fitting it to Bulgarian word-length as presented by Uhlirová (2001). The word-length data are presented in Table 1.

Table 1
Syllabic word-length in Bulgarian letters according to Uhlířová (2001)

Word–length frequency								
Text	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Rad 1	30	25	19	11	3	4		
Mumi	49	29	22	18	6	1		
Iskra 4	49	33	25	12	6	2	1	
Adam	54	31	23	18	9	5		
Genad 1	49	35	27	16	8	2		
Iskra 2	55	41	34	14	6	1		
Marg	62	38	37	12	6	1	1	
Iskra 1	65	37	40	9	7	2		
Juri	79	42	26	9	4	1		
Jorn	68	44	31	25	6	3	0	1
Iskra 5	72	43	36	22	5			
Dam 1	71	52	32	17	11	3		
Kost	56	51	55	19	14	4	3	
Sasa 1	94	73	52	29	9	2		
Sasa 2	109	60	62	21	8	2		
Boris 1	112	85	51	11	11	3		
Dam 2	134	90	58	28	10	9		
Jorn 1	120	80	64	48	17	7	0	1
Cen 1	142	75	48	41	26	5	2	1
Jan 3	154	91	87	35	13	4	1	
Jan 1	194	122	102	46	17	5	1	
Alb	198	145	90	44	17	4		
Cen 2	186	139	106	45	11	11	1	
Ziv 1	209	129	91	54	29	9	2	
Jorn 2	180	121	117	75	26	11	1	
Ziv 2	204	137	124	37	24	10	4	2
Jan 4	262	141	151	66	37	12	5	
Jan 2	302	164	133	67	34	8	1	
Boris 2	275	189	173	52	32	13	1	
Bacv 1	297	181	168	90	44	17	2	1

Test the model also fitting it to clause and sentence length as given by Uhlířová (2001). As can be seen, zero-syllabic words have been omitted. In Slavic languages they are proclitics joined phonetically with the following word.

References

Popescu, I.-I., Best, K.-H., Altmann, G. (2015). *Unified Modeling of Length in Language.* Lüdenscheid: RAM.

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Linguistics Paradigm: Levels, Constituents, Construct. Festschrift in honour of Luděk Hřebíček: 266-282.* Trier: WVT.

7.5. Clause length

Problem

Uhlířová (2001) studied clause length in Bulgarian in terms of the number of words and fitted to the empirical data the mixed negative binomial distribution. Since this distribution has 7 parameters, find a simpler (not normalized, continuous) function capturing the data.

Procedure

The data presented by Uhlířová (2001) are as follows:

Table 1
Clause lengths in three Bulgarian texts (Uhlířová 2001)

Length	Text 1	Text 2	Text 3
1	14	3	2
2	82	17	6
3	95	13	15
4	115	16	28
5	127	15	25
6	123	14	17
7	103	15	10
8	91	12	11
9	72	3	8
10	53	6	6
11	47	4	3
12	32	2	1
13	22	2	2
14	18	1	3
15	13	1	0
16	9	0	2
17	7	0	0

18	9	1	0
19	4	0	0
20	1	0	2

Find the appropriate function using software, i.e. find a model mechanically. You will obtain several good results. Then derive the functions from differential equations and interpret their components. Keep the function whose interpretation is linguistically well substantiated. Rely on the unified theory (Wimmer, Altmann 2005).

If other languages or texts are at your disposal, (1) compare them with the present results and order the languages; (2) investigate the clause length in other text-sorts and construct, step by step, a typology. If possible, use the same text translated to languages you know and can analyze.

References

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Linguistics Paradigm: Levels, Constituents, Construct. Festschrift in honour of Luděk Hřebíček: 266-282.* Trier: WVT.

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: de Gruyter.

