Problems in Quantitative Linguistics

6

by

Emmerich Kelih
Gabriel Altmann

2018

RAM-Verlag

Studies in Quantitative Linguistics

Editors

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

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
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. Mačutek, 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 pp.
19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.

20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional Occurrences in Texts: Weighted Consensus Strings.* 2015. II+178 pp.
23. E. Kelih, R. Knight, J. Mačutek, A.Wilson (eds.), *Issues in Quantitative Linguistics Vol. 4*. 2016. III + 231 pp.
24. J. Léon, S. Loiseau (eds.), *History of quantitative linguistics in France*. 2016. II + 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, III+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme*, 2017, IV + 134 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language.* 2018, VI+119 pp.

Diese Publikation wurde gefördert durch die Universität Wien.

ISBN: 978-3-942303-57-6

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

Preface

This book is the sixth volume in the series of "Problems in Quantitative Linguistics", which contains a further collection of selected open problems and questions within the paradigm of quantitative linguistics. In some cases, already formulated hypotheses about the particular behaviour of linguistic entities and phenomena are given, including some interrelations and links. In other cases only some general lines of thinking and ideas are given, which may be of interest for the further development of quantitative linguistics. An orthodox fulfilment of the given ideas and hypotheses is not required since the proposed ideas and hypotheses have to be understood as initial stimulus for an in-depth analysis and ongoing theoretical foundation.

The focus of the present book is on phonological, morphological and semantic problems and hypotheses. Additionally, some selected problems of the syntactic and lexical level, including some quantitative ideas of the analysis of borrowings in language systems, and selected problems of a quantitative analysis of poetry are proposed. The given references are in most cases only a selection of the most relevant literature available and it is strongly recommended to enlarge the bibliography if analysing the selected problem or hypotheses. Moreover, in some cases no directly relevant references could be given. However, the authors hope to give at least some basic ideas about some desiderata of quantitative linguistics. Researchers are explicitly encouraged to submit the yielded results of the analysed problems and hypotheses to the editor of *Glottometrics* (www.gabrielaltmann.de) or one of the editors of *Glottotheory* (http://homepage.univie.ac.at/emmerich.kelih/).

Vienna, Lüdenscheid February 2018

Contents

1. General

1.1. Allometric growth in language

Problem

Allometry is a well-known problem, especially in biology. It concerns the proportional increase of the size of one property depending on the size of another, e.g. organ sizes increase depending on the increase of body size. In linguistics (cf. Hřebíček/Altmann 1996, Tuldava 1998: 53), a well-known problem is that of inverse allometry, usually placed under Menzerath's law: the greater a construct, the smaller are its immediate components. Find some cases of "true" allometry, describe them and derive for each an adequate formula. Cf. also the problem *Semantic diversification* in this volume.

Procedure

Search for inspiration in linguistic synergetics where everything is related to something else. Care only for allometric cases, for example: the greater the polysemy of a word, the more compounds it produces. Compounds reduce the semantic vagueness of the polysemic word. Now, derive an appropriate formula capturing this dependence. If possible, use an interpreted differential equation, i.e. interpret the functions of which it consists and ascribe them to language, hearer and speaker; test the hypothesis using dictionaries in several languages. Order the languages according to the size of the parameters of the function and show the typological relevance of the results.

Find at least three allometric relationships, set up a control cycle. Do not care in this case for inverse allometry. Try to find some broader framework of the relevance of the allometric approach in linguistics and interrelations to other scaling laws in biology, economics, physics, etc.

References

Problems in Quantitative Linguistics Vol. 1 to 5.

Bär, J.A. (2014). Methoden historischer Semantik am Beispiel Max Webers – Teil 1. *Glottotheory 5(2), 243-296.*

Bär, J.A. (2015). Methoden historischer Semantik am Beispiel Max Webers – Teil 2. *Glottotheory 6(1), 1-92.*

Hřebíček, L., Altmann, G. (1996): The levels of order in language. In: Peter Schmidt (ed.), *Glottometrika 15. Issues in general linguistic theory and the theory of word length: 38-61.* Trier: Wissenschaftlicher Verlag Trier (Quantitative Linguistics, 57).

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

Mainzer, K. (2008). *Komplexität*. Paderborn: UTB.

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Analyse*. Trier: Wissenschaftlicher Verlag Trier (Quantitative Linguistics, 59).

1.2. Centrality

Problem

If one considers language as a collection of inventories of entities, one may ask questions about the centrality of individual inventories. The concept itself has not been defined everywhere, but one can begin. Consider for example the phonemic system: one can define centrality by means of evaluating the frequencies of individual phonemes and apply some indicator. If one considers the system of distinctive features, one can define its centrality by means of frequencies or by means of the presence of individual features in phonemes, and/or by means of relations of this system to other systems in language. If one considers the inventory of syllables, one can add the positions of individual syllables in words, the occurrence in sequences of syllables, etc. The higher the level of the observed entities, the more possibilities there are to define centrality.

One can define centrality by means of an indicator but one must consider its samplings properties, i.e. comparability by means of a test. If one sees several possibilities, all should be applied. Since the individual entities take place in the vicinity of the centre, one can define the distribution of centrality values for individual entities and evaluate it. One can compare the same systems in various languages, one can compute the parameters for several systems in a hierarchy and state whether there is a trend.

Solve at least one of these problems in one language. Ensure clearness and refer to results attained in qualitative linguistics.

Procedure

First define the systems, e.g. phonemes, distinctive features, syllables and their properties, morphemes and their properties, parts of speech, compounds, phrases, clauses, sentences, motifs, hrebs. Then for each system define several possible centrality measures and for individual members of the system define the degree of their centrality. Evaluate the centrality of the members by applying a function or a distribution. Compare all systems in the hierarchy, capture them by the same model and compare the parameters. They will change according to the level of

the system in the hierarchy. Construct a measure for this change. If you involve frequencies, use only ready data. Keep it simple.

Systems are characterized by many properties. Take other ones and state whether centrality has a relation to these properties. If so, construct a system of relations similar to that developed by R. Köhler (2005).

If you succeeded in attaining any results in your L1-language, perform the same operations in another language. Compare the results and generalize step by step.

The solution of this problem is a task for a team of researchers.

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.

1.3. Hierarchies

Problem

Language as an open dynamic system contains a number of different hierarchies. Find several ones, describe them and show what kind of regularity between the levels could be conjectured.

Elaborate at least one of the hierarchies in detail, perform measurements and test your hypothesis.

Procedure

First collect the available literature concerning at least partial hierarchies. For example, Menzerath's law makes its way in the material hierarchy: sentence – clause – word – syllable/morpheme, and controls the length of the construct and its (immediate) constituents.

Show hierarchies in grammar, semantics, lexicon and typology. Lexical hierarchies can be found for example in definition chains (c.f. Sambor, Hammerl 1991). One can take a group of lexemes belonging to some common domain and search for a more abstract lexeme of which they are special cases. Usually one finds it in every monolingual lexicon. The more abstract word is again defined by a more abstract word, etc. For example *man – animal – organism – system – thing.* Show different hierarchies concerning various parts of speech.

Grammatical hierarchies are described in usual grammars but one can also set up hierarchies concerning the grammatical behaviour of classes of words (cf. Best 2002). This is different for each language. You obtain a tree. Compare the trees of two languages using quantitative comparisons and order the languages according to one of the tree properties.

In semantics, one can take one of the parts of speech and subdivide it into classes. There are many classifications of adjectives, nouns verbs, adverbs, …. One obtains differently occupied classes. Now state a property and subdivide the classes according to the degree of this property, in order to obtain further subclasses. Then define the next property and do the same. You obtain a tree which has specific properties. Express them quantitatively.

References

Altmann, G., Schwibbe, M.H. (eds.) (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim/Zürich/New York: Olms.

Best, K.-H. (2002). *Linguistik in Kürze*. Göttingen.

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. In: Boy, J., Köhler, R. (eds.), *Glottometrika 6: 138-151*. Bochum: Brockmeyer.

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

1.4. Iconicity

Problem

Under iconicity one understands in general the similarity between the icon and the object to which it refers. In semiotics (represented by Ch.S. Peirce and C.W. Morris), the icon is a class of signs which stay in a direct perceptible relation to the indicated object by some kind of imitation of some aspects of the real object and display some similarity or commonality of properties. Unfortunately, there is no method for measuring the extent of iconicity, indexality and symbolicity of individual signs. Hence, it would be of utmost importance for semiotics to develop a quantification of these properties, which would make it easier to find their relation to other properties of language.

This is shown by Haspelmath (2008: 6): *Greater quantities in meaning are expressed by greater quantities of form.* Referring to R. Jakobson there are three domains in which iconic behaviour can be observed:

(a) With adjectives, there is an increase in the number of phonemes, starting from positive, comparative to superlative forms. The degrees may also be expressed by separate words.

(b) Compared with the singular, the forms of plural are expressed by a greater number of phonemes (cf. Cuypere 2008: 78, esp. 131 ff.). For a discussion see Mayerthaler (1980: 20), Wurzel (1989: 11).
(c) One postulates, specifically for Russian, that the perfective aspect expressing a limiting of the event is expressed by a smaller number of phonemes than the imperfective aspect (cf. Fenk-Oczlon 1990).

Haspelmath (2008: 5) discusses these observations critically; he emphasizes that these phenomena depend on the frequency of use, i.e. the given size can be explained by the frequency of occurrence. For example the positive form with adjectives occurs more frequently than the superlative; the same holds for singular as compared with plural (cf. the problem *Shortness of form*). Test his hypotheses: (a) The more predictable a sign is, the shorter it is, and (b) the more frequent a sign is, the shorter it is, a hypothesis known already to G.K. Zipf. Use the synergetic approaches.

Procedure

Take a sample of 100 frequent and rare word forms subdivided according to parts of speech: noun, adjective, verb. Measure the word length in terms of syllable numbers. The measurement can be performed in various ways according to the given declination and conjugation patterns; then measure their frequency in texts (text types, partial corpuses). Then compare the respective results statistically, i.e. test the differences between singular and plural forms, positive and comparative forms, perfective and imperfective forms. Observe whether very frequent and very seldom forms yield similar results. Take into account (besides frequency) other possible factors, e.g. suppletivism, and search for other factors.

Present the results and evaluate them. If possible, perform the analysis for several languages.

References

Cuypere, L. de (2008). *Limiting the iconic. From the metatheoretical foundations to the creative possibilities of iconicity in language*. Amsterdam: Benjamins (Iconicity in language and literature, 6).

Fenk-Oczlon, G. (1990). Ikonismus vs. Ökonomieprinzip. Am Beispiel russischer Aspekt- und Kasusbildungen. *Papiere zur Linguistik 42, 46–69.*

Haiman, J. (1983). Iconic and economic motivation. *Language 59 (4), 781–819.*

Haiman, J. (2000). Iconicity. In: G. Booij, Ch. Lehmann, J. Mugdan (eds.), *Morphologie / Morphology: 281-288*. Berlin: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 17,1).

Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics 19 (1), 1–33.*

Hiraga, M.K., Herlofsky, W.J., Shinohara, K., Akita, K. (eds.) (2015). *Iconicity. East meets West.* Amsterdam/Philadelphia: Benjamins Iconicity in language and literature, 14).

Mayerthaler, W. (1980). Ikonismus in der Morphologie. *Zeitschrift für Semiotik 2, 19–37*.

Simone, R. (ed.) (1994). *Iconicity in language.* Amsterdam/Philadelphia: Benjamins (Amsterdam studies in the theory and history of linguistic science, 110).

Schmidtke, D.S., Conrad, M., Jacobs, A.M. (2014). Phonological iconicity. *Frontiers in psychology 5, 1–6.*

Wurzel, W.U. (1989). *Inflectional morphology and naturalness*. Dordrecht: Kluwer (Studies in natural language and linguistic theory, 9).

1.5. Suppletivism in general

Problem

In linguistics the relation between suppletion and word frequency has been discussed for a long time. Suppletion is the replacement of a given word by some other word, e.g. the English "be" is supplemented by "am, are, is, was, has been", … (cf. Kruszewski 1995 and the newer literature from the domain of usage-based linguistics, Bybee 2007, 2010)

Besides an empirically useful definition of suppletion (cf. Mel'čuk 2000) we still do not have a substantiation of the relation between frequency and suppletion. On the one hand, frequent forms are created because of coding economy, on the other hand they tend to have a number of forms which are frequently irregular. Mostly there are no systematic empirical investigations on the suppletion of very frequent word forms. It is well known from several European languages that many irregular verbs (strong verbs) leave the class of strong verbs and obtain regular forms. Hence there are two domains of research: (1) The study of the relation between frequency and suppletion, (2) a historical study on the leaving of irregularity because of frequency.

Procedure

Take existing frequency dictionaries (cf. Alekseev 2005, Davis and Gardner (2010) for American English) and find between them the most frequent word forms (e.g. the first 1,000, 2,000, 3,000, depending on the capacity of the dictionary) that have suppletive forms. (1) Create a list of suppletive forms, (2) state the part of speech of these suppletion forms, (3) publish the list (Internet, Open Archives) and quote the sources, (4) develop a synergetic control circuit in

which one can see not only the frequency of words but also the length of units and their polysemy. Do frequent forms which have suppletive forms display a special behaviour? (5) Since inflectional languages have more suppletivism, compare the indicator of inflectionality with the indicator of suppletion and set up a relationship.

References

Alekseev, P.M. (2005). Frequency dictionaries. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 312-324*. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Corbett, G., Hippisley, A., Brown, D., Marriot, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure: 201-226.* Amsterdam/Philadelphia: Benjamins (Typological studies in language, 45).

Hippisley, A. Chumakina, M., Corbett, G.G., Brown, D. (2004). Suppletion: Frequency, categories and distribution of stems. *Studies in Language 28 (2), 387–418.*

Kruszewski, M. (1995). Writings in general linguistics. Edited and with an introdcution by Konrad Koerner. Amsterdam, Benjamins (Amsterdam studies in the theory and history of linguistic science, 1).

Mel'čuk, I. (2000). Suppletion. In: Geert Booij, Christian Lehmann und Joachim Mugdan (eds.), *Morphologie/ Morphology: 510-522.* Berlin,de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 17,1).

Osthoff, H. (1899). *Vom Suppletivwesen der indogermanischen Sprachen*: akademische Rede zur Feier des Geburtsfestes des höchstseligen Grossherzogs Karl Friedrich am 22. November 1899 bei dem Vortrag des Jahresberichts und der Verkündung der akademischen Preise. Heidelberg: Hörning.

Wurzel, W.U. (1990). Gedanken zur Suppletion und Natürlichkeit. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 43 (1), 86–91.*

Sources:

http://www.smg.surrey.ac.uk/suppletion/ (accessed 01/07/2018)
http://www.smg.surrey.ac.uk/bibliographies/ (accessed 01/07/2018)

1.6. Motifs

Problem

Motifs or Köhlerian motifs have been studied in *Problems Vol. 1 to 5*. Prepare a list of possible motifs and show that they behave like any other linguistic unit. Collect all laws of their behaviour and define the links between them.

Procedure

Distinguish quantitative and symbolic motifs. Define 100 linguistic properties from all domains of language – using all units that were already scrutinized – and define new ones. Take a text and set up the set of all its motifs. Quantitative motifs consist of sequences of non-decreasing numbers obtained by some kind of measurement, while qualitative or symbolic motifs consist of sequences of symbols, none of which may be repeated in the same motif.

Then begin to theorize. For those properties which are already well known, e.g. word length, word classes, sentence length, speech acts, etc., study the fitting of the same distributions or functions that were used for the frequencies or rank frequencies of the usual properties. Use simply available software and strive for obtaining the same model for all cases.

Show whether Köhler's control cycle also holds for parallel motifs, e.g. word length vs. word length motifs. Construct a control cycle in which – at least hypothetically – all motif properties are linked and test the hypotheses by applying them to your text.

Check the results using a text in another language.

If some links cannot be corroborated, study the boundary conditions, variegate the analysis of the texts and strive for a theoretical background. If necessary, consider a property as linked with two other properties at the same time, i.e. apply analogous formulas with two independent and one dependent variable.

Analyse not only the numbers resulting from the formulas but – if you analysed several texts – study also the relationships between the parameters of the resulting functions.

There is a possibility that you discover a different control cycle because motifs are hierarchically higher entities that can be constructed of any linguistic entities.

References

See *Problems Vol. 1 to 5* and the references therein.

2. Phonemics

2.1. Size of the phoneme inventory

Problem

Restate the problem of the distribution (number) of phonemes in a language. State the inventory size from the existing literature: Hockett (1955, 1958), Sigurd (1963), Lehfeldt (1975), Maddieson (1984), WALS (2013). The principle of counting must be identical in each case. Set up a model of inventory size in human languages and strive for derivation and explanation.

Procedure

First take the above-mentioned literature and evaluate the inventory size for each language. You may differentiate genetic or areal-typological aspects (cf. WALS 2013).

If you have evaluated at least 100 languages, set up the empirical distribution of inventory sizes. Strive to find a theoretical model in which the parameters are interpreted at least in an empirical way. Use physiological, acoustic, geographic and perceptual limitations as well as the minimization effort of the decoding hearer.

Set up the distributions for different subgroups (genetic, areal). Set up different models for vowel numbers and consonant numbers and try to find the causes of the given state.

To find a distribution or function use the well-known programs like NLREG, TableCurve, Origin, Altmann-Fitter, etc. If you obtain several adequate models, take that one which can be easily (linguistically) explicated.

References

Hockett, Ch. F. (1955). *A manual of phonology*. Baltimore: Waverly Press (Indiana University publications in anthropology and linguistics, 11).

Hockett, C. F. (1958). *A course in modern linguistics*. New York: Macmillan.

Kelih, E. (2016). *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main: Peter Lang, 38-42.

Lehfeldt, W. (1975), Die Verteilung der Phonemanzahl in den natürlichen Sprachen. *Phonetica 31, 274–287.*

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

Maddieson, I. (1986). The Size and Structure of Phonological Inventories: Analysis of UPSID. In J. Ohala and J.J. Jaeger (eds.), *Experimental Phonology: 105-123. Orlando: FL Academic Press*

Maddieson, I. (2006): Correlating phonological complexity: Data and validation. *Linguistic Typology 10 (1),106–123.*

Maddieson, I. (2009). Calculating phonological complexity. In: F. Pellegrino, E. Marsico, I. Chitoran und C. Coupé (eds.), *Approaches to Phonological Complexity: 85-109.* Berlin, Mouton de Gruyter (Phonology and Phonetics, 16).

Sigurd, B. (1963). A note on the number of phonemes. *Statistical Methods in Linguistics 2, 94–99.*

WALS (2013). Dryer, Matthew S., Haspelmath, M.(eds.) (2013). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, accessed on 2018-01-07.).

2.2. Diachronic perspective of the phoneme inventory size

Problem

The basic idea of language change theory is the conjecture that any process of change is to be understood as a (cyclic) simplification process (Rosemeyer 2016, Aitchison 2013, Mańczak 2004, Martinet 1981, Lüdtke 1980, for overviews). This does not hold, for example, for the dictionary, which steadily increases and becomes more and more complex. But considering the phoneme inventory (cf. Bakker 2004, Blevins 2004, Trudgill 2004) one may test whether such a hypothesis is correct. It is to be stated whether the size of the phoneme inventory changes in the course of time – concretely, whether it abides by the decreasing Piotrowski law.

Procedure

State the development of individual languages concerning their phoneme inventory. You may also consider the development of new languages like the Romanian ones from Latin or the Austronesian ones from Proto-Austronesian. Is there a change (increase, decrease, cyclic behaviour)? The possible factors like isolation of a language system, borrowings or loss of functionality can be recorded separately and used as the basis for a synergetic self-regulation circuit. Take into account Skalička's (1958) idea that changes in the number of phonemes are associated with changes in the morphology – i.e. also study the morphological changes in the given language – and Köhler's (2005) conjecture

that simplifications are the result of self-organization by the speaker's effort for ease. The last conjecture may give you a basis for the substantiation of your results.

References

See problem 4.6. in *Problems vol 3*, 77-78.

Aitchison, J. (2013). *Language change. Progress or decay? 4. ed*. Cambridge: Cambridge University Press.

Bakker, P. (2004). Phoneme inventories, language contact, and grammatical complexity: A critique of Trudgill. *Linguistic Typology 8 (3), 368–375.*

Blevins, J. (2005). *Evolutionary phonology: the emergence of sound patterns.* Cambridge, MA: Cambridge University Press.

Čech, R., Altmann, G. (2009). *Problems in quantitative linguistics 3. Dedicated to Reinhard Köhler on the occasion of his 60th birthday.* Lüdenscheid: Ram-Verlag (Studies in Quantitative Linguistics, 12).

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.

Lüdtke, H. (ed.) (1980). *Kommunikationstheoretische Grundlagen des Sprachwandels.* Berlin: de Gruyter.

Mańczak, W. (2005). Diachronie: Grammatik. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 607-627*. Berlin/ New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Martinet, A. (1981). *Sprachökonomie und Lautwandel. Eine Abhandlung über die diachronische Phonologie*. Stuttgart: Klett-Cota.

Rosemeyer, M. (2016). Modeling frequency effects in language change. In: H. Behrens, S. Pfänder (eds.), *Experience Counts: Frequency Effects in Language: 175-208*. Berlin/Boston: de Gruyter (Linguae & litterae)

Skalička, V. (1958). Typologie slovanských jazyků, zvláště ruštiny. *Československá rusistika 3, 73–84.*

Trudgill, P. (2004). Linguistic and social typology. The Austronesian migrations and phoneme inventories. *Linguistic Typology 8, 305–383.*

2.3. On the relation between the size of phoneme inventory and speaker number

Problem

One of the dispersed typological hypothesis says: "The greater the number of speakers of a language, the greater the size of the phoneme inventory" (cf. Atkinson 2011, Hay, Bauer 2007 and for a survey of the state of the discussion cf. Bybee 2011 and the special issue of *Linguistic Typology* 15,2, 2011). Test the hypothesis.