7.6. Word length and number of compounds

Problem

According to Altmann’s (1988) hypothesis there is a link between the length of a word and the number of compounds of which it is a component. Simply, the shorter a word is (in terms of syllable numbers), the more compounds are formed with it. Hammerl (1990) generalized the hypothesis. Test the hypothesis using the Polish data published by Hammerl (1990).

Procedure

Use the data presented by Hammerl (1990). He considered length the independent variable and the number of compounds the dependent variable and proposed the Hyperpoisson and the Hyperpascal distributions. Instead of a distribution, apply simply a function *number of compounds = f(length of the word)*. Propose a function which is adequate for Polish data. Then study other languages and show the difference in the parameters of the function.

Apply the function inductively (i.e. using software), then substantiate it theoretically, i.e. derive it from a differential equation relying on the unified theory (cf. Wimmer, Altmann 2005).

Bear in mind that this candidate for a law could have different forms in languages of different types, hence there must be some boundary conditions. If you succeed in applying your theory to several languages, utilize your approach for typological purposes.

References

Altmann, G. (1988). Hypotheses about compounds. In: R. Hammerl (ed.), *Glottometrika 10, 100-107*. Bochum: Brockmeyer.

Hammerl, R. (1990). Überprüfung einer Hypothese zur Kompositabildung (am polnischen Sprachmaterial). In: Hammerl, R. (ed.), *Glottometrika 12, 73-83*. Bochum: Brockmeyer.

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: de Gruyter.

7.7. Word frequency and number of compounds

Problem

It has been mentioned many times that the more frequent a word, the more compounds there are of which it is a component. This follows from the Köhlerian requirement of specification. Test the simple version using the frequency dictionary and the normal dictionary.

Procedure

Take a normal dictionary of a language and write out all the nouns beginning with the letter [a]. Then take a frequency dictionary and write out their frequency. Then take again the normal dictionary and search for all compounds containing the given word as a component. If you have an online dictionary, this step can be made mechanically. In strongly synthetic languages, take care of different morphs of the given word. Do not forget that compounds are not only stems written together but a number of various types with different degree of cohesion (cf. Fan, Altmann 2007a,b). Many of them cannot be found in a normal dictionary but one can begin in this way.

If you have all the data, order the simple words by their increasing frequency. Some of the words may belong to the same frequency class; in that case you must take the mean of the number of compounds containing them, i.e. if there are 5 words occurring each 10 times, then divide the number of compounds formed from these words by 10. In this way you obtain a function whose independent variable is frequency, and the dependent variable is the mean number of compounds.

According to the hypothesis, it can be supposed that the sequence will be increasing. Set up a model and test it. Perform the tests stepwise: first take only nouns beginning with [a], then continue up to [z]. At last, take means and test the hypothesis for nouns. Do the same for verbs and adjectives and generalize. Find the boundary conditions – if necessary.

Then perform the same operations in a second language. Compare the parameters. Do not use polynomials as fitting functions, because they cannot easily be substantiated linguistically. Derive the resulting function relying on Zipf's and Köhler's arguments, i.e. substantiate it linguistically.

References

Altmann, G. (1988). Hypotheses about compounds. In: Hammerl, R. (ed.), *Glottometrika 10: 100-107.* Bochum: Brockmeyer.

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 English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.), *Problems of typological and Quantitative Lexicology: 190-209*. Černivcy: RUTA.

Hammerl, R. (1990). Überprüfung einer Hypothese zur Kompositabildung (an polnischem Sprachmaterial). In: Hammerl, R. (ed.), *Gllottometrika 12, 73-83.* Bochum: Brockmeyer.

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

Zipf, G.K. (1935). *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Boston: Houghton Mifflin

Register of names

a Campo, F. 105
Adams, S.. 43,44
Adsett, C.R. 110
Agard, F.B. 105
Ágel, V. 58
Ahl, V. 4,9
Ahrens, A. 43,45
Aikhenvald, A.Y. 58
Aioanei, D. 106
Aitken, A.J. 46
Allen, T.F.H. 4,9
Allerton, D. 58
Altmann, E.G. 23
Altmann, G. 5,7,9-14,17,19,21,26-32,35-40,42,46,48,51,53-59, 61,62,64-68,70-74,77,79-83, 88-95,97,98,100-102,105,107, 109,111-113,115,118,120,123, 124,129,130,133-135
Altmann, V. 21,46,53,79,98
Alvarez-Lacalle, E. 23
Amaral, L.A. 24
Anderson, S.R. 62
Andreev, N.D. 70
Antosch, F. 40
Anward, J. 65
Anz, T. 43,45
Arnold, J. E. 75
Ashby, F.G. 100
Asratian, A.S. 27
Augst, G. 105
Austin, J.L. 94
Austin, W.M. 105
Baayen, H. 16
Bach, K. 94
Baerman, M. 62
Baets, B.de 101
Bailey, R.W. 40,46,57
Bakker, F.J. 40
Ballmer, T.T. 90,122-124
Bane, M. 62
Barabási, A.-L. 23
Batóg, T. 105
Beauregard, M. 118
Beddor, P. 111
Beliankou, A. 49,74,75,79,128
Bell, A. 23
Belza, M.I. 37
Bergen, B. 115
Bergenholtz, H. 65,66
Berger, T. 113
Bernstein, Y. 101
Berry-Roghe, G.I.M. 46
Best, K.-H. 12,13,30-32,40-42,53.57, 60,66,68,73,83,85,130,132
Biberman. Y. 100
Bibok, K. 99
Bird, A. 2
Bisang, W. 5
Black, M. 125
Blumenberg, H. 125
Bock, H.H. 101
Boder, D.P. 40
Bollobás, B. 27
Boriah, S. 101
Boschtan, A. 60
Bose, A. 105
Botha, R.P. 84
Boy, J. 16,118
Boyd, R. 2
Brandwood, I. 57
Brenier, J. 23
Brennenstuhl, W. 90, 122-124
Bresson, D. 118
Brock, J. 94
Broselow, E. 84
Brown, D. 16,62
Brysbaert, M. 118
Bub, D. 118
Buch, K.R. 57
Buchanan, L. 105
Bucková, M. 19,58
Budai, L. 58
Budescu, D.V. 101
Bunde, A. 101
Bunge, M. 2,12,13
Bünting, K.-D. 66
Büring, D. 75

Burkhardt, A.S. 94
Burnaby, T. 101
Busemann, A. 40
Buyukkokten, O. 101
Bybee, J. 16,64
Bzdęga, A.Z. 84
Camps, J. 85
Carlstein, E. 101
Carson-Berndsen, J. 105,106
Čech, R. 8,31,32,36,40,58,63,64,77,
79,81-83,93,94,101,102,120
Cha, Sung-Hyuk 14,16
Chafe, W. 75
Chandola, V. 101
Chang, Ch.B. 62
Changizi, M.A. 110
Chen, R. 37
Chertkow, H. 118
Chitoran, I. 105,106
Cohen, A.D. 95
Colangelo, A. 105
Cole, P. 88,95
Coloma, G. 10
Colston, H. 45
Comrie, B. 5,65
Corbett, G. 15,16,62
Corral, A. 23
Coupé, Ch. 105,106
Cover J.A. 2
Croft, W. 5,101
Croft, W.B. 90,120,124
Cruse, D.A. 90,122,124
Crystal, D. 66
Culp, Ch. 85
Curd, M. 2
Cysouw, M. 5
Damashek, M. 101
Dąmbrowska, E.99
Daneš, F. 75
Dannhauer, H.-M.46
David, J. 46
Davidson, D.J. 7,83
Davis, S. 95
DeAngelis, D. 4
Dechert, H.W. 95
Denley, T.M.J. 27
Deza E.14,16
Deza M.M. 14,16
Diaz-Guilera A. 23
Diessel, H. 79
Dik, S.C. 75
Dimitriadis, A. 84
Dixon, R. 118
Dixon, R.M.W. 58
Doležel, L. 40,57
Dolphin, C. 46
Dömötör, Z. 9,27
Donselaar, W.van 75
Dorow, B. 23
Drebet, V.V. 99
Đuraš, G. 41,93
Earman, J. 2
Eckert, G. 5
Eckmann, J.-P. 23
Eichner, J.F. 101
Elsen, I. 115
Elts, J. 118
Emons, R. 58
Engdahl, E. 77
Erteschik-Shir, N. 75
Evans, A. 118
Everaert, M. 84
Fabricius, A.H. 84
Fan, F. 14,21,31,32,59,95,107,111,
112,120,134,135
Fenk, A. 111
Fenk, L. 111
Fenk-Oczlon, G. 111
Ferrer-i-Cancho, R. 23
Féry, C. 75
Fickermann, I. 16,97
Filipović Durdević, D. 8,83
Finck, F.N. 5
Firbas, J. 76
Fischer, H. 40
Ford, C.E. 79
Fox, B. 88
Fraassen, B.C. van 2