Procedure

Before you begin to test this empirical hypothesis you must take into account the boundary conditions of this problem, especially the number of *native* speakers of a language. Then begin to construct a synergetic control circuit on the basis of speaker number and search for possible starting points (extent of language contacts which can influence the phoneme inventory, diatopic and diastratic structure of a language, redundancy depending on the number of speakers, density of social and communicative bindings of the language). It is recommended to operationalize the individual factors. Show at least the dependency graph, even if you cannot test the hypothesis. It is a very complex problem with an enormous literature.

Ensure exact definitions, e.g. of dialect, sociolect, nativity, etc.

References

Atkinson, Q.D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science 332 (6072), 346–349.*

Bybee, J. (2011). How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate. *Linguistic Typology 15 (2), 147–153*.

Cysouw, M., Dediu, D., Moran, S. (2012). Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa", *Science 335, 657.*

Hay, J., Bauer, L. (2007). Phoneme inventory size and population size. *Language 83 (2), 388–400.*

Hunley, K., Bowern, C., Healy, M. (2012). Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proceedings. Biological sciences / The Royal Society 279 (1736), 2281–2288.*

Jaeger, F.T., Pontillo, D., Graff, P. (2012). Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa". *Science 335, 1042.*

Kelih, E. (2016). *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main: Lang.

Moran, S., McCloy, D., Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language 88 (4), 877–893.*

Pericliev, V. (2011). On phonemic diversity and the origin of language in Africa. *Linguistic Typology 15, 217–221.*

Sproat, R. (2011). Phonemic diversity and the out-of-Africa theory. *Linguistic Typology 15 (2), 199–206.*

Trudgill, P. (2011). Social structure and phoneme inventories. *Linguistic Typology 15 (2), 155–160.*

Wang, Chuan-Chao, Ding, Qi-Liang, Tao, Huan, Li, Hui (2012). Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa". *Science 335, 657.*

Wichmann, S., Rama, T., Holman, E.W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology 15 (2), 177–197.*

2.4. Distribution of vowel/consonant inventory size

Problem

Test the relationship between vowel and consonant inventory. Strive to find a regularity in the languages of the world.

Procedure

First see the problem *Size of the phoneme inventory* but now count separately vowels and consonants. Set up a table (V,C) and capture the relation between them by a hypothesis. Then test it. It is important to distinguish between segmental and suprasegmental properties. At the beginning, you may use a software program, but later on derive the hypothesis from a differential or difference equation and interpret the role of individual parts of the equation in synergetic terms.

References

see *Size of the phoneme inventory*

2.5. On the relationship between the number of vowels and consonants in the inventory

Problem

Several authors have discussed the issue of a statistical relationship between the number of vowels and consonants in a language (cf. Hockett 1955: 138, Maddieson 1984, 2005a, 2005b, Justeson, Stephens 1984). There are three possibilities: (i) If the number of vowels increases, the number of consonants increases, too; (ii) if the number of vowels increases, the number of consonants decreases; (iii) there is no relationship. State which of these hypotheses might be correct – at least preliminarily.

Procedure

Consider as many languages as possible and ensure the same treatment of affricates, vowel lengths, diphthongs, etc. in all of them. The data should be taken from various language families and areal-typological groups should be separated. Set up a hypothesis substantiating the relationship based on the function of vowels and consonants in word formation (grammatical and semantic information) (cf. Kelih 2016: 48-54.) and on the compensation of vowels and consonants in various types of languages (cf. Skalička 1962, 1966). New outlooks can be won from the relationship between vowel inventory and the number of phonemes in words (cf. Coloma 2017). First look at your table, let a software program (e.g. TableCurve) give you all significant results (omit polynomials!), then set up your hypothesis and substantiate it linguistically.

References

Coloma, G. (2017). The existence of negative correlation between linguistic measures across languages. *Corpus Linguistics and Linguistic Theory 13, 1–26.*

Hockett, Ch. F. (1955). *A manual of phonology*. Baltimore: Waverly Press (= Indiana University publications in anthropology and linguistics, 11).

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

Maddieson, I. (2005a). Correlating phonological complexity: data and validation. *In: UC Berkeley Phonology Lab Annual Report, 216–229.*

Maddieson, I. (2005b). Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. *UC Berkeley Phonology Lab Annual Report: 259–268*.

Justeson, J., Stephens, L.D. (1984). On the relationship between numbers of vowels and consonants in phonological systems. *Linguistics 22, 531–545.*
Kelih, E. (2016). *Phonologische Diversität zwischen Phonologie, Morphologie und Syntax*. Frankfurt/Main: Lang.
Skalička, V. (1962). Typologie a konfrontační lingvistika. *Československá rusistika 7, 210–212.*
Skalička, V. (1966). Konsonatenkombinationen und linguistische Typologie. *Travaux Linguistiques de Prague 1, 111–114*.

2.6. Interrelation between phoneme inventory size and syllable structure

Problem

G. Fenk-Oczlon and A. Fenk (2008: 49) set up several hypotheses concerning the interrelations of the syllable structure of a language to other properties:

Hypothesis 1: The greater the inventory of phonemes, the larger the number of canonical syllables.

Hypothesis 2: The greater the number of syllable types, the greater the number of monosyllabic word forms.

Hypothesis 3: The greater the mean number of phonemes in the syllable (= mean length of syllable in terms of phoneme numbers), the greater the number of different syllable types.

Hypothesis 4. The greater the mean syllable length (in terms of phoneme numbers) the greater the number of monosyllabic words.

Test the above hypotheses.

Procedure

State the phoneme inventories in many languages with different inventories. Then state the number of canonical syllable types in texts of similar length (in ideal case in a corpus of parallel texts, in order to improve the comparability). Count the numbers of monosyllabic word forms and state their mean length in terms of phoneme numbers. The syllable boundaries in longer than monosyllabic words should be stated according to a unique principle (cf. Kelih 2012 for an overview). One can adhere to official grammars.

The above-mentioned properties can then be stated (1) by means of the usual correlation analysis, in order to find some tendencies. (2) If tendencies can be discovered, one should propose a nonlinear model in which the phoneme inventory size is the independent variable. (3) One should try to substantiate the given tendency and investigate how these mutual relations can be trade-off

results. (4) Discuss the substantiation of this kind of mutual interrelation and finally (5) show whether the choice of a language (genetically or areal-typologically) plays a particular role.

Alternatively, one can state the number of canonical syllable types, the mean syllable length and the number of monosyllables on the basis of a dictionary.

If you find several adequate functions for the dependencies, choose the simplest one and, in any case, set up a synergetic control cycle.

References

Fenk-Oczlon, G., Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In: M. Miestamo, K. Sinnemäki, F. Karlsson (eds.), *Language complexity. Typology, contact, change*: 43-65. Amsterdam/Philadelphia: Benjamins (Studies in Language Companion Series, 94).

Kelih, E. (2012). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation*. München/Berlin/Washington D.C.: Sagner. (Specimina philologiae Slavicae, 168).

2.7. Interrelation between the number of consonants and their frequency

Problem

The relations between the paradigmatic and syntagmatic levels have been mostly neglected in linguistics. Concerning phonemics one can for example examine whether the number of consonants in a language stays in mutual relation to the frequency of consonants. Starting from an older conjecture of Krámský (1941, 1946, 1948, 1959) the following hypothesis has been formulated by Kelih (2016: 55): "The greater the proportion of consonants on the paradigmatic level (system of phonemes), the smaller is their occurrence in a text." The hypothesis can be substantiated by the possibility that languages with a greater proportion of consonants have a greater articulation space leading to a smaller exploitation on the level of text. Test the hypothesis.

Procedure

State the inventory of consonants in several languages. There are ready-made lists. For each language study the relative proportions of consonants in

approximately equally long texts. (1) State the simple correlation between the relative number in system and the relative number in texts. If you see a dependence, choose a function and express it. Find a substantiation of the detected tendency and state whether one can speak here about trade-off mechanisms. (2) Perform the same operations with the vowels in system and in text. If you find exceptions to your hypothesis, search for the boundary conditions.

References

Kelih, E. (2016). *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main: Peter Lang.

Krámský, J. (1941). Je angličtina jazykem germánským. *Časopis pro moderní filologii 27, 260–268.*

Krámský, J. (1946-1948). Fonologické využití samohláskových fonémat. *Linguistica Slovaca 4(6), 39–43.*

Krámský, J. (1959): A quantitative typology of languages. *Language and Speech 2, 72–85.*

2.8. Phoneme and morpheme inventories, morphemes and word length

Problem

The relations between phoneme inventory and mean morpheme length as well as that between phoneme inventory and mean word length are well known in quantitative linguistics. Weber (2005: 224) shows further factors which are responsible for the fact that in a language with a small phoneme inventory the following conjectures can be made:

1. The smaller the number of phonemes, the smaller is the inventory of morphemes that can be formed with them.
2. The smaller the phoneme inventory, the greater is the mean morpheme length and the smaller is the mean word length.

Since here we are dealing with dependencies, it is possible to set up a synergetic control cycle with the following vertices: (1) phoneme inventory, (2) morpheme length, (3) morpheme inventory, (4) word length.

Set up the control cycle and compute the dependencies in the form of functions.

Procedure

Since dictionaries of morphemes of a language are rather rare, the above problem must be investigated at the level of texts. In the first step one must operationalize the respective entities, namely (1) the phoneme inventory, (2) the morpheme length in terms of phoneme numbers, (3) inventory of morphemes (not on the level of tokens but types, in order to exclude the frequency effect) and (4) the mean number of phonemes in word forms.

The investigation must be performed in several languages with very different phoneme inventories (see the problem: S*ize of the phoneme inventory*) and one should investigate languages with small, medium and very large phoneme inventories. Since one performs a comparative analysis of texts, one should prefer parallel texts.

As soon as one has the necessary data, one should examine first the two-dimensional dependencies (one can consider power or exponential functions). Finally, each dependent variable should be presented as a function of all the other ones. One obtains slightly complex formulas but a mathematician can help with partial differential equations.

Try to connect the discovered dependencies with known hypotheses about morphological complexity (cf. Altmann/Roelcke 2015, Anderson 2015, Juola 1998) and further problems like allomorphic complexity, frequency and allomorphy, synonymy, polysemy, morphological productivity. That means one should extend Köhler's (2005) synergetic control circuit.

References

Anderson, S. R. (2015). Dimensions of morphological complexity. In: M. Baerman, D. Brown und G. G. Corbett (eds.), *Understanding and measuring morphological complexity. First edition: 11-26*. Oxford United Kingdom: Oxford University Press.

Altmann, G., Roelcke, Th. (2015). Morphological complexity of the word. *Glottotheory 6 (1), 93–111.*

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics 5, 206–213.*

Kelih, E. (2016). *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main: Peter Lang.

Köhler, R. (2005). Synergetic linguistics. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 760-775.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 27).

Weber, S. (2005). Zusammenhänge. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 214-226.* Berlin, New

York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

2.9. Exploitation of distinctive features

Problem

A phonemic system can be described in terms of the exploitation of distinctive features. One may strive for the expression of exploitation quantitatively and set up the following hypothesis: "The larger the phonemic inventory, the smaller the degree of exploitation of individual distinctive features." Test the hypothesis.

Procedure

Consider the phonemic systems of several languages. Create the list of phonemes for each and write to each phoneme its characteristic distinctive features. There are two possibilities for processing the problem:

(1) State how many distinctive features occur only in one phoneme, how many occur in exactly two phonemes, etc. Set up the distribution of distinctive features.

(2) State how many phonemes have exactly one feature, how many have two features, etc.

The two testing methods take into account once the exploitation of features and once the characterization of phonemes.

If you obtained results from several languages, test the following hypothesis: "The more phonemes there are in the language, the smaller is the mean exploitation of features." Use simply your computations, compute the mean exploitation for every language and seek a relationship between the mean exploitation and phoneme number.

If the hypotheses are "correct", i.e. can be corroborated, find an explanation for this fact. What kinds of forces are active in the construction of the phonemic systems? If possible, insert these forces into differential or difference equations from which you can derive your distribution/function (cf. Wimmer/ Altmann 2005). You may ascribe the forces a simple constant or a function and explicate it.

References

Clements, G.N. (2009). The role of features in phonological inventories. In: E. Raimy, Ch. Cairns (eds.), *Contemporary Views on Architecture and*

Representations in Phonology: 19-68. Cambridge, MA: M.I.T. Press (Current studies in linguistics, 48)

Clements, G.N. (2003). Feature economy in sound systems. *Phonology 20, 287–333*.

Clements, G.N. (2001). Representational economy in constraint-based phonology. In: T.A. Hall (ed.), *Distinctive feature theory: 71-146.* Berlin: Mouton de Gruyter (Phonology and Phonetics, 2),

Kelih, E. (2016). *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax.* Frankfurt am Main Peter Lang, (p. 35-38).

Surendran, D., Niyogi, P. (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In: O.N. Thomsen (ed.), *Competing Models of Linguistic Change: Evolution and Beyond: 43-58.* Amsterdam: Benjamins.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 791-801*. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

2.10. Allophonic diversification

Problem

A central feature of a phonemic system is its ability to diversify the syntagmatic poverty of the phoneme system by creating allophones. It is still not known what kind of consequences the allophonic diversification has for other language levels (cf. Bowern 2011 for the relevance of the allophonic diversification in recent discussions on the origin of language). Inductively set up some hypotheses and simply test the dependencies.

Procedure

State the number of phonemes in the given languages and the number of allophones according to available phonological descriptions. Then make a list of the number of phonemes vs. number of allophones. Set up a hypothesis on the relation between phoneme number and allophone number.

Then test the relation between diversification and (a) number of vowels, (b) number of consonants, (c) number of suprasegmental properties, (d) positional restrictions of the accent, etc. For each of the questions set up hypotheses (for further information on the recent discussion of the impact of the

allophonic diversification cf. Lavoie 2002). Test also some further relations to the properties of the writing system, morphological properties, typological affiliation of the language, etc.

References

Bowern, C. (2011). Out of Africa? The logic of phoneme inventories and founder effects. *Linguistic Typology 15(2), 207–216.*

Clements, G.N. (2009): The Role of Features in Phonological Inventories. In: E. Raimy and Ch. Cairns (eds.), *Contemporary Views on Architecture and Representations in Phonology: 19-68*. Cambridge, MA: M.I.T. Press (= Current studies in linguistics, 48).

Lavoie, L.M. (2002). Subphonemic and Suballophonic Consonant Variation: The Role of the Phoneme Inventory. *ZAS Papers in Linguistics 28, 39–54.*

Ohala, J.J. (2009). Languages' sound inventories: the devil in the details. In: François Pellegrino, Egidio Marsico, Ioana Chitoran and Christophe Coupé (eds.): *Approaches to Phonological Complexity*: *47-58*. Berlin: Mouton de Gruyter (= Phonology and Phonetics, 16).

2.11. Word-initial phonemes

Problem

Shulzinger and Bormashenko (2017) studied the initial characters of words in English, Polish, French, Russian, German, Latvian, Italian, Hebrew, Spanish and Czech and stated that if the frequencies are ordered, the best fitting results can be attained by using the exponential function. Since counting of letters is not possible in languages using signs (e.g. Chinese, Japanese, Rongorongo, etc.), perform the same study but count the number of word-initial phonemes.

Procedure

Take a dictionary of a language and transform the beginning of words into phonemes. If there are ready-made data, use them. Then for each initial phoneme state the number of respective words. Then order the frequencies in decreasing order and fit the exponential function to the ranking.

Shulzinger/Bormashenko (2017) present a lot of references; take them into consideration.

Finally, order the initial phonemes into phonetic classes – as is usual – and show the rank order of phonetic classes (e.g. bilabial, dental, …). Characterize the language. Then you may begin to theorize, i.e. connect the results with other

properties, e.g. word length, proportion of vowels in the inventory, etc.; that is, begin to set up a control circuit for phonemic word beginnings. Since you now have data, find a function for each relationship.

References

Newcomb, S. (1881). Note on the frequency of use of different digits in natural numbers. *American Journal of Mathematics 4, 39-40*

Shulzinger, E., Bormashenko, E. (2017). On the universal quantitative pattern of the distribution of initial characters in general dictionaries: The exponential distribution is valid for various languages. *Journal of Quantitative Linguistics 24(4), 273-288.*

2.12. Vowel sequences

Problem

Consider only tri-syllabic words in your language and study the following problems: (1) What are the sequences of vowels? (2) What are the sequences of vowel lengths? (3) Can one weight the lengths?

Procedure

Some of the above problems have been touched on and proposed by Hayata (2017). First consider a dictionary of your language and consider only tri-syllabic words. Omit the consonants, consider only vowels. Perform the following examinations:

(a) If all vowels are equal, write the word as AAA; if two vowels are equal and one of them different, you may obtain AAB or ABA or ABB; if all are different, you obtain ABC. Order the five types according to frequency and find a well-fitting model. Now perform the same counting with various individual texts. Do not mix texts. Compare the results of the dictionary and of individual texts. For texts, you may also perform a text type analysis or a historical analysis. Compare the frequencies in individual sources using for example the chi-square test, or a simple rank test.

(b) Study the vowels from the viewpoint of length. If a vowel is short, symbolize it as S; if it is long, symbolize it as L. Now studying the trisyllables in the dictionary you obtain the possible outcomes as SSS, SSL, SLS, LSS, SLL, LSL, LLS and LLL. You must decide whether you consider diphthongs as long vowels. Rank the eight types and find a function expressing the ranking. Then perform the same analysis with individual texts. Compare the dictionary with

texts, perform a text type classification and a historical study. Compare not only the ranking but also individual classes.

(c) If you weight the length, e.g. short vowels obtain the weight 1, the long ones 2, you obtain four classes (lengths 3, 4, 5, 6). This is simply a different scaling and classification of vowel sequences. Again, find a model satisfactorily expressing the distribution, compare the dictionary with texts, perform a historical analysis.

If possible, perform the analysis in other languages too. If a language does not have long vowels, omit (b) and (c). Perform an initial typological comparison in a language family. Is there a significant tendency to prefer S in the first (or last) position? Formulate a hypothesis and test it.

References

Hayata, K. (2017). Phonological rules of present-day Japanese in sign-language dictionaries. *Journal of Quantitative Linguistics 24(4), 367-378.*

2.13. Onomatopoeia and phoneme frequency

Problem

Onomatopoeia are a linguistic phenomenon avoiding the arbitrarity of language signs. From the quantitative point of view it would be interesting to learn whether they differ in some way from the "usual" linguistic signs. Seen from this perspective, one could examine onomatopoeia on the phonetic, phonemic, morphological, syntactic and semantic levels. Restrict yourself to the phonemic frequency. The extent of literature is enormous.

Procedure

Proceed in the following way:

1. State the inventory of onomatopoeia, especially well-established ones that can be found in dictionaries.
2. If necessary, distinguish morphologically the onomatopoetic morphemes and other word forming means.
3. State the phoneme frequency in the onomatopoeia found.
4. State whether all phonemes occurring in onomatopoeia are parts of the "official" phonemic inventory of language.
5. Set up the frequency distribution of phonemes in onomatopoeia and find a theoretical model of their rank frequencies.

6. State whether the given model also holds true for non-onomatopoetic phonemes.
7. Discuss some problematic questions, e.g. whether in onomatopoeia some vowels/consonants are overloaded. To this end you must compare individual phonemes but not the ranked distribution.
8. Points 5, 6 and 7 can be examined on the basis of frequencies established from text collections.
9. Are the phonemes/sounds occurring in onomatopoeia loaded with some emotional colour?

References

Abelin, Å. (1999). *Studies in Sound Symbolism*. Göteborg: Göteborg University (Gothenburg monographs in linguistics, 17).

Graham, J. F. (1992). *Onomatopoetics. Theory of language and literature.* Cambridge: Cambridge Univ. Press (Literature, culture, theory, 4).

Sobkowiak, W. (1990). On the phonostatistics of English onomatopoeia. *Studia Anglica Posnaniensia: An International Review of English Studies, 23: 15-30.*

2.14. Free/fixed accent/stress and word length

Problem

There are languages with free or fixed accent/stress (cf. Hyman 1977). One finds the definition in the literature. From the functional point of view, the free accent can be considered a flexible additional mean for the diversification of coding. The fixed accent can play only a restricted role in coding. Test in some genetically related languages whether the different accent types influence (a) the syllable structure and (b) the word structure (length). Test the following hypotheses:

1. Languages with a free accent have a smaller syllable length than languages with a fixed accent.
2. Languages with a free accent have a smaller word length than languages with a fixed accent.

Procedure

Select some suitable languages (if possible cognate ones) with different accent systems. In the first step, state the mean syllable length in several text parts of the given languages; in the second step state the word length. The comparison of lan-

guages should be made first in terms of means. In the next step, one can compare the distributions of length directly (e.g. by using the chi-square test) or by means of various indicators (moments, entropy, repeat rate, Ord's criterion, etc.). Finally, find appropriate functions or distributions for syllable length and word length. Since in general length abides by the Zipf-Alekseev function, fit them to your data and compare only the parameters.

Perform the same investigation in a third language of the given family, state the differences and apply the model.

References

Hyman, L.M. (1977): On the nature of linguistic stress. In: L.M. Hyman (ed.), *Studies in stress and accent: 37-82.* Los Angeles: Univ. of Southern California (Southern California occasional papers in linguistics, 4).

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

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified Modeling of Length in Language.* Lüdenscheid: RAM-Verlag.

3. Morphology and Related Issues

3.1. Frequency effects in morphology

Problem

In quantitative linguistics, frequency of individual entities plays an eminent role. In many parts of linguistics it is incorporated into the Köhlerian control circuits. In previous issues of *Problems in QL* one finds a number of problems related to frequency. For the domain of morphology, Haspelmath and Sims (2010: 265-277) present a selective survey of frequency effects in morphology (cf. also Berg 2004). Frequency influences the word structure in many ways and the most striking effects are found in inflection. They explore the asymmetry of the inflectional structure, which is, of course, more expressive in strongly synthetic languages. They propose some empirically observed frequency differences, which are summarized in Table 1 (> means "is more frequent than").

Table 1
Frequency asymmetries on morphology

Features	**Values, ordered by frequency**
number	singular > plural > dual
case	nominative > accusative > dative
person	2nd > non-3rd (1st/2nd)
degree	positive > comparative > superlative
voice	active > passive
mood	indicative > subjunctive
polarity	affirmative > negative
tense	present > future

There are few empirical investigations concerning these problems. Haspelmath and Sims (2010: 266) present percentages for some selected Indo-European languages. However, the stated tendency should be analysed in as many languages as possible. Moreover, it is necessary to study this problem by sensitive distinguishing of various text types from which the frequencies are retrieved. For example in scientific texts one (usually) cannot find the second person, in stage

plays many sentences contain it. Hence use the results for characterizing also text types, discourse types, etc.

Procedure

Before one begins to perform an empirical investigation, one should set up a preliminary way of forming a net of mutual relations. To this end, one must take into account synergetic linguistic approaches.

But before you begin, consider the following issues:

(1) In what text and text types should the analysis be performed? Prepare recommendations for the choice of texts. One takes frequently everyday spoken language in the community. Omit corpora in any event, and do not mix texts. If you work in several languages, take rather parallel texts.
(2) Analyse as many different texts as possible and perform the analysis at least in three text types (e.g. press, prose, science).
(3) State the frequency of the given feature. If you take automatically annotated texts, ensure the quality of annotation and tagging.
(4) State the frequency of the feature in the domain of word-form types and word-form tokens. What is the influence of these two different analysis levels?
(5) Decide about the respective feature and its presence in some parts of speech, e.g. in Slavic languages adjectives have gender, case and number; in English or in Polynesian languages it is different. Take into account the presence of the feature with all correlated words.
(6) Since you search for frequencies but the texts are of different lengths, compare the results in the form of relative frequencies using for example an asymptotic normal test, or the binomial test. State the differences in different texts.
(7) Consider the relevance of other features (e.g. shortness with singular and plural)
(8) Strive for embedding the results obtained into a synergetic control cycle.

References

Behrens, H., Pfänder, S. (eds.) (2016). *Experience Counts: Frequency Effects in Language*. Berlin/Boston: de Gruyter (Linguae & litterae).

Berg, Th. (2004). *Linguistic structure and change. An explanation from language processing.* Oxford: Oxford University Press.

Bybee, J. (2007). *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Fenk-Oczlon, G. (2001). Familarity, information flow, and linguistics form. In: J. Bybee, P. Hopper (eds.), *Frequency and the emergence of linguistic structure*: 431-448. Amsterdam/Philadelphia: Benjamins (Typological studies in language, 45).

Haspelmath, M., Sims, A.D. (2010). *Understanding morphology.* London: Hodder.

Köhler, R. (2005). Synergetic linguistics. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 760-774*. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Pfänder, S., Behrens, H. (2016). Experience counts: An introduction to frequency effects in language. In: H. Behrens, S. Pfänder (eds.), *Experience Counts: Frequency Effects in Language: 1-20.* Berlin, Boston: de Gruyter (Linguae & litterae).

3.2. Frequency and irregularity

Problem

Haspelmath and Sims (2010: 274) state that if there are irregularities in inflection, these primarily affect the most frequent lexemes. This is explained by the cited authors in two different ways. On the one hand, frequency leads to phonological reduction, because frequent expressions are relatively predictable, or that speakers can afford to articulate them clearly. On the other hand, frequency leads to better memory settlement and fast lexical access, so that frequent items are less susceptible to analogical levelling and other regularizations. So, while frequency causes faster phonological change, with respect to morphology it has a conserving, decelerating function.

Examine the hypothesis and take into account the size of the regular class, for example weak verbs in German. Their weight is so strong that from time to time verbs from the strong class change their form. For example in Goethe's time we had "Der Hund boll", but today it is "Der Hund bellte" (*the dog barked*). State the trend for regularity in English or your own language.

Procedure

Take a historical dictionary and find 100 irregular verbs. Then take a modern dictionary and find the forms of these verbs. State how many of them became regular. You can use dictionaries of the given language of multiple centuries. In each century state the number of those that changed their class membership. The

resulting numbers form a line whose form can be derived theoretically. Usually it is the variant of the Piotrowski function but maybe you find another one.

Then perform the same operation with another word class or grammatical category. Study also the reduction of categories, e.g. German has three genders in articles, Dutch only two. Slavic languages have three genders for adjectives, Hungarian has none. Compare languages and, if possible, study also the areal situation.

The various influences of frequency should be studied in the framework of synergetic linguistics. The main influence is known but there may be boundary conditions leading to deviating results.

Do not restrict yourself to verbs, study also other phenomena.

References

See: *Frequency effects in morphology*

Haspelmath, M., Sims, A.D. (2010). *Understanding morphology*. London: Hodder.

References on irregularity (selected)

Baronian, L., Kulinich, E. (2012). Paradigm gaps in Whole Word Morphology. In: Th. Stolz (ed.), *Irregularity in morphology (and beyond): 81-100.* Berlin: Akademie Verlag (Studia typologica, 11).

Corbett, G., Hippisley, A., Brown, D., Marriot, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: J. Bybee, P. Hopper (eds.), *Frequency and the emergence of linguistic structure: 201-226.* Amsterdam/ Philadelphia: Benjamins (Typological studies in language, 45).

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? In: *Linguistics: An Interdisciplinary Journal of the Language Sciences 39 (6), 1041–1070.*

Nübling, D. (2000). *Prinzipien der Irregularisierung. Eine kontrastive Analyse von zehn Verben in zehn germanischen Sprachen.* Tübingen: Max Niemeyer (Linguistische Arbeiten, 415).

Stolz, Th., Otsuka, H., Urdze, A., Auwera, J. van der (2012). Introduction: Irregularity glimpses of a ubiquitous phenomenon. In: Th. Stolz (ed.), *Irregularity in morphology (and beyond): 7-38.* Berlin: Akademie Verlag (Studia typologica, 11).

Stolz, Th. (ed.) (2012). *Irregularity in morphology (and beyond).* Berlin: Akademie Verlag (Studia typologica, 11).

Plank, F. (1981). *Morphologische (Ir-)Regularitäten. Aspekte der Wortstrukturtheorie.* Tübingen: Narr (Studien zur deutschen Grammatik, 13).

Ramat, P. (2012). Sturtevant's paradox revisited. In: Th. Stolz (ed.), *Irregularity in morphology (and beyond): 61-80*. Berlin: Akademie Verlag (Studia typologica, 11).

3.3. Interrelation between frequency and differentiation

Problem

Haspelmath/Sims (2012: 268-269) discuss the empirical observation that "generally frequently used values tend to be more differentiated".

In particular, they propose that frequent values show less syncretism than rare values. As one example the partial paradigm of the Old English verb *bindam* ('bind') is given:

	Present Ind	Present Sbjv	Past Ind	Past Sbjv
1 SG	binde	binde	band	bunde
2 SG	bintst	binde	bunde	bunde
3 SG	bint	binde	band	bunde
1-3 PL	bindap	binden	bundon	bunden

This paradigm shows that there is more syncretism in the plural than in the singular, more syncretism in the subjunctive than in the indicative, and more syncretism in the past indicative than in the present indicative.

A second claim is that inflection classes differ primarily with respect to the frequent values and differ less in respect to rare values. Thus, it appears that frequently used values have fewer shared exponents. For the illustration of this tendency an example from Russian noun inflection is given (I-IV are the inflectional classes).

	Singular				Plural			
	IV	**I**	**III**	**II**	**IV**	**I**	**III**	**II**
NOM	-o	- Ø		-a	a		-i	
ACC				-u				
GEN		-a	-i		-Ø	-ov	-ej	- Ø
DAT		-u	-i	-e		-am		
LOC		-e				-ax		
INSTR		-om	-ju	-oj		-ami		

As can be seen, the singular has 12 distinct endings, whereas the plural has only eight. Moreover, according to Haspelmath/Sims (2010: 269), the rare cases (dative, locative, instrumental) show fewer allomorphs than the more frequent cases. Thus, as already discussed by R.O. Jakobson, the greater syncretism can be found in the plural (cf. Brown 2000).

Test the above-mentioned hypotheses in at least one strongly synthetic language.

Procedure

Since the observations presented above are not systematically tested, one has to start with an operationalization of the required linguistic units and phenomena (syncretism, inflection classes, exponents).

In the next step, determine a set of at least 100 verbs and nouns (with two subsets of high- and low-frequency items). State the relations between their frequency and syncretism, express it by a function and compare the parameters of verbs and nouns. Then compare the situations in individual grammatical categories. Draw consequences, generalize the problem.

References

Browne, W. (2000). Serbo-Croatian adjective-declension nouns and Viggo Børndal's principle of compensation. In: O. Tomić Mišeska, M. Radovanović (eds.), *History and Perspectives of Language Study. Papers in honor of Ranko Bugarski: 133-139.* Amsterdam: Benjamins (Amsterdam studies in the theory and history of linguistic science, Series 4, Current issues in linguistic theory, 186).

Haspelmath, M., Sims, A.D. (2010). *Understanding morphology*. London: Hodder.

Luraghi, S. (2000). Synkretismus. In: G. Booij, Ch. Lehmann, J. Mugdan (eds.), *Morphologie / Morphology: 638-647*. Berlin: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 17,1).

Saam, Ch. (2001). *Untersuchungen zur Flexionsmorphologie im Rahmen der Synergetischen Linguistik.* Trier: Magisterarbeit.

3.4. Function words and analytism

Hypothesis

The more function words occur in the text, the more analytic is the language. Test the hypothesis.

Procedure

Take the same text in several languages, e.g. *Le Petit Prince* by Exupéry, consider only the first chapter and compute (a) the number of function words, (b) the number of words containing some kind of inflection, affixation, or composition. Consider as function words all those given by a grammar (e.g. article, conjunction, particle, pronoun, preposition, modal verb, auxiliary verb). Some of them may also contain inflection or affixing, in which case they are also inserted into group (b), e.g. the prepositions in Italian. For some languages you may find texts analysed manually in national corpora.

The comparison is optimal when using the same text but if you do not have any at your disposal, you can also apply the counting to different texts – but ensure you take them from the same text type and approximately same text length.

Consider x = number of function words, y = number of inflected or affixed words. If you do not compare identical texts, consider rather the relative numbers. Then propose the kind of link or dependence, find a function that describes it adequately, first inductively, and finally strive for its derivation, linguistic substantiation, and subsumption in a theory, i.e. strive to find its place in a synergetic control cycle.

This is rather a typological investigation expressed quantitatively. The function words are a sign of analytism; inflections and affixing are signs of synthetism.

Consider in this way the development of child language. Compare the state of Roman languages with that of Latin. Express the development from this point of view quantitatively.

References

Bentz, C., Kiela, D., Hill, F., Buttery, P. (2014): Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. In: *Corpus Linguistics and Linguistic Theory 10 (2), 175–211.*

Braußе, U. (1994). *Lexikalische Funktionen der Synsemantika.* (Forschungsberichte des Instituts für Deutsche Sprache, 71). Tübingen: Narr.

Kelih, E. (2011). Zum Analytismus und Synthetismus in slawischen Sprachen: Morphologische Wortstruktur in slawischen Sprachen. In: K. Karl, G. Krumbholz, M. Lazar (eds.), *Beiträge der europäischen slavistischen Linguistik (Polyslav). Band 14, 99-107.* München/Berlin: Sagner (Die Welt der Slaven, Sammelbände/Sborniki, 43),

Lehmann, Ch. (1995). Synsemantika. In: Joachim Jacobs (ed.), *Syntax. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of contemporary research: 1251-1266.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 9,2),

Popescu, I.-I. et al. (2009). *Word Frequency Studies.* Berlin/New York: Mouton de Gruyter (Quantitative Linguistics, 64).

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

Skalička, V. (2004). *Souborné dílo I-III*. Praha: Karolinum.

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Analyse.* Trier: Wissenschaftlicher Verlag Trier (Quantitative Linguistics, 59).

http://www.christianlehmann.eu/publ/synsemantika.pdf (accessed 01/12/2018)

3.5. Shortness of forms

Problem

Haspelmath and Sims (2010: 267) claim that in inflectional languages more frequently used grammatical cases are shorter than rarely used ones. This hypothesis is part of the Zipfian view but it was not yet systematically tested concerning inflectional languages. Test the hypothesis, e.g. in a Slavic language or Latin, etc.

Procedure

Perform the following operations:

1. State the case system of a language using official grammars.
2. Set up a corpus containing different text types and use it for constructing a frequency dictionary. You may use ready-made corpuses.

3. Take a sample of 100 nouns from the domain with high frequency and 100 from the domain of low frequency.
4. State the frequency of individual grammatical cases for each word form. Set up a rank-frequency table and study whether there is some tendency concerning the frequency of use of forms. Is a certain case the most frequent everywhere?
5. Compare the word forms with high frequency with word forms with low frequency.
6. State the length of the examined word form in terms of phonemes, syllables and morphemes (including zero morphemes) (but not letters!).
7. Study the postulated relation between frequency and length of units at all levels.
8. Insert the phenomenon in a synergetic control cycle but first find a formula expressing the given dependency.
9. Study the problem of iconicity of the forms.

References

Haspelmath, M., Sims, A. D. (2010). *Understanding morphology.* London: Hodder.

3.6. Word length in Czech

Problem

In a thorough investigation, L. Uhlířová (1996) analysed word length in Czech in 24 short stories by B. Hrabal and six journalistic texts (Uhlířová 1994, 1996, 1997). She used the extended binomial distribution with different modifications, testifying to the fact that there are several boundary conditions which are akin either to Hrabal or to the text type or to Czech. In spite of this fact, try to find a unique model for all and extend the investigation to other Czech texts.

Procedure

Consider all data obtained by Uhlířová and do not use for fitting a distribution but a usual function. Use a software program, e.g. TableCurve, which automatically fits many functions. Take that one which is common to the most cases of the data but avoid polynomials. Apply the function with the smallest number of parameters. Admit exceptions and find for them other functions which are related to the main one by an added factor in the differential equation leading

to the main one. In this way you can construct a theory containing also boundary conditions.

Though general theories already exist (cf. Popescu et al. 2014), Czech may exhibit boundary conditions or trends which are quite specific. A theory is never finished, it develops.

References

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

Uhlířová, L. (1994). O jednom rozložení délky slov. *Slovo a slovesnost 56, 8-14.*

Uhlířová, L. (1996). How long are words in Czech? In: P. Schmidt (ed.), *Glotto-metrika 15: 134-146*. Trier: Wissenschaftlicher Verlag.

Uhlířová, L. (1997). Word length distribution in Czech: On the generality of linguistic laws and individuality of texts. In: K.-H. Best (ed.), *Glottometrika 16: 163-173.* Trier: Wissenschaftlicher Verlag.

3.7. Word length in Semitic languages

Problem

S. Abbe (2000) studied word length in Arabic letters and stated that the so-called Cohen-Poisson distribution is appropriate in all 21 cases. For Old Hebrew psalms, C. Balschun (1997) found that the Hyper-Poisson distribution is appropriate in all 23 studied cases. The empirical distributions are very different. Find either a common distribution or a common function for both. There will be problems with Arabic data, many of which have two maxima.

Procedure

First collect the given data and substantiate linguistically the possible outcomes. Take into account that the Arabic texts consist of letters, the Old Hebrew ones represent psalms.

Then find a common function inductively using software. Do not apply functions with more than three parameters. Interpret one of the parameters as a realized boundary condition, e.g. text type or age.

If you found a common function, add texts from other Semitic languages and study the divergence. As far as possible, use modern texts but no religious ones. Study the divergence typologically, geographically, historically (if necessary), concerning text types, etc.

Always measure word length in terms of syllable numbers.

Perform a separate analysis measuring word length in terms of morpheme numbers. Take into account each case of intro-flexion, variation of consonants, etc. as separate morphemes.

If possible, take very old texts available in some languages and compare the development of Semitic languages from this point of view. Set up also a typology of Semitic languages based on word length.

Compare your results with those performed in other language families and position the Semitic ones.

References

Abbe, S. (2000). Word length distribution in Arabic letters. *Journal of Quantitative Linguistics 7(2), 121-127.*

Balschun, S. (1997). Wortlängenhäufigkeiten in althebräischen Texten. In: K.-H. Best (ed*.), Glottometrika 16. The Distribution of word and sentence length*: *174-179*. Trier: WVT.

3.8. Parts of speech in text types

Problem

Is the distribution of parts of speech (POS) equal in all text types? That is, do POS occur with the same relative frequency everywhere? It is not quite easy to believe that they occur with the same frequency in both fairy tales and in scientific articles? Test the hypothesis, find a model, and compare the frequencies.

Procedure

Take approximately equally long texts from the following text types: novel, short story, travel book, scientific text, journalistic text, fairy tale, letters and poetry. Compute the number of individual POS. If there is software for the automatic analysis, you can use it, but check the kind of part-of-speech tagging.

With the data in table form, perform the following operations:

(1) Compare the frequencies of equal POS in two text types. Use the chi-square test and compare all text types with each other. You obtain a table of chi-square values and the respective probabilities. Take note of the fact that the number of POS in all classes is greater than 5. If they are not equal (i.e. the probability of the chi square is smaller than 0.05) continue obtaining further data from these text types. If they really differ, show why it is so. Refer to the type of language.

(2) For each text type separately, order the frequencies in decreasing order and ascribe ranks to the frequencies. Then find a simple model of ranking. You may use Zipf's, Zipf-Mandelbrot's or Zipf-Alekseev's formula but you may also use the exponential function. Find a model which holds true for all your data.

(3) Perform – if possible – the same analysis in some other language. Finally, compare the results with the first language. If you analyse several languages, compare at least the parameters of the resulting functions.

For comparisons, you may use the data obtained for Czech by M. Kubát (2016: 137). Even if you do not know Czech, his table 8.2 is made very clearly and the words can be easily translated.

(4) If the rank order in one text type differs from that in another, try to interpret this phenomenon.

References

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

Best, K.-H. (1998). Zur Interaktion der Wortarten in Texten. *Papiere zur Linguistik 58 (1), 83-95.*

Hardie, A. (2007). Part-of-speech ratios in English corpora. *International Journal of Corpus Linguistics 12 (1), 55–81.*

Hudson, R. (1994). About 37% of all word-tokens are nouns. *Language 70, 331–339.*

Kubát, M. (2016). *Kvantitativní analýza žánrů.* Ostrava: Ostravská univerzita.

Mair, Ch., Hundt, M., Leech, G.N., Smith, N. (2003). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics 7 (2), 245–264.*

Mizutani, Sh. (1989). Ohno's lexical law: its data adjustment by linear regression. In: Sh. Mizutani (ed.), *Japanese Quantitative Linguistics: 1-13.* Bochum: Brockmeyer.

Rayson, P., Leech, G.N., Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics 2 (1), 133–152.*

Säily, T., Nevalainen, T., Siirtola, H. (2011): Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing 26 (2), 167–188.*

Seifart, F. (2011). Cross-linguistic variation in the noun-to-verb ratio: the role of verb morphology and narrative strategies. *Poster presented at the Association for Linguistic Typology 9th Biennial Conference, The University of Hong Kong, July 21-24, 2011.*

Seifart, F., Meyer, R., Zakharko, T., Bickel, B., Danielsen, S., Nordhoff, S., Witzlack-Makarevich, A. (2010). Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis. *Paper presented at the DobeS Workshop "Advances in Documentary Linguistics" Nijmegen, 14-15 October 2010.*

Thompson, P., Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics 12 (1), 1–23.*

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Analyse.* Trier: Wissenschaftlicher Verlag Trier (Quantitative Linguistics, 59).

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

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L, Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček: 295-312.* Trier: Wissenschaftlicher Verlag Trier.

3.9. Parts of speech – Modelling

Problem

State whether the ordered frequencies of parts of speech in any language are distributed according to the exponential function defined as $y = 1 + a*\exp^{(-bx)}$. If not, show a modified model.

Procedure

Take a text and order the individual words to the respective parts of speech. For some languages POS taggers are available. You can use them, but in any case manually check the results. Count the frequencies of individual classes and order them in decreasing order. The order should be simply ranked. Then apply the above formula in the form of a function (this time omit Zipf's power distribution, Mandelbrot's distribution, etc.). If the determination coefficient is $R^2 > 0.8$, you may accept the fitting.

Then take several texts of the same text type and perform the same analysis. Order the texts according to the parameter *b*. Then take another text type and do the same. Continue the analysis using various text types and the ordering of texts according to parameter *b*. Finally, draw a figure in which all parameters *b* are presented, for each text type with another sign (e.g. cross, empty circle, full circle, etc.).

You can consider the parameters *b* of individual texts as a variable and compare the text types. Or you can compare the parameters *b* of two texts using the normal test.

Do not forget that the ordering may differ both for two texts as well as for two text types. Ensure statistical correctness. If you compare two texts, then you must compare the frequencies of the same classes.

References

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

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: L. Hřebíček (ed.), *Glottometrika 11, 142-146.* Bochum: Brockmeyer.

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

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

3.10. Parts of speech development in American presidential speeches

Problem

J. Savoy (2017) studied various aspects of American presidential speeches taking into account ten presidents and published the percentages of individual parts of speech as they are presented in Table 1

Table 1
Percentage of various POS for some selected presidents

	Wash.	**Linc.**	**Wilson**	**Roos.**	**Eisen.**	**JFK**	**Reagan**	**Clinton**	**Obama**	**Trump**
noun	19.9	18.2	19.8	20.9	22.7	21.5	20.1	19.5	19.6	18.7
name	3.0	4.1	1.8	3.0	3.6	3.4	3.9	3.9	3.5	5.8
pron.	5.7	4.8	7.8	6.7	5.5	6.5	8.2	9.5	9.1	8.9
adj.	7.4	7.8	7.9	8.5	9.4	8.4	7.5	7.0	6.5	6.6
verb	14.9	14.6	15.0	13.8	13.3	12.8	14.9	15.4	16.5	15.1
adverb	3.8	5.0	4.7	4.1	3.8	4.2	4.8	4.6	5.2	4.7
det.	12.9	12.3	11.0	11.0	10.3	10.1	9.0	8.9	8.8	8.1
prep.	19.3	17.6	17.5	16.7	15.7	14.7	14.0	14.0	13.3	12.5
coor.	3.5	4.0	4.9	4.1	3.7	4.7	4.1	3.8	4.1	4.5
other	9.5	11.5	9.5	11.3	12.0	13.7	13.6	13.5	13.3	15.1

Study the evolution of individual POS and if you obtain for some parts of speech a not strongly oscillating sequence, fit a function to the given course.