Frazier, L. 100
Frey, W. 76
Gabelentz, G.v.d. 76
Garabík, R. 7,83
Garvin, P. 75
Gasper, P. 2
Gaume, B. 4
Genzor, J. 19,58,120
Geofroy, A. 46
Geršić, S. 105
Gibbons, J.D. 21
Gibbs, R.W. 43-45,125
Ginstrom, A. 75
Giraud, C. 23
Givón, T. 60,76,88
Glymour, C. 2
Gnatchuk, A. 115
Goatly, A. 125
Godbole, A. P. 21
Godfrey-Smith, P. 2
Goh, K.-I. 23
Goldman-Eisler, F. 40
Goodall, D.W. 101
Goodwin, G. 23
Goschler, J. 125
Götz-Vottler, K. 58
Greenberg, J.H. 5,85
Gregory, M. 23
Grice, H.P. 95
Gries, St.Th. 108,109
Grimes, J.E. 105
Groot, A.W. de 28
Grosz, B. 23
Grotjahn, R. 105
Grzybek, P. 19,26,28,29,41,42.59,
80,113
Gundel, J.K. 74,76
Günderson, K. 95
Gunkel, L. 61,88
Gunzenhäuser, R. 48,108
Haarmann, H..6
Haeberlin, H. 84
Häggkvist, R. 27
Hajičová, E. 76
Halliday, M.A.K. 61,76,88,122
Hamilton-Smith, N. 46
Hammerl, R. 23,66,68,133,134
Hammond, M. 76
Hanulíková, A. 7,83
Harary, F. 106
Harjung, J.D. 43,45
Harnish, R.M. 94
Haspelmath, M.H. 6
Haverkamp, A. 125
Havlin, S. 101
Haynie, H.J. 62
Hedberg, N. 74,76
Helbig, G. 58,66
Herbst, T. 58
Herrera, J.P. 23
Hindelang, G. 129
Hinrichs, U. 6
Hintikka, J. 125
Hinton, L. 115
Hippisley, A. 16
Hockett, Ch.F. 76
Hoekstra, T. 4
Hoffmann, L. 61
Hopkins, E..105
Hopper, P. 16
Hornberger, N.H. 95
Householder, F.W.Jr. 70,76
Hřebíček, L. 14,16,28,35,36,40,63,
66,71,77,109
Hurch, B. 84,85
Huyssteen, G.B.v. 85
Hymes, D. 106
Ineichen, G. 6
Ishihara, S. 75
Ivanová, M. 58
Jacobs, J. 76,77
Jäger, G. 76
Janda, L.A. 79,99,108,109
Jastrzembski, J.E. 98
Jayaram, B.D. 28
Jayawardhana, A. 21
Jirsová, A. 59
Johnson, M. 126