Procedure

First look at the numbers in each line separately. You may put them into Excel and make a graphical representation of each line. If you find a rather smooth curve, find a function expressing it. Consider only those fittings which have a determination coefficient of at least 0.8. Avoid polynomials. If you have found a "good" function, state the proportion of the given POS with other presidents – not mentioned here. If the relation remains as it is, take other texts, separate text types and make a statement about the evolution of the given POS in the given American text type.

If the problem yields some "good" results, consider other languages. Presidential speeches are frequently the object of investigation (cf. Tuzzi, Popescu, Altmann 2010, Čech 2011, 2014) but you can take any texts for the study of this problem.

References

Čech, R. (2011). Frequency structure of New Year's presidential speeches in Czech. The authorship analysis. In: Kelih, E. et al. (eds.), *Issues in Quantitative Linguistics 2, 82-94.* Lüdenscheid: RAM-Verlag.

Čech, R. (2014). Language and ideology. Quantitative semantic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949-2011). *Quality and Quantity 48(2), 899-910.*

Li, W., Miramontes, P. (2011). Fitting ranked English and Spanish Letter Frequency. Distribution in US and Mexican presidental speeches. *Journal of Quantitative Linguistics 18 (4), 359–380.*

Overbeck, A., Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). Analysis of Italian word classes. *Glottometrics (20), 12–28*.

Savoy, J. (2010). Lexical analysis of US political speeches. *Journal of Quantitative Linguistics 17 (2), 123–141.*

Savoy, J. (2016). Vocabulary growth study. An example with the State of the Union Addresses. *Journal of Quantitative Linguistics 22 (4), 289–310*.

Savoy, J. (2017). Analysis of the style and the rhetoric of the American presidents over two centuries. *Glottometrics 38, 55-76.*

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

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

3.11. Adjectives: Semantics

Problem

Adjectives can be ordered in many classes, e.g. Yesypenko (2009) defined 18 for English. Now each class can be subdivided into more determined classes ordered according to some principle, e.g. *less – more.* If you can perform this operation, quantify the ordering and find the distribution of this order in some texts.

Procedure

First take a classification of adjectives from an official grammar or use the classification proposed by Yesypenko (2009). In English one can find for example the classes of adjectives of quantity, quality, size, shape, age, colour, purpose, origin, material, nationality, etc. Then subdivide the adjectives in each class according to some principle in such a way that they are scalable. Take into account also the grammatical way of scaling, e.g. *nicer* is "more" than *nice*; *beautiful* is more than *nicer*, etc. If you have a ready scale, take a text and set up a sequence containing only the degrees of adjectives found.

Find the empirical distribution of degrees, then derive a theoretical distribution or simply find a function capturing the empirical distribution. Compute various properties of the empirical distribution, e.g. mean, variance, excess, Ord's criterion, Gini's coefficient, entropy, repeat rate, etc.

Compare several texts of the same text type and order them according to some of the above-mentioned indicators. Then compare texts of different text types. Is it possible to show the differences between text types using this kind of scaling? Compute confidence intervals for each text type and show the intersections.

Take the same text in two languages. Perform an analogical scaling in the second language, perform all computations and show the differences between the two languages. What can be ascribed to the translator and what is a property of language?

References

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: Černiveckij nacional'nij universitet.

3.12. Adjectives: Formal aspects

Problem

The form of adjectives depends on the type of language. In some of them they are not marked with an adjectival morpheme (e.g. Polynesian languages, Hungarian), in other ones they may basically have an additional morpheme representing some category, e.g. gender in Slavic languages. Some of them do not have a base from which they are derived, other ones may be called denominal (e.g. *Flucht* > *flüchtig*), deverbal (e.g. *gehen* > *vergangener*), etc. Some of them contain a derivational morpheme, other ones may be compound, some are participial forms of verbs, etc. Study the trend in the given language, set up frequency distributions, rank them and find a model.

Procedure

Take a dictionary of a language and analyse one letter. Write all adjectives down and observe their behaviour and form. Set up a classification of adjectives. An adjective may belong to several classes simultaneously. If it is in the same classification then construct a new class of adjectives belonging to class A and B, to A and C and D, etc. Perform the same classification only with compound adjectives.

After you have obtained at least one classification, rank the classes according to their frequency content and for each of the classifications find a model.

Usually one can apply the Zipf-power function, or the Zipf-Mandelbrot function or the Zipf-Alekseev function. Striving for simplicity and unity apply also the exponential function.

If you obtain positive results with one initial letter, perform the same operation with adjectives beginning with another letter. Perform the operations for at least five initial letters.

Now do the same in another language and compare the results.

Perform the same operations with a longer text. In texts, the position of the adjective may also play a role. In some languages they always stay in front of the noun, in other ones they stay behind the noun, and there are texts in which the position plays a textual role – for example in Slovak the position of an adjective behind the noun creates a poetic colouring.

Analyse two texts and compute the proportion of adjectives. The length of the text consists of all word tokens. Is the difference in proportion significant? That is, perform a significance test for the difference of two proportions. You may manually analyse several texts – in corpuses distinguish all texts; do not take the corpus as a whole. Then strive for a classification of texts based on adjectives.

References

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: Černiveckij nacional'nij universitet.

3.13. Adjectives: Composition

Problem

In some languages, adjectives have the same form as other parts of speech. In other ones, e.g. (Indo)-European languages, the adjective may have an identification affix, e.g. in German there are several ones: *-ig, -end, -ern*, but there are many without the above adjectival identification affix, e.g. *schön, brav, kalt.* The identification affix may be weighted or not. At the same time, an adjective may be a compound and, as such, it may also have affixes. The problem is how to measure the complexity of an adjective. Further, if complexity is fixed, one should find a model capturing the frequencies of adjectives with the given weight. Find a method for weighting the complexity of adjectives.

Procedure

In the first step take a dictionary and write out all adjectives beginning with the same letter. Unfortunately, if they have a prefix, in many dictionaries they are placed under the beginning letter of the prefix. Hence one should rather collect 100 adjectives in the basic form and then search for all adjectives that have the same base.

It may be conjectured that simple adjectives are more frequent than complex ones, hence one obtains a decreasing distribution of weights. For example in the Slovak poem *Kykymora* by A. Sládkovič one finds 61 simple adjectives, 13 affixal ones and two compound ones. Find a model for this phenomenon. Use simple functions, in no case polynomials. Test the model on data in your language. Then perform two further operations:

(1) Take another language, state the weights and test the model. Interpret the differences.

(2) Take a longer text and analyse all adjectives. Repeated occurrence of adjectives must be counted. Test the same model. If it is sufficient, analyse various text types and compare them. Set up a typology of texts based on adjective types.

In the dictionaries of some languages, one finds all basic forms, and all other forms formed by affixation or composition are presented under the given basic form. This is usually the case in weekly synthetic languages. In strongly syn-

thetic languages one must take a basis and form all adjectives that are possible. Unfortunately, compounds are again written separately.

References

None

3.14. Compound formation in English

Problem

H. Gnatchuk (2015) studied compound forming in English in terms of parts of speech using "The New York Times" (Monday, 2 February 2015). She found 28 types of compounds, the most frequent being the structure *Noun+Noun* (one of them is named twice, No. 12 should be *Adj+Noun+Noun*). She ranked the frequencies of individual types and obtained the table presented below. In order to capture the rank order, she fitted the power function in form $y = 1 + ax^b$ and obtained an excellent result ($R^2 = 0.9952$). Having in mind this excellent result a number of further hypotheses can be formulated: (1) Is it possible to capture the ranking by applying other formula? (2) What is the development of compound type building like in English in general? (3) What are the domains in which a specific compound type occurs, i.e. how could one define the environment of compounds? (4) Can one reorganize the types of compounds and define them in a different way? For example to introduce a semantic criterion? (5) Can one scale them according to a property which must be specially defined?

Find methods for solving all of these problems.

Procedure

Here we shall present only Gnatchuk's table and the reader should try to find several functions expressing the rank order. The abbreviations are: N = noun, V = verb, Adj = adjective, Ph = phrase, Pr = preposition, Nu = numeral. P2 = participle 2, Pn = pronoun

Table
Types of English compounds (Gnatchuk 2015)

Rank	Pattern	Number	Rank	Pattern	Number
1	N+N	64	15	Pr+Pr	2
2	N+N+N	12	16	N+V(ing)	2

3	Adj+N	7	17	Nu+N+N+N+N	1
4	N+N+N+N	4	18	N+Adj+N	1
5	Ph	4	19	N+P2	1
6	Adj+V(ing)	3	20	Adj+P2	1
7	Pr+N	3	21	N+V+Pr	1
8	Nu+N+N	3	22	Pr+Pn	1
9	Adj+Pr+V	3	23	Nu+N	1
10	Adj+N+N+N	2	24	Pr+V	1
11	Pr+V+V	2	25	Pr+P2	1
12	Adj+N+N	2	26	N+Adj+N+N	1
13	V+Pr	2	27	N+Pr+V	1
14	Nu+Nu	2	28	N+Adj	1

For each function you use add 1 because there are no smaller frequencies; otherwise any function would converge to zero. Set up a table of possible functions (avoid polynomials) and obtain further data. Test your best function again and again. Then derive it from a differential equation and substantiate its constants/parameters.

After you have obtained a satisfactory result for several collected data, begin to solve the problems (2) to (5). Compare English with other languages and based on the results obtained construct a kind of theory.

References

Gnatchuk, H. (2014). A statistical analysis of English compounds in the newspaper style. *Mathematical Linguistics 1(1), 81-90.*

3.15. Denominal verbs in German

Problem

In many languages there are verbs derived from nouns by means of affixes. In German, one uses prefixes (while the suffix *-en* must stay everywhere).

Descriptions can be found for example in Kaliuščenko (1988). There are books for every language describing this morphological process. Besides, some of them name also the class of nouns from which it may be derived, so that one can also speak about motivation. Study the ranked distribution of the respective prefixes in German, set up hypotheses and test them.

Procedure

Consider the distributions that can also be found in U. Rothe (1990). Rothe fitted the modified Zipf-Alekseev distribution with good results. Show that a simpler fitting is, perhaps, possible using the exponential function defined as $y = a * exp(-b*x) + 1$. Table 1 contains all data. The particular motives of derivation (all together 29) will not be presented here.

Table 1
Ranked frequencies of denominal verbs in German

							Middle High German		
Rank	*ab-*	*aus-*	*be-*	*ein-*	*ent-*	*ver-*	*be-*	*ent-*	*ver-*
1	16	24	86	18	71	42	36	46	14
2	7	10	13	4	12	40	4	5	10
3	3	10	10	4	3	32	3	3	4
4	3	9	8	3	3	17	3	2	3
5	2	5	6	2	3	9	2	1	3
6	1	4	3	2	1	7	2		2
7	1	4	2	2		4	2		2
8	1	3	2	1		4	2		2
9	1	3	2	1		4	2		2
10	1	2	1			3	1		1
11	1	2	1			2	1		1
12		1	1			1	1		1
13			1			1	1		1
14			1						1
15			1						1
16			1						

If the exponential function does not hold, find a simple function that holds for all data and does not have more than two parameters.

Consider also the data brought by K.-H. Best (1990) using the mixed negative binomial distribution.

References

Best, K.-H. (1990). Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen. In: Hřebíček, L. (ed.), *Glottometrika 11: 107-110*. Bochum: Brockmeyer.

Kaliuščenko, V.D. (1988). *Deutsche denominale Verben*. Tübingen: Narr

Rothe, U, (ed.) (1989). *Diversification processes in language: Grammar.* Bochum: Brockmeyer.

Rothe, U. (1990). Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina. In: Hřebíček, L. (ed.), *Glottometrika 11: 111-121*. Bochum: Brockmeyer.

3.16. Nominal affixes in a language

Problem

Study the frequency of all nominal affixes in a language and prepare a ranked list. Consider only those affixes that make a word a noun. You may distinguish pre-, in- and suffixes. Strive to find a theoretical function expressing the ranking. Collect data and test the hypothesis.

Procedure

Take a text in a given language and find all nouns. Then eliminate those which do not contain an affix making the stem to a noun. The affixes may be found in the respective linguistic literature. The rest should be ranked according to the frequency of individual affixes. The ranking shows the inclination of affixes to build new words. That means you count the frequency of individual affixes.

Find a function expressing the trend of ranking. You may begin with the usual Zipfian function $y = ax^b$; if it does not give satisfactory results, continue with the Zipf-Mandelbrot function, generalize to Zipf-Alekseev or apply simply the exponential function given as $y = 1 + ae^{-bx}$.

First continue analysing different texts and set up the ranking for each text separately. Then compare the individual text types. Do poetic texts use the building of nouns differently than for example scientific texts? Do not compare individual affixes but the complete rank-frequency sequence.

In the next step take a specific text type and study its history. Does the forming of nouns by means of affixes change over the course of years? You may consider for example journalistic texts and compare them with poetic texts. You may compare the evolution of a specific writer.

If possible, take another language and perform the same analyses. Compare the resulting numbers of the two languages. Strive for a unique formula for all. Strive for the simplification of the formula; derive it from a differential equation and strive for a theory.

References

Nemcová, E. (2009). Nominal suffixes in German press texts. *Glottometrics 18, 70-77.*

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

3.17. Consensus strings 1

Problem

Set up the so-called consensus string of a text defined over parts of speech and compute its weight.

Procedure

Classify the parts of speech according to an official grammar. Then take a text and transcribe it in terms of parts of speech. Consider each sentence separately. You have, say, nine different classes. For each sentence you obtain a vector whose elements are parts of speech in the sentence. Transcribing the complete text, state which part of speech occurs in the first position most frequently, then in the second, third … positions. If the sentences are not equally long, fill the last places with zeroes. You obtain a vector representing the consensus string of the text.

Now divide in each position the frequency of the most frequent part of speech found by the number of all sentences. Now you obtain the weight of the consensus string for the given text in the form of proportions. Study this vector: do the numbers represent a horizontal line (with small deviations) or does it develop in some way? If it displays some trend, capture the trend by a formula. Proceed inductively: apply software to obtain a good result; choose a function with a small number of parameters. Then continue deductively: define the dif-

ferential equation of the formula and find its place in the unified theory. Interpret the individual parameters linguistically.

References

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

Wimmer, G., Altmann, G. (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.

Zörnig, P., Stachowski, K., Popescu, I.-I., Mosavi Miangah, T., Chen, R., Altmann, G. (2016). *Positional Occurrences in Texts: Weighted Consensus Strings*. Lüdenscheid: RAM-Verlag.

3.18. Consensus strings 2

Problem

Are the weighted consensus strings different on different language levels?

Procedure

First solve the problem *Consensus strings 1* for parts of speech. Then begin to set up for each sentence strings of other entities. Begin with phonology and define some units on all relevant levels. You may use symbolic or numerical sequences, you may consider any properties of classical units or modern motifs.

Having analysed several texts belonging to a specific text sort or language variety, study the behaviour of the weighted consensus strings: scrutinize their course in the sentence. Weighting means the proportion of the most frequent entity at the given position divided by the number of strings. Mostly you obtain a decreasing sequence. Derive a formula capturing all empirical forms and interpret the parameters in relation to the text sort, history of language, development of speech of children, mental disorders, style, etc. For some kinds of strings you may also use translated texts.

Define explicitly the boundaries of the sentence, i.e. the syntactic unit analysed. It does not cause greater problems in written texts, but for oral communication (e.g. telephone conversation, etc.) one has to find appropriate delimiting signals. One can rely on intonation or pauses even if the sentences are not complete. If you analyse written poetry, you may consider verses or

sentences. For rhythm, verses/lines are relevant, but for higher language levels you must define exactly the boundaries of the frame string.

If some data do not follow the ascertained formula, you must modify it adding some boundary conditions which must be substantiated linguistically.

Strive for a theory, i.e. set up hypotheses, derive them from background knowledge, test it and interpret the results.

References

Cf. the references for *Consensus strings 1*.

3.19. Consensus strings 3

Problem

State the polysemy of individual words in a sonnet. Compute the mean polysemy of each verse/line separately and state whether there is some non-linearity. Then compute the consensus string of polysemy and observe whether there is a trend in it.

Procedure

Choose a sonnet in your language. Consider each word separately and state its polysemy using a monolingual dictionary. Describe the way you counted polysemy. For each verse separately set up a vector of polysemies. Compute for each verse the mean polysemy. In this way you obtain a sequence of means. Now, formulate a hypothesis which would capture the course of this sequence. Even if it is a horizontal line, substantiate it linguistically. If it is not linear and horizontal, substantiate its course linguistically, then find a function expressing it and test it.

Perform the same operations for another sonnet – it need not be the same author or even the same language. Compare the results, test the differences between them. Continue with other sonnets or take another type of poem

Now, in the first sonnet you have for each verse a vector whose elements are numbers. Define the consensus vector in such a way that you take the most frequent number (polysemy) in each position. Having the consensus vector, divide each number by 14. Study the resulting curve, formulate a hypothesis, test it and interpret. Do it for all texts you analysed.

If there is some commonality in the curves you obtained, generalize your hypothesis. Derive the function from a differential equation and interpret the differential equation.

Analyse other poetic texts belonging to the same type and set up, stepwise, a theory of polysemy sequences in poetry.

The problem is not easy but for many languages there are monolingual dictionaries in which one can quickly find the numbers of meaning for each word.

References

Cf. the references for *Consensus strings 1*.

3.20. Sequential approach

Problem

Study all forms of occurrence of the English preposition “at”. Use only texts and compute: (1) The Belza chains of its occurrence, i.e. the number of subsequent sentences in which it occurs. This sequence of sentences yields the length of chains. (2) Frumkina 100-word passages and the occurrence of “at” in them. Then differentiate the meaning of “at” and count the sentences in which it has the same meaning.

Procedure

Take a longer text. In order to obtain Belza chains, one has to determine the end of sentences/phrases analysed. Preferably one has to develop a software program which is able to retrieve the required information for this analysis automatically. Then it should count the number of those subsequent sentences in which “at” occurs in order to obtain Belza chains. You obtain a distribution of lengths. Prepare a table and fit to the numbers a probability distribution. You need not differentiate between discrete and continuous distributions; moreover, you can simply take a (non-normalized) function. First, try to fit the exponential function, then the power function, then the Zipf-Mandelbrot function and finally the Zipf-Alekseev function. Most probably you will consider the results attained by the exponential function as satisfactory. If you were successful, perform the same operation with all prepositions individually. Then generalize the results and choose the simplest fitting function.

Now tell the program to look at the first 100 words of the text and count the number of “at” in it. Then continue with the next 100, etc. In each passage you find a certain number of “at”; here even zero is possible. Having the result, count the passages containing 0, 1, 2 … times the preposition “at”. Set up a distribution and find a well-fitting function, as above. Then do the same with the

other English prepositions, find a common function and step by step form a theory.

Derive the resulting functions from a differential equation (cf. Wimmer-Altmann's unified theory 2005).

If possible, take further texts and do the same. Search for corroboration of your conjectures.

References

Altmann, G., Burdinski, V. (1982). Towards a law of word repetitions in text-blocks. In: W. Lehfeldt, U. Strauss (eds.), *Glottometrika 4: 147-167.* Bochum: Brockmeyer.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 94-109.*

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

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

3.21. Structural centrality: Parts of speech

Problem

Study the structural centrality of a text using Zörnig's method applied to sentence and parts of speech. Evaluate 10 texts belonging to the same text type, show the central structure of each text and compare the texts.

Procedure

Take a text and partition it in sentences. Decide which mark signals sentence end (semicolon, colon, exclamation mark, question mark, full stop) and transcribe the text in terms of abbreviations of parts of speech. Set up the vector of abbreviations representing the given sentence. Take the longest sentence and complete the empty places in the other vectors by zeroes.

Now place the vectors in a column and for each position find the most frequent symbol (in the given column). The sequence of the most frequent symbols is the central sentence structure, the consensus string, in the given text.

In order to express the weight of the common vector, divide the number of most frequent signs in a column by all signs in the column (i.e. number of sentences). Study the course of the numbers. Do they increase or decrease from the beginning to the end? Explain the fact found.

In stage plays, consider the speech of each person separately. Can you characterize the persons using the weighted vector of parts of speech?

References

Zörnig, P., Stachowski, K., Popescu, I.-I., Mosavi Miangah, T., Chen, R., Altmann, G. (2016). *Positional Occurrences in Texts: Weighted Consensus Strings*. Lüdenscheid: RAM-Verlag.

4. Syntax and Syntactical Functions

4.1. Frequency of noun phrase patterns

Problem

Wang Hua (2012) studied the relationship between the length of noun phrase patterns and the number of different patterns of the given length in English. He stated that the number of patterns depends on the length of the pattern measured in terms of word numbers. Fitting $y = ax^be^{cx}$ (y = number of different patterns of the given length, x = length of a pattern) yielded a satisfactory result. Check the results in another language.

Procedure

Consider your L1-language. If you use a corpus, process every text separately and state whether the above function is sufficient. If not, use the function y = a*exp(b + c ln x), i.e. the Zipf-Alekseev function generally used for length problems (cf. Popescu, Best, Altmann 2014). If you obtain satisfactory results, order the texts according to the parameters and strive for a possible explication.

In the second step combine all texts belonging to a specific text type and perform the fitting anew. Can you perform an ordering of text types according to one of the parameters? How do text types differ from one another?

In the third step join all texts and consider all as representatives of the given language. Fit again the above function following Wang Hua's procedure.

Study the mutual relationship between the exponential parameters of the above-mentioned functions. If you find some regularity, interpret it linguistically. The easiest way is to find the differential equation of the above functions and interpret the individual components of the right hand side linguistically.

Can you find some difference between English and your language from this point of view? Consider not only the frequencies but also the simple existence of noun phrase patterns. Could you find in your language all noun phrase patterns occurring in English?

Compare the frequencies of individual patterns with those in English by applying a simple statistical test.

References

Hua, W. (2012). Length and complexity of NPs in Written English. *Glottometrics 24, 79-88.*

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

4.2. Attributes

Problem

Boshtan and Best (2010) found in German the following types of attributes:

Adjective attribute (f)
Compositional attribute (f)
Participial attribute (f)
Apposition (b)
Attributive sentence (b)
Genitive attribute (b)
Prepositional attribute (b)
Genitive attribute (f)
Attributive infinitive with "zu"

where *f* = staying in front of, *b* = staying behind. German examples can be found in Boshtan, Best (2010). Define the attributes in your language, analyse ten texts belonging to the same text type and evaluate the results.

Procedure

First define the possible attributes in your language. The results will surely differ from those in German. Adhere to the official grammar. Then take ten longer journalistic texts and count the frequency of individual attribute types.

(1) Set up the rank-frequency distribution and state whether it abides by the power function. If not, search for another function.

(2) State whether the individual classes tend to a certain length, i.e. compute the length of each attribute (in terms of word numbers) and compute for each class its mean. Compare the results in the ten texts. You may use various tests for comparison.

(3) Find the theoretical distribution of length of the attributes occurring in a text. Substantiate the model linguistically. You may use a distribution or a sequence, discrete or continuous.

(4) For each text set up the ranks of the classes (taking into account the frequencies), construct the rank table and perform Kendall's W-test for comparing the ranks.

In the next step take each class of attributes individually and classify the attributes you found (in the given class) into new, more specific classes, e.g.

types of compositional attributes. Then perform the same procedures as above. If the texts you used were short, you can consider all texts as one sample and compare the results with text of another text type.

Perform a scaling of the original classes according to some property like simplicity-complexity, length, concreteness-abstractness, generality-specificity, kinds of adjectives in them, etc. Compare the texts on the basis of this scaling. Unfortunately, you must propose, quantify and measure a property which is, perhaps, not yet current in linguistics.

Study the development of a writer on the basis of attributes. Study the development of child language.

Compare journalistic texts with other text types.

Compare journalistic texts in one language with similar texts in some other language.

Study the development of attributes in journalistic texts. Here, the years are clearly given.

Show that the above properties are linked with other ones and construct a partial Köhlerian control cycle.

Take a specific text and its translations into various languages. State the attributes in the original and its translations. Compare all results in all reasonable ways. Order the languages from various standpoints. This task is very complex and should be processed as the last one. As soon as you stay at this level, you have entered the territory of theory. You may be sure that not all languages have the same kinds of attributes. Hence, you can also study the forms of translating them from one language into another.

References

Best, K.-H., Altmann, G. (2015). On the frequency of simple attributes in German. *Mathematical Linguistics 1(1), 1-8.*

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

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.

4.3. Verb valency motifs

Problem

Define verb valency motifs both qualitatively and quantitatively and study the properties of the two sequences. Capture them by well-substantiated formulas and compare texts.

Procedure

Take a text and notify qualitatively the sequence of verb valencies for each sentence or verse/line separately. Symbolize the parts of speech belonging to the verb in the usual way, e.g. for the sentence "*Today, I saw in the cinema a good film*" we obtain (Adverb; Pronoun; Place; Object) as the valencies of the verb *saw*, and the numerical valency is [4].

Prepare a matrix of transition frequencies from one valency to the next. Each sentence must be considered separately. After having the matrix of transitions, evaluate all of its properties you know: study the monotony, symmetry of cells, the symmetry of the whole table, the behaviour of the diagonal, Markov properties, distribution of the marginal cells (containing the sums of rows or columns), etc.

Now, for each sentence you have a number representing its numerical valency. Write the whole text as a sequence and evaluate all of its properties known from various linguistic investigations. Study for example the distribution and its properties, compute the Hurst coefficient, study the distances between equal numbers, construct Köhlerian motifs and evaluate them, etc.

Analyse several texts and compare them using all your results. Since this work is enormous, restrict the investigation rather to some selected properties and perform comparisons based on statistical tests. Draw consequences from the results, e.g. concerning the author, his evolution, text type, language, etc.

If possible, begin to theorize: derive the formulas, interpret the parameters, answer intuitively some "why"-questions, join the results into a control cycle, present the valency "theory" in quantitative terms.

References

Allerton, D. J. (2005). Valency grammar. In: K. Brown (ed.), *The Encyclopedia of Language and Linguistics*: Amsterdam: Elsevier, 4878–4886.

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.

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

Köhler, R. (2006). The frequency distribution of the length of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Viktor Krupa: 145-152.* Bratislava: Slovak Academy Press.

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 Gruyter.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, Ch., Burkhardt, H., Schmidt-Thieme, L., Becker, R. (eds.), *Data Analysis. Machine Learning and Applications: 635-646.* Berlin/ Heidelberg: 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.

Liu, Haitao (2011). Quantitative properties of English verb valency. *Journal of Quantitative Linguistics 18 (3), S. 207–233.*

Mačutek, J. (2009). Motif richness. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 51-60.* Lüdenscheid: RAM-Verlag.

Milička, J. (2015). Is the distribution of L-motifs inherited from the word length distribution? In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 133-145.* Berlin/Boston: de Gruyter.

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: 183-194.* Wien: Praesens.

Vincze, V. (2014). Valency frames in a Hungarian corpus. *Journal of Quantitative Linguistics 21(2), 153–176.*

4.4. Order of adjectives

Problem

If in the pre-nominal position there is more than one adjective, then in every language there is a preferred order, e.g. value > size > dimension > various physical properties > colour, as can be read in psycholinguistic and neurolinguistic works. Since adjectives can be classified in various ways, the task of linguists is testing the given hypothesis using texts.

Procedure

First choose a semantic classification of adjectives as shown in *Adjectives: Formal aspects*. Then take a longer text and retrieve all immediate (!) sequences of two or more adjectives in any position in the sentence. Write the sequences in a table whose second column represents the sequence of classes. Then add the equal sequences, e.g. state how many times the sequence *value – size* occurs.

Take into account all sequences and all combinations of two, three, etc. adjectives. Then compare the numbers of individual pairs, e.g. *value-size* against *size – value* and test the difference. You may use the simple chi-square test. If you find sequences of three adjectives, then there are 3! = 6 possibilities. All must be tested.

Now take several texts belonging to the same text type, study the situation and compare the text types as wholes. Here, you may use also text-type corpuses.

Study the problem historically, e.g. compare the sequences in Latin and in Italian or French. Did something change? You may also use translations and compare the sequences in these texts.

Finally, study the work of one author and observe the years of publication of his works. Did something change in his placing of adjectives?

References

Bache, C. (1978). *The order of premodifying adjectives in present-day English.* Odense: Odense University Press.

Deese, J. (1964). The associative structure of some English adjectives. *Journal of Verbal Learning and Verbal Behavior 3, 347–357.*

Hetzron, R. (1978). On the relative order of adjectives. In: Seiler H.J., (ed.), *Language universals:* 165–184. Tübingen: Narr.

Kemmerer, D., Tranel, D., Zdanczyk, C. (2009). Knowledge of the semantic constraints on adjective order can be selectively impaired. *Journal of Neurolinguistics 22(1), 91–108.*

Kemmerer, D., Weber-Fox, C., Price, K., Zdanczyk, C., Way, H. (2007). *Big brown dog* or *brown big dog?* An electrophysiological study of semantic constraints on prenominal adjective order. *Brain and Language 100, 238–256.*

Martin, J. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior 8, 697–704.*

Richards, M.M. (1975). The pragmatic rule of adjective ordering: A critique. *American Journal of Psychology 88, 201–215.*

Wulff, Stefanie (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of American Linguistics* 8, 2, *245–282*.

4.5. Referential adjectives

Problem

L. Uhlířová (2007: 659) states: “In various languages, like Czech, Russian, English, German and many others, there is a relatively small, but open class of

adjectives which can be used as lexical means of anaphora in text structure. Below, we shall call such adjectives referential adjectives."

Such adjectives include *quoted, discussed, mentioned, outlined, suggested, described*, etc. but in Czech they have the form of an adjective (*citovaný, diskutovaný, jmenovaný*, …) mostly derived from a verb. Uhlířová differentiates between referential use and other uses and computes the frequency in a corpus. She applies to the ranked frequencies the Zipf-Alekseev distribution. Do the same in another language.

Procedure

If you use a corpus then it should be specialized to a certain text type, e.g. journalistic texts. Consider the complete corpus and find all adjectives used referentially at the given place. Set up a ranked sequence and find a model. Fit the model to the data. You may use either a function or a distribution in the role of a model.

Now perform the same operation in long individual texts and characterize the texts using an indicator. Order the texts using this indicator. The indicator must not depend on the total number of referential adjectives but can be defined as for example entropy, repeat rate, lambda indicator, parameters of the applied function, etc.

Classify the texts into text types, take the means of the given indicators in the text type and order the text types according to adjectival referentiality.

This is, of course, not the complete referentiality. In order to obtain a more complete picture, count all referential entities, e.g. pronouns, verbs containing a personal affix, etc. Prepare a complete picture of referentiality for each text, or better for each text type, and order them.

Name all referential elements like words, phrases, even clauses explicitly (i.e. make a list) and present the referential picture of a text or language.

If you use a function, use for fitting either a manual program like TableCurve which computes thousands of functions, or if you know which function is the best, use for example NLREG. Strive for a theory of referentiality, that is, find also some other properties which are related to referentiality and derive the function applied from a differential equation or a stochastic process.

References

Uhlířová, L. (2007). Using Altmann-Fitter for text analysis: An example from Czech. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 659-664*. Berlin/New York: Mouton de Gruyter.

4.6. Adnominal modifiers: Construction of hierarchies

Problem

In *Problems Vol. 4* (2014: 26-35) many aspects of adnominal modifiers have been described. As is well known, if one entity constitutes classes, then the members of the class may be further classified, i.e. there is a property which may have manifold forms, degrees, subclasses, etc. Construct at least one hierarchically lower level and evaluate its properties.

Procedure

From the problem *1.19. Adnominal modifiers – 0. Classification* in *Vol. 4*, take for example the category *5. Apposition*. Then take a longer text and write out all appositions to nouns. Set up a classification of these appositions with as many classes as necessary. If possible, do it in two languages in order to obtain a deeper insight.

The individual classes have their own properties, e.g. lengths, complexity, parts of speech, etc. But in the first step, do not worry about a still deeper level. Count the frequency of individual subclasses and express the rank-frequency dependence by a formula. It may be a known formula or you may propose it yourself. It may be a probability distribution or a function.

Test the adequacy of the formula. If it is corroborated, take one of the subclasses you created and find a new classification of its members. Repeat again the modelling and the testing step. Adhere to the same formula if possible.

After you have described some levels of the hierarchy and obtained positive results, perform the same operation with other members of the apposition class. Then scrutinize the other classes mentioned in *Problems Vol. 4.* Continue until you obtain at least one step in the hierarchy for all classes.

Generalize your procedure and your results, show its consequences and, if possible, link the results with other properties of language in order to obtain a control cycle.

Later on, extend your search for hierarchies to other linguistic or textual entities. Strive for a hierarchy theory.

References

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

4.7. Adnominals and the sentence

Problem

Adnominals are those parts of the sentence which serve the specification of some noun in the sentence. There are many types and the more of them occur in a text, the better it is specified. From the logical point of view, one can consider them as predicates and as such they can also be scaled. Study the number of adnominals in a sentence, find their distribution and set up a model.

Procedure

First consider the adnominals which can be found in the literature. Then take a text and in each sentence underwrite the adnominals. The number of adnominals in a sentence represents the predicative length of the sentence. Now, set up the distribution of the predicative sentence length. You obtain a short distribution.

The distribution is surely monotone decreasing. Find a function or distribution capturing this sequence with at least $R^2 > 0.8$ or an appropriate chi-square for the given degrees of freedom. Consider the parameters of the function as characteristic features of the text. If one of the parameters represent the text size, ignore it.

Analyse the problem specified to various text types: analyse poems, journalistic texts, scientific texts, prose of one author, etc. Set up a preliminary typology of text types. If possible, perform the same analysis of texts in another language.

References

Andreev, S., Lupea, M., Altmann, G. (2017). Belza chains of adnominals. *Glottometrics 39, 72-86.*

Andreev, S., Popescu, I.-I., Altmann, G. (2016). On Russian adnominals. *Glottometrics 35, 64-84,*

Andreev, S., Popescu, I.-I., Altmann, G. (2017). Motifs of adnominals in Russian. *Glottometrics 38, 77-106.*

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

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

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

4.8. Development of adnominals

Problem

Adnominal modifiers are determining parts of nouns in sentence. Their classification and various properties (types, weight, distribution, complexity, cohesion, scaling and motifs) have been discussed in Köhler, Altmann (2014), where one can also find the relevant literature. Newer publications are devoted for example to the study of Russian (Andreev, Popescu, Altmann 2017). Though adnominal modifiers are necessary parts of sentence, one may ask two questions: (1) Do individual text types differ in the use of certain types? Automatically, one can ask the same question concerning languages and contribute to typology, e.g. comparing a text and its translations into other languages. (2) Can one discover some development in the use of adnominal types in time? In each text there are specific proportions of individual adnominal types; if one compares these proportions with respect to the time of creation of texts, one can, perhaps, discover some tendencies. Analyse this problem.

Procedure

Take some texts in your language and note the exact date of their creation. You can select them for example in five-year intervals but in any case ensure the same text type. Compute the percentage of individual adnominal types in texts of the same year. For each adnominal type you obtain a sequence of percentages. Study these sequences and state whether there is some clear development (monotone increase or decrease or a bell form).

Now, describe them using some formulas and find the causes, i.e. the relationship of these tendencies to the development of other ones. That means, set up a Köhlerian control circuit in which the adnominals play the role of one of the vertices of the graph.

If possible, compare this development with that in some other languages. One could discover both areal and genetic similarities but in any case one would show the role of adnominals in linguistic synergetics.

References

Andreev, S., Popescu, I.-I., Altmann, G. (2017). Some properties of adnominals in Russian texts. *Glottometrics 38, 77-106.*

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

4.9. Clause types

Hypothesis

We can assume that the position of specific clauses within a sentence depends on

1. the type (function) of the given clause;
2. the language or cultural background in which the given text is formed;
3. the relative length of the clause with respect to the lengths of the other clauses (Behaghel's law; early immediate constituent hypothesis Hawkins 1994).

Some clause types may be positioned preferably preceding its matrix clause, others may frequently follow the matrix clause. Test the hypothesis of preferences and find rank-frequency models.

Procedure

Set up a list of clause types such as causal, consecutive, final, concessive, adversative, etc. Count the number of clauses in initial and final positions separately for each clause type, language and text type. Counting can only partly be performed automatically: conjunctions may help to identify clause types in languages which employ conjunctions. Their use is limited because at least some of them will be ambiguous. First state the proportions of a type in the two positions and state whether there are preferences in the positions. You may use the binomial or the asymptotic normal test.

Calculate the mean (relative) clause length for each of these groups in both positions and determine whether the differences of positional means are significant by applying the asymptotic normal test.

Then state the frequencies of individual clause types in both positions separately, rank the frequencies and find a theoretical model for the rank frequencies.

References

Behaghel, O. (1930). Von deutscher Wortstellung. *Zeitschrift für Deutschkunde, 4, 81–89.*

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

4.10. Clause type motifs

Problem

Form categorical motifs from sequences of clauses (cf. *Clause types,* this volume) in several texts and text sorts. Determine the statistical characteristics of these motifs (frequency distributions and their parameters, indices combining the number of lexemes and text length, etc.; cf. Köhler & Naumann 2010). Derive the theoretical probability distribution(s) from corresponding hypotheses. Test whether the characteristics of the motifs can be used for text classification.

Procedure

Set up a list of clause types such as causal, consecutive, final, concessive, adversative, etc. and tag each clause in the selected texts according to its type. Form motifs as described in Köhler (2015) and determine their empirical rank-frequency distributions.

Set up a hypothesis for each kind of motif which explains the specific distribution.

By means of fitting software such as Altmann Fitter, obtain the parameters of the distributions. Calculate additional characteristics such as L (number of different words in a text) and N (number of running words). Classify the texts using the characteristics (Köhler & Naumann 2010) and evaluate the results: is this method promising?

References

Altmann-Fitter, Lüdenscheid, RAM-Verlag.

Beliankou, A., Köhler, R., Naumann, S. (2012). Quantitative Properties of Argumentation Motifs. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and applications of quantitative linguistics: 33-42.* Belgrade: Academic Mind.

Beliankou, A., Köhler, R., Naumann, S. (2013). Distribution of the depth of argumentation relations. In: Köhler, R., Altmann, G., (eds.): *Issues in Quantitative Linguistics 3: 195-205*. Lüdenscheid: RAM-Verlag.

Köhler, R. (2015). Linguistic motifs. In: Mikros, G., Mačutek, J. (eds.): *Sequences in Text and Language: 89-108.* Berlin/Boston: de Gruyter.

Köhler, R., Tuzzi, A, (2015). Linguistic modelling of sequential phenomena. In: Mikros, G., Mačutek, J. (eds.): *Sequences in text and language. Structures, functions, interrelations, quantitative perspectives: 109-124.* Berlin/Boston: de Gruyter.

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.

4.11 Sentence length

Problem

In a short article, M. Roukk (2007) studied sentence length in German texts and their translations into Russian as well as a Russian text translated into English. Study the equality of the distributions using the tables presented by M. Roukk.

Procedure

Consider the four tables presented by Roukk, consider the frequencies as a function expressing all. At the beginning, use the Zipf-Alekseev function (not the distribution), then simplify it stepwise and fit the simple exponential function to all cases.

Then take pairwise other texts whose translation in Russian, German or English is known and perform the same operations. If you are successful, continue the research using other translations. Do not use poetic texts since in this text type sentence (length) is not determinable in the usual way.

Strive for the simplest function applicable to all cases. If you find exceptions strive to find the boundary conditions causing the deviation.

If possible, extend your investigation to many languages. You may take short texts containing approximately 50 sentences. Extend the theory, if necessary.

References

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

Kelih, E., Grzybek, P. (2005). Satzlänge: Definitionen, Häufigkeiten, Modelle (Am Beispielslowenischer Pressetexte). *Quantitative Methoden in Computerlinguistik und Sprachtechnologie 20, 31-35.*

Kelih, E., Grzybek, P., Antić, G., Stadlober, E. (2006). Quantitative text typology. The impact of sentence length, In: Piliopoulou, M., Kruse, R., Nürnberger, A., Borgelt, Ch, Gaul, W. (eds.), *From Data and Information*

Analysis to Knowledge Engineering: 382-389. Heidelberg/Berlin: Springer.
Niehaus, B. (997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In: Best, K, -H. (ed.), *Glottometrika 16, 213-275.* Trier: WTV.
Popescu, I.-I., Best, K.-H., Altmann, G. (2014), *Unified Modeling the Length in Language.* Lüdenscheid: RAM-Verlag.
Roukk, M. (2007). The Menzerath-Altmann law in translated texts as compared to original texts. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 605-610.* Berlin/New York: Mouton de Gruyter.

4.12. Sentence specification

Problem

Under specification of individual words in sentence one can understand the distance of a word from the "main" word (e.g. verb in dependence grammar). The computation can easily be performed if one has the respective graphs of each sentence. It can be expected that the distance 1 is the most frequent and the number of words in greater distances decreases monotonically. Of course, there can be differences in text types – at least in parameters. The task is to compute the distribution of specifications (distances from the verb) for individual texts and to find a simple model. Begin with one text. At the beginning, it is better to analyse the same text type than to consider different ones and set up different models.

Procedure

Take a text in which one can easily state the end of a sentence. Then prepare a dependence graph of the first sentence and compute the distances/specifications. Each sentence is a vector of distances. As a first problem, count the number of distances 1, 2, 3, …, and set up the distribution of distances. Find a function or distribution expressing well the situation. The interpretation will not be difficult because we have merely the style of the author and his braking by the future reader, who will understand everything easily. That means we have a simple synergetic situation.

Now do it for several texts of the same text type and strive for simplification and unification of the model.

Now, for each text you have the specification models of sentences in the form of vectors. Compute the distance of the first vector to the second, the second to the third, etc. Use the cosine between the vectors, then take arcos (cos τ) for each two subsequent sentences. You obtain a sequence for the whole text.

If possible, capture this sequence, even if you find cyclic repetitions. State how many immediate subsequences have the same direction, i.e. state the sequential dependency structure of the text. You may compare the sequences of two texts.

Here you compare not only the length of sentences but also their dependency structure. Elaborate a typology of text types and, if possible, perform the same analysis for a second language, e.g. take translations.

References

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and Codes of Text*. Lüdenscheid: RAM-Verlag (esp. Chapter 3).

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

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative Analysis of Italian Texts.* Lüdenscheid: RAM-Verlag (esp. Chapter 7).

4.13. Structural centrality: Clauses

Problem

Study problem *Structural centrality: Parts of speech.* Perform the same operations on the text partitioning each sentence in clauses.

Procedure

Take a text and subdivide it into sentences. Then for each sentence mark its clauses using abbreviations of types. The clause type can be taken from any reliable grammar of the language. Set up a vector of abbreviations (= clause types) for each sentence. Then write the vectors in a column and compare all with the longest one. Add zeroes in order to make them equally long.

Then study the most frequent clause type in individual positions. Take the most frequent one in the given position and divide the (greatest) number by the number of vectors. In this way you obtain a sequence of relative numbers representing the most frequent structuring of the given text.

Study the sequence, state whether there is a tendency from the beginning to the end of sentences and, if possible, make a conjecture about the form, cause, style, person in a stage play, text type, language, etc.

Quantify the classes of clauses, i.e. ascribe each type of clause a degree. The quantification must be performed using a specific selected aspect, e.g. dependence, restriction, integration, etc. One can find all kinds in grammars but you must develop your own quantification.

Having mastered the creation of a kind of quantification, transcribe the text in terms of degrees. (a) Study the frequency distribution of degrees (rank and spectrum); (b) construct numerical motifs out of degrees and study again their frequency forms; (c) study the length of the motifs and ascribe a property of length (e.g. its mean) to the given text type; (d) study the development of a text type historically, e.g. journalistic texts, children development, the work of an author.

References

Zörnig, P., Stachowski, K., Popescu, I.I., Mosavi Miangah, T., Chen, R., Altmann, G. (2010). *Positional Occurrences in Texts: Weighted Consensus Strings.* Lüdenscheid: RAM-Verlag.