Joshi, A. 23
Juola, P. 62
Jurafsky, D. 23
Jurčenko, G.E. 90,122-124
Kadmon, N. 76
Kaliuščenko, V. 135
Kaltz, B. 66
Kane, M. 105
Kantelhardt, J.W. 101
Karlík, P. 58
Karlsson, F. 111
Katz, S.M. 23
Kelih, E. 8,29,31,32,48,49,53,59,64, 75,79,80,82,83,91,113,120,123, 124,128,130
Kempgen, S. 113
Kerre, E. 101
Key, H. 85
Kiiko, J.J. 99,122
Killian-Hatz, Ch. 116
Kisro-Völker, S. 118
Klemke, E. 2
Knauer, K. 108
Knowles, M. 125
Kohl, K. 125
Köhler, R. 5-8,10-14,16,17,19-21, 24,28,29,31,32,35,38-40,46,48, 49,51-59,61,64,66-68,71,72,74-76,79,80,83,85,86,88,89,93-95, 98,101,102,112,113,115,117, 118,120,123,128-130,133-135
Kohlhase, J. 130
König, E. 6
Kostić, A. 7,83
Kövecses, Z. 125
Krajewski, G. 7,83
Krappe, A.H. 108
Kreuzer, H. 40,108
Krifka, M. 76
Krstev, C. 113
Krug, M. 64
Krupa, V. 6,28
Kruyt, J.G. 77
Kubczak, J. 118
Kullenberg, H. 48
Kumar, V. 101
Künne, W. 118
Kunz, M. 24
Kuperman, V. 118
Kurz, G. 125
Ladefoged, P. 105
Ladyman, J. 2
Lafon, F. 46
Lakoff, G. 126
Lambrecht, K. 76
Lang, E. 6
Lange, M. 2
Lausberg, H. 43,45
Leacock, C. 99
Lee, Ch.-Ch. 105
Lehfeldt, W. 5,70,105,113
Lehmann, W.P. 6
Lenno, J.G. 2
Lentz, J. 75
Levická, J. 7,83
Levickij, V. 31,32,48,71,72,82,91, 98,99,113,115,122-124,135
Levin, B. 90,122-124
Levin, S.R. 126
Lewy, E. 6
Li, Ch.N. 75,76
Lieven, E.V.M. 7,83
Lin, Yi-Ling 21
Lindner, G. 105
Liu, H. 24
Longacre, R.E. 80
Lopatková, M. 59
Losongco, A. 75
Lučak, M. 122
Lupea, M. 12,14
Machamer, P. 2
Mačutek, J. 8,24,28-32,35,36,40,53, 58,59,63,64,80,83,90,102,113, 120,124,128
Maddieson, I. 105,111
Maduka-Durunze, O.N. 116
Malmgren, R.D. 24
Marchand, Y. 110

Marriot, P. 16
Marsico, E. 105,106
Martin, R. 46
Mathesius, V. 77
Matskuliak, Y. 31,32,102
Mattes, V. 84,85
Mauclair, J. 105,106
McCarthy, J.J. 54
McGuire, J.E. 2
McKay, S.L. 95
Melka, T.S. 107
Memar, P. 4
Menon, G. 116
Mesarović, M.D. 4
Metzler, D. 101
Meyer, U. 43,45
Miestano, M. 111
Mikk, J. 118
Mikros, G. 24,59,128
Milin, P. 8,83
Mirković, J. 7,83
Moffat, A.101
Molnár, V. 77
Montemurro, M.A. 24
Moon, R. 125
Moravcsik, E. 76,85
Morgan, J. 88
Morgan, J.L. 95
Moscoso del Prado Martin, F. 8,83
Moses, E. 23
Most, R.B. 77
Motter, A.F. 23,24
Mugdan, J. 65
Müller-Hasemann, W. 130
Murtha, S. 118
Musgrave, S. 84
Naumann, C.L. 105
Naumann, S. 12,19,26,49,59,74,75,
79,80,128
Németh, T.E. 99
Nichols, J. 6,62,115
Niehaus, B. 56,57
Nižníková, J. 59
Nooteboom, S.G. 77
Norton, J.D. 2
Obradović, I. 8,29,49,64,75,79,83,
112,113,120,128
Obuljen, A. 113
Oesterreicher, W. 6
Ohala, J.J. 115
Ohala, M. 111
Olshtain, E. 95
Olson, I.R. 118
Ondrejovič, S. 40,71,77,109,120
O'Neill, R.V. 4
Ord, J.K. 120
Ortony, A. 126
Osgood, Ch. 114,116
Ossner, J. 66
Pajas, P. 58
Panevová, J. 59
Panther, K-U. 85
Papastavridis, G. 21
Partee, B.H. 76
Pattee, H.H. 4
Pavlyčko, O.O. 71,72
Payne, Th.E. 77
Pellegrino, F. 105,106
Perrin, N.A. 100
Peterson, G.H. 106
Pethö, G. 99
Pierrehumbert, J.B. 23
Piotrowski, R.G. 5,7,10,11,14,17,21,
30,32,35,39,40,46,48,54,55,61,
64,66-68,71,72,74,79,89,93,98,
102,112,115,118,120,133-135
Plaut, D.C. 118
Plett, H.F. 43,45
Poddubnyy, V. 119,120
Polikarpov, A. 119,120
Politi, M. 24
Polyakov, V. 111
Popescu, I.-I. 12-14,19,26,28,31,32,
38-40,42,46,51,53,66,70,73,80,
83,102,130,132
Portner, P. 77
Posner, R. 17
Preisach, C. 12,59,79

Primus, B. 77
Prince, E.F. 77
Prouzová, H. 59
Průcha, J. 28
Pszczolowska, L. 108
Pumain, D. 4
Pury, P.A. 23
Pustet, R. 28,70
Radden, G. 85
Rádl, Z. 24
Radulović, V. 113
Raible, W. 6
Ramat, P. 6
Ravin, Y. 99
Reinhart, T. 77.
Rice, K. 62
Rieger, B. 46,113,123
Rijkhoff, J. 61,88
Riška, A. 9,27
Roelcke, Th. 6,17,62,81
Roeper, Th. 110
Rolf, E. 126
Rooth, M. 77
Rosenberg, A. 2
Rothe, U. 16,66,69,83
Rottmann, O. 57
Rowland, C. 99
Rubino, C. 85
Sacks, S 126
Sahami, M. 101
Salmon, W.C. 2
Salthe, S. 10
Saltz, E. 77
Sanada, H. 59
Sander, Th. 95
Scales, E. 24
Schaeder, B. 66
Schaffner, K.F. 2
Scheibman, J. 123
Schenkel, W. 58
Schierholz, S. 117,118
Schindler, W. 85
Schlissmann, A. 40
Schmale, G. 48
Schmidt-Thieme, L. 12,59,79
Schreuder, R. 16
Schulz, K.P. 56,80,98
Schumacher, H. 59
Schüttpelz, E. 43,45
Schwarz, M. 123
Schwarz-Friesel, M. 126
Schweers, A. 66,68,69
Schweier, U. 113
Scullen, M.E. 85
Searle, J.R. 95
Sebeok, Th.A. 17
Seidel, G. 46
Seiler, H. 6
Semenyuk, T.G. 71,72
Sgall, P. 76,77
Shallice, T. 118
Shewan, A. 109
Shopen, T. 6,80
Shisted, R. 62
Silnickij, G.G. 90
Silnickij, V.V. 122-124
Simon, H.A. 4
Sinnemäki, K. 111
Skalička, V. 85
Skinner, B.F. 13,14,16,109,110
Skipper, L.M. 118
Skirl, H. 126
Skorochod'ko, E.F. 37
Sokolová, M. 59
Solé, M-J. 111
Solovyev, V. 111
Spertus, E. 101
Spolnicka, S.V. 122
Stachowski, K. 85
Staffeldt, S. 95
Starr, T.B. 4
Steen, G. 44
Steffen-Batogowa, M. 105
Stouferm, D.B. 24
Strauss, U. 14,21,31,32,59,95,111,
112,129
Strecker, B. 61
Stuhlpfarrer, M. 130

Štukovský, R. 28
Suci, G.J. 114,116
Svozilová, N. 59
Swadesh, M. 106
Tannenbaum P.H. 114,116
Tatar, D. 12,14
Teyssier, J. 48
Theakston, A.L. 7,83,99
Thompson, S.A. 76,80
Thun, N. 85
Tolstaja, S.M. 106
Tournier, M. 46
Trout, J.D. 2
Tsohatzidis, S.L. 95,129
Tuldava, J. 40
Tuzzi, A. 24,46,59,66
Uhlířová, L. 28,36,53,57,69,71-74, 77,79-82,102,131-133
Uibo, H. 118
Ulkan, M. 95, 129
Unuk, D. 113
Vainikka, A. 100
Vallduvi. E. 77
Vennemann, T. 106,111
Verstraete, J.-C. 80
Vidya, M.N. 28
Vielliers, J. de 100
Vincze, V. 66
Voeltz, F.K.E.116
Vogel, P.M. 65
Vulanović, R. 66
Waide, J. 4
Wang Hua 88,89
Wang, X. 101
Ward, G.L. 77
Warren, B. 48,49
Warriner, A.B. 118
Wasow, T. 75,77
Weinstein, S. 23
Weiss, P.A. 4
Welke, K. 59
West, D.B. 27
Whyte, L.L. 4
Wickmann, D. 46
Wierzbicka, A. 90,123,123
Wiese, R. 85
Wiltshire, C. 85
Wimmer, G. 29,30,35,36,30-40,57, 65-68,71-74,77,92,93,101,102, 107,109,113,118,120,132-134
Wimmerová, S. 40,71,77,109
Wirth, J. 76
Yabushita. K. 77
Yesypenko, N. 48,49,81,82,90,91, 121,123,124
Yorkston, E. 116
Zacharski, R. 74,76
Zadotozhna, L. 102
Zanette, D.H. 24
Zhu, J. 66,68,69,85
Ziegler, A. 35,36,40,46,53,66-69
Zifonun, G. 6,61,88
Zimmermann, M. 95
Zipf, G.K. 11,17,135
Zobel, J. 109
Zörnig, P. 16,19,24,26,51,80,112, 113
Zwick, R. 101

Register of subjects

abstractness 117-119
activity 54,102,103,121,122
adjectival motif 49
adjective 47,48,54,
adjective-verb ratio 39,40,51
adnominal modifier 60
adverb 8,,9
adverbial 76-78,80-82
aggregation 100
alliteration 107,109
assonance 107,108
asymmetry 23
Belza-chain 36
binomial test 60
borrowing 129,130
centrality 50,51,69,70
clause length 72,73,132-134
clause type 71
climax 27,28
compactness 101,102
complexity 81,88,104,107,110
compound 133,134
concentration 71,72
conceptual inertia 36-38
control cycle 38,48,54,72,89,115
denotative-connotative concentration 34,35
distance 13-15,23,25,42,63,65,87, 89,92,96,99
distribution
- binomial 61,109
- Dacey 92
- Fucks-Poisson 92
- Hyperpascal 133
- Hyperpoisson 133
- mixed negative binomial 132
- mixed Poisson 66
- negative binomial 113
- negative hypergeometric 66
- of distances 25
diversification 4-7,82
economy 107
entropy 43,47,51,81,83,86,89,98
euphony 111,112
evolution 7
excess 43,87
explanation 1
frequency 20,54,114,125,127,134, 135
frequency motif 18
frequency sequence 25
Gini's coeffcient 51,83,87
grammar 60-91
grammatical category 82,83
hexameter 111
hierarchical net 3
hierarchy 2
homogeneity 68
h-point 30,32,89
hreb 36
Hurst exponent 63
hypothesis 1,3,5,12,13
iconicity 114,115
irregularity 15,21
Köhlerian control cycle 17,64,68, 115,117
Köhlerian motif, s. motif
Köhlerian requirement 9,17
Language
- Brazilian-Portuguese 66
- Bulgarian 131,132
- Chinese 107
- Czech 63
- English 49,109,112,119,129, 130
- French 104,108,112
- German 71,111
- Hindi 108
- Hungarian 108
- Indonesian 112
- Latin 8,41,104,111
- Old Church Slavic 41

- Polish 133
- Portuguese 68
- Romance languages 41,104
- Russian 15,29,119
- Serbian 113
- Slavic languages 41,104,108, 130

law 1,13,14,114
length 10,11,19,27,29,48,54,72, 73,78, 83,87,124,127
length--frequency 11
link 127
Markov chain 65,78,87,97
measurement 1
Menzerath's law 3
metaphor 124,125
model 1
morphological complexity 20,27, 54,61-63
morphological change 53
motif 58,65,78,96,106,107,127-130
motif range 19
nominativity 52,53,54
noun phrase 87,88
Ord's criterion 8,23,43,47,51,81,83, 86,87, 89,120,122
parallelism 107
parts of speech 7,8,11,64-68,82,84, 89,90,123,124
phonetic symbolism 114
Piotrowski law 129,130
poetif figure 42,43,44,
Poisson process 23
polysemy 20,27,54,98,117,119,120
polytexty 8,
power function 26
predication 54,55
problem 1,16,17
property 3,7
quantification 1
range 96
rank-frequency 8,9,47,58,66,71,72, 78, 83, 86,87
reaction 99
reduplication 83,84,
repeat rate 43,47,51,81,83,87,89
rhetoric figure 42,43,44
R-motif 49,128,
run 19,20,65,81,87,96
scaling 43,44,47,54
sentence length 29,55,56,92
sequence 21,22
similarity 5,13,14,100,107
skewness, 23
Skinner effect 22,100
Skinner hypothesis 13,15,109,
sound symbolism 84,
spectrum 42,
speech act 93-96,122
stage play 92-103
steepness 83,86
syllable 104,110,112,113
symmetry 60,78,87,97
synergetic linguistics 13
synonymy 20
synthetism 63
test 1
text 18-59
text compactness 35,36
text sort 51,52
thematic concentration 30,31,32,33, 34,38
theory 1,3,16,22,
time series 48,
topic-comment 74,75
transition probability 65,97
uniformity 72
valency 57,58,
verb 121,122
vocabulary richness 51
Weber-Fechner law 23
word class, s. parts of speech
word length 41,130,131,133,134
world view 26,27,45,46,
Zipf-Alekseev function 10,25,26,86
Zipf's law 25,86

Contents Glottometrics 32, 2015

Hanna Gnatchuk

A quantitative investigation of English compounds in prose texts 1-8

Cong Zhang, Haitao Liu

A quantitative investigation of the genre development of modern Chinese novels 9-20

Peter Zörnig, Ioan-Iovitz Popescu, Gabriel Altmann

Statistical approach to measure stylistic centrality 21-54

Xiaxing Pan, Hui Qiu, Haitao Liu

Golden section in Chinese contemporary poetry 55-62

Yu Fang, Haitao Liu

Probability distribution of interlingual lexical divergences in Chinese and English: 道 (*dao*) and *said* in *Hongloumeng* 63-87

Christopher Michels

The relationship between word length and compounding activity in English 88-98

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies.**

Herausgeber – Editors

Actual external academic peers for Glottometrics

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de
Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. –32 (2015). – Lüdenscheid: RAM-Verlag, 2015
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.
Bibliographische Deskription nach 32 (2015)
ISSN 1617-8351