4.14. Fitting of ranked hrebs

Problem

Analysing a text for hrebs and ranking them according to the number of sentences in them or the number of individual words in all their forms one usually obtains the Zipf-Alekseev or Zipf-Mandelbrot distribution. Show that the ranked sequences can also be captured by means of a simpler exponential function.

Procedure

First take the data from Ziegler and Altmann (2002: 76-83), and fit to the given ranked frequencies the exponential function given as $y = 1 + a*\exp^{(-bx)}$ where the 1 is added because the relative rate of change of *y* reacts to the frequency in the y-1 class. Besides, when you fit a function, you may omit all zero frequencies and rank the sequences anew. Study the parameter *b*.

Now take texts of interest to you and analyse them for hrebs. You may consider as hreb the original conception of "sentence aggregate", i.e. sentences containing the same concept or its synonyms, or its references, or you can simply consider hreb as all words signifying the same object, e.g. "person", "he", "a man", "who", "George" (if it concerns the same person), etc. All concepts of the text must be taken into account. Finally you consider the numbers characterizing the same hreb and fit the exponential function. If the experiment is successful, take a set of texts belonging to the same text type and analyse all. Classify the texts according to the increasing parameter *b*. Then compare the results with other text types.

If you have enough working capacity, perform the same operation on texts in other languages. Show the differences between a highly synthetic and highly analytic language. Synthetic languages will have richer hrebs than analytic ones because in the first class even the verbs and adjectives may belong to the same hreb as the noun to which they refer with their affixes.

Strive for a theory. Find other properties of texts associated with the parameter *b* of the exponential function.

References

Hřebíček, L. (1997). *Lectures on Text Theory.* Prague: Oriental Institute.
Hřebíček, L. (2000). *Variation in Sequences*. Prague: Oriental Institute.
Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

5. Semantics and Lexical Issues

5.1. Semantic diversification

Problem

Find a model for the semantic diversification of selected words. Consider first the overwhelming possibilities shown by J.A. Bär (2014, 2015) analysing the Max Weber corpus and searching for the word *Geist*, its derivates and compounds of which it is part. He found 106 lexical units and 1,251 occurrences. Using his data, find a function fitting well to the rank-order frequencies and generalize.

Procedure

Consider below the ordered data presented by J.A. Bär concerning the word *Geist,* its derivatives and compounds:

Rank	Freq.	Rank	Freq.	Rank	Freq.	Rank	Freq.
1	525	28	3	55	1	82	1
2	246	29	3	56	1	83	1
3	116	30	3	57	1	84	1
4	31	31	3	58	1	85	1
5	28	32	2	59	1	86	1
6	26	33	2	60	1	87	1
7	23	34	2	61	1	88	1
8	23	35	2	62	1	89	1
9	19	36	2	63	1	90	1
10	18	37	2	64	1	91	1
11	9	38	2	65	1	92	1
12	9	39	2	66	1	93	1
13	9	40	2	67	1	94	1
14	7	41	2	68	1	95	1

15	7	42	2	69	1	96	1
16	7	43	2	70	1	97	1
17	6	44	2	71	1	98	1
18	6	45	2	72	1	99	1
19	5	46	1	73	1	100	1
20	4	47	1	74	1	101	1
21	4	48	1	75	1	102	1
22	4	49	1	76	1	103	1
23	4	50	1	77	1	104	1
24	4	51	1	78	1	105	1
25	4	52	1	79	1	106	1
26	3	53	1	80	1		
27	3	54	1	81	1		

First find a simple function expressing this rank-ordering. Apply the exponential function with added 1. Then perform an analogous investigation in a corpus with other words. You may distinguish between derivates and compounds and find a model for both.

Perform the research historically, i.e. take respective corpora and study the change of parameters in the fitted functions. Generalize the results to diversification of any kind.

References

Bär, J.A. (2014). Methoden historischer Semantik am Beispiel Max Webers – Teil 1. *Glottotheory 5(2), 243-296.*
Bär, J.A. (2015). Methoden historischer Semantik am Beispiel Max Webers – Teil 2. *Glottotheory 6(1), 1-92.*

5.2. Semantic classification of compounds

Problem

Usually a compound is classified according to the parts of speech of which it consists or according to the number of compounding words, according to the kinds of joining (separated, hyphenized, joined, using prepositions, conjunctions, cases, etc.), according to the position of the main part, etc. There are of course other possibilities, one of which will be tested here.

Take a compound and interpret it, i.e. "explain" its meaning in such a way that the components remain, e.g. *error-free* as *free of errors* or *without errors,* etc. Now, classify the result according to some criteria you set up. You may apply both grammatical and semantic criteria but ensure all the possibilities are captured. Then perform a classification, state the situation in individual texts and set up a hypothesis about the frequencies. Test it.

Procedure

Take a text, for example a newspaper issue (one is sufficient), and write out all compounds you find together with their frequency. You obtain a very long list. Now, "interpret" each compound and set up the classification of "interpretations", which must be a kind of synonym of the compound. They must be made in the same language as the compound itself. Now state the frequencies of individual classes of "interpretations" and order them. You obtain a usual rank-frequency distribution which must be modelled. At the beginning, you can apply Zipf's power functions, Zipf-Mandelbrot's generalization, Zipf-Alekseev's generalization, exponential function, etc. Strive for a minimal number of parameters and maximal determination coefficient. You may use available programs but avoid polynomials. If you do not obtain a satisfactory result, continue searching and check the data, your classes, your hypothesis.

If you obtain satisfactory tests, perform the same operation for another issue of the newspaper. It is appropriate to take the same issue edited one year later and continue until one obtains a history. You may compare journalistic texts with other text types. If you have analysed several text types, you may compare the situation with that in other languages and step by step set up a theory of composition in which both grammatical and semantic backgrounds are captured. As a matter of fact, up to now only formal types of compounds have been classified. The present view opens a new field of research.

Reference

Gnatchuk, H. (2015). A statistical study of English compounds in the newspaper style. *Mathematical Linguistics 1(1), 81-90.*

5.3. Semantic function of adverbs

Problem

Take any language and study the semantic function of adverbs occurring in a text, i.e. set up the classes and count the occurrences. Then find a distribution or a function expressing the ranking of the classes.

Procedure

Laufer and Nemcová (2009) stated the following classes for German:

Temporal, Modal, Interrogative, Local, Consecutive, Concessive, Causal, Instrumental.

Using these classes, take individual texts and compute the frequencies of adverbs of the above type and prepare a rank table of frequencies. Then begin to search for an appropriate model. Evaluate at least ten longer texts. Apply, as is usual in quantitative linguistics, Zipf's power function, the exponential function, the Zipf-Mandelbrot function, etc.

After you have prepared the tables, apply the comparison of data or functions also to different text types. Here, do not compare the ranking but the individual classes. If you evaluated several languages, always compare the same text types in the different languages. A good possibility is given by using translations of literary works.

If you obtain the same model for all cases, you are on the trace of a law.

References

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

5.4. Adverbs of place

Problem

Perform a classification and a quantification of adverbs of place, direction, origin, etc. in at least one language.

Procedure

Take a dictionary and collect all adverbs of place and direction. As a matter of fact, one can find them on the Internet in various languages. Now, order them according to some principle. Some of them are mentioned in official grammars. Consider other criteria. Find an orientation in the space. Place the speaker in the centre of space and quantify the adverbs in their relation to this centre.

Express the results of quantification by an indicator expressing various properties of this space, e.g. entropy, centrality, nearness, motion; show the relation of the indicators to some other formal properties of adverbs, e.g. their length, position in sentence, etc.

If possible, show the given system graphically too.

Then take a text and express its place-adverbial structure by some of the above indicators.

Find the rank-frequency distribution of the adverbs in text and connect it with other properties of adverbs.

Study the semantic and the grammatical diversification of individual adverbs. Semantic diversification means polysemy; grammatical diversification means the ability of adverbs to be members of other parts of speech (e.g. prepositions).

Measure the degree of coalescence of adverbs with other words.

Begin to compare texts, text types and even languages.

References

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

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

Helbig, G., Buscha, J. (2001). *Deutsche Grammatik*. Berlin: Langenscheidt.

Hoffmann, L. (2009). Adverb. In: Hoffmann, L. (ed.), *Handbuch der deutschen Wortarten: 223-264.* Berlin: de Gruyter.

5.5. Divergence of prepositions

Problem

Prepositions are entities with much diversified meanings. In addition to their local meaning they are used in temporal domains, and in grammatical domains to express for example reason, aim, and many other non-local meanings. Each preposition diversifies semantically. Perform a thorough examination based on a comparison of two languages.

Procedure

Take the most comprehensive dictionaries of two languages for example German-English and English-German. For both languages make a complete list of prepositions. Then show the semantic correspondences in the form of a graph. Some prepositions will be very "powerful", so you must prepare several graphs for your purposes. Finally, take one of the languages and for each preposition write the number of those in the other language which may be used in translations. You obtain the distribution of the number of translations. Propose a model which tells us that "there are x prepositions in language A having y translations in language B". Test the model in both directions. Substantiate the model linguistically and use semantic, grammatical, semiotic, historical, etc. forces.

Then perform a study of translated texts. You may use works like *The Little Prince* by Exupéry or a translation of a work by Shakespeare, etc. Now, for each occurrence of a preposition write down the number of its individual translations. In this way you obtain for each preposition in one language a distribution of the number of various translations. If in the other language not a preposition but a phrase or a clause has been used, you have to consider this fact. One can conjecture that in the background there is a kind of law regulating the correspondences. Set up a hypothesis concerning the distribution, derive it, if possible, and test it on several texts.

Finally characterize the strength of the correspondence using some indicator, e.g. the entropy, the repeat rate, the moments of the distribution, etc.

You can also perform an analogical study for other parts of speech, e.g. conjunctions. One can classify the adverbs and consider only one class of them and search for semantic correspondence in a second language. You can study, for example, the diversification of languages of the same family if you find a text translated to all of them. The problem seems to yield a very extensive field of research.

References

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

5.6. Diversification of prepositions

Problem

1. Every preposition has more than one meaning. Study the frequency of use of individual meanings and state whether it is the same for every preposition.
2. The individual meanings of a given preposition have different frequencies. Find the distribution of meanings of each preposition separately, then set up a common model.

Procedure

First find a list of prepositions in the given language. Then define the meanings of each preposition. Use a reliable grammar, a dictionary, or your own system. If necessary, state the meanings using the translations of a preposition in another language.

Then take a longer text and prepare the frequency distribution of the meanings of each preposition separately. You obtain as many distributions as there are prepositions in the language. Now perform a test, e.g. a chi-square test for the equality of the distributions; you may set up several groups in which the distributions do not differ significantly.

Characterize the distributions using various indicators (e.g. entropy, repeat rate, moments, excess, Ord's criterion) and compare the indicators.

In the last step, find a theoretical distribution capturing either all cases or at least those in the given class. Substantiate the difference – if there is any – linguistically and perform the derivation of the distribution or function theoretically. Interpret the parameters of the function(s) leaning against synergetic linguistics. If your distribution has two parameters, show that they are linked: find the function linking them.

If your language uses postpositions instead of prepositions, perform the same procedures.

Show that cognate languages are different. Show that the translation of a text into another language yields differences but the theoretical approach is always the same.

Compare texts taking the prepositions individually. That is, state the semantic distribution of one preposition in each text separately and compare the

texts. Can you find differences between text types? Or, can you define text types using the differences between the semantic distributions of prepositions?

References

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

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 69-81*.

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

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

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.

Nemcová, E. (1991). Semantic diversification of Slovak verbal prefixes. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 67-74.* Hagen: Rottmann.

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

Rukk, M. (2016). Context-specific distribution of word meanings. In: Kelih. E., Knight, R., Mačutek, J., Wilson, A. (eds.), *Issues in Quantitative Linguistics 4: 110-112.* Lüdenscheid: RAM-Verlag.

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

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

5.7. Quantification of adjectives

Problem

Adjectives can be classified semantically, though it may differ in various languages, and there is a possibility that an adjective may belong to several classes at once, e.g. *hard*. Yesypenko (2009) used 18 semantic adjectival classes: 1. Traits of characterization. 2. Physical/natural condition. 3. Intellectual

capacity. 4. Appearance. 5. Senses. 6. Age/time. 7. Temperature/sound. 8. Shape/ size. 9. Flavour. 10. Weight. 11. Degree/intensity. 12. Colour. 13. Actions done to the object. 14. Positive evaluation. 15. Evaluation of length/distance/position of the object. 16. Evaluation of value/function of the object. 17. Material. 18. Negative evaluation. One can use them or consider other classifications. This is, of course, not scaling. In order to perform scaling, take one of the classes, find *all* adjectives belonging to it and set up a scale. Then order the adjectives belonging to this class according to the intensity of the property they express. Finally analyse texts and characterize them using this scaling.

Procedure

First choose a class, then set up the scaling and examine a long text taking into account all adjectives belonging to this subclass. Set up the distribution of individual degrees and search for a model expressing it.

You will have different difficulties: (1) The quantification in some classes is sometimes possible only using a second property, e.g. in class "17. Material" one must use another criterion. One can apply results from the literature. (2) Modelling is a problem because one does not know what the "forces" working in the given subclass are. In the derivation of models one must identify the parameters and functions at least preliminarily. (3) Do not use corpora but individual texts.

Now, if you obtained a model of the specific quantified property in the text, interpret the model and the character of the text. Scientific texts will clearly differ from poetic texts.

Do the same procedure with the other subclasses and show for a specific long text its adjectival structure.

References

Hundsnurscher, F., Splett, J. (1982). *Semantik der Adjektive des Deutschen. Analyse der semantischen Relationen*. Opladen: Westdeutscher Verlag.

Krause, M. (2011). Adjektive multifunktional *vs.* monofunktional. graduierbar *vs.* nicht graduierbar: Fragen. In: G. Schmale (ed.). *Das Adjektiv im heutigen Deutsch. Syntax. Semantik. Pragmatik* (Eurogermanistik 29): *15-27*. Tübingen: Stauffenburg.

Oguy, A., Mgeladze, M. (1993). *Sistema prilagatelnych v „Pesne o Nibelungach": rekonstrukcija srednevekovoj ozenocnoj sistemy*. Chernivtsy: Chernivtsy UP.

Schmale, G. (ed.) (2011). *Das Adjektiv im heutigen Deutsch – Syntax. Semantik. Pragmatik* (= Eurogermanistik 29). Tübingen: Stauffenburg.

Trost, I. (2006). *Das deutsche Adjektiv. Untersuchungen zur Semantik. Komparation. Wortbildung und Syntax.* Hamburg: Buske.

Warren, B. (1984). *Classifying Adjectives*. Gothenburg: Acta Universitatis Gothoburgensis.
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*. Chernivcy: CNU.

5.8. Modal verbs and text type

Problem

Study the connection of modal verbs with text types. You can use any language.

Procedure

In German one finds the following modal verbs: *können, müssen, wollen, sollen, mögen, dürfen.* In other languages the situation may be different. Take texts of different text types and state the frequency of modal verbs. Perform the following investigations:

(1) Compare the text types using a simple chi-square test and state whether the frequencies have the same proportions. If not, test each text type against each other.

(2) Propose a quantity characterizing the text type and order the text types according to this indicator. Scale the modalities of the above verbs – if possible.

(3) Order the frequencies of modal verbs in each text type in decreasing order. You obtain a rank-frequency dependence. Find a model for this sequence. Begin with the simplest functions like geometric, exponential, Zipf (power), etc.

(4) If you succeeded in scaling the modalities, order the frequencies according to them and find an appropriate function expressing this dependence. Show that there are differences between text types.

(5) Perform a classification of text types using the modal verbs. Show the dependence of the parameters of the resulting functions on this classification.

Study especially stage plays in which there are more modal verbs than in other texts. Characterize stage plays.

Finally, scale the modalities on the basis of some property, and characterize the texts by an indicator. Derive the properties of the indicator, compare the texts and order them.

References

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

5.9. Meaning specificity

Problem

The synergetic model of language, in particular the lexical control circuit, models the relation between word length and polysemy as a direct bond. However, the explanation given is based on an indirect argumentation: length and specificity of meaning are functions of each other, and polysemy follows specificity. Specificity is not considered explicitly because measurement of this property is not as easy as of the other two variables. Extend the model by introducing the two relations explicitly.

Procedure

Set up two individual hypotheses covering the direct relations

specificity = f(length)

polysemy = g(specificity)

and connect them. Justify the specific form of the functions and calculate the combined function. Test the resulting hypothesis on data from a dictionary. The observed variables for each word are length, specificity and polysemy. Length should be measured in terms of the number of syllables, polysemy in terms of the number of meanings as indicated in a dictionary, and specificity can be measured on a rank scale, e.g. *dining room* is more specific than *room*. A better solution is measuring generality of meaning, the opposite property, by its extension, i.e. the number of objects which belong to the class. One can assume that there exist more rooms than dining rooms, and counting the number of objects in a class would yield a natural number, i.e. a measure on a metrical scale. For practical reasons, however, and because we would have to expect many classes with an infinite number of elements, this method does not seem to be performable. Another method of measurement is determining the number of distinctive semantic features of a word in a classificatory system. More properties must be specified to identify, for example, *racing car* than are needed for *vehicle*.

Specificity can be measured for example as the position on the generality scale, e.g. *kitchen table – table – furniture – fixture – thing – (system)*. Here, *kitchen table* has position 5(6) in the specification chain. Lexical chains and networks are available for several languages; a good dictionary will do otherwise.

In any case, be warned not to confuse the specificity/generality dimension with the abstractness/concreteness axis. Concrete words denote elements of the class of physical entities whereas abstract ones denote ideas or concepts. Both kinds can have any degree of specificity.

References

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol. 3*, (p.70 f., 75). Lüdenscheid: RAM.

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

Kisro-Völker, S. (1984). On the measurement of abstractness in lexicon. In: Boy, J., Köhler, R. (eds.), *Glottometrika 5, 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.

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

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

5.10. Lexical productivity

Problem

In some languages one writes the compounds together, in others there are other ways, e.g. with a space or hyphen. The derived words are usually written together. Study the lexical productivity of 100 simple stems. Distinguish compounding and derivation and distinguish the original and the resulting parts of speech.

Procedure

Take a dictionary of a language and choose any simple word. You may use some Internet dictionaries and seek manually. Then find all words in which it occurs. In strongly synthetic languages, e.g. Slavic ones, you must take into account all morphs of the given word; in languages with antiquated writing – for example English – you must also pay attention to the way of writing, e.g. *body* and *bodily.* Having found all derivatives and compounds (written in any way), (1) count the derivational and the compositional productivity of each word. Set up the distribution of the degree of productivity for the 100 words; consider separately derivation and composition and prepare also the complete productivity (sum of both). Find a function expressing how many words (x) have productivity y. If possible, substantiate the formula linguistically and derive it from a differential equation.

(2) For each word make a list of parts of speech to which the result belongs, e.g. a word yields y_1 nouns, y_2 adjectives, y_3 verbs, etc. Set up the distribution of resulting POS for all simple words you investigated, find a fitting function and substantiate it.

(3) For each simple word state the number of morphemes in the resulting words, i.e. compute the morphemic length of resulting words. Then prepare a distribution of morphemic lengths of resulting words for all 100 words together. An indicator on these numbers shows you the lexical productivity of the given language.

(4) For each resulting compound make a list of POS of parts, e.g. one obtains noun+noun, noun+preposition+noun, noun+verb, verb+noun, etc. Add all identical compositions, prepare a distribution and find a function expressing it. The result represents the compounding productivity of the given language. You may use an indicator for characterization.

(5) If possible, compare the results of one language with those of another. Very interesting would be a comparison of cognate languages or of languages occurring in the same area (areal comparison). A very interesting comparison could be based on the translation of a work in several languages (e.g. *Le Petit Prince* by Exupéry).

(6) Search for other properties of the given language and set up a Köhlerian control circuit. What is the "cause" of the given result? How is the specific result related to other properties?

Solve all problems separately. If possible, take more than 100 words in order to make the results (distributions) more empirically grounded.

References

Bär, J.A. (2014). Methoden historischer Semantik am Beispiel Max Webers – Teil 1. *Glottotheory 5(2), 243-296.*

5.11. Imagery in texts

Problem

If in a text there is something concerning human senses or abilities, it may evoke various images in us. Usually one distinguishes seven kinds of imagery: visual, auditory, olfactory, gustatory, tactile, kinaesthetic and organic. The kinaesthetic ones concern movements or actions, the organic ones concern feelings of the body (pain, hunger, thirst, fatigue, exhausting, etc.). The imageries may be expressed by multiple parts of speech, not only by the three main ones (verb, noun, adjective). Even interjections, onomatopoeias, adverbs, prepositions or metaphors may contain a kind of imagery. Some texts, especially poetic ones,

contain many concepts evoking some kind of imagery. Analyse texts belonging to the same text type (e.g. poetic ones), state the number of imageries in the individual classes, then rank them and find a function capturing this ranking. Compare texts, study the development of a writer and of the language.

Perform a quantification (scaling) of imagery in individual classes and show the character of the classes. One may also ask test persons to order the concepts in individual classes. Perhaps you find a very strongly expressed situation in some classes. If so, analyse texts of the given type and make the first steps towards theory.

Procedure

The simplest method is to take short poetic texts with the same form, e.g. the sonnet. One can find them on the Internet in several languages. Do not use translations. Then take one text and state the imageries belonging to the individual classes mentioned above. First rank the classes according to the number of imageries in them and find a function expressing this ranking. Then take the individual classes and quantify the concepts, e.g. visual imagery may be ordered from dark to bright or according to the given situation. If there are enough concepts in a class, order the concepts according to the scale you used and find a function expressing it. The test persons will scale the concepts differently; you may take the averages.

After having analysed several texts compare them from the viewpoint of imagery. Strive for hypotheses that can be derived from some theory (cf. Wimmer, Altmann 2005) and search for the relations of the given result to other properties of the texts, i.e. strive to include imagery in a synergetic control cycle as proposed by Köhler (2005).

One can find imagery on the Internet and in many books but one cannot find a quantification.

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.

5.12. Colours

Problem

Levickij and Kantemir (2005) found in German 406 colour terms and studying many texts they obtained the frequencies of colours. Find the ranked distribution.

The technical list of colours is much longer but if one does not find some of them in the texts, one can simply omit them.

Procedure

First, use the given data and apply to the numbers a usual distribution/function, e.g. Zipf's power function, Zipf-Mandelbrot's function, Zipf-Alekseev's function, etc. If you do not obtain satisfactory results, apply also the exponential function in form y = a*exp(-b*x) +1. The added 1 helps to obtain better fitting because there are no frequencies smaller than 1. If none of the functions was satisfactory, use a software program to compute the best fit automatically. Do not use polynomials.

Then perform the same analysis in a second language. Find the colour names on the Internet and let the program seek the frequencies of individual colours in a corpus. You will have some difficulties in strongly synthetic languages because the adjectives may acquire a number of declinational affixes and internal changes. Apply again the above functions.

Then let the program seek the individual colours and state the field of associations of adjectives with nouns. For each adjective find the number of different nouns with which it can be associated. You obtain, again, a frequency distribution for each colour. Of course, if some colour occurs only once, there may be no association field. Hence take only the first ten colours.

Compare the association distributions and generalize. Make conjectures about the relationship of the colour with human visual abilities.

Perform also a historical analysis: take texts from other centuries and study both the development of frequencies as well as the development of associations with nouns. You obtain a number of results forming part of cognitive linguistics.

Set up a bibliography of studying colours quantitatively.

References

Berlin B., Kay P. (1969). *Basic color terms. Their universality and evolution*. Berkeley and Los Angeles: Univ. of California Press.

Bernhart, Toni (2003). „Adfection derer Cörper". *Empirische Studie zu den Farben in der Prosa von Hans Henny Jahnn*. Wiesbaden: Deutscher Universitätsverlag.

Levickij, V., Kantemir, S. (2005). Die statistische Analyse des semantischen Feldes der Farbbezeichnungen im Deutschen. *Glottometrics 11, 64-97.*

Pawlowski, A. (1999). The quantitative approach in cultural anthropology: application of linguistic corpora in the analysis of basic color terms. *Journal of Quantitative Linguistics 6, 222-234.*

Wyler, S. (1992). *Colour and language. Colour Terms in English*. Tübingen: Gunter Narr Verlag.

5.13. Word frequency representation

Problem

Words occur in text with a certain frequency. Show that there are some regularities, prepare models of them and substantiate them linguistically. This is one of the oldest problems of quantitative linguistics.

Procedure

Perform two kinds of counting: (1) Lemmatize all words in the given text and prepare a lemma count. (2) Count all words in their morphological forms.

Having determined the distribution of entities, replace the words of the text (for the first count the lemmas, for the second the word forms) by their frequency. In this way you obtain a string of numbers. Study the given string in the following ways:

(a) Compute the distances between subsequent equal numbers; you may use any type of distance measure. Show the distribution of distances, i.e. state how many times you found distance x. Prepare a model of the distribution, even if you do it inductively (using software). Then substantiate the model linguistically, e.g. apply Skinner's hypothesis.

(b) Compare the distributions in several texts of the same language, i.e. show whether there are similarities, differences, etc. and, if possible, apply the results to characterize text types. You may perform the comparison by taking into account the parameters of the obtained distribution/function.

(c) Partition the text into original sentences and prepare the sequences of occurrence for each text in the form of sets. Since not all sentences contain the same number of words, replace missing positions with zeroes. Apply a similarity measure or a test for similarity of sequences. Prepare a table of differences and draw consequences. Do technical texts differ from prosaic texts? Use the mean similarity and compare the means applying the asymptotic normal test.

The easiest way is to compare poetic and non-poetic texts.

Omit computing the rank-frequency distribution and the spectrum.

References

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.

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

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

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

Zörnig, P. et al. (2015). *Descriptivenes, Actvity and Nominality in Formalized Text Sequences.* Lüdenscheid: RAM-Verlag.

6. Borrowings

6.1. Borrowings: Sources

Problem

Borrowings and their frequencies are usually studied in association with the various forms of Piotrowski's law (cf. *Problems vol. 1,* 35-36.) but one can also find simpler models. Since borrowings are mostly words, and words have an infinite number of properties, they can be studied in many different ways. There are two basic perspectives: (1) Study the sources and the ways of loanwords through languages, and (2) study the individuals. For the individuals one can study the degree of their incorporation, the semantic fields they occupy, their polysemy, synonymy, time of borrowing, morphological productivity, association with parts of speech, stylistic use and text types in which they occur, presence in idiomatic use, etc. Some of these problems will be presented in the next chapters. The first problem considers the number of words coming from donor languages.

Procedure

Take an etymological dictionary in which the source language of each word is noted. If you have the dictionary in electronic form, you can save much time. Count how many words have the same origin (donor language). You must decide whether you distinguish general loanwords and so-called migratory loanwords, where the source language cannot be identified definitely. Order the source languages according to the number of loanwords (this can be called the "etymological spectrum"), i.e. set up the ranking (cf. the references) of languages. Then seek a function capturing well the given rank order. You may use any known procedure, e.g. Zipf's power function, Zipf-Mandelbrot function, Zipf-Alekseev function. Strive to find a common expression.

References

Best, K.-H. (2006). *Quantitative Linguistik. Eine Annäherung. Göttingen:* Peust & Gutschmidt.

Best, K.-H. (2008). Das Fremdwortspektrum im Türkischen. *Glottometrics 17, 8-11.*

Best, K.-H. (2010). Zum Fremdwortspektrum im Japanischen. *Glottotheory 3(1), 5-8.*

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

Kelih, E. (2014): Zur quantitativen Lehn- und Fremdwortforschung: Eine Einleitung. In: Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter – Quantitative Aspekte: 1-6.* Lüdenscheid: RAM-Verlag (Studies in Quantitative Linguistics, 15),

Wolff, D. (1969). *Statistische Untersuchungen zum Wortschatz englischer Zeitungen.* Diss. Saarbrücken.

6.2. Borrowings: Time

Problem

In some etymological dictionaries one can find the year or the century in which a borrowed word came into a language. The problem is whether the borrowing activity of studied language changes with time or whether it is always the same. Test the hypothesis that borrowing increases with time. Since the cultures and languages of the world are becoming closer to one another, one can conjecture such a development. Find the function describing the development.

Procedure

You may restrict your investigation to one source language or to one text type (e.g. newspapers) and count the number of borrowed words in each year. If the beginning lies too far in the past, you may subdivide the time into decades or even into centuries. But if you use a continuous function, it does not play any role. In order to make the computation easier, consider the first year of borrowing as 0 or 1 (i.e. subtract from all years the first year and/or add 1). You obtain in any case either a monotone increasing function or a function with a maximum and a subsequent decrease. The latter case is well known and is usually caused by purists or by certain political changes. Find a general formula, test it in several cases and interpret the found differential equation on the background of sociolinguistic and synergetic linguistic approaches.

References

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

Leopold, E. (2005). Das Piotrowski-Gesetz. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein*

internationales Handbuch. An International Handbook: 627-633. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse.* Hamburg: Kovač (Philologia. Sprachwissenschaftliche Forschungsergebnisse, 30).

Stachowski, K. (2014): The volume of Ottoman lexical influence on Romanian. In: G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to Text and Language Analysis. Dedicated to Luděk Hřebíček on the occasion of his 80th birthday: 207-228.* Lüdenscheid: RAM-Verlag (Studies in Quantitative Linguistics, 17).

6.3. Borrowings: Polysemy

Problem

A loanword is transferred into a language usually as a monosemic one. Usually it expresses a unique concrete phenomenon. But in the course of time a semantic diversification may take place. Hence there are two problems concerning the polysemy: (1) How many meanings does a borrowed word have at the present time? (2) Is polysemy correlated with time? New meanings may be added or not; it depends on the scope of the word. Hence there are two hypotheses to be tested.

Procedure

Take a dictionary in which the origin of all words is noted. Take the borrowings and for each of them count the number of its meanings. You obtain a quite simple probability distribution (x = polysemy, y = frequency). Find an appropriate function expressing this relation. Restrict, if necessary, your investigation to "technical" words. You will also obtain a distribution but perhaps the parameters will be different. That means the parameters represent another variable associated with the text type.

Perform, if possible, the same investigation in another language and generalize your findings.

Now, to test the second hypothesis, note the year (or decade or century) of the borrowing and evaluate the present polysemy. If it changes (= increased), take 100 borrowings and study the relation between the borrowing data and polysemy. You will, most probably, obtain a monotonically increasing function. Model this function. Then perform, if possible, the same investigation in another language and make the first steps in generalization. Specify the vocabularies you examine, e.g. technique, medicine, arts, journalism, etc. Show that the function is not linear but convex. Since language is a self-regulating system, find perhaps a

limit of convergence of polysemy, so that, finally, you obtain a version of Piotrowski's law.

References

Since to our best knowledge no quantitative studies about polysemy of loan words exist we refer to general quantitative works on polysemy of particular parts of speech.

Levickij, V.V. (2005): Polysemie. In: R. Köhler, G. Altmann, R. G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 458-464*. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

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

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.*

Strauß, U. Altmann, G. (2003). Age and polysemy of words. *Glottometrics 6, 61–64.*

6.4. Borrowings: Productivity

Problem

A loanword can remain in its original borrowed form or it can take part of the morphological processes usual in the given language and morphological productivity in general (for a discussion about the operationalization of the morphological productivity, which is highly relevant in this context, cf. Baayen 2005, Bauer 2001, Plag 1999).

There are several possibilities and all must be scrutinized separately: (a) The borrowing may accept affixes, (b) it may enter different classes of parts of speech, (c) it may take part in compounding processes given in the language, (d) it may appear in specific registers (standard language, colloquial language, technical terminology, journalistic language, etc.) or it may undergo stylistic dispersion. Test at least one of the above problems, count the phenomena and set up a model.

Procedure

First state the categories you want to investigate, e.g. state whether you consider merely source-language affixes or also target-language ones. Which morphemes (affixes) do you take into account? What types of compounds exist in the target language? What type of discourses do you want to consider? Then take a dictionary in which the origin of the word is noted. You may use a corpus here, find all occurrences of the borrowing – in all its forms – and prepare a separate counting for each aspect. Take at least 100 borrowings and study them. Analyse each problem separately. Consider a vernacular word and compare its behaviour with that of the borrowing. Finally you obtain numbers either in the form of frequencies, or frequencies related to certain properties. Set up models for each phenomenon.

Finally, if possible, show that the behaviour of the borrowings is related to some other properties which are already present in the Köhlerian synergetic control cycle. Show the place of the borrowings and observe the differences.

References

Baayen, R. H. (2005). Morphological productivity. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein Internationales Handbuch. An International Handbook: 243-255.* Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Bauer, L. (2001). *Morphological productivity.* Cambridge: Cambridge Univ. Press (Cambridge Studies in Linguistics, 95).

Gries, S.Th., Ellis, N.C. (2015). Statistical measures for usage-based linguistics. In: *Language Learning* 65 (S1), S. 228–255.

Köhler, R. (1986). Zur *linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer (Quantitative Linguistics, 31).

Plag, I. (1999). *Morphological Productivity. Structural Constraints in English Derivation*. Berlin: de Gruyter (Topics in English Linguistics 28).

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 791-801.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

6.5. Borrowings: Phraseology

Problem

A valid indicator of the degree of the integration of a loanword is its presence in idiomatic expressions (fixed collocations, phrases, etc.). Define them a priori or a posteriori in such a way that one can use the results for a cross-linguistic study. Propose a measure or the degree of this kind of integration and evaluate many texts. State whether the integration is related to the time of borrowing, frequency of use of the borrowing and polytextuality.

Procedure

First make a collection of 100 borrowings from an etymological dictionary, preferable items for which the time of borrowing is more or less undisputed. Then take a collection of texts, e.g. a corpus, and consider all phrases in which the given borrowing occurs. Try to determine whether the analysed borrowing is part of an idiomatic/fixed expression or not.

For the complete collection of texts notify the number of occurrences of the borrowing and the number of texts in which it occurs.

Now prepare a table containing the borrowing, its time of integration and its frequency. Test the following hypotheses:

(1) The integration is related to the age of the borrowing.

(2) The integration is related to the frequency of occurrence of the borrowing.

(3) The integration is related to the polytexty (number of individual texts in which it occurs).

For each relation – if there are any – set up a model expressing the degree of integration as the dependent variable. Perform, perhaps, the same operations using another text collection with different text and discourse types. Present everything in such a way that it is possible to use it for the analysis of other languages.

References

To our best knowledge no quantitative studies about borrowings and their occurrence in fixed/idiomatic expressions has been performed. However as general starting point we refer to recent approaches in quantitative sociolinguistics and usage-based linguistics.

Backus, A. (2013). A usage-based approach to borrowability. In: E. Zenner, G. Kristiansen (eds.), *New Perspectives on Lexical Borrowing. Ono-*

masiological, Methodological and Phraseological Innovations. Boston: de Gruyter (Language Contact and Bilingualism 7), S. 19–39.
Chesley, P., Baayen, R.H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics 48(6), 1343–1374.*
Zenner, E., Speelman, D., Geeraerts, D. (2013). Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. In: L. A. Janda (ed.), *Cognitive linguistics. The quantitative turn. The essential reader: 251-294.* Berlin: de Gruyter Mouton (Mouton reader),
Zenner, E., Speelman, D., Geeraerts, D. (2013). What makes a catchphrase catchy? Possible determinants in the borrowability of English catchphrases in Dutch. In: E. Zenner, G. Kristiansen (eds.), *New Perspectives on Lexical Borrowing. Onomasiological, Methodological and Phraseological Innovations*: *41-64.* Boston: de Gruyter (Language Contact and Bilingualism 7),
Zenner, E., Kristiansen, G. (2013). Introduction: Onomasiological, methodological and phraseological perspectives on lexical borrowing. In: E. Zenner, G. Kristiansen (eds.), *New Perspectives on Lexical Borrowing. Onomasiological, Methodological and Phraseological Innovations: 1-18.* Boston: de Gruyter (Language Contact and Bilingualism 7).
Hout, R.v., Muysken, P. (1994). Modelling lexical borrowability. *Language Variation and Change 1 (6), 39–62.*

6.6. Borrowings: Survival

Problem

In the same way as vernacular words, borrowings may be eliminated or survive. Test the hypothesis "the shorter a loanword, the greater its chance of surviving". To this end you must collect many data, especially from the past, and test how many years each of the given borrowings survived.

Procedure

First take a historical dictionary and state for 100 borrowings their first appearance and their death (for German loanwords in Polish cf. Hentschel 2001). Then set up a scale containing the approximate number of years a borrowing survived. Compute the mean length of each borrowing in the time classes. Finally, you obtain a table in which the first column contains the years of life, the second column the mean length of borrowings. If the above hypothesis holds true then with increasing number of years the mean length will be shorter. Model this decreasing function – you may find a "good" formula using programs but later on

substantiate the formula and its parameters. The modelling can be embedded in general approaches of "birth-and-death" models in quantitative linguistics (cf. Wimmer/Altmann 2005).

There can be problems with the text types. Do not use scientific texts, use only prose, especially journalistic texts if you have a respective corpus. You may also check the hypothesis in a complete newspaper. For each issue you obtain the list of borrowings and may perform their length classification. The hypothesis may also be tested in such a way that one compares the borrowings died in a certain age and their length. In this way one must consider many more borrowings, however considering the means this would be simpler.

References

Zenner, E., Speelman, D., Geeraerts, D. (2013). Cognitive sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. In: L.A. Janda (ed.), *Cognitive linguistics. The quantitative turn. The essential reader: 251-294.* Berlin: de Gruyter Mouton.

Hentschel, G. (2001). Deutsche Lehnwörter im Polnischen als Reflexe von tausend Jahren deutsch-polnischer Sprachkontakte. In: Franciszek Grucza (Ed.), *Tausend Jahre polnisch-deutsche Beziehungen. Sprache, Literatur, Kultur, Politik*, Warszawa. Warszawa: Graf-Punkt, *300–310*.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: R. Köhler, G. Altmann, R.G. Piotrowski (eds,), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*: *791-801.* Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

7. Poetry

7.1. Hexameters

Problem

Test whether the first four feet types in hexameters (i.e. SSSS, SSSD, SSDS, SDSS, DSSS, SSDD, SDSD, SDDS, DSSD, DSDS, DDSS, DDDS, DDSD, DSDD, SDDD, DDDD), where S = spondee, D = dactyl, ranked according to their frequency in individual works, behave according to the exponential function $y = 1 + a*\exp^{(-bx)}$. If not, show an alternative model.

Procedure

Take as many poems as possible, in as many languages as you know. Omit the last two feet in each verse because they are identical. Compute the number of sequence types given above, order them according to frequency and fit the above function.

Perform the investigation first in your L1-language, then extend it to the poetry of other ones. If necessary, use also published numerical results. The analyses have been performed in all languages in which one uses hexameters (cf. References).

Collect the hexameters of one language or the hexameters written by one author and compute the mean of the parameters *a* and *b*. Do they differ from those in other languages or those of other authors? Consider the means as usual variables. If you compare two hexameters from the same language, then the software gives you the standard error of *a* and *b*, and you can perform a normal test.

References

Best, K.-H. (2008a). Zur Diversifikation deutscher Hexameter. *Naukovyj Vysnik Černivec'koho Universitetu. Herman'ska filolohia. Vypusk 41, 172-180.*

Best, K.-H. (2008b). Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics 17, 45-53.*

Drobisch, M.V. (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Classe 18, 73-139.*

Drobisch, M.V. (1868a). Weitere Untersuchungen über die Formen des Hexameters der Vergil, Horaz und Homer. *Berichte über die Verhandlungen der*

Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Classe 20, 16-53.

Drobisch, M.V. (1868b). Über die Formen des deutschen Hexameters bei Klopstock, Voss und Goethe. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Klasse 20, 138-160.*

Drobisch, M.W. (1872). Statistische Untersuchungen des Distichon (von Hrn. Dr. Hultgren). *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe 24, 1-33.*

Drobisch, M.W. (1875). Ueber die Gesetzmässigkeit in Goethe's und Schiller's Distichen. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 27, 8-34-146.*

7.2. Distances of rhythmic patterns in hexameters

Problem

Find a model for the distances between equal rhythmic patterns of a hexameter. There are 16 patterns (SSSS, SSSD, SSDS, …) composed of spondees (S) and dactyls (D) because the last two patterns in the verse are always identical. Study the distance between individual patterns and adding all equal distances find the general distance.

Procedure

Take a longer, already analysed poem written in hexameters in any language. Count the number of steps necessary to come from a given pattern to the same pattern. You obtain a sequence of numbers which can be ordered according to the size of distance (x = 1, 2, 3, …). Then for each rhythmic pattern separately search for a common function or distribution. Finally, add the frequencies of all equal distances and find the general model.

According to Skinner's hypothesis there will be many small distances and fewer great distances; that is, the frequencies will decrease. Find an appropriate function. Begin with the exponential function, try to use the power function, Zipf-Mandelbrot's function, etc. Strive for unification, i.e. for the same function in at least one language. Comparing the frequencies of distances, e.g. using the chi-square test, perform a classification of hexameters. The same can be done by using the parameters of the function found.

References

Best, K.-H. (2008a). Zur Diversifikation deutscher Hexameter. *Naukovyj Vysnik Černivec'koho Universitetu. Herman'ska filolohia. Vypusk 41, 172-180.*

Best, K.-H. (2008b). Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics 17, 45-53.*

Drobisch, M.V. (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Classe 18, 73-139.*

Drobisch, M.V. (1868a). Weitere Untersuchungen über die Formen des Hexameters der Vergil, Horaz und Homer. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Classe 20, 16-53.*

Drobisch, M.V. (1868b). Über die Formen des deutschen Hexameters bei Klopstock, Voss und Goethe. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Klasse 20, 138-160.*

Drobisch, M.W. (1872). Statistische Untersuchungen des Distichon (von Hrn. Dr. Hultgren). *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe 24, 1-33.*

Drobisch, M.W. (1875). Ueber die Gesetzmässigkeit in Goethe's und Schiller's Distichen. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 27, 8-34-*

Job, U. (1981). Annotated bibliography on the statistical study of hexameter verse. In: Grotjahn, R. (ed.), *Hexameter Studies: 226-262.* Bochum: Brockmeyer.

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

Strauss, U., Sappok, Ch., Diller, H.J., Altmann, G. (1982). Zur Theorie der Klumpung von Textentitäten. In: Rothe, U. (ed.), *Glottometrika 7, 73-100.* Bochum: Brockmeyer.

7.3. Syllabic verse structure

Problem

Study the sequence of syllable types in a poem. Define syllable structure as CV, VC, CCV, etc., where C is a consonant and V is a vowel. Do not forget that you may interpret for example diphthongs as vowels (or define a third category), and

that some consonants may be syllabic, e.g. in Slavic languages. The famous Czech sentence "Strč prst skrz krk" (*Push the finger through the throat*) does not contain any vowels, but it contains syllables with the structure CCCC, CCCC, CCCC, CCC. Especially liquids (/l/, /ĺ/, /r/, etc.) may be syllabic. The interpretation of problematic cases depends on your decision but you may adhere to available grammars. After having transcribed the poem verse-wise perform the following investigations.

Procedure

(1) Count the numbers of individual syllable types in the complete poem. Rank the frequencies in decreasing order and find a well-fitting function. Take as first the exponential function; if it does not fit well (the determination coefficient should be greater than 0.8), use the Zipfian power function, Zipf-Mandelbrot's function and Zipf-Alekseev's function.

(2) Compare the analysed poems, using for example the chi-square test. Do not rank the frequencies; compare equal types. If in some languages there are no types that can be compared, replace their frequency with zero. This is the case, for example, in Polynesian languages, which do not have consonant clusters. If the frequencies are too small, perform a comparison of ranks.

(3) Study the development of an individual poem verse-wise. How do neighbouring verses differ? You may perform a positional comparison, i.e. you always compare the same positions and write 1 for identity of syllable types, and 0 for a difference. If the verses do not have the same number of syllables, the difference at the last positions is zero (= no identity). Then you may express the similarity as the proportion of identical positional pairs; the number of syllables in the longer verse is the number of comparisons. After having compared the neighbouring verses, study the course of similarities. Do they increase or decrease?

(4) Study the distances between syllabically equal verses. The distance can be defined simply as the number of verses between the identical ones. There is the possibility that no verses are syllabically equal; in that case the distance is infinite.

(5) Study Skinner's hypothesis applied in several other problems in this series that there is greater similarity between near verses than between distant ones.

(6) Propose an indicator of syllabic similarity or variety which takes into account all syllabic verse structures. Find the variance of the indicator and propose an asymptotic normal test for the comparison of poems.

(7) Now look at the poem vertically. State the distribution of types in each position separately taking into account the complete poem. Compute the consensus vector leaning against the results in Zörnig et al. (2016).

(8) Solve the problems presented in Zörnig et al. (2016) concerning consensus strings where a number of problems is solved taking into account parts of speech in texts.

(9) Take a specific rhythmic poem, e.g. written in hexameters, and solve all problems mentioned above. Then compare individual authors.

(10) Study the positional development of hexameters and compare different languages too.

References

Zörnig, P., Stachowski, K., Popescu, I.-I., Mosavi Miangah, T. Chen, R., Altmann, G. (2016). *Positional Occurrences in Texts: Weighted Consensus Strings.* Lüdenscheid: RAM-Verlag.

7.4. Study of feet

Problem

Feet consisting of stressed and unstressed syllables can be found not only in poetry but in any other text. Unfortunately, few investigations have been performed in this domain (cf. Tomaševskij 1929, Grzybek, Kelih 2005). One may ask (1) what is the distribution, e.g. ranked, of the feet, (2) how can be the text characterized, (3) how can texts be compared, (4) is there a development in a specific text type. Solve at least one of these problems.

Procedure

Take a longer prosaic (!) text and analyse it counting all different feet. You obtain a list of frequencies. Rank them in decreasing order and find a simple function expressing this ranking. Do not forget that a foot may also consist of two words, mostly in cases where there are clitics. Pay particular attention to prefixes and suffixes.

If you obtained the function, perform the same procedure with another text belonging to another text type. Compare the two ranked distributions, first applying the chi-square test, then comparing the parameters of the given functions. Any software program will always give the variances of the parameters. Construct a simple normal test.

Continue analysing texts and perform a classification, ordering, etc.

Having made it in one language, make the same in another. Restrict yourself to the same text type. After having performed the analysis, compare the two languages. Propose some statistical tests.

Now, begin to construct a theory. Why is the distribution of feet in one language of the given form and in another somewhat different? That means, find

some reasons in the phonetics of the languages, and find another property which is correlated with the feet distribution.

Consider in any case also poetic texts. Even in poetry in which each line has the same number of feet, there may be frequency differences. Analyse the work of one famous poet and analyse data separately for each poem.

References

Grzybek, P., Kelih, E. (2005). Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 23-64.* Berlin: de Gruyter.

Tomaševskij, B.V. (1929). *O stiche.* Leningrad (München: Fink 1970).

7.5. Abstract assonance in poetry

Problem

Poems can have evident or latent regularities. Evident regularities may be displayed by rhyme, number of syllables in the verse, sequence of vowels, e.g. in Old Javanese poetry, number of verses in a strophe, a regular rhythm, a specific form of strophes (e.g. in sonnet), etc. Latent phenomena need not be quite regular; they may display a tendency, e.g. alliteration in the verse or in the first words of verses, assonance in the parallel words of verses, etc. One can only find them using statistical methods.

Search for a regularity in the placing of syllable types in poems. Syllable types are syllables reduced to combination of vowels (V) and consonants (C), and have the form C, CV, VC, CCV, CVC, VCC, … .

Procedure

Take a poem in your language and transcribe it in terms of syllable types. For each verse, you obtain a sequence of types. Study the following possibilities:

(a) Is there a tendency to place words of the same syllabic structure in the same position? If you cannot discover a regularity, state at least whether there are similar verses. That is, set up the distribution of verse types (consisting of syllable sequences). If the distribution is uniform or slightly oscillating, there is no tendency. But if some types occur more frequently, study their position in the strophes. If the poem is long enough, you may discover a tendency represented by the distribution of syllabic verse types.

(b) Count the numbers of all syllable types in the whole poem. Order them according to frequency and find a satisfactory function expressing this rank order.

(c) Compare the frequencies of syllable types in individual poems using for example the chi-square test. If the types do not have the same order, perform a test for the equality of orders (use a nonparametric test).

(d) Write the poem in terms of syllable types and compute the consensus strings. State whether there is a tendency, or whether the individual poems are different.

(e) Define syllable complexity. A syllable CCV is more complex than a CV syllable. Propose a kind of quantification. Then study whether there is an increase/ decrease of syllable complexity and a trend from verse beginning to verse end. By defining syllable complexity quantitatively the previous problems can also be solved more easily.

(f) Study the transition probabilities (within verse) of the syllable types. In practice, set up a contingency table and in each cell write the number of transitions from the type in the first column to the type in the first line of the table. Evaluate the table using a chi-square test. Evaluate the symmetry of transitions. Evaluate the preference of the diagonal.

(g) If you analysed the work of two writers, perform all comparisons, characterizations, etc. You may also perform a historical study but you must analyse several authors of the same time and compare merely the resulting numbers.

References

Zörnig, P., Stachowski, K., Popescu, I.-I., Mosavi Miangah, T., Chen, R., Altmann, G. (2016). *Positional Occurrence in Texts: Weighted Consensus Srtings*. Lüdenscheid: RAM-Verlag.

7.6. Rhyme words: Length

Problem

The words used in rhymes cannot always be the same because their frequent use in this position makes them non-effective. Hence there must be a development in different aspects. Study first the length of the rhyme word.

Procedure

Take a collection of poems in one language. Usually one finds everything in corpora like "Project Gutenberg" but for many languages there are collections presented on the Internet.

Collect rhymed poems historically, i.e. for each poem note the year in which it was written or published for the first time. Then for each poem compute the length of the last (rhymed) word in the verse.

Set up a table containing a column with the length scale (1, 2, 3, …), and a column with the number of rhyme words with the given length. For each poem you obtain a separate table. In a final step the tables are joined. Or you write in the first column the title of the poem and the year, and in the second column simply the number of lengths 1, 2, 3, …, for example 7, 15, 0, 4, 2. In this way you can insert all poems in one table.

Now, in order to study the development, consider one of the functions of the given distribution, for example the mean length, its excess or kurtosis. Write it in the third column.

Order the poems according to years. Since these numbers are usually very great, subtract from each year the smallest year (+1) and you obtain smaller numbers.

Now study the development of length in the history. First put the numbers in Excel and use it to make a figure of means. If the curve is not horizontal but displays a trend, capture the trend first by applying an empirical function. There are programs yielding thousands of functions. Choose a very simple function (= having few parameters), interpret the parameters linguistically, use the unified theory (Wimmer, Altmann 2005) in order to say how mean length changes over the course of years, and describe the result.

You can perform the investigation for only one writer and describe his development seen from this view. You can also compare several writers or different languages.

If you measure length, do not take into consideration the written but the phonetic form!

For each poem or period you can also compute the distribution of rhyme-word lengths, but only if the poem is long enough, and state its theoretical form. You probably mostly obtain a variant of the Poisson distribution.

References

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol. 3*. Lüdenscheid: RAM-Verlag (esp. p.120 f.)

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. Rhyme words: Accent

Problem

Study the position of the last accent in the verses of poems. Compute the tendency and state whether there is a trend.

Procedure

The last syllable of the rhyme word of the verse in a poem can be accented or not. Take a poem, compute the proportion of verses with this property and perform a two-sided test for the existence of a tendency. The test may be binomial or asymptotic normal. Omit poems with stereotype meters, e.g. hexameters.

Now, study this property either in the work of one poet chronologically or in the given language. If you find several poems created/edited in the same year, you can compute their common proportion.

Since in a language everything changes, set up a hypothesis concerning the development of this property. If the poetry began with a great proportion of "last-accented" syllables, then the development could go in the opposite direction (and vice versa). Study this development in the poetry of a single author and in the poetry of the analysed language.

Compare authors and compare languages. Express the development (if there is any) by a function and substantiate it linguistically.

References

None

7.8. Rhyme words: Parts of speech

Problem

(1) Study the distribution of parts of speech to which the rhyme words belong. Set up the rank-frequency distribution of rhyme-POS and substantiate it linguistically.
(2) Study the equality of the rank-frequency distributions in various poems. Is it always the same in a language?

(3) Study the development of the rank-frequency distribution in time concerning the POS of rhyme words.
(4) Study the POS in the pairs of rhyme words, e.g. N-V, Adj-N, i.e. state all pairs of rhyme words in terms of their POS and search for a tendency.

Procedure

In some languages some parts of speech need not have different forms, hence they must be ascertained syntactically from the context. In German "warm" (warm) can be either an adjective or adverb, in English "look" can be either a verb or a noun, etc. Hence define exactly your ascription of a rhyme word to a POS.
(1) Take a longer poem and state the numbers of POS of all rhyme words. Order the frequencies in decreasing order and state the theoretical rank-frequency distribution. Apply a rank-frequency distribution from quantitative linguistics and substantiate it. If you have tested a hypothesis positively in several languages, you may speak of a possible law. The placing of POS in the rhyme position is not done consciously by the writer; he simply looks for a rhyming word. But subconscious processes need not be chaotic. There may be a strong trend behind this process. Try to find it.
(2) If you performed this analysis with several authors in one language, order the pairs of POS and test whether in all poems there is the same tendency. Perform the chi-square test for homogeneity and if the columns are not homogeneous, find those which display the greatest difference. Search for boundary conditions which may be situated in the kind of poetry, in the time of creation, in the theme, etc.
(3) Use all rank-frequency distributions you obtained in problem (1), compute some of their properties, e.g. moments, entropy, repeat rate, excess, kurtosis, Gini's coefficient, Ord's criterion, etc. and study the development of some of them. Classify the poems according to well-known poetic criteria or content and study the development of the given indicators in each class separately. If you obtain some positive results, generalize it analysing rhymed poetry in other languages. Set up the first hypotheses in this domain.
(4) Now, construct POS pairs of rhyming words, e.g. if the rhymed words are a noun and an adjective, you obtain N-Adj. The order of POS is relevant, i.e. N-Adj is something different from Adj-N. State the numbers of such pairs in the first poem. Then

(a) Compute the rank-frequency distribution of pairs and find a model of it

(b) Compare individual poems using a chi-square test and set up classes with similar frequencies of POS pairs.

(c) For each poem study the symmetry of rhymed POS, i.e. compare N-V with V-N, N-Adj with Adj-N, etc. Express the strength of symmetry with an appropriate indicator. It may be for example the proportion of symmetric pairs, a chi-square of similarity, etc.

(d) Study the development of this kind of symmetry with one writer, with one class of poems, within one language. If you obtain noteworthy results, study the generality of the given state of affairs by analysing other languages.

References

None

7.9. Rhyme words: Runs

Problem

Take a longer rhymed poem in your language and consider the rhyme words. Compute the runs of rhyme-word lengths, runs of open and closed rhymes and runs of parts of speech to which the rhyme words belong.

Procedure

First solve the problems concerning rhyme words in the previous three chapters. For each poem write the respective numbers or symbols in the form of a sequence, e.g. lengths [2,1,3,1,2,…], open and closed rhyme words such as [o,o,o,c,o,c,c,…], and for POS use your own abbreviations, e.g. [N,N,Adj,N,V,…] where N = noun, V = verb, Adj = adjective, etc.

Now for each of these categories and each poem separately, study the runs of equal numbers or symbols. Write also the length of the run; for example in the sequence [o,o,o,c,o,c,c] there are two runs of length 1, one run of length 2 and one run of length 3. Characterize each poem by the distribution of run lengths of the above categories. Then for each category find a model. It may be a distribution or a simple function.

Having analysed several poems, you may perform the following investigation: (1) Study individual writers concerning the year of the creation of the poem. (2) Study the development of these properties in the poetry of your language. (3) Characterize each poem in each category by an indicator and study the development of this indicator. (4) Show the variability of a poet or his adherence to this way of writing. (5) Test for each poem and each category whether the number of runs is as expected or more extreme. (6) If possible, perform the analogous analysis in another language and compare the results with the first language.

References

None

7.10. Rhyme words: Concreteness – abstractness

Problem

Rhyme words are usually taken from the main parts of speech like nouns, verbs, adjectives. Now each of them has more or less an abstract/concrete meaning. Use various psycholinguistic studies enabling you at least to classify the rhyme words according to this property. If possible, propose a quantification of abstractness, order the words to individual degrees and characterize the poem.

Procedure

Take a longer poem and study only the rhyme words belonging to the main parts of speech. First perform a simple dichotomic classification to concrete – abstract, count the frequencies in the two classes and state whether there is a significant difference between them. You may use for example the binomial test with p = 0.5 and n = number of rhyme words, or you may use the chi-square test, etc.

Then perform a quantification, e.g. in each of the two classes set up degrees, five for concrete (1 to 5), five for abstract (6 to 10). For example *to sing* belongs to a degree in the concrete class, *to admire* has a degree in the abstract class. The same must be done with nouns and adjectives. Then count the frequencies of the individual degrees in all rhyme words of the poem and find a function or a distribution expressing them.

Having gathered the first results of counting, inductively try various functions or distributions. Take the best one, derive it from a theory, substantiate the hypothesis linguistically, and test it on data. Then perform the same operations with other poems.

The individual poetic genres differ by the parameters of the resulting function/distribution. Study poems in some other language and show some general results.

References

Hill, F., Korhonen, A., Bentz, Ch. (2014). A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive science 38 (1), 162–177*.

Kablau, M. (2002). Vom Konkreten zum Abstrakten? Bedeutungswandel von Wörtern. In: R. Köhler (ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: 253-270.* Trier: Universität Trier.

7.11. Hreb analysis of sonnets

Problem

The decomposing of a text in hrebs involves the ordering of all words and morphemes with the same meaning into sets. Each text is then a collection of sets which have their properties and whose distribution can be modelled (cf. Ziegler, Altmann 2002). Since sonnets are quite short one can analyse the complete creation of a poet and the development of sonnets in time, compare old and new sonnets, etc. This is, as a matter of fact, a semantic analysis of sonnets caring only for denotations. Perform some operations analysing sonnets in your language.

Procedure

Take a collection of sonnets in your language and perform separately the hreb analysis of each sonnet. Then count the number of elements in each set and set up the distribution of the number of sets (f_x) with x = 1, 2, 3, … elements. Derive a model of this distribution. You can also use a simple continuous function (without normalization). Test the adequacy of your model in all cases.

Compute for each sonnet the concentration of the text using the repeat rate of the above distribution. Order the sonnets according to the size of repeat rate and examine whether it correlates with the years of creation, i.e. whether there is a development in this sense.

In analysing a sonnet write for each element its position in the poem. Finally, compute the diffuseness (D_H) of each hreb by applying formula 3.8 in Ziegler, Altmann (2002: 55), i.e. the difference between the highest and the lowest position divided by the number of elements in the hreb. Consider only hrebs with at least two elements; omit the rest. You obtain for each hreb (containing at least two elements) a number. Order the hrebs according to increasing diffuseness and find a model for this sequence. The model will have multiple parameters – strive for simplicity (!). State whether some of the parameters correlate with the year of creation. Since the sonnets have the same length (in terms of line numbers) you may use the compactness indicator (cf. Ziegler, Altmann 2002: 60 ff.) – compute it for each sonnet separately and study its correlation with the year of creation.

Strive to find other properties of text correlating with the hreb-like characterization and take the first steps towards a theory.

References

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

7.12. Sonnet: Phonetic coincidence

Problem

A sonnet is a short poem, so it can be examined quickly. Study the phonetic coincidences given by equal sequences, set up the graphs of this phenomenon and evaluate its properties. The coincident part must contain at least one consonant and one vowel in immediate succession.

Compare the results concerning the development of an author, then those concerning the years of creation in the given language, and examine the historical development of phonetic coincidence.

Syllabic coincidence can be initial, central or final in the word. For example, the last syllables or the parts of two rhyme words have the same phonetic whole, e.g. in Goethe: *Wind* vs. *Kind*. In this case we have a coincidence "ind" and search for other words containing it.

Evaluate the weight of coincidences and order them according to their decreasing weight. Propose a model of the sequence, characterize the sonnet by an indicator and study the development of phonic coincidence in sonnets in a language.

Procedure

Take the first sonnet of a collection and your program should search for all words containing the same sequence of phonemes. For each entity you may obtain either a list of words or only the word from which you took it. Count for each individual entity the number of words in which it occurs. Prepare the distribution of these numbers. You may evaluate the phonetic coincidence of the sonnet by any indicator constructed on the basis of these distributions. Construct a model and compare the parameters of the model with all other sonnets. Show the change in the values of parameters over the course of years.

Now study the distances between the same phonetic entities. You obtain, again, a distribution of distances. In order to evaluate them, study the problems in the first five volumes of *Problems in Quantitative Linguistics.*

In this study one omits the assonances, alliterations and rhymed words studied elsewhere. It is a specific view of phonic structuring of the text.

References

None.

7.13. Adjectives: Poeticity

Problem

Adjectives usually have the function of modifying nouns, to express their properties, to colour a poem in a poetical sense. It can be conjectured that the number of adjectives in individual strophes of a long poem abides by some law which can be expressed formally by a function or a distribution. One can, at least, suppose that the poet had a particular intention and intuitively followed some regularity. The poeticity of the poem can be expressed by various indicators. Since the strophes may have different lengths, a direct comparison of indicators is not possible; they must be related to the number of lines in the strophe. In any event, one can subdivide the poem according to its rhyme structure.

The task is to state the number of strophes f_x containing x adjectives, derive a formula, test it and compare the texts.

Procedure

Take a long poem and count the number of adjectives in each strophe. Then set up the distribution of the number of strophes containing x adjectives. Seeing the numbers, find a distribution or function expressing this regularity. Here, one cannot apply the ranking laws; one must find another possibility. Use manual means, e.g. NLREG or TableCurve, or Origin or Fitter and find a model which is common to all.

Then take longer poems in another language, perform the same operations and strive for equality of modelling.

Now, consider again the original language and compare the distribution of adjectives in strophes in epical and lyrical poetry. Is there a difference? The statistical tests must be performed in a complex way; one must take into account the length of the poems and strophes, etc.

The same procedures may also be performed with adverbs which modify the nouns, the verbs and the adjectives. One obtains, again, distributions whose form tells something about the background mechanisms of poem writing.

References

None

7.14. Adjectives: Sonnet

Problem

Sonnets have 14 lines subdivided into 4,4,3,3. They are too short for stating the type of distribution of adjectives in individual strophes. But there may be other methods for quantifying and measuring the poeticity (= in this case presence of adjectives) of individual strophes. Devise such a method and perform analyses of (a) the development of the individual poet, (b) the development of sonnet writing in a language considering sonnets of different authors but in historical succession, (c) perform analyses of sonnets in several languages.

Procedure

Take a collection of sonnets of one poet in your language. First count the number of adjectives in the first strophe and divide it by 4; the same for the second strophe. The numbers in the last two strophes containing only three lines should be divided by 3. You obtain four numbers which are not proportions but weights. Take the mean of these four numbers as a weight of the given sonnet. Then do the same for other sonnets of the same author and note the time of creation (or publication). You obtain a sequence of numbers which may be captured by a horizontal straight line or by an increasing/decreasing function, or by an oscillating function. In any case you can characterize the development of the poet.

In order to describe problem (b), make a list of authors who wrote sonnets in the given language. Then take from each author exactly one sonnet, order the computation results according to years and express formally the development of adjective placing in sonnets in the given language. Again, you obtain some kind of function.

For problem (c) you have the numbers of one language. Perform the same operations in another language and compare the sequences. You may compare various properties of the resulting curves: heights, number of maxima for oscillating curves, directions expressed by a parameter, etc.

If you succeeded in solving some aspect of this problem, consider other parts of speech, e.g. adverbs or verbs and their behaviour in the strophes. You may perform other quantifications of properties which must be exactly defined.

Collections of sonnets can be found on the Internet.

References

Böhn, A. (1999). *Das zeitgenössische deutschsprachige Sonett. Vielfalt und Aktualität einer literarischen Form*. Stuttgart: Metzler.

Borgstedt, Th. (2009). *Topik des Sonetts. Gattungstheorie und Gattungsgeschichte*. Tübingen: Niemeyer.

Fechner, J.-U. (ed.) (1969). *Das Deutsche Sonett. Dichtungen – Gattungspoetik – Dokumente*. München: Fink.

Fuller, J. (1972). *The Sonnet.* London: Methuen.

Greber, E., Zemanek, E. (eds.) (2012). *Sonett-Künste: Mediale Transformationen einer klassischen Gattung.* Dozwil: Edition Signathur.

Jasinski, M. (1903). *Histoire du Sonnet en France.* Douai (Genève 1970).

Kemp, F. (2006). *Das europäische Sonett*. 2 Bde., Göttingen: Wallstein.

Mitlacher, H. (1932). *Moderne Sonettgestaltung.* Leipzig: Noske.

Mönch, W. (1955). *Das Sonett. Gestalt und Geschichte.* Heidelberg: Kehrle.

Neubauer, P. (2001). *Zwischen Tradition und Innovation. Das Sonett in der amerikanischen Literatur des zwanzigsten Jahrhunderts.* Heidelberg: Winter.

Schlütter, H.J. (ed.) (1979). *Sonett*. Sammlung Metzler, 177: Abt. E, Poetik. Stuttgart: Metzler.

Stemmler, Th., Horlacher, St. (eds.) (1999). *Erscheinungsformen des Sonetts.* Tübingen: Narr.

Welti, H. (1884). *Geschichte des Sonettes in der deutschen Dichtung. Mit einer Einleitung über Heimat, Entstehung und Wesen der Sonettform.* Leipzig: Veit.

Subject Index

Author Index

The RAM-Verlag Publishing House edits since 2001 also the journal ***Glottometrics*** **–** up to now 40 issues – containing articles treating similar themes. The abstracts can be found at http://www.ram-verlag.eu/journals-e-journals/glottometrics/.

The contents of the last issue (40, 2018) is as follows: