Problems
in
Quantitative Linguistics
4

by

Reinhard Köhler
Gabriel Altmann

2014
RAM-Verlag

Studies in quantitative linguistics

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

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.

ISBN: 978-3-942303-22-4

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

Preface

The fourth volume of *Problems in Quantitative Linguistics* provides once more new evidence for the truth of the statement that research will never run out of challenges. Any aspect of a unit, a property, a level of linguistic analysis, a language, an individual text, a text sort, etc. can be specified and deepened or generalized. New ways of looking at language are found, new methods are developed and new questions are asked. Established units and well-known properties are connected in new ways and incorporated in control cycles which furnish us new hypotheses and even theories. Besides, new vistas taken from other sciences can be introduced and the linguistic reality can be seen as something that has analogies with the "rest" of the world. Today, we use a great part of "quantitative" mathematics, see the language from psychological, communicative, pragmatic, social, grammatical, textological, evolutionary, diversificational, stochastic points of view and introduce ever further views like that of systems theory, graph theory, fractals, time dependent processes, etc.

Testing the presented hypotheses does not only concern their corroboration but rather the search for more general hypotheses, or, more specific ones with inclusion of some boundary conditions.

The present volume contains again diversified problems which can be used for writing contributions to journals, dissertations or for organizing projects in quantitative linguistics. There are no exercises in the book, but problems whose solution would contribute to the development of this science. The readers are invited to write articles and send them to the journals *Glottometrics, Glottotheory* or *Journal of Quantitative Linguistics.*

We are aware of the fact that some problems represent complex projects. Do not try, in these cases, to solve all details at once and set up a complex theory at the first attempt and in a single step. Solve only a first, partial problem, collect data from many texts or languages; then solve the second partial aspect of the problem and generalize stepwise. The linguistic aspects and data collection must be made by a linguist (not by a mathematician), the mathematician should help solving the mathematical problems. A programmer can be consulted only if the linguist is able to present all definitions in a formal way.

In the present volume there is more syntax than in the previous ones and many problems are more complex. The authors are ready to help researchers who are interested in this kind of investigations.

Reinhard Köhler
Gabriel Altmann

Contents

Preface

1. Syntax 1

1.1. Event integration 1
1.2. Cline of grammaticality 2
1.3. Complementation scale and co-lexicalization 3
1.4. Complementation scale and subordinating morphemes 4
1.5. Voice diversification 5
1.6. Remote referent 7
1.7. Cohesion, coherence and thematic continuity 8
1.8. Anaphoric distance 9
1.9. Cataphoric persistence 11
1.10. Causality in text 12
1.11. Coherence/Cohesion of conjunctions 15
1.12. Degrees of finiteness 16
1.13. Study of POS bigrams 17
1.14. Position of function words in sentence 19
1.15. Noun phrase 19
1.16. Length of R-motifs 21
1.17. Length of D-motifs 23
1.18. Syntactic complexity 25
1.19. Adnominal modifiers - 0. Classification 26
1.20. Adnominal modifiers - 1. Weight 28
1.21. Adnominal modifiers - 2. Distribution 29
1.22. Adnominal modifiers - 3. Complexity 31
1.23. Adnominal modifiers - 4. Cohesion 32
1.24. Adnominal modifiers - 5. Scaling 33
1.25. Adnominal modifiers - 6. Motifs 34

2. Semantics 36

2.1. Polysemy in German 36
2.2. Polysemy in the Swadesh list 37
2.3. The course of polysemy in sentence 38
2.4. Polysemic similarity and distance 39
2.5. Polysemic text construction 40
2.6. Vectors of polysemy 42
2.7. Synonymy 43
2.8. Semantic classes of nouns 44
2.9. Semantic diversification of prefixes 46

2.10. Semantics of prepositions 1 48
2.11. Semantics of prepositions 2 49
2.12. Degree of metaphoricality 50
2.13. Distribution of metaphoricality 52
2.14. Metaphoricality and frequency 53
2.15. Metaphoricality and length 54
2.16. The weight of metaphoricality in text 55
2.17. Metaphoricality motifs 56
2.18. Modality marking 57
2.19. Modality degree and frequency 59
2.20. Sound symbolism 60
2.21. Association of antonyms 62
2.22. Adjectival antonyms 64
2.23. B-motifs 66

3. Textology 69

3.1. Style 69
3.2. Style evolution 70
3.3. Entropy deployment 71
3.4. The type-token relation 73
3.5. Stage play analysis 1 75
3.6. Stage play analysis 2 76
3.7. Text properties 77
3.8. Thematic words and Frumkina's law 78
3.9. Distances in text 80
3.10. Text cohesion 83
3.11. Arc length of frequencies 85
3.12. Sentence sequences 86
3.13. Verbal antonymy 88
3.14. Hurst exponent 89
3.15. Hreb construction 92

4. Pragmatics 94

4.1. Speech act distribution 94
4.2. Speech act motifs 95
4.3. Speech act length 97
4.4. R-motifs of speech acts 98
4.5. Scaling of speech acts 100

5. Synergetics 102

5.1. Word length and polysemy 102

5.2. Kelih's Repeat Rate hypothesis 103
5.3. Word length and compositionality 104
5.4. Allomorphic complexity 105
5.5. Control cycle 106

6. Various issues 108

6.1. Scaling of dogmatism 108
6.2. Word frequency and initial clusters 109
6.3. Frequency and position in text 110
6.4. Hapax legomena and synthetism 110
6.5. Compound degree 112
6.6. Diversification theory 113
6.7. Givón's hypothesis 115
6.8. Laws in language 115
6.9. Morphological complexity of words 118
6.10. Syllable length 119
6.11. Nominal compounding tendencies in German 121
6.12. Quantification exercise 123
6.13. Corpus linguistics and theory 125
6.14. Small inventories 126
6.15. Borrowing 129
6.16. Frequency and irregular verbs 130
6.17. Ord's plane 132
6.18. Block distribution of modal expressions 132
6.19. Sonority sequences 135
6.20. Verb classes 137

Author index 139
Subject index 146

1. Syntax

1.1. Event Integration

Hypotheses

The present hypothesis concerns languages which allow clauses as objects of verbs such as in „I can see that she is coming“. Many of such verbs can take either a NP or a clause. Givón (2001, 39ff) considers the interdependence of semantic event integration and syntactic clause union. Specifically, he sets up the hypothesis:

> *The stronger is the semantic bond between the two events, the more extensive will be the semantic integration of the two clauses into a single though complex clause.*

The degree of event integration can be expressed on an ordinal scale as shown in Table 1, where the degree of event integration and the degree of clause union decrease from top to bottom.

Table 1
Complementation scale (after Givón 2001, 43; ranks added)

	Semantic scale of verbs	**Syntax of Comp-clause**	
1	She let go of the knife	co-lexicalized comp	1
2	She made him shave	bare-stem comp	
3	She let him go		2
4	She had him arrested		
5	She caused him to switched jobs	infinitive comp	
6	She told him to leave		
7	She asked him to leave		3
8	She allowed him to leave		
9	She wanted him to leave		
10	She’d like him to leave		
11	She’d like for him to leave	for-to comp	
12	She’s suggested that he leave		4
13	She wished that he would leave		
14	She agreed that he should leave	subjunctive comp	5
15	She knew that he left		
16	She said that he might leave later	indirect quote comp	6
17	She said: “He might leave later”	direct quote comp	7

a) Test the hypothesis that the degree of event integration abides by a probability distribution.
b) Test the hypothesis that the degree of clause union is also distributed according to a regular probability distribution.
c) Find theoretical justifications for the distributions obtained.
d) Determine the functional dependency between event integration and clause union.

Procedure

Data for the tests can easily be acquired from text corpora. For English, the procedure is straightforward because Givón's classification and ranking can be used; for other languages, the occurring clause types must be identified and ranked first.

Find modal, perception and manipulation verbs to identify the relevant clauses and assign to them the corresponding ranks for event integration and clause union. Appropriate probability distributions can most easily be found by means of the Altmann-Fitter software. Function fitting can be done also by several other available programs.

References

Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.
Altmann-Fitter 3.1. Lüdenscheid: RAM-Verlag (ram-verlag@t-online.de).

1.2. Cline of grammaticality

Problem

Hopper and Traugott (1993: 7) show that grammaticality is a gradual property. They show that there is a difference between linguistic entities that can be ordered, e.g.

content item > grammatical word > clitic > inflectional affix

and say that "forms do not shift abruptly from one category to another, but go through series of gradual transitions, transitions that tend to be similar in type across languages" and "it is often difficult to establish firm boundaries between the categories represented on clines" (1993: 6 f., cf. also Krug 2001: 325 f.). Quantify the problem.

Procedure

First find all possible categories of morphemes from several languages. Then define the property qualitatively, operationalize and quantify it. Set up at least one hypothesis, e.g. "historical change is the quicker, the lower the degree of the above defined property", or "the higher the degree of the above mentioned property, the longer are the entities", etc.

Explain the concept of "gradual transition", make it exact. Scrutinize the concept "similar in type" and quantify it.

Meditate about the necessity of stating exact boundaries between some types of linguistic entities which are, as is well known, fuzzy, and if quantified, placed on a continuous scale. Do not try to set up classes as is usual in qualitative linguistics.

Study the changes of the status of an entity as a historical process, set up a hypothesis and test it. Do not consider individual examples as cases of corroboration.

References

Bybee, J., Pagliuca, W., Perkins, R. (1994). *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: University of Chicago press.

Hopper, P., Traugott, E. (1993). *Grammaticalization.* Cambridge: Cambridge University Press.

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

1.3. Complementation scale and co-lexicalization

Hypotheses

The present hypotheses can be tested on data from languages which allow clauses as objects of verbs such as in „I can see that she is coming" (cf. problem *1.1. Event Integration*). Many of such verbs can take either a NP or a clause. Givón (2001, 39ff and 63ff) sets up the hypotheses:

a) The higher a verb is on the semantic-cognitive scale of event integration, the more likely it is to co-lexicalize with its complement verb.

b) If a complement-taking verb is co-lexicalized, all the verbs above it on the scale will also be co-lexicalized.

Consider the contrast between the examples (1) and (2):

(1) Mary let-go of John's arm.	(co-lexicalized)
(2) Mary made John go.	(not co-lexicalized).

Givón interprets these assumptions as iconic coding which underlies the general principle of proximity:

> *The closer two linguistic entities are functionally the more contiguously they will be coded.*

Test the hypotheses (a) and (b) on data from various languages.

Procedure

Hypothesis (a) makes a probabilistic assumption. Therefore, it has to be tested statistically and the results have to be evaluated by means of a significance test.

Annotate the verbs in texts with tags for their rank on a complementation scale and for their co-lexicalization status. Arrange the data which result from this procedure as values of a frequency distribution showing the dependence of co-lexicalisation probability on the degree of complementation. Find a mathematical model of the dependency and test it using e.g., the Altmann-Fitter.

Hypothesis (b) can also be interpreted as a probabilistic statement if not all the verbs found in the texts under study confirm with Givón's prediction. In such a case, you might be able to come to a corroboration of a probabilistic variant of the assumption.

References

Altmann Fitter 3.1: Lüdenscheid: RAM-Verlag (ram-verlag@t-online.de).
Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.

1.4. Complementation scale and subordinating morphemes

Hypotheses

The present hypotheses can be tested on data from languages which allow clauses as objects of verbs such as in „I can see that she is coming" (cf. problem 1.1. *Event integration*). Many of such verbs can take either a NP or a clause. Givón (2001, 39ff and 71ff) sets up the following hypotheses:

c) *The lower a verb is on the semantic-cognitive scale of event integration, and thus the less integrated the main and the complement events are cognitively-semantically, the more likely it is that a subordinating morpheme be used to separate the two clauses.*
d) *If a language uses a subordinating morpheme at a certain point on the scale, it will also use it at all points lower on the scale.*

Givón remarks that most languages use no subordinating morpheme for direct quotes but a pause (intonation break), a fact that seems to be a counter-example.

Test the hypotheses (a) and (b) on data from several languages.

Procedure

Hypothesis (a) makes a probabilistic assumption. Therefore, it has to be tested statistically and the results have to be evaluated by means of a significance test.

Annotate the verbs in texts with tags for their rank on a complementation scale (cf. problem 1.1. *Event integration*) and for the use of a subordinating morpheme. Arrange the data which results from this procedure as values of a frequency distribution showing the dependence of the probability of separation by means of a subordinating morpheme on the degree of complementation. Find a mathematical model of the dependency and test it using e.g. the *Altmann-Fitter.*

Hypothesis (b) can also be interpreted as a probabilistic statement if not all the verbs found in the texts under study confirm with Givón's prediction. In such a case, you might be able to come to a corroboration of a probabilistic variant of the assumption.

References

Altmann Fitter 3.1, Lüdenscheid: RAM-Verlag (ram-verlag@t-online.de).
Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.

1.5. Voice diversification

Hypotheses

a) *The text frequencies of the voices used in a language are distributed according to one of the diversification distributions* (Altmann 2005).
b) *The parameters of the distribution vary with text sort.*

Test the hypotheses.

Procedure

Grammars differentiate a number of voices, such as

Adjutative voice
Antipassive voice
Applicative voice
Active voice
Causative voice
Circumstantial voice
Impersonal passive voice
Mediopassive voice
Middle voice
Passive voice
Pseudo-passive
Reciprocal voice
Reflexive voice.

Givón (2001, 91f) presents a categorisation with respect to their function into semantic and pragmatic voices:

Primarily semantic
- Reflexive
- Reciprocal
- Middle-voice
- Adjectival-resultative

Primarily pragmatic
- Passive
- Antipassive
- Inverse

Set up a list of the voices you observe in a language and count their applications in texts. You should register their occurrences separately for each text and group them according to text sort.

Determine the theoretical probability distribution and the parameters, again, separately for each text. Do the parameters differ significantly between individual texts or between the groups of texts?

Analyze a stage play and show the sequence of voices. If you scaled the voices, study the time series.

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.

Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.
Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

1.6. Remote referent

Hypothesis

Besides the Behaghel-Hawkins Law of word order (*long after short*, cf. Köhler 1999), other principles are known in linguistics: One of them is the principle '*remote'* referents *after 'near' ones*, where *near* and *remote* may be understood not only with respect to space or time but also to mental and emotional dimensions (cf. e.g., Givón 1985, 1988). In Givón (2001), a specific hypothesis is set up concerning the word order of frozen word pairs such as *now and then, here and there, large and small* or German *nah und fern "near and remote", heute und morgen "today and tomorrow"*. Although this hypothesis seems at a first look to be limited to frozen pairs, Givón reports on languages where the reverse order would yield an ungrammatical expression (cf. Givón 2001, 17). Test the hypothesis, for which by now only examples are available, by means of a statistical test.

Procedure

Collect data from one or more languages: word pairs or other sequences preferably of the same length resp. complexity to exclude the influence of the Behaghel-Hawkins Law. Assign to each of them a code for either the order near–remote resp. remote–near. Perform a significance test to determine whether the first order is in fact preferred. You might also report the ratio and the absolute numbers of the individual values for comparability of results from several languages. In that case, a simple test for difference of two proportion would be sufficient.

References

Givón, T. (1985). Iconicity, isomorphism and non-arbitrarity coding in syntax. In: Haiman, J. (ed.), *Iconicity in syntax: 187-220*. Amsterdam: Benjamins.
Givón, T. (1988). Tale of two passives in Ute. In: Shibatani, M. (ed.), *Passive and voice*: *417-440*. Amsterdam: John Benjamins.
Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.
Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 27, 517-556.*

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergence of linguistic structure: 309-335*. Amsterdam: Benjamins.

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

1.7. Cohesion, coherence, and thematic continuity

Hypothesis

Conjunction morphemes can be ranked according to the degree of continuity with the following clause within a conjunction class (continuative, contrastive etc.). The ranks will show a lawful functional interrelation with the continuity degree. Test the hypothesis.

Procedure

The degree of continuity can be measured as the inverse probability of a subject change in the following clause. Consider the following examples (after Givón 2001, 348 ff), where punctuation marks are taken into account as reflex of intonation:

Conjunction type (continuative)	**% Subject change across the conjunction**
and	15
, and	70
. And	81
and then	16
, and then	36
. And then	100
, then	50
. Then	56
. PARAG/Then	100
comma (alone)	10
period (alone)	72

Conjunction type (contrastive)	Rank	% Subject change across the conjunction
and (all punctuations)	1	29
while	2	77
but (all punctuations)	3	85
, though	4	100
. Yet	5	100

Set up a list of conjunctions in a language and locate them in a text corpus. Determine the number of clause chains with a subject change between the clauses. Arrange the conjunctions according to the observed numbers. You may consider the result as a rank distribution; in this case find the theoretical probability distribution and test for goodness-of-fit. You may also use a function as model, a solution which is advantageous if you have only a few conjunctions in a category and therefore too few degrees of freedom.

References

Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.
Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English*. London: Longman.

1.8. Anaphoric distance

Problems

A measure of mental activity or accessibility of a referent in a text was introduced in several works by Givón (e.g. 1983, 2001). He used it, among others, to determine the topicality of agents and patients in clauses. For his purposes, he simplified the quantitative results via a threshold and formed a dichotomic variable topical/non-topical. However, anaphoric distance can, of course, be used as a fully metric quantity to characterize phrases, clauses, and even texts.

a) Determine the mean anaphoric distance for each of the referents in a text and find its frequency distribution; i.e. how many co-references are there in the text with anaphoric distance 0,1,2,…,*n*?
b) If the frequency distribution of distances is not random (cf. Zörnig 1984a,b; 1987), find an adequate distribution and substantiate it linguistically.
c) How can the significance of small or large anaphoric distances be tested?

d) Determine the mean anaphoric distance in texts and characterize each text with respect to the dimensions introduced in Ziegler and Altmann (2002): A text can be more or less diffuse, focused, dense, etc. Thus, a text with a significantly large mean anaphoric distance will appear more diffuse than one with a small anaphoric distance. The results can be used for text comparison and classification.

Procedure

Anaphoric distance can be measured as the number of phrases between the current clause in which a reference occurs and the last previous occurrence of the same referent. In the 'text' (1),

(1) In *a small garden* in *our neighborhood* lived *a pony*. *The animal* was fed *each day* by a *little girl. I* could observe *the garden* through *my window*. *She* seemed to live *far away from* *the place*.

the phrases which refer to something are highlighted. *the place* is double-marked because it refers to the garden and is also part of the reference to another place *far from this one.*

The first co-reference occurs with *the animal*, which co-refers with *a pony*. The anaphoric distance is zero; there is not a single clause between these two references. The distance between *the garden* and its previous co-reference *a small garden* is one clause. The same distance can be counted between *she* and *a little girl*. Of course, the distance can be measured also by the number of steps necessary for reaching the reference. In that case, between *pony* and *the animal* there would be the distance 1.

Problem (a) can be solved simply by stating all anaphoric distances in the text and computing their mean. Problem (b) can be considered either as an urn process with attracting urns yielding the negative binomial distribution. However, if one wants to use a simple function, then the Zipf-Alekseev function would yield the best results. It is bases on the logarithmic increase of distances and can easily be derived from an interpreted differential equation. Problem (c) is solvable for the difference of two distance distributions. Either ones tests the difference between means or the homogeneity of the empirical distributions. Problem (d) may be at least graphically solved if one computes for each text Ord's criterion (Ord 1972) and presents the results in an <I,S> coordinate system. The diffusity or density of the text can be described by some indicators such as the repeat rate or entropy, but <I,S> gives a sufficient possibility to classify the texts.

References

Givón, T. (1983). Topic continuity in discourse: an introduction. In: Givón, T. (ed.), *Topic continuity in discourse, a quantitative cross-language study: 5-37*. Amsterdam-Philadelphia: J Benjamins.

Givón, T. (2001*). Syntax. An introduction. Vol. II.* Amsterdam: J. Benjamins.

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

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

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

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

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

1.9. Cataphoric persistence

Problems

This measure of topicality of a referent in a text was introduced in several works by Givón e.g. 1983, 2001). He used it, among others, to determine the topicality of agents and patients in clauses in connection with questions of de-transitivating constructions. For his purposes, he simplified the quantitative results via a threshold and formed a dichotomic variable topical/non-topical. However, cataphoric persistence (CP) can, of course, be used as a fully metric quantity to characterize phrases, clauses, and even texts.

e) Determine the mean CP for each of the referents in a text and find its frequency distribution; i.e. how many co-references are there in the text with CP = 0,1,2, ..,*n*? If possible, set up a model of this distribution.
f) How can the significance of small or large CP be tested?
g) Determine the mean CP in texts and characterize each text with respect to the dimensions introduced in Ziegler and Altmann (2002): A text can be more or less diffuse, focused, dense etc. Thus, a text with a significantly large mean CP will appear more focused than one with a small CP. The results can be used for text comparison and classification.

Procedure

First read the problem *1.8. Anaphoric distance*.

CP can be measured as the number of occurrences of a referent within the next 10 sentences following the given occurrence. In 'text' (1),

(2) In *a small garden* in *our neighborhood* lived *a pony*. *The animal* was fed *each day* by a *little girl. I* could observe *the garden* through *my window*. *She* seemed to live *far away from* <u>*the place*</u>.

the phrases which refer to an individual person, object, space, or time are highlighted. <u>*the place*</u> is double-marked because it refers to the garden and is also part of the reference to another place far from this one.

The first referring phrase is *a small garden*. Provided the text ends with the example (1) so that we do not have 10 sentences, its CP is two: the first recurrence (*the garden*) is found in the third sentence, the second (*the place*) in the next one. For each phrase define its cataphoric mean in relation to the text length. Needless to say, there will be many phrases with cataphoric mean zero.

Compare text-books, poems, newspaper articles and scientific texts and show that they differ in this sense. Can the given problem be put in connection with Skinners principle of formal strengthening?

References

Givón, T. (1983). Topic continuity in discourse: an introduction. In: Givón, T. (ed.), *Topic continuity in discourse, a quantitative cross-language study: 5-37*. Amsterdam-Philadelphia: J Benjamins.

Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.

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

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

1.10. Causality in texts

Problem

In text or speech, causality can be expressed overtly or covertly. An overt expression contains some causal conjunctions or adverbs like *because, since, hence,* G. *weil, daher, weswegen,* covert expressions can have various forms. Expressions like

You cannot enter. The door is closed.
Heute ist schönes Wetter. Ich gehe spazieren.

express causality (or motivation) covertly.

Set up a scale expressing the strength of causality contained in texts. Analyze texts and set up hypotheses about text-sorts. Analyze especially stage plays and compare them with poetic texts.

Procedure

Analyse a text and mark all cases of causality, explicit or implicit. Set up groups of causal expressions. Scale them. This is not a simple task. In the first step, order simply the detected cases and find a linguistic or non-linguistic substantiation for this order. Several expressions can have the same "strength". Then ascribe the given degrees to the respective places in texts and compute the distribution of causality expressions.

Characterize the distribution using usual indicators.

Compare texts and group them in classes.

Establish hypotheses concerning the expressed causality and some other properties of texts.

Is causality stronger in spoken speech than in written texts? Perform a significance test.

Set up the sequence of causal expressions as they occurred in text using abbreviations or degrees and study this sequence. If you use abbreviations, find the R-motifs and their distribution, distances between equal expressions and the distribution of distances. If you use degrees, consider the result as time series and perform some characterizations (distribution, distances, autocorrelation, Fourier analysis, Hurst exponent, etc.). In stage plays, study the change in individual scenes from the beginning to the end.

References

Au, T. K. (1986). A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory & Language 25, 104-122.*

Brown, R., Fish, D. (1983). The psychological causality implicit in language. *Cognition 14, 237–273*.

Brown, R., Fish, D. (1983). Are there universal schemas of psychological causality? *Archives de Psychologie 51, 145-153.*

Burwick, F. (1982). The language of causality in "Prometheus Unbound". *Keats-Shelley Journal 31, 36-158*.

Corrigan, R. (1992). The relationship between causal attributions and judgments of the typicality of events described by sentences. *British Journal of Social Psychology 31, 351–368*.

Corrigan, R. (1993). Causal attributions to states and events described by different classes of verbs. *British Journal of Social Psychology 32, 335–348*.

Corrigan, R. (2001). Implicit causality in language: Event participants and their interactions. *Journal of Language & Social Psychology 20, 285-320.*

Girdenienė, S. (2001). *Syntaktische Ausdrucksmittel der Kausalität im Deutschen, Litauischen und Russischen der Gegenwart*. Diss. Vilnius.

Girke, W. (1979). Motivationelle Restriktionen. *Slavistische Linguistik 1978, 17-38.*

Girke, W. (ed.) (1999). *Aspekte der Kausalität im Slavischen.* München: Sagner.

Goikoetxea, E., Pascual, G., Acha, J. (2008). Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods 40(3), 760-772.*

Greene, S.B., McKoon, G. (1995). Can't know: Experimental approaches to verbs exhibiting implicit causality. *Psychological Science, 6,* 262–270.

Henschelmann, K. (1977). *Kausalität in Satz und Text.* Heidelberg: C. Winter Verlag.

Hermodsson, L. (1978). *Semantische Strukturen der Satzgefüge im kausalen und konditionalen Bereich*. Stockholm: Almqvist Wiksel.

Kang, Ch. (1996). *Die sogenannten Kausalsätze des Deutschen*. Münster: Waxmann.

Kasof, J., Lee, J.Y. (1993). Implicit causality as implicit salience. *Journal of Personality and Social Psychology 65, 877–891.*

Kelley, H.H. (1972). *Causal schemata and the attribution process.* Morristown: General Learning Press.

Küper, Ch. (1989). Die Leistung der kausalen Satzverknüpfer für Textkonstitution und Erzählperspektive. In: Weydt, H. (ed.), *Sprechen mit Partikeln: 488-497.* Berlin: de Gruyter.

LaFrance, M., Brownell, H., Hahn, E. (1997). Interpersonal verbs, gender, and implicit causality. *Social Psychology Quarterly 60, 138–152.*

Mannetti, L., De Grada, E. (1991). Interpersonal verbs: Implicit causality of action verbs and contextual factors. *European Journal of Social Psychology 21, 429-443.*

McArthur, L.Z. (1972). The how and what of why: Some determinants and consequences of causal attributions. *Journal of Personality and Social Psychology* 22, 171–188.

McKoon, G., Greene, S.B., Ratcliff, R. (1993). Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory, & Cognition 19, 1040-1052.*

Michajlov, M.N. (1993). K voprosu o kosvennych sredstvach vyraženija kauzal'nych otnošenij. In: *Kauzal'nost' i struktury rozsuždenij v russkom jazyke: 49-53.* Moskva.

Novikova, N.L. (1989). Leksičeskie i sintaksičeskie sredstva vyraženija kauzal'noj implikacii v vyraženii. In: *Leksičeskaja i sintaksičeskaja semantika: 16-23*. Saransk.

Patzke, U. (2005). Implizite Kausalität. *Anzeige für Slavische Philologie 33, 151-162.*

Rudolph, E. (1976). Zusammenhänge von Kausalität und kausalen Satzgefügen. *Deutsche Sprache 3, 193-206.*

Rudolph, E. (1979). Zur Austauschbarkeit von Kausalsätzen mit Kausalphrasen. In: Van de Velde, M. (ed.), *Sprachstruktur, Individuum und Gesellschaft: 123-132.* Tübingen: Niemeyer..

Rudolph, U. (1997). Implicit verb causality: Verbal schemas and covariation information. *Journal of Language and Social Psychology 16, 132–158.*

Rudolph, U. (2008). Covariation, causality, and language developing a causal structure of the social world. *Social Psychology 39(3), 174–181.*

Rudolph, U., Försterling, F. (1997). The psychological causality implicit in verbs: A review. *Psychological Bulletin 121, 192-218.*

Rudolph, U., Försterling, F. (1997). Zur impliziten Kausalität in Sprache: Kriterien zur Selektion von Stimulusmaterial in Studien zur Verbkausalität. *Zeitschrift für Experimentelle Psychologie 44, 293–304.*

Rudolph, U., von Hecker, U. (1997). Die Erklärung interpersonaler Ereignisse: Zur Bedeutung von Balanciertheit und Kausalität. *Zeitschrift für Experimentelle Psychologie 44, 246–265.*

Schmidthauser, B. (1995). *Kausalität als linguistische Kategorie.* Tübingen: Niemeyer.

Stewart, A.J., Pickering, M.J., Sanford, A.J. (2000). The time course of the influ--ence of implicit causality information: Focusing versus integration accounts. *Journal of Memory & Language 42, 423-443.*

Stroyny, K. (1997). *Die Entwicklung des Ausdrucks von Kausalität im Spanischen.* Frankfurt a. Main: Lang.

Weiss, D. (1993). Begründungserwartung und implizite Kausalität. *Slavistische Linguistik 1981, 234-262.*

1.11. Coherence/Cohesion of conjunctions

Problem

First read the problem 1.7. *Coherence, cohesion, and thematic continuity* for orientation. Then use a standard textbook of a grammar and make a list of all simple conjunctions. Quantify the force with which these conjunctions join (a) two objects, (b) two clauses, (c) two sentences. Use the scale for text characterization.

Procedure

(i) First analyze the force of a conjunction joining two objects/persons and set up an ordinal scale. Substantiate your decisions linguistically. Then do the same separately for (b) and (c). The position of a conjunction in (a) may be different from its position in (b), etc. Two different conjunctions can have the same linking force.

(ii) Construct a scale containing the averages of each conjunction. This scale will not be ordinal but continuous.

Then analyse a text using the result in (i), and separately in (ii). Set up the distribution of joining forces and find a discrete and continuous distribution respectively. It is sufficient to find an appropriate simple function.

Analyse several texts and compare them. Is the joining force of conjunctions linked with other properties of the texts? Differentiate text sorts. Differentiate the same text sort in two languages.

Set up a contingency table for the given text. The variables are X = force degrees, Y = joined entities (two words, two clauses, two sentences). Write in the cells the frequencies observed in a given text. (1) Test for independence of the two classifications using the chi-square test; (2) test each cell separately for prominence or avoidance using the normal test or the chi-square test.

Generalize your results on the basis of an analysis of (at least one) non-Indo-European language.

References

Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.
Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English*. London: Longman.
Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: OUP.

1.12. Degrees of Finiteness

Problem

Finiteness is a dichotomic concept in traditional linguistics; verbs have, in this view, finite and infinite forms. Cross-linguistic studies show, however, that languages tend to more or less finite clause types. Some languages such as Tibeto-Burman, Turkic, Carib, Quechuan, some Papua and Uto-Aztecan languages can be characterized as extremely nominalizing while e.g., Iroquois, Arawak, Atabashkan and many others are extremely finite, i.e. all clause types are finite (cf. Givón 2001: 25ff and the references therein). Most languages seem to be mixed types somewhere in between.

A rank scale of finiteness can be set up as shown in Table 1 (Givón, ibid.):

Table 1
Finiteness scale

	most finite
1	Her good knowledge of math helped
2	Her knowing math well helped

3	For her to know math so well surely …
4	She wanted to know math well
5	Having known math well since highschool, she …
6	She should have known math well
	least finite

e) Define a measure of finiteness of a language which can be used for cross-linguistic studies.
f) Determine the distribution of finiteness of languages in a large typological sample.
g) Find the probability distributions of clause types (i.e. of the degree of finiteness) within individual languages on a sufficiently large number of texts.

Procedure

The problems can be approached in two ways: 1. on the basis of the inventory of clause types according to a scientific grammar; 2. on data from large text corpora. If material for both ways is available, do both variants.

Since not all grammars describe finiteness, analyze each sentence in a long text. Classify the individual sentences according to kinds of finiteness which may be quite different from English. Order the sentences in different groups. At last, scale the groups and compute the numbers of sentences in individual groups (degrees). Then state the distribution of these degrees and search for a theoretical distribution. At last analyze the distribution and substantiate it linguistically.

Reference

Givón, T. (2001). *Syntax. An Introduction. Vol. II.* Amsterdam: Benjamins.

1.13. Study of POS bigrams

Hypotheses

Bigrams of parts of speech have very characteristic rank distributions. Find the distribution and test it. Finally, derive the distribution based on linguistic arguments.

Procedure

Extract the POS from a sufficiently large text as a sequence of POS symbols (tags). Count the individual POS and find their rank-frequency distribution or

simply a ranked sequence of frequencies. Then count the frequency of bigrams of POS and show the properties of this sequence. For bigrams, do not surpass the ends of the sentences! Compute some properties of the simple sequences (monograms) of POS (e.g. entropy, repeat rate, roughness, Ord's criterion, excess, etc.) and compare them with those of the bigrams. If something changed, continue with trigrams,…, n-grams and study the development of the above properties.

As a by-product, find a method for discerning POS-n-gram collocations, i.e. the strongest association of n-grams.

Compare texts in different language types. How does a strongly synthetic language differ from a strongly analytic one? For comparison use the above mentioned properties.

Apply several POS definitions (tagsets).

Is it possible to distinguish text-sorts according to POS bigrams? Since in some languages there is a very strict control of POS sequences, it is better to consider for this purpose also trigrams.

References

Bekkerman, R. El-Yaniv, R., Tishby, N., Winter, Y. (2003). Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research 3, 1183–1208.*

Collins, M. (1966). A new statistical parser based on bigram lexical dependencies. In: *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics, Santa Cruz, CA. 184-191.*

Damerau, F.J. (1971). *Markov Models and Linguistic Theory*. The Hague: Mouton.

Jones, M.N., Mewhort, D.J.K. (2004-08). Case-sensitive letter and bigram frequency counts from large-scale English corpora. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society 36(3), 388–396*

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law - another view. *Quality and Quantity 44(4), 713-731.*

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2), *257–286*. doi:10.1109/5.18626. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf.

1.14. Position of function words in sentence

Hypothesis

"*There is a decrease of function words and an increase of content words from the beginning to the end of sentence*" (Fenk-Oczlon, Fenk 2002a,b; Müller 2004). Test the hypothesis.

Procedure

First state exactly which words in a given language should be considered as function resp. content words. Usually, on the basis of traditional grammar, function words are: pronouns, conjunctions, prepositions, postpositions, interjections, negations, auxiliary verbs; the content words are: nouns, verbs, adjectives, adverbs, numerals. But even here there are (language-specific) problems which must be decided by the linguist, e.g. the German detachable prefixes which are mostly considered adverbs, if detached; prepositions written together with the noun, etc.

Then group sentences of equal length (= equal number of words) and compute the proportion of function words in individual positions. If the hypothesis is correct, you should obtain a significantly decreasing trend. If so, set up a model of this trend (some type of regression) and analyse languages of different types. Avoid poetic texts. Compare your results with those of B. Müller (2004).

References

Müller, B. (2004), *Die statistische Verteilung von Wortklassen und Wortlängen in lateinischen, italienischen und französischen Sätzen*. Diss. Klagenfurt 2004

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

Fenk, A., Fenk-Oczlon, G. (2002b). Funktions- und Inhaltswörter in der statistischen Binnenstruktur von Sätzen. *Paper presented at the 30. Österreichische Linguistiktagung, December 6-8 in Innsbruck.*
Abstract, http://www.uibk.ac.at/c/c6/c604/abstract.html

1.15. Noun phrase

Problem

Set up a classification of noun phrases using whatever type of grammar, i.e. set up a list of different kinds of noun phrases. Do not mix the basic criteria (e.g. semantic, morphological, speech act properties, length, etc.).

Characterize texts by means of a vector of frequencies of the elements in individual sets.

Compare texts.

Show the interrelations among the vectors. If possible, set up a control cycle of these properties (or whole vectors).

Perform a scaling of classes.

Procedure

Classify the noun phrases in a given language on the basis of some well known works, (e.g. Givón 2001; Fox 1987; Cole, Morgan 1975; Rijkhoff 2004; Gunkel, Zifonun 2009) and apply one of the classification schemes. Annotate the noun phrases in a text and extract the resulting tags as a sequence of symbols. Then compute the frequency of each class and set up the vector of frequencies. First rank the frequencies and find a theoretical distribution capturing it.

Use another classification and perform the first step with the same text. Since you already have the sequence of noun phrases, each type of the first classification (e.g. [a,b,c,…]) obtains a value of the second classification. Thus the second classification yields a matrix

a1, b1, c1, …
a2, b2, c2, …
a3, b3, c3, …
……..

State whether there is some link between the first and second classification. Are the classes of the second classification uniformly distributed or are there some associations between the classes. Test each cell separately for significance.

Repeat your analysis with another text. Compare the two texts in different ways (their rank-frequencies, the strength of correlation, chi-square, etc.) and make a statement about the difference. Use some indicators like moments, Ord's <I,S>, skewness and excess, etc.

Consider the first classification and the sequence of noun phrases. Compute the transition frequencies from one type to another and set up a matrix of transitions (with relative frequencies). Compute the order of the Markov chain.

Do the same with the second text and compare the orders of the Markov chains.

References

Cole, P., Morgan, J. (eds.) (1975). *Speech acts, syntax and semantics*. New York: Academic Press.

Fox, B. (1987). The noun phrase accessibility hierarchy revisited. *Language* 63(4).

Givón, T. (2001). *Syntax I, II.* Amsterdam-Philadelphia: Benjamins.
Gunkel, L., Zifonun, G. (2009). Classifying modifiers in common names. *Word Structure 2 (2), 205-218.*
Halliday, M.A.K. (2004). *Introduction to functional grammar*, 3rd ed, London, Hodder Arnold,
Rijkhoff, J. (2004). *The Noun Phrase.* Oxford: Oxford University Press.
Zifonun, G., Hoffmann, L., Strecker, B. (1997). *Grammatik der deutschen Sprache*. Berlin: de Gruyter.

1.16. Length of R-motifs

Problem

Motifs in the sense of Köhler, Naumann (2010) were originally defined on the basis of numerical (or at least ordinal) variables such as frequency, length, polysemy, or polytextuality of linguistic units. There are, however, also categorical variables in linguistics which resist (temporarily) metrification. An example of such a variable is the type of relation in argumentation structures, e.g. *justification, elaboration, concession, circumstance,* etc. or speech acts of different kind, or parts of speech, etc. Motif studies are motivated by the wish to investigate texts with quantitative methods not only with respect to unordered sets of elements (such as vocabularies and other inventories) but also with respect to the sequences of linguistic elements. Therefore, the analysis of the syntagmatic dimension of argumentation elements in texts seems to be worthwhile as well. There are several ways to form motifs from categorical data without scaling them. One of the possibilities is the following definition:

An R-motif is an uninterrupted sequence of unrepeated elements.

An example of the segmentation of a text fragment (represented as a sequence of argumentative relations) into R-motifs is the following:

["elaboration"], ["elaboration", "concession"], ["elaboration", "evidence", "list", "preparation", "evaluation", "concession"], ["evidence", "elaboration", "evaluation]

The first R-motif consists of a single element because the following relation is a repetition of the first; the second one ends also where one of its elements would occur again etc.

State the character of motifs that are known or establish some new classes and study their sequential, motif-like properties.

Procedure

Restrict yourself only to the expressions of the selected class and form a corresponding sequence in the order in which these expressions appear in the text under analysis. Replace these elements by some abbreviations. Ignore everything else in the text.

Now mark the boundaries of R-motifs according to the above principle: a new R-motif begins at the first repetition of an element from the preceding R-motif.

Having prepared the data, evaluate as many (quantitative) properties of the R-motifs as possible. First set up the frequency distribution of their lengths and compute its properties (mean, variance, repeat rate, entropy, Ord's criteria, asymmetry, peakedness); for sequences of R-motifs find autocorrelation, Hurst's exponent, Lyapunov's coefficient, or other properties of time series. Find Markov dependencies, the order of the Markov chain and Hidden Markov chains.

Find models of the above result. Consult the references given below.

References

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I.. Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics. Selected Papers of the 8th International Conference on Quantitative Linguisttics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 35-43*. Belgrade: Academic Mind**.**

Köhler, Reinhard (2014, to appear). Linguistic Motifs. In: Mačutek, J., Mikros, G. (eds), *Sequential Analysis*. Berlin, New York: de Gruyter.

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 Academic Press.

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

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

Köhler, Reinhard (2014, to appear). Linguistic Motifs. In: Mačutek, J., Mikros,G. (eds). *Sequential Analysis*. Berlin, New York: de Gruyter.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-sequences. In: Preisach, C., Burghardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data analysis, machine learning and applications: 637-646*. 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.* Vienna: Praesens.

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

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

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

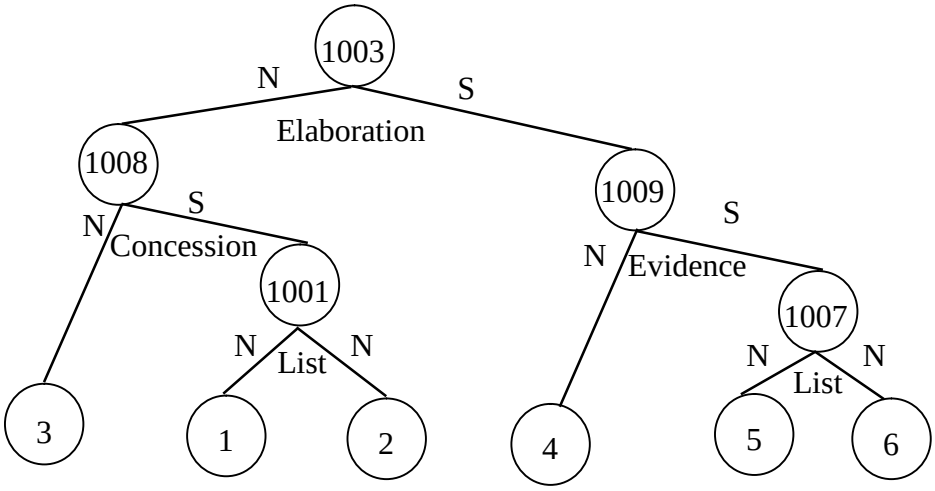
1.17. Length of D-motifs

Problem

Another way of defining motifs on the basis of categorical data is:

> *A D-motif is an uninterrupted depth-first path of elements in a tree structure.*

In a tree structure such as in the following diagram,



the node sequence [1003, 1008, 3] forms a D-motif. Another one is [1003, 1008, 1001,1] etc. Whether the node numbers are replaced by the determiners of the argument types or not does not play any role for the properties of the resulting motifs, of course.

The length of motifs determined in this way displays a behavior that differs considerable from that of the R-motifs. A linguistically interpretable theoretical probability distribution which can be fitted to the empirical frequency dis-

tribution is the mixed negative binomial distribution (cf. Fig. 2 and Table 2). The example was taken from Beliankou, Köhler, Naumann (2013), where German newspaper commentaries were analysed.

In the same way, any kind of tree structure consisitng of categorical data can be transformed into D-motifs.

Procedure

Form D-motifs on the basis of any kind of categorical data which are structured as trees, such as syntactic construction types or hierarchical semantic relations. Find an appropriate distribution as a model of the frequencies of the motif length and calculate the statistical properties of the distributions for several texts. Proceed in analogy to the preceeding problem (R-motifs).

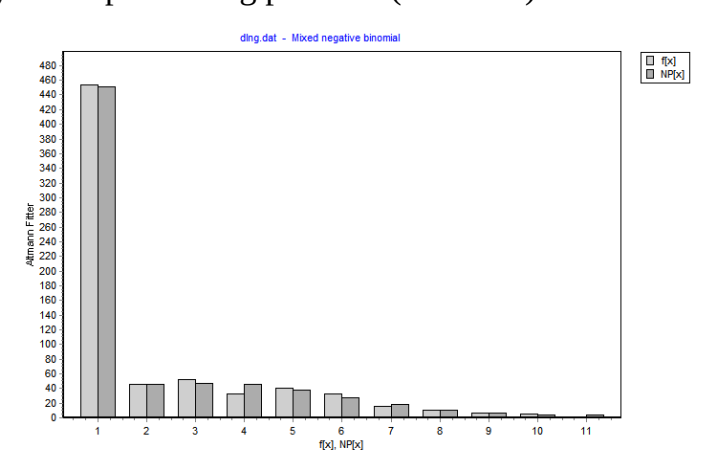


Fig. 2. Fitting the mixed negative binomial distribution to the D-motif data

Table 2
Result of fitting the mixed negative binomial distribution to the D-motif data using *Altmann-Fitter* (3.1)

Distribution	Mixed negative binomial (k,p1,p2,α)			
x[i]	f[i]	NP[i]		
1	454	451.32		
2	46	45.88		
3	52	46.29		
4	33	46.03		
5	40	38.04		
6	33	27.52		
7	16	18.03		
8	11	10.93		
9	6	6.22		

10	5	3.36		
11	1	3.39		
Parameters: k = 7.5963; p1 = 0.9965; p2 = 0.6880; α = 0.6434				
X^2 = 8.3176; P(X^2) = 0.2157; DF = 6; C = 0.0119; R^2 = 0.9985				
N = 697; m_1 = 2.2195; m_2 = 4.2804; m_3 = 15.5021; m_4 = 97.6665				
Ord I = 1.9285; Ord S = 3.6217; Skewness = 1.7505; Excess = 44.2069; Entropy = 0.5601; Repeat rate = 0.2932				

References

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

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 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 35-43.* Belgrade: Academic Mind**.**

1.18. Syntactic complexity

Problem

In Köhler (2012, 186ff), syntactic complexity was defined, on the basis of a constutient strcture, as the number of immediate constituents of a syntactic construction. A number of hypotheses concerning dependencies between complexity and other variables such as frequency, position in the mother constituent, depth of embedding was set up and tested on data from two languages (English and German).

(a) Test these hypotheses on data from other languages.

(b) Give an alternative definition of complexity. We can call Köhler's definition *'local complexity'* whereas a definition with respect to the number of all constituents in the tree under the given construction could be called *'global'*. Test the hypotheses on this basis.

(c) Give an alternative definition of complexity for the application to other representations of syntactic structure such as dependency grammar or categorical grammar.

(d) Test Köhler's hypotheses on data from tree-banks which are annotated with respect to various kinds of grammar.

Procedure

Read the chapter on synergetic modelling of syntactic properties and interrelations (2012, Chpt. 4.2, p. 169ff.). Evaluate texts or corpora with syntactic annotation. Find corresponding tree-banks from as many languages as possible.

(a) Analyse the data with respect to the frequency distribution of syntactic constructions found in the material.

(b) Determine the complexity, frequency, position, and depth values recursively for all the constituents in the data.

(c) Test the hypotheses and publish the results including the estimated parameters.

(d) Determine the frequency distributions of the evaluated variables complexity, frequency, position, and depth and publish the results including the estimated parameters.

Reference

Köhler, R. (2012). *Quantitative Syntax Analysis.* Berlin, Boston: W. de Gryuter.

1.19 Adnominal modifiers - 0. Classification

Problem

Best and Boschtan (2010) studied the diversification of simple attributes of nouns in sentences. They found several classes of attributes in German that can preliminarily be extended to the following ones (for German):

1. Adjective attribute preceding the noun
2. Adjective attribute following the noun
3. Participial attribute preceding the noun
4. Participial attribute following the noun
5. Apposition
6. Attributive sentence preceding the noun
7. Attributive sentence following the noun
8. Genitive attribute preceding the noun
9. Genitive attribute following the noun
10. Prepositional attribute preceding the noun

11. Prepositional attribute following the noun
12. Attributive infinitive with “zu” preceding the noun
13. Attributive infinitive with “zu” following the noun
14. Compositional attribute according to its position in the compound

The above set is merely a subset of the set of *adnominal modifiers* studied thoroughly in qualitative linguistics (cf. Gunkel, Zifonun 2009; Halliday 2004; Rijkhoff 2004; Givón 2001).

Some of these classes may occur also in other languages while still other classes may be observed in a language under study. Find linguistic criteria for arranging the classes or groups (intervals) of classes in the given language along an ordinal (rank) scale. If linguistic criteria which could be applied to all of the classes cannot be found, you might order them according to a frequency ranking. Perform the scaling and compute the frequencies of classes or groups of classes (intervals). Differentiate text sorts according to these classes (degrees of your scales). Differentiate languages on the basis of this scaling.

Procedure

Finding criteria for distinguishing these classes and for ordering them on a scale is the most pretentious task. Hence, quantification is not easy but not impossible. As appropriate criteria could be employed e.g., relevance or degree of a modification for the modified head (e.g. 'attributive' vs. 'restrictive'), conceptual affinity between the meanings of head and modifier, decoding ease of determining the meaning of the adnominal construction, the probability of common occurrence in a corpus, degree of semantic combinability, the requirements of the speaker and hearer, etc.

If this part is successfully accomplished, the solution of the other problems is easy. At last, using all possible reasons and forces which work in language, set up a basis for a construction of a theory of adnominal constructions Not all steps will succeed at the first trial on the basis of merely one language; hence extend the investigation to several languages just from the beginning.

Since the problem has not been investigated as yet, you might adopt the principles of synergetic linguistics (cf. Köhler 2005).

The problem is only a special case of a more general problem of predication.

You may introduce a scaling also with the aid of graphs representing the noun and its attributes, namely by using the properties of the graphs.

Begin with analyses of individual texts in one language and then extend your research.

Best and Boschtan (2010) presented the frequencies of individual classes and fitted a function to them. Use the frequencies of individual classes in the works of the three authors analyzed by them and ascribe them ranks. Then test the equality of texts (from this point of view) using some rank test for the three

pairs of texts and finally, use a multiple correlation test for the homogeneity of theses texts. If the result signals significant divergence, how would you characterize the individual texts?

Analyze not only German but also several other languages and set up an "attributive" od "adnominal" typology of texts.

References

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

Bortz, J., Lienert, G.A., Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

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

Givón, T. (2001). *Syntax II.* Santa Monica, CA: Systems development corporation.

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

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

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

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

1.20. Adnominal modifiers - 1. Weight

Problem

Adnominal modifiers or simply, adnominals, are those words, phrases and clauses which modify, amplify, reduce or make more exact the meaning of the given noun. The most common ones are adjectives. Show that texts, text-sorts, and languages differ in the distribution of the weight of adnominal modifiers.

Show that the older an author, the greater weight obtain the adnominal modifiers.

Procedure

The syntactic weight of a noun can be measured in terms of the number of all words that occur in the noun phrase, except for the noun itself. It can be mechanically stated from dependence representations.

In order to obtain the distribution of noun weight, count the number of elements in each noun phrase (skipping the noun). Frequently, a noun itself can be a modifier of another noun but can in turn have itself other modifiers. If a noun has not any modifier, it has weight zero. The distribution of weights may be characteristic not only of the given text but also of the given language. This could be scrutinized by analyzing the same text in several languages. If Busemann's hypothesis is correct, then even the development of the author can be traced down: the older the author, the greater become the syntactic weights of nouns because he diminishes the "activity" of the text and increases its "ornamentality" by means of adnominal modifiers.

Consider always at least a pair of texts, i.e. two poetic texts by the same author, or two texts of the given text sort, or the same language, in order to obtain at least an elementary picture of the situation. Then perform preliminary comparisons. The differences between two distributions can be tested in many ways. If you find a significant difference, perform a thorough investigation in many texts.

Examples:

G. *das auf dem Spielplatz spielende Kind* (*Kind*: 5; *Spiel* 0; *Platz*: 3)

E. *the child playing on the playground* (*child*: 5; *play*: 0; *ground* 3)

Hu. *a pályán játszó gyerek* (*gyerek*: 3; *pálya*: 1)

Slk. *dieťa hrajúce sa na hrišti* (*dieťa*: 4; *hrište*: 1)

yielding even for simple cases different weights. As can be seen, also compounds containing a noun can be evaluated.

References

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

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

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

Rijkhoff, J. (2004). *The noun phrase.* Oxford: Oxford University Press.

1.21. Adnominal modifiers - 2. Distribution

Problem

First consider the problem „Attribute quantification 1". State the different kinds of nominal modifiers in the language(s) you analyze, then set up sequences of modifiers for every text separately. Then solve the following problems:

1. Find a common distribution for the ranked-frequencies of all texts.

2. Compare the texts and state whether they are homogeneous.
3. Characterize the texts by using the repeat rate and the excess of the empirical distribution.
4. If you analyzed texts belonging to different text sorts, show the differences.
5. If you analyzed texts of one author, show the temporal change in his technique of placing adnominal modifiers.
6. In the individual sequences you obtained, state the distribution of distances between equal adnominal modifiers.
7. If you analyzed the translations of the same text in different languages (e.g. *Le petit prince*) show both qualitatively and quantitatively the differences between these languages.

Procedure

1. Find all adnominal modifiers, assign them to classes (categories) and state the numbers of entities in individual classes. Then rank them simply and find their distributions. Most probably you will find several well fitting distributions. Choose one of them and substantiate it linguistically. Alternatively, find a continuous function capturing the ranked frequencies and substantiate it using a differential equation.

2. Fix the modifier classes (no ranking), assign them the frequencies and rank the classes for each text separately. Then perform a non-parametric test for homogeneity of all texts, e.g. compute Kendall's W indicator. Since some frequencies will be very low, avoid the chi-square test for homogeneity.

3. Consider the frequencies and characterize the text, e.g. computing the repeat rate; or, for ranked frequencies, compute the excess or the arc length of the distribution. Test the differences between the texts using these indicators.

4. Show that texts belonging to the same text sort have a more similar modifier structure than those belonging to different ones. This can be done in several ways: (a) Compute an indicator for all texts, order them according to the value of the indicator and state whether all belonging to one text sort are similar. (b) First classify, then define the text sort. (b) Perform discrimination analysis using various indicators of modifiers.

5. Compute the indicators of individual texts of a given author and compute the correlation between indicator and year of origin of the text. Describe what changed.

6. Compute the distances between equal adnominal modifier classes and set up the distribution of distances. State the kind of distribution. In most cases, the so called Zipf-Alekseev function captures the data adequately. Fit this function to data (it can be transformed into a distribution by normalizing) and compare the parameters in texts belonging to the same text sort. It can be supposed that texts belonging to the same text sort will have more similar parameters. If you use a

ready-made software, you obtain automatically the variance of the parameters. Using it set up a normal test for the difference of the parameters in different texts. Alternatively, use the parameters as indicators and perform discrimination analysis. State whether different text sorts have different sets of parameters.

7. If you analyzed texts from several languages, set up for every language a mean rank for each attribute class. Using these means compare two languages using any appropriate test. Describe the differences.

References

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

1.22. Adnominal modifiers - 3. Complexity

Problem

Adnominal modifiers or simply adnominals are those entities which modify, describe, explain the topical meaning of the given noun. The adnominal may be represented by (i) a part of the word (affixes, non-head components of compounds), (ii) separate words, (iii) phrases, and (iv) clauses. Devise a method for measuring the complexity of an adnominal.

Procedure

First define "complexity". Do you consider, for your study, complexity as a formal, morphological, syntactic, semantic or a mixed property?

Then one can proceed "top down", i.e. take into account all descriptions of grammar in the given language beginning with word formation and ending with syntax; each different form must obtain its degree. It is to be decided whether morphology creates more complex forms than syntax. In German, we have the famous word "Donaudampfschiffartsgesellschaftskapitän" which is in English a phrase "captain of the Danube steam shipping company".

Or one proceeds "bottom up", i.e. and excerpts all nouns with their adnominals from a long text. Comparing and ordering them one obtains a sequence with increasing degree of complexity. Each place in this sequence may be occupied by several adnominals. One devises a way of ascribing the elements of the sequence a number. Later on, one can normalize the degree.

A third, important step would be the setting up the distribution of the degrees for a given text. After having analyzed several texts in one language, find the theoretical distribution of degrees, e.g. using the unified theory (Wimmer,

Altmann 2005) which should capture all data in all languages. The parameter values play the role of boundary conditions.

Further, elaborate on the characterization of this distribution, e.g. using means, variances, Ord's criteria, asymmetry, steepness, etc. Propose a test for comparing texts from this point of view.

Finally, link the characteristics of this distribution with other properties of language, e.g. synthetism (measured in various ways), mean word length, etc.

References

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

Givón, T. (2001). *Syntax II.* Santa Monica, CA: Systems development corporation.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 627-633.* Berlin-New York: de Gruyter.

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

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

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

Roelcke, Th. et al. (2014). *Adnominal modifiers*. In preparation.

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.

1.23. Adnominal modifiers - 4. Cohesion

Problem

The adnominal modifiers in the sentence can not only be classified (i.e. categorized) but the individual classes can be scaled according to the cohesion with the given noun. For example an adjectival attribute on the left hand side, e.g. "*nice* girl", is "nearer" to the noun than a relative sentence, e.g. "the boy *who played basketball*" because the content in the relative sentence is distributed over several words. This, in turn is nearer to the noun than an infinitive attribute, e.g. "the risk *to be killed*" because there is no reference.

Collect all classes of adnominal modifiers and scale them according to different aspects. Just as words or phrases, adnominal modifiers have a great

number of properties. Do not strive for finding "all" of them - this is impossible - but consider only one of them and study its behaviour. Read the other problems concerning adnominal modifiers.

Procedure

First look at the problems *Adnominal modifiers - 0. Classification* and *- 2. Distribution*, use the pertinent results and extend the class of adnominal modifiers in your language (cf. Roelcke et al. 2014). Then find a property which allows you to arrange the classes on a scale according to cohesion. Use the ordinal scale.

Then analyze texts and for every text set up the frequency distribution whose variable is the rank on your scale.

Now perform the "full program": (a) find the theoretical probability distribution for all results and apply it to individual texts; (b) use the empirical frequencies and compute for each text Ord's criterion; (c) present the texts graphically using Ord's criterion; (d) compare individual texts and test the differences; (e) compare text sorts; (f) compare languages.

Draw consequences from all these procedures and interpret them from a linguistic or the point of view of literary criticism. Show the typological impact of this kind of measurement. If possible, transfer the argumentation to the measurement of empirical valency of verbs.

References

See the other problems concerning adnominal modifiers.

1.24. Adnominal modifiers - 5. Scaling

Problem

Develop scaling procedures for the properties of adnominal modifiers, analyze several texts and set up hypotheses concerning the mutual relation of some pairs of properties. Test the hypotheses statistically.

Procedure

Consult the book by Givón (2001) in which one can find many classifications of adnominal modifiers. For each chapter develop a scaling procedure; since the book is very voluminous, restrict yourself to at least two chapters or two properties. Then analyze several texts (at least 10), find the degrees of the properties of adnominal modifiers in each text on the two scales and test whether the de-

grees are correlated. If this inductive procedure yields positive results, set up a more specified hypothesis in the form $y = f(x)$, where x and y are the two scaled properties.

Then proceed to the next property (z), state the degrees of the adnominal modifiers found and correlate the new results with the two scalings performed in the first step. You may search for $z = f(x)$ or $z = g(y)$ or $z = h(x,y)$.

Continue in this way and construct step by step a control cycle. Apply the results to texts in other languages and if you obtain positive results, formulate an elementary theory. Take boundary conditions into consideration.

Draw a diagram of the control cycle (cf. Köhler 2005) and add stepwise other properties and languages.

References

See the references with the other problems concerning adnominal modifiers.

Givón, T. (2001). *Syntax II.* Santa Monica, CA: Systems development corporation.

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.

1.25. Adnominal modifiers - 6. Motifs

Problem

Motifs are numerical non-decreasing sequences of some values (cf. Köhler 2006, 2008, Köhler, Naumann 2008, 2010). Considering the problems concerning adnominals (0 to 5), find the sequence of adnominals in a text, transcribe them in terms of some of their properties and form the motifs. Thereafter study the behaviour of the motifs, e.g. their distribution, length distribution, autocorrelation, range, etc. Find some regularities in their behaviour.

Procedure

Let A, B, C,… be some adnominals and x = 1,2,… be their length. Extract the adnominals in the order in which they occur in the text, i.e. create a sequence of adnominals. Replace the name of each adnominal (A,B,C,…) by its length (x = 1,2,3,…).

(1) Then underline the motifs and replace each motif by its length. State the distribution of these lengths and find its properties. Derive the distribution theoretically.

(2) Study the autocorrelation of motif lengths in the given text. Interpret the meaning of autocorrelation in linguistic terms.
(3) Study the range of the motifs. Each motif has the minimum value as its first element, the greatest value as its last one. For each motif compute the difference (last element minus first element). State the distribution of these ranges and compare some texts.
(4) Study the runs of motifs, or more correctly, the runs of their values. Is the number of runs too small or too large? Test the result.
(5) Having computed some of the above numbers, compare texts, text sorts and languages. Strive for a theoretical overview.

References

Beliankou, A., Köhler, R., Naumann, S. (2013). Quantitative properties of argumentation motifs. In: Obradović, I.. Kelih, E., Köhler, R. (eds.). *Methods and Applications of Quantitative Linguistics. Selected Papers of the 8th International Conference on Quantitative Linguisttics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 35-43.* Belgrade: Academic Mind**.**

Köhler, R. (2014, to appear). Linguistic Motifs. In: Mačutek, J., Mikros, G. (eds), *Sequential Analysis.* Berlin, New York: de Gruyter.

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 Academic Press.

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

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-sequences. In: Preisach, C., Burghardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data analysis, machine learning and applications: 637-646.* 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.* Vienna: Praesens.

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

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

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

Semantics

2.1. Polysemy in German

Problem

Find the distribution of polysemy of nouns, verbs and adjectives in German. The basic idea is that there exists a common probability distribution of polysemy for all the three part-of-speech classes. Find such a distribution using the *Altmann-Fitter* software. There are several sources of polysemy counts, among them Levickij, Drebet, Kijko (1999: 174f.).

Procedure

Take the data from the article by Levickij, Drebet, Kijko (1999: 174f.), apply the *Altmann-Fitter* software and find a distribution common to all these classes. If you find several well-fitting distributions ($P > 0.05$), substantiate deductively one of them or show that all belong to the same family of distributions. Give linguistic reasons leaning against the unified theory (Wimmer, Altmann 2005).

Since one can also use simple continuous functions (with or without normalizing them) and transform them into discrete distributions (cf. Mačutek, Altmann 2007) or series, apply a fitting software, e.g. *NLREG* or *TableCurves* or *Origin* or the statistics package *R* and find the appropriate function(s) (with $R^2 > 0.9$) with the smallest number of parameters. Substantiate linguistically such a function, find its differential equation and interpret the parameters as forces of speaker and hearer. Adopt the synergetic approach (Köhler 2005).

Acquire data from other languages and generalize both the problem and the modelling. Set up a family of functions or distributions because one cannot expect that one obtains in all languages the same result. Consider the size of the vocabulary as one of the factors influencing the choice of the function or its parameters, i.e. do not ignore boundary conditions.

If you obtained sufficiently large samples of nouns, adjectives and verbs and data concerning their polysemy, link polysemy with other linguistic properties of words. To this end use all problems published in the first three volumes of "Problems".

Apply the same method to other languages, compare all your results and draw consequences on the character of language.

References

Falkum, I.L. (2011). *The Semantics and Pragmatics of Polysemy: A Relevance-Theoretic Account.* London: University College, Diss.

Jastrzembski, J.E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology 13, 278–305.*

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.

Leacock, C. (ed.) (2000). *Polysemy: Theoretical and computational approaches*. Oxford: Oxford University Press.

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.*

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 100-109.*

Polikarpov, A.A. (1987). Polisemija: sistemno-kvantitativnye aspekty. In: *Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov. Učenye zapiski Tartuskogo Universiteta 744, 135-54.*

Schierholz, S.J. (1991). *Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive.* Tübingen: Niemeyer.

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

2.2. Polysemy in the Swadesh list

Problem

State the distribution of polysemy in the common stock of languages.

Procedure

Let us suppose that the words of the greater (200 words) Swadesh list form the daily vocabulary of a language. Consult a large dictionary of one language and state the number of words in the Swadesh list (f_x) with exactly x meanings (x = 1,2,…). Take into account also the meaning of the given words if they occur in compounds. If here the meaning is slightly changed, consider it as a component, e.g. the German "Mutter" (mother) can be found also as "Schraubenmutter" (*female screw, nut*), etc. Consider all cases and set up the distribution of the 200 words.

Then derive the distribution of the number of meanings on the basis of a birth-and-death process: new meanings of a word are steadily born but many of them are eliminated immediately or later on. If not possible, simply adopt a for-

mula from the literature in the references or the result of the application of a software package to your data and find some well fitting distributions. Find a linguistic foundation.

Analyze a second language and compare the results with the first one. Use either the chi-square test for homogeneity of the two distributions or consider each word separately, ascribe it the appropriate polysemy and compare the two data using e.g. a non-parametric test based on ranks.

Can we find arguments for language diversification in this domain - even if the Swadesh list would be insufficient or incorrect?

References

Andreevskaja, A.V. (1990). Kvantitativnoe issledovanie polisemii kornevych slov russkogo jazyka XI-XX vekov. *Učenye zapiski tartuskogo universiteta 912, 3-11.*

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 Quatitative Linguistics 6(2), 172-187.*

Levickij, V.V. (2005). Polysemie. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 458-464*. Berlin: de Gruyter.

Tuldava, J. (1979). O nekotorych kvantitativno-sistemnych charakteristikach polisemii. *Učenye zapiski TGU 520, 107-141*.

Wimmer, G., Altmann, G. (1999). Rozdelenie polysémie v maorijštine. In: Genzor, J., Ondrejovič, S. (eds.). *Pange lingua. Zborník na počesť Viktora Krupu: 17-25.* Bratislava: Veda.

2.3. The course of polysemy in sentences

Problem

The sequence of autosemantics in the sentence displays increasing/decreasing polysemy. Test which of the two alternatives holds.

Procedure

Word polysemy can be determined either including the often multiple grammatical meanings of function words or considering exclusively the lexical meanings of autosemantic words. Both methods yield reasonable but differing results. Decide which method you prefer for your current study.

Analyse a text sentence by sentence following one of the mentioned methods. Determine the numerical value of polysemy for each word. The most comfortable way is to set up a word list of the text under study and annotate the number of meanings of each word in this list. The easiest way to do that is consulting an appropriate electronic dictionary or a 'wordnet'; which can be done via an automated procedure if the text is available in lemmatized form. Next, you can annotate the running words in the text with the polysemy values in the list. For the next step, only the sequence of these values is needed.

Group the sentences with equal lengths in terms of the number of words. Then, for each group separately, state the *mean* polysemy in position 1, in position 2, etc. Investigate the course of polysemy and find a function expressing this course. Is it possible to use an identical function for all groups? Do the parameters change for groups of different lengths? If so, find another function capturing the change of parameters with increasing group length.

Perform the same operations with other texts and compare the results.

Do technical or scientific texts differ from novels?

Perform the same operations with a text in another language. Compare the results. Can you tell something about the difference between languages? Use for example the translations of Exupéry's novel *The Little Prince* and perform the investigation in form of team work.

It is important to take the mean polysemy in the given position into consideration, otherwise one would obtain very great variances and no trend. Our aim is to find some background mechanisms which control the semantic structure of sentences.

References

None.

2.4. Polysemic similarity and distance

Problem

Are sentences in short distances semantically more similar than those which are more distant?

Procedure

The problem can be solved in many different ways. For the first step, we propose the following procedure: For each sentence of a text set up the vector of the polysemies of individual words. Use a monolingual dictionary. The size of the

vectors is determined by the longest sentence of the text: if it contains e.g. 10 words, then all vectors must contain 10 elements. If a sentence is shorter than 10 words, all missing elements have the value 0. For example <6,1,5,7,2,0,0,0,0,0>.

Now compute the similarity of all pairs of sentences which are in immediate neighbourhood i.e. in distance 1. One can use any indicator of similarity, we propose here the radian of the angle between the two vectors defined as

$$d = \arccos\left(\frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} x_i^2}\sqrt{\sum_{i=1}^{k} y_i^2}} \right).$$

With k sentences, you will obtain k-1 similarity values. Compute the mean of these values and mark it as d_1. Now perform the same procedure for all sentence pairs which are in distance 2 and compute the mean of all the values (this time you get only 18 values). Continue in this way until there are at least five pairs of sentences.

Now you have a sequence of mean similarities for the individual distances. If you see a trend in the values, find an appropriate algebraic expression for it. If it can be positively tested, substantiate this fact linguistically or psychologically and set up a corresponding differential equation.

Reformulate your hypothesis in that sense that you conjecture that "with increasing distance between sentences then mean polysemic similarity increases/decreases/does not change".

Perform the computation for several texts and if possible, for several languages.

References

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

2.5. Polysemic text construction

Problem

The sequence presented below represents the polysemy of the first individual words in Pushkin's poem Mednyj vsadnik (including function words). Set up sequences of polysemies of this kind for several texts and study their properties.

[8,2,2,7,23,3,4,5,7,8,1,8,7,3,2,3,5,3,2,19,3,4,3,19,2,1,2,2,3,2,2,6,1,8,3,1,4,12,6,3,
6,3,3,8,5,3,1,4,3,12,1,3,12,3,9,11,3,1,1,3,5,3,3,1,12,3,3,3,2,9,12,12,5,2,19,…]

Procedure

We will mention here merely a part of the enormous number of problems. They are all easily computable and should be performed on other texts, too.

(1) Compute the sequence of autocorrelations of individual lags. Study the possibility of the existence of a trend and express it formally.

(2) Compute the Hurst exponent, interpret it and compute with its help the fractal dimension of the sequence. (Cf. the Problem: Hurst exponent)

(3) Compute the sequence of mean arc lengths of the given sequence and study its properties.

(4) Set up the distribution of individual polysemy values and find a discrete distribution capturing it adequately. Substantiate the distribution linguistically.

(5) Compute the entropy of the resulting empirical distribution. Then set up the distribution of pairs of neighbouring polysemy values (bigrams) and compute its entropy. Continue up to decagrams and follow the course of entropy change from 1 to 10. Express the change of entropy by a function.

(6) For all ten distributions compute Ord's criteria I and S and plot <I,S>. State the course of this sequence.

(7) Study the transition frequencies from one number to the next one. Since the greatest polysemy in Pushkin's text is 30, prepare a 30 x 30 matrix of transitions. Continue computing the transitions of second, third etc. order. At last, state the order of the Markov chain. Does this result hold for other texts, too? Or for other languages?

(8) Set up the Minkowski sausage of polysemy values and find a function expressing the decrease of breaks with increasing radius.

References

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

Brainerd, B. (1976). On the Markov nature of the text. *Linguistics 176, 5-30.*

Gottman, J.M., Roy, A.K. (1989). *Sequential analysis. A guide for behavioral researchers.* Cambridge: Cambridge University Press.

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin

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

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

2.6. Vectors of polysemy

Problem

For each word of a text state its polysemy (x = 1,2,…) using a very good dictionary or a semantically annotated corpus. You obtain a vector whose elements are numbers, as shown in the problems *The course of polysemy in sentence*, *Polysemic similarity of sentences*, *Polysemic text construction*. Now for each of the words set up a separate vector consisting of individual semantic categories which may be also graduated. For example: concrete-abstract, specific-general, direct-metaphoric, etc. You may use also grammatical categories, parts-of-speech etc. There is an enormous number of classifications of this kind. On the basis of the obtained succession of vectors compute the course of similarity and express it formally.

Procedure

First, study several works concerning semantics and elaborate on possible meaning classes. You may (should) restrict the number of elements because probably every linguist has a different classification. In case of necessity construct different vectors. The individual elements of the vector are numbers, e.g. concrete = 0, abstract = 1, from *specific* to *general* there are degrees represented e.g. by definition chains, e.g. John(0) - human(1) - animal(2) - organism(3) - system(4) - …

Now you have a series of vectors. The number of elements in each vector should be the same. If a given category/class is not present, evaluate it as 0. Now compare the similarity of the vectors using whatever similarity measure. The simplest measure is the cosine of the angle between the vectors. To obtain similarity compute the arccos function: the smaller the radian, the greater the similarity.

You may, of course, abbreviate semantic categories by letters (a = abstract, c = concrete). There is no problem if some letters are used for different categories, because in vectors also the position of the element is relevant. But in this case you should use another indicator of similarity. The number of available measures is overwhelming, it is better if you use several different ones.

Now study the course of semantic similarity among neighbouring words ($S_{i,i+1}$). Express the course of similarity by a function. Study its properties. If it is a simple function, can you detect a regular oscillation? If it is an irregular fractal, what is the fractal dimension of the sequence?

Next, compute the mean of all neighbouring similarities and continue computing the similarities $S_{i,i+2}$ for all i and take their mean. Continue with the means of (i,i+3), (i,i+4),… That is, you obtain a sequence of mean semantic similarities of words in distance d = 1,2,3,… It is sufficient to continue up to distance 20 in long texts, or 10 in short texts. Find a function expressing this course of similarities.

On the basis of these initial data, substantiate the functions theoretically. Derive them from linguistic assumptions or from the existing theories taking into account Zipfian and other forces.

Are there differences between text-sorts and languages?

References

All works concerning polysemy.

2.7. Synonymy

Problem

Levickij and Wenhrynowytsch (2009) published data on synonymy concerning semantic word classes in German. Find the discrete distribution which is adequate for capturing all classes.

Procedure

In their Table 2 (2009: 76 f.) the authors show the number of synonyms of words in 24 semantic classes. Consider the numbers in each row separately and if necessary pool some rows. Each pooling must be substantiated linguistically. The authors pooled e.g. natural phenomena and diseases. The last row contains the sums of columns. First find a distribution for this last row and apply it to individual rows. The last row contains the following numbers:

Number of synonyms						
1	2	3	4	5	6	7 and more
27810	2948	1050	756	581	472	2217

As can be seen, the last class is a very strong pooling of at least 5 classes.

Consult dictionaries of synonyms in other languages, do not pool the classes with 7 or more synonyms but take all numbers separately and compare the result with German. You can use another classification of nouns - stick to the national tradition.

(1) Elaborate on the possibility of substantiating the discovered regularity linguistically.

(2) Derive the distribution theoretically or subsume it under the unified theory.

(3) Do not distinguish classes of nouns, count all synonyms. You need not analyze the complete dictionary, consider only one or more initial letters.
(4) Departing from your results, you can easily solve some other problems concerning synonymy as presented in the first three volumes of "Problems in Quantitative Linguistics".

References

Bulitta, E., Bulitta, H. (1993) *Das Krüger Lexikon der Synonyme*. Fankfurt: Fischer.
Cruse, D.A. (1987). *Lexical Semantics.* Cambridge: Cambridge University Press.
Levickij, V.V., Wenhrynowytsch, A.A. (2009). Quantitative Charakteristika der substantivischen Synonymie im heutigen Deutsch. *Glottotheory 1(2), 75-85.*
Lyons, J. (1995). *Linguistic Semantics. An Introduction.* Cambridge: Cambridge University Press.
Murphy, M.L. (2003). *Semantic Relations and the Lexicon.* Cambridge: Cambridge University Press.
Stopelli, P. (ed.) (1998). *Sinonimi e contrar*i. Garzanti.
Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Lingistics. An International Handbook: 791-807.* Berlin: de Gruyter.

Czech:
http://www.slovnik-synonym.cz/web.php/hledat?typ_hledani=prefix&cizi_slovo
English:
http://www.synonym.com/

2.8. Semantic classes of nouns

Problem

Evaluate the frequencies of semantic classes of nouns in texts, find the rank-frequency distributions, propose a model, and compare texts and text sorts.

Procedure

One of the classifications based on a German dictionary is that of V.V. Levickij and A.A. Wenhrynowytsch (2009), Levickij, Drebet, Kijko (1999) containing the following classes:
1. Abstract notions,

2. Language, speech and text,
3. Names of institutions and organisation,
4. Form and structure,
5. Objects, equipments, outfits,
6. Activities, processes,
7. Space,
8. Time concepts,
9. Human or animal organs or body parts,
10. Position, state, condition,
11. Materials and substances,
12. Terms,
13. Plants,
14. Clothes,
15. Humans and mythological beings, professions and family relations,
16. Natural phenomena, diseases,
17. Animal world,
18. Quality of objects,
19. Measurement and units of emasurement,
20. Emotions, feelings, perceptions,
21. Character traits and customs of Man,
22. Foods and drinks,
23. Other.

Every possible classification has disadvantages, none is exhaustive, classes tend to be non-disjunctive etc. Therefore, researchers will almost always have to modify or extend a given classification for their purposes or even begin from scratch with an own one.

Some languages have morphemes for semantic classification of nouns and the counts may be performed mechanically. Other languages have numeratives, classifiers, etc., but only for some of the nouns.

Count the frequencies of nouns in a text in individual classes. Arrange the frequencies in decreasing order and find a theoretical distribution for this ranking. Substantiate it linguistically. Then do the same with other texts. Compare individual texts. You may compare either the frequencies or some functions of frequencies in order to obtain some classification of texts into text-sorts.

Then evaluate the individual texts according to another property and compare some function of this evaluation with a function of nominal classes. If you obtain some significant link, compare the both previous results with one more new property. Continue in this way until you obtain a control cycle in which the links can be represented by some simple functions.

Perform the same investigation using any other noun classification. Perform your analysis also for another language.

References

Aikhenvald, A.Y. (2000). *Classifiers: A typology of noun categorization devices.* Oxford: Oxford University Press.

Contini-Morava, E. (2012). Noun classification in Swahili. http://www2.iath.virginia.edu/swahili/ (Oct. 10, 2012)

Craig, C.G. (ed.) (1986). *Noun classes and categorization.* Amsterdam: John Benjamins.

Demuth, K. (2000). Bantu noun class systems: Loan word and acquisition evidence of semantic productivity. In: G. Senft (ed.), *Classification Systems*: *270-292*. Cambridge University Press.

Dixon, R.M.W. (1968). Noun classes. *Lingua 21, 104-125.*

Dornseiff, F. (2004). *Der deutsche Wortschatz nach Sachgruppen*. 8th edition. Berlin-New York: de Gruyter

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).

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.*

Levickij, V.V., Wenhrynowytch, A.A. (2009). Quantitative Charakteristika der substantivischen Synonymie im heutigen Deutsch. *Glottotheory 1(2), 75-85.* https://en.wikipedia.org/wiki/Noun_class

2.9. Semantic diversification of prefixes

Problem

Show that the meaning of verbal or nominal prefixes in a given language is diversified. The diversification is usually described in school grammars. Study several texts, count the individual meanings of a prefix and show that the rank-order of frequencies of individual meanings is distributed according to the negative binomial distribution, i.e., test this hypothesis. If you find an exception, find another distribution. Substantiate the negative binomial and the new distribution linguistically.

Procedure

Evaluate a not too short text and write out all verbs/nouns with the same prefix. Find the meanings of the prefix in an official grammar text-book. Count the

numbers of words with the same meaning of the prefix and set up the empirical rank-frequency distribution of these numbers. Fit the one-displaced negative binomial distribution to the frequencies

$$P_x = \binom{k+x-2}{x-1} p^k q^{x-1}, \quad x = 1,2,3,...$$

where p and k are parameters, $q = 1 - p$. If the fitting result is adequate, show that the above formula is the result of a Poisson process as illustrated by the experiment of throwing balls into urns (= prefixed words into the meaning class) in which the attraction of an urn (= meaning class) is the greater, the more balls (= words) are in it. If the fitting result is not adequate, find another distribution and substantiate it.

Evaluate a long text and its translation in some other language. Ascribe all translation variants to the original meanings of the prefix and count the frequencies. If the text is long enough, you obtain a two-dimensional distribution (X = original meanings, Y = translation means). Show that the distribution is a bivariate negative binomial d. or find another bivariate distribution. In any case, you can transform the field into several univariate distributions.

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.

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

Johnson, N.L., Kotz, S. (1969). *Discrete distributions*. Boston: Houghton Mifflin.

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

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

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

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

2.10. Semantics of prepositions 1

Problem

Prepositions express in many cases the conceptualisation of space and the direction of action. Show as many properties of the prepositional locational and directive system of a language as possible. Use the results in languages from which data are available (Nimboran and Slovak) and show the differences.

Procedure

Collect all prepositions of your language expressing location or direction. In English, e.g. *on, at, in, from, to* are such prepositions. Then using the previous literature (cf. References) define and enumerate the following properties defined in the references:

Efficiency of the location system;
discriminativity of the location and directive system
discriminative entropy
directive transitivity
directive symmetry
planar symmetry.

Express them by means of indicators. Find the sampling distribution of these indicators. Since most of them will be proportions, use the binomial distribution or find the asymptotic expectation and variance of the indicators. Set up e.g. the asymptotic normal test to compare individual languages. Show that languages partition the space and directions in it differently. Express the extent of the difference.

Ignore other meanings of prepositions than location or directive ones.

Perform an identical analysis in a language "replacing" prepositions by affixes, e.g. Ugro-Finnish languages, or by postpositions (e.g. Japanese).

The same can be done with adverbs ignoring those that arose from adjectives. Show other spatial systems in the given language.

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). Die Darstellung des Raumes im System slowakischer Präpositionen. *Jazykovedný časopis 19, 1968, 25-40* (Originally in Slovak. The German translation can be sent on demand)

2.11. Semantics of prepositions 2

Problem

Describe the semantic diversification of prepositions in a language, i.e. find the distribution of polysemy of prepositions. This is a problem of semantic diversification.

Procedure

Collect all prepositions which can be found in a good grammar of a language or in a big dictionary. Today, one can find them also on the Internet. For each preposition write out all its meanings as given by the grammar or dictionary. Then construct the distribution of the number of prepositions (f_x) with exactly *x* meanings. Find the theoretical distribution fitting well to the empirical data and substantiate the distribution linguistically. If possible, base the model on a birth-and-death process and derive the formula from it, or at least describe the diversification process.

Then find the extent of synonymy of prepositions. Under which conditions can a preposition replace another one? Are there crisp boundaries between the meanings of prepositions? Use fuzzy (or other non-crisp) sets or Venn diagrams to express the dynamics of the prepositional system. Find the properties of this system and express them by means of indicators. Find the variances of the indicators.

Set up a graph of prepositions joining those that may be synonymous in at least one case. Evaluate the quantitative properties of this graph and make statements about the distinctness of this system. Compare it with an earlier stage of the given language or compare it with another language.

Care for clear a definition of the class of prepositions. The problem can be solved mutatis mutandis for any small semantic system.

Working with languages without prepositions study instead the respective morphological means, e.g. affixes or postpositions.

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, 2008, 79-86.*

Ferrer-i-Cancho, R. (2013). Hubiness, length and crossings in syntactic dependencies. *Glottometrics 25, 2013, 1-21*

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

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

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.

2.12. Degree of metaphoricality

Problem

Words and other lexical units can be used in their basic meaning such as "back" for the corresponding part of the body or metaphorically such as "back" in "back issues of a journal". It is useful to transform this dichotomy into a quantitative concept or at least into a comparative one. Define such a scale of metaphoricality for lexical items.

Procedure

A first idea as to how a scale could be set up, the conceptual distance of a metaphoric meaning from the basic one can be used. Thus, "back" as a spatial expression can be considered closer to the basic "part of the body" meaning than "back" in its temporal meaning ("back in the 70th"). "Back" in connection with development ("the child is a bit back in its mental development") is even more distant and hence more metaphoric.

As a measurement procedure, you might look up the word under study in a dictionary and arrange the meanings given in the entry according to their distance from the most basic one. In this way, you can assign the degree of metaphoricality to the meaning of the word in the given context.

Now perform two kinds of investigations.

(1) Take a large random sample of words from the dictionary and to each of them write the highest degree of its metaphoricality as found in the dictionary. Let the maximal degree of metaphoricality be the new variable. Count the number of cases for each maximal degree and set up the distribution of lexicon meta-

phoricality. First find an empirical function capturing this distribution, then set up such a function theoretically using either difference or differential equations. Test the theoretical result. Compare the result with those in another language.

(2) Evaluate a not too short text and for each word annote its topical (local) degree of metaphoricality. You obtain a sequence of numbers whose properties may be studied by different methods. First, study the distribution of degrees found in the text leading to some functions or distributions, and characterize the text; then study the sequence itself, e.g. by methods of time series. Compare texts and elaborate on the differences between text-sorts or even languages. Compare e.g., a text and its translation into several languages (e.g. *The Little Prince* by Exupéry), let specialists perform the analyses in individual languages and you order languages according to their mean metaphoricality. Is a language with high degree of metaphoricality adequate for international communication?

In order to alleviate the problem, distinguish parts of speech and consider only one of them, for example verbs.

References

Bergem, W., Blum, L., Marx, F. (eds.) (1966). *Metapher und Modell. Ein Wuppertaler Kolloquium zu literarischen und wissenschaftlichen Formen der Wirklichkeitskonstruktion*. Trier: Wissenschaftlicher Verlag Trier.

Blumenberg, H. (1997). *Paradigmen zu einer Metaphorologie.* Bonn: Bouvier 1960, Neuausgabe Frankfurt/Main: Suhrkamp.

Bohunická, A., Orgoňová, O. (2006). Druhy podobnosti v metafore. In: Myslovičová, S. (ed.), *Jazyk a jazykoveda v pohybe: 390-407*. Bratislava: Veda.

Eder, T., Czernin, F.J. (eds.) (2007). *Zur Metapher. Die Metapher in Philosophie, Wissenschaft und Literatur*. München: Fink.

Haverkamp, A. (ed.) (1966). *Theorie der Metapher*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Hintikka, J. (ed.). *Aspects of Metaphor*. Dordrecht: Kluwer.

Krupa, V. (1990). Conceptual distance within metaphor. *Asian and African Studies 25, 89-94.*

Krupa, V. (1997). Podobnosť ako základ metafory. *Jazykovedný časopis 48, 81-88.*

Krupa, V. (2003). *Cognitive aspects of language and the creativity of metaphor*. Bratislava: SAV.

Kurz, G. (1982). *Metapher, Allegorie, Symbol*. Göttingen: Vandenhoeck und Ruprecht.

Levin, S.R. (1977). *The Semantics of Metaphor*. Baltimore: Johns Hopkins University Press.

Mahlmann, R. (2010). *Sprachbilder, Metaphern & Co.* Weinheim: Beltz Verlag.

Nordquist, R.: http://grammar.about.com/od/mo/g/metaphorterm.htm (4.2.2013)

Ortony, A. (ed.) (1993). *Metaphor and Thought*. Cambridge, U.K.: Cambridge University Press.

Pavelka, J. (1982). *Anatomie metafory*. Brno: Blok.
Ricœur, P. (2002). *La métaphore vive.* Paris: Seuil.
Sacks, S. (ed.) (1979). *On Metaphor*. Chicago: University of Chicago Press
Skirl, H., Schwarz-Friesel, M. (2007). *Metapher*. Heidelberg: Winter.

2.13. Distribution of metaphoricality

Hypothesis

The distribution of the degree of metaphoricality of the lexical units in a text abides by a law.

That means, the metaphors of individual degrees are lawfully distributed in a text. The text applies metaphors of different degrees in a (stochastically) regular way. If all metaphors would be of very high degree, the text could become unintelligible, and with merely low level metaphors it might become too specific. Even mathematical texts contain numerous metaphors of different degrees (e.g., tree, lattice, graph, ring, group, mapping...).

Procedure

Determine the degree of metaphoricality (cf. the Problem: *Degree of metaphoricality* in this volume) of each word in a text in a language of your choice, and then state the frequency distribution of the resulting numbers. Find a theoretical probability distribution that fits the data and can be linguistically interpreted. Interpret the parameters (or the form) of the distribution and ascribe them to different text-sorts.

Classify or order the texts on the basis of the parameters of the distribution, set up confidence intervals. Perform tests for difference in metaphoricality of texts, text-sorts, etc. by means of standard procedures.

References

Arduini, S. (ed.) (2007). *Metaphors*. Roma: Edizioni di Storia e Letteratura.
Black, M. (1954). Metaphor. *Proceedings of the Aristotelian Society, 55, 273–294.*
Black, M. (1962). *Models and metaphors: Studies in language and philosophy*, Ithaca: Cornell University Press.
Cazeaux, C. (2007). *Metaphor and Continental Philosophy: From Kant to Derrida.* New York: Routledge.
Davidson, D. (1978). What Metaphors Mean. Reprinted in *Inquiries Into Truth and Interpretation*. (1984). Oxford: Oxford University Press.

Derrida, J. (1982). White Mythology: Metaphor in the Text of Philosophy.In: *Margins of Philosophy*. Trans. Alan Bass. Chicago: University of Chicago Press.

Dirvens, R., Pörings, R. (eds.) (2002). *Metaphor and Metonymy in Contrast*. Berlin: Mouton de Gruyter

Fass, D. (1988). Metonymy and metaphor: what's the difference?. *Proceedings of the 12th conference on Computational linguistics 1, 177–181*. doi:10.3115/991635.991671.

Kövecses, Z. (2002). *Metaphor: a practical introduction*. Oxford University Press US.

Lakoff, G., Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Nordquist,R.:http://grammar.about.com/od/mo/g/metaphorterm.htm (04.02.2013)

Punter, D. (2007). *Metaphor*. London: Routledge.

Richards, I.A. (1936). *The Philosophy of Rhetoric*. Oxford: Oxford University Press.

Ricoeur, P. (1975). *The Rule of Metaphor: Multi-Disciplinary Studies in the Creation of Meaning in Language*, trans. Robert Czerny with Kathleen McLaughlin and John Costello, S.J., London: Routledge and Kegan Paul 1978. (Toronto: University of Toronto Press 1977)

Underhill, James W. (2011). *Creating Worldviews: Metaphor, Ideology & Language*. Edinburgh: UP.

2.14. Metaphoricality and frequency

Hypothesis

It can be assumed that the degree of metaphoricality of a lexical unit is related to the frequency of the unit in the following way

(1) the more frequent a word, the higher the tendency to be used as a metaphor and

(2) the more frequent a word which is used as a metaphor the higher the mean degree of metaphoricality of its tokens in text.

Procedure

Determine the degree of metaphoricality of each word of a text and its frequency (c.f. problem *Degree of metaphoricality* in this volume). You can do this either on the basis of the frequency of the words in the given text or (probably better) of the frequency as given in a frequency dictionary or taken from corpus data. Find

a simple function (such as a power law) which fits the relation between the two variables and test the goodness-of-fit.

Though even this result would be a good contribution to a future theory, use the simple function you obtained inductively, transform it into a differential equation and show its place in the unified theory (Wimmer, Altmann 2005). Interpret the parameters linguistically. Incorporate the result into Köhler's control cycle (2005).

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.

cf. Problems: *2.12. Degree of metaphoricality; 2.13. Distribution of metaphoricality* in this volume.

2.15. Metaphoricality and length

Hypotheses

It can be assumed that the degree of metaphoricality of a lexical unit depends on the length of the unit, and that the dependency shows a decreasing tendency, i.e.

(3) *the longer a word the less the tendency to be used as a metaphor* and
(4) *the longer a word which is used as a metaphor the less the degree of metaphoricality*.

Substantiate these hypotheses linguistically and test them statistically.

Procedure

First read the problem *Degree of metaphoricality* in this volume.

Determine the degree of metaphoricality of each word token in a text and its length in terms of the number of syllables. Find a simple function (such as a power law) which fits the relation between the two variables and test for goodness-of-fit.

If you obtained an empirical function using a software, transform the function into a differential equation and interpret it linguistically. Then incurporate

the result into Köhler's control cycle (2005) and link it both with length and with frequency.

Looking at the control cycle propose further possible links between metaphoricality and other properties. Strive for a first theoretical statement in this domain.

References

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

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

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.

See the problems: *2.12. Degree of metaphoricality*;
2.13. Distribution of metaphoricality;
2.14. Metaphoricality and frequency;
in this volume.

2.16. The weight of metaphoricality in text

Problem

Metaphors occur perhaps in any text. Express the extent (weight) of metaphoricality in a text, compare several texts and classify them into text sorts.

Procedure

Find those sentences in a not too short text that contain a metaphor of any kind. State also the number of all sentences and compute the proportion of sentences with metaphor. You obtain a simple proportion which can be used for text comparisons. Establish groups of texts with a non-significantly different metaphoricality. Do they belong to different text-sorts?

If you already read the problem *2.12. Degree of metaphoricality* and can ascribe some degrees to individual metaphors, characterize the text using an indicator expressing also these weights. Derive the variance of the indicator and compare again several texts using the normal test. Set up a new, "weighted" ordering of texts into groups or a rank-order of texts. Compare your results with texts in other languages.

References

See the problems: *2.12. Degree of metaphoricality; 2.13.Distribution of metaphoricality; 2.14. Metaphoricality and frequency; 2.15. Metaphoricality and length* in this volume.

2.17. Metaphoricality motifs

Problem

Form “Köhler Motifs” from the numbers representing the degree of metaphoricality of the word tokens in a text and perform the usual studies on them.

Procedure

First read the problem *Degree of metaphoricality* in this volume and form the sequence of metaphor degrees for a text. Then segment the sequence into motifs.

A motif is defined as a non-decreasing sequence of numbers. For example, the sequence 1,1,2,3,1,2 contains two motifs: 1,1,2,3 and 1,2. Having transcribed the text in form of motifs:

(1) Find the distribution of motif lengths. Length is the number of elements in the motif. Show that the lengths abide by a regular probability distribution.

(2) Analyse the distribution, derive it theoretically and determine the role of the parameters, i.e. interpret the approach and substantiate it linguistically.

(3) Study the distribution of the difference between the first and the last member of a motif. In the above example, the first difference is 3 - 1 = 2, in the second it is 1. You obtain a relatively small interval of values. Nevertheless, study also the sequence of these differences, compute the autocorrelation and other indicators used for the evaluation of time series.

(4) Use all these indicators for characterising texts, authors and text sorts.

(5) Study the distance between metaphors of the same degree. Set up the empirical distribution of these distances and find a theoretical distribution. Perform experiments with the Zipf-Alekseev distribution or function.

(6) Use the parameters of the distribution/function of distances to characterize texts, authors, text-sorts.

References

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

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

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learning and Applications: 637-646.* 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: 81-89.* Wien: Praesens.

2.18. Modality marking

Hypothesis

Show that the expression of modality is an instance of linguistic diversification and abides by a typical probability distribution.

Procedure

Modality can be expressed in different ways. In English, e.g. the following equivalents of "must" are available:

must do	(modal verb)
shall do	(modal verb)
do	(no marking; indicative)
has to do	(verbal paraphrase)
it is compulsory/mandatory to do	(lexicalised)
–	(ellipsis)

a) Set up lists of equivalents of modal expressions in a language under study and determine their rank-frequency distributions in texts. Find the corresponding probability distributions and parameter estimations. Perform goodness-of-fit tests.
b) Compare different texts and text kinds with respect to the rank-frequency distributions.
c) Compare the modal expressions and their distributions in different languages using parallel texts (corpora).

d) Analyze a stage play. For each person find the number of different modal expressions and their frequencies. Prepare a table of all persons and compare the individual classes. State whether the preference for modal expressions is equal with all persons; use a non-parametric test based on ranks, e.g. Kendall's W coefficient.
e) Observe the course of modality in a stage play (cf.problem *2.19. Modality degree and frequency*)

References

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 646-657*. Berlin/New York: de Gruyter.

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

Blakemore, D. (1994). *Evidence and modality*. In: R.E. Asher (Ed.), *The Encyclopedia of language and linguistics*: *1183–1186.* Oxford: Pergamon Press.

Bybee, J., Perkins, R., Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Calbert, J.P. (1975). *Toward the semantics of modality*. In: Calbert, J.P., Vater, H. (Eds.), *Aspekte der Modalität*. Tübingen: Gunter Narr.

Chung, S., Timberlake, A. (1985). Tense, aspect and mood. In: Shopen, T. (ed.), *Language typology and syntactic description: Grammatical categories and the lexicon: Vol. 3, 202-258.* Cambridge: Cambridge University Press.

Felici, A., Pal, P. (2012). Predicting translation equivalents and norm formulations: a study on some EU legislative features. *Journal of Quantitative Linguistics 19(3), 181-204.*

Kaufmann, S., Condoravdi, C., Harizanov, V. (2006), Formal approaches to modality. In: Frawley, W. (ed.), *The Expression of Modality*. Berlin, New York: Mouton de Gruyter.

Kratzer, A. (1981). *The notional category of modality*. In H.-J. Eikmeyer, H. Rieser (eds.), *Words, worlds, and contexts: New approaches in word semantics*. Berlin: Walter de Gruyter.

Kratzer, A. (1991). Modality. In: von Stechow, A.,Wunderlich, D. (eds.) *Semantics: An International Handbook of Contemporary Research.* Berlin: Walter de Gruyter.

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

Palmer, F.R. (1979). *Modality and the English modals*. London: Longman.

Palmer, F.R. (1994). *Mood and modality*. Cambridge Univ. Press. Second edition 2001.

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

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

Portner, P. (2009). *Modality.* Oxford: Oxford University Press.

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

Saeed, J.I. (2003). *Sentence semantics 1: Situations: Modality and evidentiality*. In: J.I. Saeed, *Semantics* (2nd. ed): *135–143*. Malden, MA: Blackwell Publishing.

Sweetser, E.E. (1982). Root and epistemic modality: Causality in two worlds. *Berkeley Linguistic Papers, 8, 484-507.*

2.19. Modality degree and frequency

Hypothesis

The relative frequency of a modal expression is a function of the degree of the expressed modality.

Procedure

Modality can be expressed in several ways (cf. the Problem *Modality Marking* in this volume), which often differ in their degree of the expressed modality. Consider the English expressions “must”, “should”, “ought to”, “may”, etc. and their respective degrees of commitment.

(1) Set up rank scales of modal expressions according to degrees (of their “strength”) on appropriate scales in one or more languages. Determine the text frequencies of the expressions and test the interdependence of degrees and frequencies.

(2) Set up an indicator of modality for a text and derive its variance. In order to compare two texts, use the asymptotic normal criterion. As an indicator, e.g. the mean degree, the entropy, the repeat rate, etc. can be used.

(3) Study the course of modality in the deployment of a text and capture it using at least empirical formulas. Later on, derive the formulas form theoretical considerations taking into account some potential influential factors which can be associated with economy or other effects from the speaker/reader and hearer/ writer.

(4) Do texts differ with respect to the course of modality? Analyze texts belonging to different text-sorts.

(5) Consider the extent of modality in a text and find other properties with which it is associated. Combine modality with dogmatism (cf. *Problems Vol. 3*: *65*), with descriptivity/activity (cf. *Problems Vol 3: 113-116*), emotionality, etc. Then set up a control cycle for the relation of modality to other properties. Express the relations with simple formulas.

References

See problem *2.18. Modality marking*

2.20. Sound symbolism

Problem

Sound symbolism is a well known phenomenon linked with some aspects of the origins of language. Sounds or combinations of sounds may evoke some meaning or association that can be studied systematically. Common terms for these phenomena are symbolism, phonesthemes or submorphemes (Firth 1930; Fenz 1940; Householder 1946; Bolinger 1965; Peterfalvi 1970; Wescott 1980; Malkiel 1990; Hinton, Nichols, Ohala 1994; Levickij 1973, 2009, 2010).

Study the existence of a correlation between initial consonant groups in words and a semantic commonality of groups of words with equal or similar initial sounds. Devise a measure of semantic similarity.

Two other investigation possibilities are presented below.

Procedure

(1) Compile a list of all words in a big dictionary with the same initial consonant cluster. Study whether the words have some meaning commonality. In the positive case you may find several verbs in which the initial cluster may correlate with some sound arising at performing the activity expressed by the verb. Do not decide intuitively, find a measure of similarity between the given verbs. But especially, compute the probability of common meaning and compare it with the discovered reality.

(2) Select a language out of the Indonesian languages which still has its word final consonants, especially voiced stops (as compared with the Austronesian reconstruction). The best choice is, perhaps, Sundanese. Indonesian words consist mostly of a monosyllabic root placed in the second syllable, and a preformative in the first. Meaning similarity of the same root is very conspicuous.

Set up a dictionary of roots with all respective words. First set up a dictionary of Sundanese (or other) independent *verbal interjections* and search for words containing them as a root. Compare the Sundanese sound symbolism with that of English phonesthemes. Is there a similarity of meaning? Here, too, state all words with the same root and compute the probability that *x* of them have the "same" or "similar" meaning. Use combinatorial methods for computing the probability.

(3) Find in your language all words containing the same sequence of vowels and state whether there is some common meaning. Usually, the vowels are placed in two subsequent syllables. Here, too, compute the probability of a common meaning or common association.

Do not simply collect and classify data, devise quantitative methods of evaluation.

References

Abelin, Å. (1999). *Studies in sound symbolism*. Göteborg University, Diss.

Bolinger, D. (1965). The atomization of meaning. *Language 41, 555-573.*

Fenz, E. (1940). *Laut, Wort, Sprache und ihre Deutung. Grundlegung einer Lautbedeutungslehre.* Wien: Deuticke.

Firth, J.R. (1930). *The Tongues of Men, and Speech*. London: Oxford University Press.

French, P. (1977). Toward an explanation of phonetic symbolism. *Word 28*, 305-322.

Hinton, I., Nichols, J., Ohala, J. (eds.) (1994). *Sound symbolism*. Cambridge: Cambridge University Press.

Householder, F. (1946). On the problem of sound and meaning, an English phonestheme. *Word 2, 83-94.*

Kim, K.-O. (1977). Sound symbolism in Korean. *Journal of Linguistics 13, 67-75.*

Levickij, V.V. (1973). *Semantika i fonetika. Posobije, podgotovlennoe na materiale experimental'nych issledovanij.* Černovcy: Černovickij gosudarstvennyj universitet.

Levickij, V.V. (2009). *Zvukovyj simvolizm: mify i realnosť.* Černovcy: Černovickij gosudarstvennyj universitet.

Levickij, V.V. (2010). Deutsche Phonemverbindungen im Anlaut des Wortstammes. *Glottotheory 3(2), 35-42.*

Malkiel, Y. (1990). *Diachronic problems in phonosymbolism. Edita and Inedita 1979-1988.* Amsterdam, Philadelphia: Cambridge University Press.

Newman, S.S. (1933). Further experiments in phonetic symbolism. *American Journal of Psychology, 45,* 53-75.

Peterfalvi, J.M. (1970). *Recherches expérimentales sur le symbolisme phonétique.* Paris: Centre National de la Recherche Scientifique.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology 12, 225-239.*

Wescott, R.W. (1980). *Sound and sense. Linguistic essays on phonosemantic subjects.* Lake Bluff, Illinois: Jupiter Press.

Wichmann, S., Holman, E.W., Brown, C.H. (2010). Sound symbolism in basic vocabulary. *Entropy 12, 844-858.*

2.21. Association of antonyms

Problem

O. Nuzban and S. Kantemir (2013) subdivided nouns into twenty classes and studied the association of the adjectives „soft“ and „hard“ with the elements of these classes. They obtained results for English as presented in Table 1.

Table 1
Association of the adjectives “soft” and “hard” to subclasses

Subclasses of nouns	**„soft“**	**„hard“**
Human appearance	87	57
Names of humans	12	4
Social status	1	5
Proper nouns	2	1
Flora	13	2
Fauna	3	0
Nature, space	35	6
Clothes, footwear	14	6
Edifices, premises	3	13
Interior objects	24	4
Inanimate objects	7	7
Substance, materials	19	17
Food, beverages	6	5
Time notions	7	26
Character traits and humans’ features	14	18
Feelings, emotions, relationships	6	18
Abstractions	16	42
Actions, arrangements	12	20
Acoustic phenomena	94	16
Olfactory phenomena	2	1
Light phenomena	46	3
Motion, movement	15	9
Language and speech units	5	6
Shapes, figures	12	15
Other notions	4	8

(a) Find the distribution of the ranked frequencies separately for „soft“ and „hard“.
(b) Test whether the rankings are equal.
(c) Test the strength of association of each nominal class with one of the two antonyms.

Procedure

(a) First set up the rank-frequency distribution for the two adjectives separately, i.e. order them according to decreasing frequency. Then find the distribution or the function capturing this decrease. If you obtain a positive result, derive the distribution or the function from the unified theory (Wimmer, Altmann 2005) and interpret the parameters.

(b) Consider the frequencies as they are and test whether the two sets of data are equal. Since there are many small frequencies, use, instead of raw frequencies, their ranks and perform a test for equality of the two rankings. Use Spearman's or Kendall's tests and interpret the result. If the ranking is equal (= not significantly different), then the ascription of these adjectives is equal for all noun classes; otherwise there are some preferences.

(c) Which classes have a greater propensity to be associated more with "soft" than with "hard", and vice versa? Perform a test for each category of nouns. Compare simply the two numbers using the binomial distribution (performing an exact test) or, if the numbers are larger, use the Poisson distribution. In practice, consider N the sum of the two classes, p the proportion in one of the classes and x_s the frequency in the class of the given antonym. The probability that the antonym occurs x_s times or more frequently (more seldom) can be obtained using the binomial distribution or in case of great N using the Poisson distribution. One can apply also the asymptotic normal test (not forgetting the covariance).

(d) Perform this research using corpora in other languages and strive for a more general statement. Then select some other adjectival antonyms and perform the same investigations. Compare your own language with English. If possible, study a non-Indo-European language, too.

If you have analyzed several languages for the same antonym pair, can you generalize the result in form "*soft* is associated significantly with the following classes… in all languages"?

Analyze the extent of synaesthesia associated with the given antonyms.

References

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

Nuzban, O., Kantemir, S. (2013). Statistical analysis of perception adjectives 'soft' – 'hard' in English. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 4: 174-190*. Lüdenscheid: RAM-Verlag.

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

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

2.22. Adjectival antonyms

Problem

Study the use of adjectival antonyms in general, i.e. state whether their use in texts is equilibrated or one of the poles is preferred. Distinguish antonyms according to kinds (gradual, incompatible, complementary, converse, reverse,…). Evaluate the data found quantitatively.

Procedure

Count the frequency of each adjective in a not too short text. Some text collections and corpora provide part-of-speech annotations (tagging), which would make it easier to find all instances of adjective tokens. In any case, individual and complete texts should be used, neither corpora or corpus parts nor text fragments. Order the adjectives to sets of antonyms. When both poles of an adjective pair exist in the given language but one of them does not occur in the text under study, ascribe it the frequency 0. There are adjectives without direct antonyms, e.g. some colour terms; do not define an antonym using "not"; e.g. for "mediaeval" or "American" there is no direct antonym but you can take into account the oppositions presented in texts.

Classify the pairs according to more generic properties, e.g. intensity (*strong - weak*), aesthetics (*pretty - ugly*), size (*big - small*), perception (*hard - soft*), etc. You will obtain several classifications because some pairs cannot be subsumed under all classes. In each generic class, state the number of adjectives and their frequencies, and the level of asymmetry of the representation of one pole using a test, i.e. characterize a text as an entity with special propensities.

Classify texts into text-sorts using this criterion, i.e. add this criterion to the classificatory criteria of text-sorts.

There may be texts without any antonyms. This can simply be expressed by the respective probability resulting from the test.

Since this property is not language-specific, one cannot compare languages but only text-sorts or individual texts. Study the complete work of a writer and show that with regard to adjectival antonyms (s)he develops in the course of time. Capture the course of this development and express it by an empirical formula. Study different authors and show these developmental differences.

Study the nouns characterized by these adjectival antonyms and show whether the association (of the noun with the adjective) is significant for the given text (cf. Strauss, Fan, Altmann 2008: Chapter 4).

References

Agricola, C., Agricola, E. (1984). *Wörter und Gegenwörter. Antonyme der deutschen Sprache.* Leipzig: Bibliografisches Institut.

Bulitta, E., Bulitta, H. (2003). *Wörterbuch der Synonyme und Antonyme.* Frankfurt: Fischer.

Cruse, D.A. (1992). Antonymy revisited: Some thoughts on the relationship between words and concepts. In: A.J. Lehrer, E.F. Kittay (eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization*: *289–306.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cruse, D.A, Togia, P. (1995). Towards a cognitive model of antonymy. *Journal of Lexicology 1, 113-141*.

Jeffries, L. (2009). *Opposition in Discourse: The Construction of Oppositional Meaning* London: Continuum.

Jones, S. (2002). *Antonymy: A Corpus-based Perspective*. London and New York: Routledge.

Lehrer, A.J. (1985). Markedness and antonymy. *Journal of Linguistics 21, 397-421***.**

Lehrer, A.J. (2002). Paradigmatic relations of exclusion and opposition I: Gradable antonymy and complementarity. In D.A. Cruse, F. Hundsnurscher, M. Job, P.-R. Lutzeier (eds.), *Lexicology: An international handbook on the nature and structure of words and vocabularies*: *Vol. 1, 498–507.* Berlin: de Gruyter.

Lehrer, A.J., Lehrer, K. (1982). Antonymy. *Linguistics and Philosophy 5, 483-501*.

Mettinger, A. (1994). *Aspects of semantic opposition in English*. Oxford: Clarendon Press.

Müller, W. (1998). *Das Gegenwort-Wörterbuch. Ein Kontrastwörterbuch mit Gebrauchshinweisen.* Berlin: de Gruyter.

Murphy, M.L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge: Cambridge University Press.

Nuzban,O., Kantemir, S. (2013). Statistical analysis of perception adjectives 'soft' – 'hard' in English. In: R. Köhler, G. Altmann (eds.), *Issues in Quan-*

titative Linguistics 3, To honor Karl-Heinz Best on the occasion of his 70th birthday: 177-194. Lüdenscheid: RAM.
Strauss, U., Fan, F., Altmann, G. (^{2}2008). *Problems in Quantitative Linguistics Vol 1*. Lüdenscheid: RAM.

2.23. B-motifs

Problem

Study the processes in texts: (1) State their frequencies and their distribution. (2) Set up R-motifs of processes (called here B-motifs) and study their properties. (3) Characterize texts using the results obtained.

Procedure

Use Ballmer's (1982a: 73ff.) classification of processes in a general way (i.e. not restricted to Man) and set up the sequence of these units in text. You can use the following abbreviations (the author used German abbreviations):

MO: Motion
EX: Experience
PP: Passive Perception
PE: Psychological effect
CO: Cognition
EF: Effecting
AC: Action
PF: Performance
AM: Active movement
LM: Locomotion
AP: Active perception
IF: Information
WK: Working
EN: Execution
UT: Utterance
PA: Psychological activity
DF: Danger-Fear-Risk
IN: Influence
PC: Process control
RP: Reproduction
GR: Grasping
GU: Guiding
SU: Supporting

FR: Freedom
TP: Transport
DR: Driving
CD: Composing/Decomposing (Manipulate)
MF: Modification
PR: Production
CO: Consume
GT: Give and take
RG: Regeneration
TA: Transaction
EX: Expressives
EN: Enaction
IA: Interaction
DC: Discourse

First find examples either in the given reference or use a dictionary in order to be able to make a correct decision. It is to be noted that the author of this classification tried to follow grosso modo the evolutional complexity but this does not play any role in text analysis. Do not consider only verbs; if necessary, combine the processes. You may also group some processes and set up a smaller set. Do not absolutize any classification!

Then analyse a text taking down the sequences of processes in form of abbreviations. First, state the rank-frequency distribution of processes. Find a model for this distribution and substantiate it linguistically. Then segment the sequence into B-motifs (in analogy to R-motifs) and compute their properties. Is it possible to tell something about the text-sort? For setting up B-motifs, see the problem *Length of R-motifs* in this volume.

Study the texts of an individual writer in chronological order and perform the same analyzes. Can you recognize a tendency?? You can use the vector of processes and compare two vectors either by means of a statistical test or simply by computing the angle between the vectors. Study the distances of the current result from a number of texts scrutinized in the same way. Compare your result with that obtained from the evaluation of texts by means of the modified Busemann's ratio.

Another insight is offered by the system of Levickij and Lučak (2005), cf. *Problems Vol. 1*

References

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

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

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

Ballmer, Th.T. (1982a). *Biological foundations of linguistic communication*. Amsterdam-Philadelphhia: Benjamins.

Ballmer, Th.T. (1982b). The interaction between ontogeny and phylogeny: A theoretical reconstruction of the evolution of mind and language. In: Koch, W.A. (ed.), *Semiogenesis: 481-544*. Frankfurt: Lang.

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik.* Jena.

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 Academic Press.

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-sequences. In: Preisach, C., Burghardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data analysis, machine learning and applications: 637-646.* 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.* Vienna: Praesens.

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

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

3. Textology

3.1. Style

Problem

Style is a feature of individuality of the individual author. A stylistic feature can be measured either as a simple quantitative property such as lexical diversity or as a complex of properties, often represented by an index, i.e. a compound property. An author need not differ from all other writers by the given feature(s) but at least from some other ones. Find as many features of texts as possible and determine the methods of comparison.

Procedure

Begin to evaluate phonic features (e.g. assonance, euphony, alliteration, word length, etc.), continue with morphological ones, then syntactic ones, lexical ones, semantic and pragmatic ones, text-sort, up to associations. Set up a list of all possible text features that you can find in the literature.

Then evaluate two texts of the same author and one text of another one. Analyse the three texts according to your criteria and state which of the properties is common only to the two texts of the given author and differs from that of the second author. Do not forget that every property must be quantified, then measured and its quantity/extent must be given in terms of frequencies or degrees. That means, determine exactly the kind of the test you must use in order to state a significant difference of properties.

If you find a property in the two texts of the first author which is significantly different form that in the third text, evaluate further texts both by the same author and of many other ones. Test stepwise the differences. If you find authors that do not differ from the texts of the first author, determine a class of texts. Then select another property of the first author and repeat the procedure.

At last you should obtain a set of features which distinguish the given author from at least one of the other authors.

In the next step list the values of all given properties in all texts of the given author and find some kinds of relations among these properties, i.e. set up a control cycle.

Extend the analysis taking into account all texts you considered and present each property as a function of another one(s). In the first step use empirical formulas obtained with the aid of a software (e.g. TableCurves, Origin, etc). If you obtain positive results, derive the formulas from differential or difference equations. Avoid polynomials, use functions with a minimal number of parameters. Substantiate the parameters linguistically.

References

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

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

Köhler, R. (1990). Elemente der synergetischen Linguistik. In: Hammerl, R. (ed.), *Glottometrika 12, 179-188.* Bochum: Brockmeyer.

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

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

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

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

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.

3.2. Style evolution

Problem

There are as many definitions of style as there are scientists engaged in its investigation. Here, we consider style a property or several well defined properties of text or its components. Since the number of text properties is not limited, one must restrict oneself to a small number of them.

Study the evolution of an author or a poet on the basis of the following properties:

(a) *Prose*: (i) sentence length measured in terms of clause numbers; (ii) word length measured in terms of syllable numbers; (iii) distribution of parts of speech; (iv) distribution of adnominal modifiers (cf. Best, Boschtan 2010 and the problem *Adnominal modifiers: Classification*)

(b) *Poetry*: (i) word length distribution; (ii) rhyme type distribution; (iii) parts-of-speech distribution.

Procedure

Select a (regular) chronological sequence of texts of one author. Compute the above properties for each text separately. For the rank-frequency distributions, compare their means and state whether they develop in some direction. Set up the

Ord-scheme <I,S> and state whether the points on the straight line you obtain correspond to the year of creation.

Rhyme types can be masculine-feminine, closed-open, perfect-not perfect, length of the rhyming part, loss of rhyme. Study the proportion of these types in a chronological order of poems.

If the above properties do not display any evolution, use other text properties (cf. e.g. Waddell 2012; Popescu et al. 2009; Popescu, Mačutek, Altmann 2009) as long as you find some regular change in style. Hence, the investigation consists of three parts: (1) Is there evolution at all? (2) Which entities display evolution? (3) How does the evolution proceed?

Generalize the problem and evaluate texts from an individual language but from different historical epochs. Apply only one property, quantify it and compute it for all texts. Then show the given trend numerically and fit an empirical function (no polynomial!) to the data. You obtain a positive result even if the property remained constant. Define a set of properties and study their development - usually it is described in a slightly fuzzy way in text-books on historical linguistics.

If you have a historical dictionary at your disposal, study the development of the meaning of individual words, its diversification or change.

References

Andreev, S. (2009). Evolution of style. Young and versatile periods of F.I. Tutchev. *Glottotheory 2(1), 1-11*.

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

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

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

Waddell, C. *Basic Prose Style and Mechanics*.
http://www.hu.mtu.edu/~cwaddell/Basic_Prose_Style.html (Febr. 1, 2012)

3.3. Entropy deployment

Problem

Does entropy computed from the distribution of some entities change in the course of text or is it constant? Describe the problem for at least one kind of units and set up a hypothesis. Then test the hypothesis using other kind of units. At last, generalize the problem and pronounce a very general hypothesis.

Procedure

Analyse a text which is "naturally" partitioned in several parts, e.g. a poem is partitioned in strophes, a stage play is partitioned in acts, a novel in chapters, or even a press text in paragraphs which are adequate for the study of units of a very low level, e.g. phonemes or letters.

Compute the Shannon entropy for the given unit in all parts separately. Use the well known formula

$$H = -\sum_k \hat{p}_k \; ld \; \hat{p}_k = ld\, N - \frac{1}{N}\sum_k f_k \, ld \; f_k$$

where f_k is the absolute frequency of the given unit k; $p_k = f_k/N$; ld is the dual logarithm, i.e.. $ld\, x = log_2\, x$, and N is the sum of all f_k, i.e. the umber of units in the given text part.

For each part separately compute the variance of H

$$V(H) = \frac{1}{N}\left(\sum_k p_k \, ld^2 \; p_k - H^2 \right).$$

Then set up the stepwise asymptotic criterion showing the difference between the previous r parts and the $(r+1)th$ part as

$$u = \frac{(H_1 + H_2 + ... + H_r - rH_{r+1})}{r\sqrt{V(H_{r+1})}}.$$

which is asymptotically normally distributed. If $u < -1{,}96$ or $u > 1{,}96$, there is a break between the parts. Trace down all breaks in as many different units as possible and sketch the phonic, morphological, lexical etc. structure of the text. Begin with a poem with not too few strophes.

Perform the analysis on texts of different sorts (press texts, private letters, short stories, fables, poems, etc.) and state the kind of entropy deployment.

Entropy is not the only indicator of a state. Apply different indicators, set up analogical formulas and elaborate on the deployment of the given text(s). Strive for a more general view and substantiate the results linguistically.

Select texts of the same author written in different years. Compute the entropy of a unit and examine the development of entropy in time.

Another way of evaluating the deployment of entropy is simply its computation for the subsequent parts of the text and searching for the change of its value. Find a function describing the deployment or set up indicators of the change, apply Hurst's exponent, compute the dimension of the curve.

Not all units will display change and not all text sorts will display the same change. Since entropy is a measure of dispersion, conjecture some courses of this indicator for some units and texts. It is rather a linguistic than a mathematical problem.

References

Bennett, C.H., Ming Li, Bin Ma (2003). Chain letters and evolutionary histories. *Scientific American* 288(6), 76-81

Reza, F.M. (1961, 1994). *An Introduction to Information Theory*. New York: Dover.

Cover, T.M., Thomas, J.A. (1991). *Elements of information theory.* New York: Wiley-Interscience.

Lin, Shu-Kun (1999). Diversity and entropy. *Entropy 1(1), 1–3*

3.4. The type-token relation

Problem

There are a great number of formulas describing the growth of the type-token relation <#Types, #Tokens> in texts. Some of them are reasonable, other ones are ad hoc inventions. But the authors very seldom propose a test for deciding about the difference between two TTR functions. Devise such a test and exemplify it on texts from different languages.

Procedure

Apply at least one of the TTR functions from the extensive literature, for example that of

Brunet (1978) $V = a\left(\ln N\right)^{\beta}$

Dugast (1979) $V = N\exp(-\ln^2 N / a)$

Ejiri, Smith (1993) $V = aN^{1-b}$

Guiraud (1954) $V = a\sqrt{N}$

etc. and derive, after linearization, the variance of individual parameters. Then set up an asymptotic normal test for the difference of two (identical) parameters from two different texts and test the difference between these two texts.

Do not perform tests for corpora as closed units, i.e. do not mix texts. If you want to measure the general trend in a language, then compute the parameters for all individual texts, let the averages of them represent the expected value and set up a confidence interval for (the mean) parameter in the given language. Afterwards you can compare the means in two languages and prepare a typology.

Test the hypothesis that the mean parameter of a strongly synthetic language significantly differs from that of a strongly analytic language using your criterion.

Apply other functions and prepare tests for their parameters. In particular, take account of the results in Wimmer et al. (2001).

References

Baayen, H. (1996). The effects of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics 22(4), 455-480.*

Baayen, R.H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

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

Dugast, D. (1979). *Vocabulaire et discourse. Essai de lexicométrie quantitative*. Genève: Slatkine.

Ejiri, K., Smith, A.E. (1993). A proposal of a new 'constraint measure' for text. In: Köhler, R., Rieger, B. (eds.), *Contributions to Quantitative Linguistics: 195-211*. Dordrecht: Kluwer.

Fan, F. (2006). Models for dynamic inter-textual type-token relationship. *Glottometrics 12, 1-10.*

Fan, F. (2008a). A corpus-based study on random textual vocabulary coverage. *Corpus Linguistics and Linguistic Theory 4(1), 1-17.*

Fan, F. (2008b). Hapax legomena and language typology, a case study. *Journal of Quantitative Linguistics 4*, 370-378.

Fan, F. (2010). An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics 36(4), 631-637.*

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-token mathematics*. 's-Gravenhage: Mouton.

Herdan, G. (1964). *Quantitative Linguistics*. London: Butterworth.

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

Köhler, R., Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In: Altmann, G., Koch, W.A. (Eds.), *Systems. New paradigms for the human sciences*: *514-546.* Berlin: de Gruyter.

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

Laufer, B., Nation, I. (1995). Vocabulary size and lexical richness in L2 written production. *Applied Linguistics 16(4), 307-322.*

Orlov, J. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (1982), *Sprache, Text, Kunst. Quantitative Analysen*: *118-192*. Bochum: Brockmeyer.

Popescu, I.-I., Altmann, G. (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics 15(4), 370-378.*

Sichel, H. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist 11, 45-72.*

Tweedie, F., Baayen, H. (1998). How variable may a constant be? Measure of lexical richness in perspective. *Journal of Quantitative Linguistics 32(5), 323-352.*

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S, Wimmerová, S. (2001). *Úvod do analýzy textov*. Bratislava: Veda.

3.5. Stage play analysis 1

Problem

Characterize the individual persons of a stage play quantitatively from as many points of view as you can. Then set up rank-orders of persons in all dimensions you scrutinized. Find the correlations of all pairs of properties or dimensions and construct a control cycle.

Procedure

Start with simple properties such as the number of sentences of a person, the sentence length distribution (measured in terms of clause numbers), the distribution of degrees of speech acts (cf. *Problems Vol. 2: 124-126*), the distribution of word length with respect to a person (measured in terms of syllable numbers), activity or ornamentality of the speech (verb-adjective ratio), vocabulary richness, thematic concentration, internal or external references (reference to own speech or to that of other persons), observed verb valency, Carroll's vector (1960), etc.

Evaluate all these properties by means of established methods (cf. the References below), and characterize each person by a vector of these properties. Then begin to set up hypotheses about the mutual relations of these properties. Combine these properties and establish a new, complex dimension, e.g. dogmatism, dominance, rhetoric ability, etc. Construct indicators or vectors of these dimensions, derive their variances and propose tests for differences.

Perform some tests for comparison of persons and classify the persons according to the dimensions you defined. You can use classification programs or

discriminant analysis. Finally, interpret the results linguistically. Do not remain on the level of classification.

References

Carlson, M. (1993). *Theories of the Theatre: A Historical and Critical Survey from the Greeks to the Present.* Ithaca-London: Cornell University Press.

Carroll, J.B. (1960). Vectors of prose style. In: Sebeok, T.A. (ed.), *Style in language: 283-292*. Cambridge, Mass.: The M.I.T. Press.

Elam, K. (1980). *The Semiotics of Theatre and Drama*. London-New York: Methuen.

Fergusson, F. (1949). *The Idea of a Theater: A Study of Ten Plays, The Art of Drama in a Changing Perspective.* Princeton, N.J: Princeton UP.

Freytag, G. (1969). *Die Technik des Dramas*. Nachdruck. Darmstadt.

Harsh, P.W. (1944). *A Handbook of Classical Drama*. Stanford: Stanford UP.

Rehm, R. (1992). *Greek Tragic Theatre.* London-New York: Routledge.

Taxidou, O. (2004). *Tragedy, Modernity and Mourning*. Edinburgh: Edinburgh UP.

3.6. Stage play analysis 2

Problem

Study the development of properties of the speech of individual actors in the course of a stage play. Does a property change or is it constant? If it changes, find the course of the change.

Procedure

First, solve some questions in the preceding problem, read *Problems Vol. 2: 124-126*, then search for the deployment of some of the above properties. Select a property and describe its change from act to act. Does the development of a drama coincide with Freytag's analysis? Do some sequences in comedies differ from those in dramas? If so, how do they differ? Perform statistical tests and analyse the results theoretically, i.e. sketch a theoretical background of the course of drama in terms of the analyzed property. Capture the courses by some functions, perform - if necessary - Fourier analysis, wavelets, evaluate the results as time series, characterize the individual courses by Hurst's exponent or Lyapunov's coefficients. Employ only procedures that are easy to interpret linguistically.

Compare the same properties and courses with those in a comedy. Compare classical stage plays with modern ones. Develop a “teorita” and derive from it some hypotheses about drama.

References

See the previous problem.

3.7. Text properties

Problem

Choose a property from the immense repertoire of text properties, e.g. activity, nominality, ornamentality, descriptivity, lyricity, humour, sadness, speech acts, etc. Study the given property from two points of view: (a) statically, i.e. concerning its proportion or extent in text; (b) dynamically, i.e. concerning the sequence of entities carrying the given property. Show the significance of the phenomenon in the static approach, and show some regularity in the sequence of the given entities.

Procedure

Define a property operationally, do not rely on the judgement of informants. That is, define exactly which linguistic entities express the given property. If possible, devise even a scaling procedure for the property as expressed by the individual entities. For example, “book” is more concrete than the abstract “beauty”; or “revolver” is more specific than the more general “weapon”.

If you examine the text without a scaling procedure, state the proportion of entities with the given property and comparing it with other texts apply a statistical testing procedure. Order texts or text-sorts according to the proportion found.

If you use a scaling procedure, find the number of entities with the given degree and find the probability function of the property.

Study the distances between the occurrence of relevant entities and state whether they are random or display some tendency.

Form the sequence of sentences containing the pertinent element (A) and those without the pertinent element (B). Transcribe the text as a sequence of A's and B's and study the runs in the text. Find the empirical distribution of runs, find the probability of the longest run, compare texts and set up a classification of text sorts using the given criterion. Finally, substantiate the association of the text sort with the given property and its presence in the text.

Do not stop at this point of classification but study several properties and set up a control cycle of properties. Derive the individual associations using either the control cycles or another procedure, e.g. differential equations. Strive for a theory.

Cf. also problem 5.8 in Čech, Altmann (2011).

References

Carroll, J.B. (1960). Vectors of prose style. In: Sebeok, T.A. (ed.), *Style in language: 283-292*. Cambridge, Mass.: The M.I.T. Press.

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

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

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

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

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

Wimmer, G., Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Language. Word Length Studies and Related Issues: 207-316*. Boston: Kluwer.

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

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

Zörnig, P. (2013). A continuous model for the distances between coextensive words in a text. *Glottometrics 25, 54-68.*

3.8. Thematic words and Frumkina's law

Problem

(a) Are thematic words distributed uniformly in text passages or do they abide by Frumkina's law? (b) Do sequences of text passages display a monotonous occurrence of thematic words or does the arising time series display any special tendency?

Find some regularities.

Procedure

(1) Analyse a not too short text and compute the rank-frequency distribution of lemmas. Then compute the h-point (cf. Popescu et al. 2009: 18, formula (3.2)) and isolate all autosemantics placed at ranks smaller than *h*. These are the thematic words.

On the basis of these words (either separately or as a set), perform the computation of Frumkina's law: Partition the text in passages of, say, 100 lemmas and for each part compute the number of thematic words occurring in it. Show that the number of passages containing exactly *X* thematic words abides by Frumkina's law (c.f. *Problems Vol. 1: 117; Problems Vol. 2: 71*) given in form of the negative hypergeometric distribution

$$P_x = \frac{\binom{-M}{x}\binom{-K+M}{n-x}}{\binom{-K}{n}}, \; x = 0,1,...,n; \; K > M > 0,$$

which can be written in several ways (cf. Wimmer, Altmann 1999: 464ff.). The distribution has a number of special or limiting cases. The applicability of a special case of this distribution is a sign of a specific style. Test whether text-sorts display different patterns.

(2) Set up the sequence of passages (of 100 lemmas each) and study the sequence of occurrences of thematic words in them i.e. the number of thematic words in subsequent passages as a sequence. You obtain some kind of monotonous trend or an irregular oscillation or, in extreme cases, a fractal. State the properties of this sequence; compare texts; compare text sorts. Give reasons for the form of the sequence, i.e. formulate some hypotheses on the thematic sequence in texts.

(3) Take all synsemantic pre-h words and perform the same procedure as with the autosemantics. Show the differences and similarities, set up hypotheses and begin to construct a theory. That means, derive the discovered regularities from linguistic assumption, use the difference equation leading to the negative hypergeometric distribution as the basis of your construction and interpret the constants or functions in the equations on the basis of linguistic requirements.

See also the problem *6.18. Block distribution of modal expressions*.

References

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

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

Best, K.-H. (2001/2003). *Quantitative Linguistik. Eine Annäherung.* 2., überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Best, K.-H. (2005). Sprachliche Einheiten in Textblöcken. *Glottometrics 9, 1-12.*

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

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

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

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

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

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

Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. (1985). *Mathematische Linguistik.* Bochum: Brockmeyer.

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

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

Suhren, S. (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen „De lütte Prinz“*. Staatsexamensarbeit, Göttingen.

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

3.9. Distances in text

Problem

Any linguistic unit whose inventory is not infinite (as e.g. sentences) occurs several times within a given text. However, the repetitions may underlie various regularities, trends, rules, oscillations or they may be chaotic or random. Analyze the regularities (or irregularities) of distances between equal elements in text.

Procedure

Select one of the following entities

1. Sound types: according to place or manner of articulation or both
2. Syllable types: V, VC, CV, CVC, VCC, VCV,…
3. Syllable lengths (in terms of phonemes)
4. Morpheme types: proclitic, prefix, stem, infix, suffix, postclitic, ...
5. Morph length
6. Word classes: traditional parts of speech: Noun, Verb, Pronoun, Adverb, Adjective, Preposition, Postposition, Interjection, Conjunction, Article, Particle; – or apply any other system of parts of speech (tagset); – alternatively any other classification into word classes (on ontological, semantic, pragmatic criteria etc.)
7. Word length
8. Clause types
9. Clause length (in words)
10. Sentence types (according to different criteria)
11. Sentence lengths (in terms of clause numbers)
12. Hreb members
13. Types of speech acts
14. Equal frequencies of (also different) words, i.e. sequence of frequencies
15. Alliteration (both in prose and poetry)
16. Assonance
17. Verb valency
18. Degree of verb activity (scaling!)
19. Types of modificators/attributes of nouns
20. Grammatical categories
21. Individual markers of a category (e.g. individual cases; times; numbers; persons,…)
22. Polysemy (= number of meanings of the given word in a dictionary)
23. Types of reference (lexical identity, pronoun, hypero- or hyponym, synonym, ...)
24. Length of co-reference chains
25. Syntactic construction type
26. Complexity of the syntactic constructions
27. Frequency of the syntactic constructions
28. Depth of embedding.

First describe and capture quantitatively the different phenomena by evaluating many texts. Then begin to generalize. Next, set up the first hypotheses and test them. Approach a theory from different sides.

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

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

i. Are there any distance tendencies concerning special words, author, text sort, age, education, historical time of text creation, language, etc.?
ii. Which entities display an evident Skinner effect?
iii. If you consider merely the class of nouns, how can e.g. "nominal style" be expressed?
iv. Describe the properties of the distribution of distances (moments, Ord's indicators, skewness, asymmetry, etc.)
v. Can any laws be conjectured?
vi. How to set up a theory?
vii. Does the Weber-Fechner law intervene?
viii. Can a concrete hypothesis be derived from an existing distance theory?
ix. If a tendency is found, how can it be interpreted, linguistically substantiated and derived from the background theory?
x. Show which of the entities display random distances (applying Zörnig's model or the Poisson process) and which are not random.

References

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

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

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

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

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

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

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

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

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

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

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

Serrano, M.A., Flammini, A., Menczer, F. (2009). Modeling statistical properties of written text. *PLoS ONE4, e5372.* [PMC free article] [PubMed]

Vázquez A. (2005). Exact results for the Barabási model of human dynamics. *Phys. Rev. Letters 95(24):248701.*

Vázquez, A., Oliveira, J.G., Dezsö, Z., et al. (2006). Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E. 73, 036127.*

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

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

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *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. (2013). A continuous model for the distances between coextensive words in a text. *Glottometrics 25, 54-68.*

3.10. Text cohesion

Problem

Define a measure of text cohesion and give a general operationalisation, i.e. a measurment instruction. Illustrate the measure by means of examples. Find the mathematical properties of the measure. Interpret the result linguistically or psychologically. If necessary, introduce a scaling method for cohesion.

Procedure

Cohesion and coherence are indispensable ingredients of texts. We lack, however, a well-defined concept and a corresponding measure of both properties. Languages possess various ways to generate cohesion, and every language applies a mixture of them. Therefore, we cannot expect a scalar measure to be appropriate. Set up an inventory of linguistic means which increase the cohesion in a text and form a vector which will represent the multi-dimensional extent of cohesion in an individual text.

Such an inventory could consist of e.g.,
<connectives (conjunctions and pronominal adverbs), co-references (in form of recurrent lexical elements or their substitutes and pro-forms; cf. also the problems *1.8. Anaphoric distance* and *1.9. Cataphoric persistence*), deictic elements, tense, modality, aspectuality>.

Each of the vector elements corresponds to one dimension of cohesion and should be measured on a metrical scale. Describe the mathematical properties of these measures.

a) Can texts and even languages be compared with respect to cohesion means using your cohesion measure?
b) How can observed differences in cohesion be tested for significance?
c) Does text classification based on the cohesion measure make sense?

Ad a). If you define a vector as shown above and all elements are restricted e.g. to the interval <0,1> or normalized in some way, you can compare languages simply by the angle between the vectors. There is a great number of similarity measures that can be used for this purpose.

Ad b). The elements of the vector can be used for defining an indicator whose sampling distribution must be derivable. If at least the variance is known, two texts can be tested for significance using the normal approximation.

Ad c). Before you perform a classification, the background of ordering should be known. If the classes are not interpretable, the value of the classification is problematic. But a precedent classification of texts in text-sorts may be useful and the classification by cohesion may corroborate or reject it. Nevertheless, classification in social sciences yields only fuzzy sets.

A more prolific approach would be the search for links between the elements of the above mentioned vector. This could lead to a control cycle of cohesion and the first steps towards a theory.

References

Carrell, P. (1982). Cohesion is not coherence. *TESOL Quarterly 16(4), 479-488.*

Beaugrande, R. de, Dressler, W. (1981). *Einführung in die Textlinguistik.* Tübingen: Niemeyer.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers 36(2), 193-202.*

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

Halliday, M.A.K., Hasan, R. (1976). *Cohesion in English*. London: Longman.

Kaakinen, J.K., Salonen, J., Venäläinen, P., Hyönä, J. (2011). Influence of text cohesion on the persuasive power of expository text. *Scandinavian Journal of Psychology 52, 3, 201-208.* DOI: 10.1111/j.1467-9450.2010.00863.x

Kaufmann, S. (2013). Cohesion and collocation using context vectors in text seg--mentation. http://faculty.wcas.northwestern.edu/kaufmann/ Papers/ pacling99.pdf (Febr. 20,2013).

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics. Vol 2: 71-73.* Lüdenscheid: RAM-Verlag.

Morris, J., Hurst, G. (2012). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. http://acl.ldc.upenn.edu/J/J91/J91-1002.pdf (March 1, 2012)

Yde, Ph., Spoelders, M. (1985). Text cohesion: An exploratory study with beginning writers. *Applied Psycholinguistics 6, 407-415.* DOI: 10.1017/S0142716400006330 (March 1, 2012)

3.11. Arc length of frequencies

Problem

Determine the empirical rank-frequency distribution of words or, if a variable is measured in some way, the distribution of the values, e.g. word lengths, sentence length, motif lengths, distances between equal words, etc. Then compute the relative arc length between the neighbouring frequencies and interpret the result linguistically. Compare several texts and languages.

Procedure

First determine the empirical distribution of a linguistic variable obtained from a text. Compute the arc length L between neighbouring frequencies f_i and f_{i+1} according to the formula

$$L = \sum_{i=1}^{V-1} [(f_i - f_{i+1})^2 + 1]^{1/2}$$

and divide it by the maximum arc given as

$$L_{max} = \sqrt{(N-V)^2 + 1} + V - 2,$$

where N is the sum of all frequencies and V is the highest rank (vocabulary). Show that for large N the maximum L_{max} can be simplified.

Interpret $L_{rel} = L/L_{max}$ as an expression of stereotypy, structural uniformity, etc., according to the evaluated unit.

You can define new units taken from psycholiguistics, dialectology, etc. Evaluate in any case several texts and compare them. If you use L_{rel}, you can consider it a proportion and perform the usual normal tests; if you use only L, apply the variance of L defined in Popescu, Mačutek, Altmann (2009) and construct a test based on normality.

Show that L_{rel} does not depend on text size. Perform a linear regression (between L_{rel} and N) and show that the regression parameter is almost zero.

Find a difference between text sorts based on relative L, i.e. find a difference between languages, e.g. by comparing the translations of a work (e.g. *The Little Prince* by Exupéry). Compare means of many texts and adapt the variance.

Interpret the meaning of the arc.

References

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

Popescu, I.-I., Zörnig, P., Altmann, G. (2013). Arc length, vocabulary richness and text size. *Glottometrics 25, 43-53.*

3.12. Sentence sequences

Problems and procedures

Just as sequences of repeated parts of speech can be found in every text, one can find also sequences of repeated sentence types. One can apply any kind of classification of sentence types. In this way, one obtains a sequence of symbols and can study the properties of this sequence.

(1) Count the frequency of individual symbols and set up the rank-frequency distribution. What kind of distribution is it? Is it equal for all texts and are there only parameter differences; or do the texts differ basically?

(2) Compute the entropy, repeat rate and Ord's criteria I and S and compare the texts.

(3) Compile a table of transition frequencies from one type to another. In this way one you obtain the frequency of pairs of sentences. Alternatively, state the frequencies of bigrams, then that of trigrams up to pentagrams.

(4) Order the n-grams (separately) according to decreasing frequency and compute again the entropy, the repeat rate and Ord's criteria for each.

(5) Show that with increasing number of elements in sequences (n-grams) something changes. If entropy, repeat rate and Ord's criteria are not suffi-

cient, find some other possibility of characterization. Strive for corroborating the following regularities:

(a) With increasing *n* of n-grams the entropy increases, the repeat rate decreases. Find the regularities and express them using empirical formulas.
(b) Show that Ord's criterion takes a very regular course, at least in one language.
(c) Show that the quantitative regularities are different in different text-sorts. Hence characterize a text-sort as a special phenomenon having a special sequential structure.

(6) Compute now the distances between equal sentence types A. Find a distribution of distances or at least the mean distance. Then count the distances between sentence type A and sentence type B, i.e. select two types and count the number of steps from A to B. If there are two A's before the next B comes, omit the first A; then take the next A (behind the given B) and count the distance to the next B. You obtain a distance distribution. Perform this procedure for all sentence types and find a common distribution for all of them. If it does not exist, search for a common model from which all these distributions can be deduced.

(a) Compare the distributions of D(A,B) and D(B,A) for all symbols. Are they symmetric (homogeneous), asymmetric, etc.?
(b) Study all stage plays by an author and elaborate on his technique of sentence sequences. Strive for finding regularities.
(c) Replace the symbols denoting sentence types by their frequencies and you obtain a time series. Search for any type of dependence in this series. Study correlations, Hurst's exponents, possible oscillations, arc length, motifs.
(d) Compare different text-sorts from the view of problem (5c).

References

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 Academic Press.

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-sequences. In: Preisach, C., Burghardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data analysis, machine learning and applications: 637-646.* 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.* Vienna: Praesens.

Köhler, R. (2012). *Quantitative syntax analysis*. Berlin-Boston: de Gryuter.

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

Schmill, M.D., Oates, T., Cohen, P.R. (1995). Tools for Detecting Dependencies in AI Systems. *Proceedings of the 7th IEEE international conference on tools with artificial intelligence*: *148-155.* http://dx.doi.org/10.1109/TAI.1995.479507

Kokol, P., Podgorelec, V. (2000). Complexity and human writings. *Complexity International 7, 1-6.*

Oates, T, Schmill, M.D. Gregory, E.D., Cohen, P.R. (1996). Detecting complex dependencies in categorical data. *Lecture Notes in Statistics 112, 185-195.*

Oates, T., Schmill, M.D., Gregory, D.E., Cohen, P.R. (1995). Detecting complex dependencies in categorical data. In: *Finding Structure in Data: Artificial Intelligence and Statistics V.* Heidelberg: Springer.

Narasimhan, S.L., Nathan, J.A., Murthy, K.P.N. (2005). Can coarse-graining introduce long-range correlations in a symbolic sequence? *Europhysics Letters 69(1), 22-28.*

Paluoš, M. (1996). Coarse-grained entropy rates for characterization of complex time series. *Phys. D, 93, 64–77.*

Pompe, B. (1993). Measuring statistical dependencies in a time series, *J. Stat. Phys. 73, 587-610.*

Buiatti, M., Grigolini, P., Palatella, L. (1999). A non extensive approach to the entropy of symbolic sequences. *Physica A: Statistical Mechanics and its Applications 268, 214-224.*

3.13. Verbal antonymy

Problem

Texts may be differently „active“. One can compare activity with descriptivity (verbs vs. adjectives) or scale the activity of verbs (cf. *Problems Vol 3. 113-116*). However, there is still another, easier way to describe activity by means of active verbs, their antonyms or negations.

Set up hypotheses such as e.g. epic poetry is more “active” than lyrics, scale the activity of verbs and test the differences.

Procedure

Choose an epic text and make a list of the verbs. State how many times each of them occurs. Omit modal and auxiliary verbs. Then state which of them has an

antonym or only a negation. Thus you obtain two groups: verbs in texts, existing antonyms, only negations. Using these two numbers, characterize the given text by an indicator and derive its variance in order to be able to compare texts with respect to your new measure. Or consider the numbers as classes of a binomial distribution in order to be able to apply exact tests. To this end define an indicator in form of a proportion.

Then analyze a not too short lyric poem and perform the same operations. You can compare it with individual classes of the epos (for homogeneity) or you can compare values of the established indicator or you can compare the parameters of the binomial distribution.

At last, perform this analysis for several texts and strive for a theoretical insight. Compare also newspaper and scientific texts.

Practically, mark each verb in the text (except for modal and auxiliary verbs) with PA (which posses an antonym) or PN (possesing only a negation) and perform a count. The resulting number can be expoited for characterizations and comparisons. Examples are G. “erblassen” has no antonym, only a negation; “move” has an antonym (“stay”).

References

Agricola, C., Agricola, E. (1984). *Wörter und Gegenwörter. Antonyme der deutschen Sprache.* Leipzig: Bibliografisches Institut.

Bulitta, E., Bulitta, H. (2003). *Wörterbuch der Synonyme und Antonyme.* Frankfurt: Fischer.

Cruse, D.A. (1992). Antonymy revisited: Some thoughts on the relationship between words and concepts. In: A.J. Lehrer, E.F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization*: *289–306*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lehrer, A.J., Lehrer, K. (1982). Antonymy. *Linguistics and Philosophy 5, 483-501*.

Mettinger, A. (1994). *Aspects of semantic opposition in English*. Oxford: Clarendon Press.

Müller, W. (1998). *Das Gegenwort-Wörterbuch. Ein Kontrastwörterbuch mit Gebrauchshinweisen.* Berlin: de Gruyter.

Murphy, M.L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge: Cambridge University Press.

3.14. Hurst Exponent

Hypothesis

The higher the abstraction level of a property of a linguistic unit, the more volatile is the given text sequence. Test the hypothesis.

Procedure

In order to state the volatility or persistence of a series, one usually computes the Hurst exponent. If H < 0.5, the series is volatile; if H > 0.5, the series is persistent. Here we propose to compute the rescaled range performing the following steps as described in many sources:

First compute the mean of the series containing *n* elements as

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i \quad .$$

Then adjust the individual values subtracting this mean from each element, i.e.

$$Y_i = X_i - m, \qquad i = 1,2,\ldots,n.$$

Now, compute the partial sums of these adjusted values, i.e.

$$Z_k = \sum_{i=1}^{k} Y_i \text{ for } k = 1,2,\ldots,n$$

i.e. $Z_1 = Y_1$,
$Z_2 = Y_1 + Y_2$
…….

In the next step, for each *k* we take the range of Z-values, i.e.

$$R_k = \max(Z_1, Z_2,\ldots,Z_k) - \min(Z_1, Z_2,\ldots,Z_k) \text{ for } k = 1,2,\ldots,n.$$

Finally, we compute the standard deviation of the X_i-values up to the *k*-th element, i.e.

$$S_k = \left[\frac{1}{k}\sum_{i=1}^{k}(X_i - \bar{x}_k)^2\right]^{1/2}$$

and set up the series $(R/S)_k$ representing the sequence we search for. It will be assumed that

$$(R/S)_k = ak^H$$

in which the exponent *H* represents the Hurst exponent.

H can be used to compute the Hausdorff-Besicovitch dimension of the series (cf. Mandelbrot 1982: 249ff) which is defined as

$D = 2 - H$.

Many researchers compute *R/S* using simply the *X*-values without considering the difference from the mean (*Y*) and setting up sums (*Z*).

Perform this analysis for word length (in terms of syllable numbers), text length (in terms of word numbers), verb valency and polysemy. That is, evaluate a text and form separately sequences of word length, sentence length, verb valency and polysemy of individual words. Compute the Hurst exponent for all properties and show that the higher the abstraction, the more volatile is the sequence (i.e. the smaller is *H*).

Apply this procedure to form sequences of other types (between purely material ones and semantic ones) and set up an order of properties.

Compare different text-sorts and scrutinize the form of the hierarchy. Finally, perform similar analyses in texts of several languages. Are you tracing down a language law? If so, set up a new hypothesis, use the background proposed by the unified theory (Wimmer, Altmann 2005) and derive the hierarchy formula deductively. Care for strict definitions and exact measurement.

Employ another definition of a chaotic series and compute e.g. the relative arc length (cf. Popescu et al. 2009). Study the behaviour of relative arc length on different levels.

References

Anis, A.A., Lloyd, E.H. (1976). The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika 63, 111-116*.

Chamoli, A., Bansal, A.R., Dimri, V.P. (2007). Wavelet and rescaled range approach for the Hurst coefficient for short and long time series. *Computers & Geosciences 33(1), 83-93.*

Feder, J. (1991). *Fractals.* New York: Plenum Press.

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

Hurst, H.E. (1951). Long-term storage of reservoirs: an experimental study, *Transactions of the American society of civil engineers 116, 770-799.*

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

Peters, E.E. (1994). *Fractal market analysis: applying chaos theory toinvestment and economics.* New York: Wiley.

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

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

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.

3.15. Hreb construction

Problem

Hrebs have been defined in several problem descriptions (cf. Problems Vol. 1: 46 ff., Problems Vol. 2: 66, 126 f.; Problems Vol. 3: 111, 134) and in several publications (s. the References section). The definitions differ because the authors took into account different units and their properties.

Order the existing definitions, prepare a hierarchy (if there is any) of hrebs and define the properties of hrebs at individual levels. Model the behaviour of properties in a unique way, i.e. begin with a model of the simplest hreb and derive the complete system deductively by adding new parameters or changing the basic presuppositions.

Procedure

Define at least the following hrebs: morpheme-, word/lemma/synonymy-, phrase-, clause-, sentence-, reference- and co-reference-, anaphora and cataphora-, speech act- and other. It is better to begin with those you consider the simplest ones. Analyse several texts for each and set up models for the distribution of hreb sizes, distances between individual elements of the hreb, Ord's criteria, text concentration based on hrebs, h-point, hreb diffusity, compactness, connectivity, etc.

Link all properties with some functions and make draw a chart of the control cycle. Strive for at least a kernel of a possible theory. Publish even individual text analyses, i.e. analyse texts according to all hreb definitions, compute some properties of the given hrebs and show how they differ according to their place in the hierarchy.

Derive for every indicator its variance and compare texts using the normal distribution.

References

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

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

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

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

Hřebíček, L. (2007). *Lectures on text theory*. Prague: Oriental Institute**.**

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

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

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter. (esp. Chapter 11).

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

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

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

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

4. Pragmatics

4.1. Speech act distribution

Problem

Determine the distribution type of speech act categories in texts. Compare the parameters of the frequency distributions with respect to text sorts and other text parameters such as text length, age of the author, etc. Is the distribution type common among all text sorts? If not, find a plausible relation between text-sort and distribution type (cf. also Problems, Vol 1, *Polylogue analysis*; this volume: *Small inventories*)

Procedure

Base your study on an inventory of speech acts derived from one of those proposed in the corresponding literature (Austin, Searle, or modern authors). If you study a small inventory consisting of the basic categories such as {assertive, directive, commissive, expressive, declaration} there won't be enough degrees of freedom to successfully fit a probability distribution to the ranked data. Instead, you can use a function such as the Popescu model (Popescu et al. 2010) or the Zipf-Alekseev model. Annotate the speech acts in the texts and count the frequencies of speech act category occurrences.

If you set up a larger inventory on the basis of a finer differentiation of speech act categories, the classical approach using probability distributions is likely to work. You should make sure that the categories in the inventory do not vary with respect to their level in the classificatory hierarchy, i.e. that they share the same level of generality and granularity.

Fit the model (distribution or function) to the data. Test the differences between the parameter values for the text-sorts, authors, etc. for significance.

Having stated the distribution of speech acts, compute the entropy, the repeat rate and Ord's criteria. Show that different authors may differ as to the size of these indicators.

If you analyzed several texts of the same text-sort, show that as to speech acts they are homogeneous. Test either the differences of frequencies or those of rankings. If there are differences, redefine the text-sorts or show that in texts which differ from the general image there are some boundary conditions. Search for the boundary conditions and if possible insert them in the formulas.

References

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piot-

rowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 646-658.* Berlin-New York: de Gruyter.
Austin, J.L. (1962). ***How to do things with words***. Cambridge (Mass.): Cambridge University Press.
Ballmer, T.T., Brennenstuhl, W. (1981). *Speech act classification: a study of the lexical analysis of English speech activity verbs.* Berlin: Springer.
Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law – another view. *Quality and Quantity 44(4), 713-731*.
Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.
Searle, J. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

4.2. Speech act motifs

Problem

A motif is a new linguistic unit introduced by Köhler (2006, 2008) in analogy to Boroda's F-motif in musicology. It is defined as the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit. Such motifs can be formed on the basis of any linguistic unit and any quantified property which can be applied to the selected unit.

An example of an L-motif segmentation, i.e. of a motif on the basis of length, is the following. The sentence "Word length studies are almost exclusively devoted to the problem of distributions." is, according to the above-given definition, represented by a sequence of five L-motifs: (1-1-2) (1-2-4) (3) (1-1-2) (1-4), if the definition is applied to word length measured in the number of syllables. Similarly, motifs can be defined for phones, phrases [types], clauses [types], etc.) and for any quantified linguistic property (polysemy, polyfunctionality, polytextuality, etc.).

The given definition applies only to quantitative properties, i.e. to those properties which are measured on a metric or interval scale and yield numbers as values. Give a definition of motif on the basis of categorical variables such as the speech acts. Form motifs using this definition and investigate the frequency distribution of the motifs in texts, the distribution of their length etc. in analogy to the studies on numerical motifs in the literature. Cf. the problems concerning R-motifs and D-motifs in this volume.

Procedure

Set up an inventory of speech acts and annotate texts accordingly (cf. the problem *4.1. Speech Act Distribution*).

Motifs on the basis of categorical properties can be defined in various ways, e.g. as uninterrupted sequences of equal categories or as sequences of unrepeated acts etc.

All kinds of quantitative studies can be conducted on data collected using one of these methods in analogy to the study of properties of words (vocabulary, TTR, frequency, length etc.) and other units. Moreover, categorical motifs become quantitative (numerical) ones as soon as their quantitative properties are determined, e.g. by forming motifs on the basis of the length, frequencies etc. of the speech act motifs.

Further properties can be found in Strauss, Fan, Altmann (2008) and in the problem 4.3 *Speech acts in stage plays.*

References

Austin, J. L. (1962). *How to Do Things with Words.* Cambridge (Mass.): Cambridge University Press.

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 8th International Conference on Quantitative Linguisttics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 35-43*. Belgrade: Academic Mind**.**

Köhler, Reinhard (2014, to appear). Linguistic Motifs. In: Mačutek, J., Mikros, G. (eds.), *Sequential Analysis*. Berlin, New York: de Gruyter.

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

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

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learning and Applications: 637-646.* 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: 81-89.* Wien: Praesens,.

Searle, J. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

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

4.3. Speech act length

Problem

The length of speech acts can be measured in terms of word numbers. Analyze a stage play and show that the length of speech acts is distributed according to

$$P_x = \binom{k+x-1}{x} \frac{p^k q^x}{1-p^k}, \quad x = 1,2,3,...$$

where x is length, k and p are parameters, $q = 1-p$, i.e. according to the *positive negative binomial distribution.*

Procedure

Consider only illocutive speech acts. Ascribe each act its length and set up the empirical distribution of lengths. Fit the above distribution to your data.

Derive the recurrence function and interpret it as the basic difference equation. Interpret the parameters in terms of linguistic concepts or principles.

Study the individual persons separately. Are there differences of the parameters?

Study the sequence of lengths. Is there a visible tendency, e.g. do the acts become longer (or shorter) from the beginning of the text to the end? You can pool e.g. ten subsequent acts and compute their average length. Can one find any autocorrelation of the length values? Compute Hurst's exponent and make general statements.

Construct motifs of lengths (cf. Köhler 2006, 2008a,b; Köhler Naumann 2008, 2010) and study the distribution of motif lengths. Find the appropriate distribution (cf. the problem *4.2. Speech act motifs*).

References

Austin, J.L. (1962). *How to do things with words*. Cambridge (Mass.): Cambridge UP.

Günderson, K. (ed.) (1975). *Language, Mind, and Knowledge*. Minneapolis: University of Minneapolis Press.

Holdcroft, D. (1978). *Words and Deeds: Problems in the Theory of Speech Acts.* Oxford: Clarendon Press.

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

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

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learning and Applications: 637-646.* 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: 81-89.* Wien: Praesens.

Rothe, U., Altmann, G., Wagner, K. (1992). Verteilung der Länge von Sprechakten in der Kindersprache. In: Wagner, K.R. (Hrsg.), *Kindesprachstatistik: 47-56.* Essen: Die blaue Eule.

Searle, J. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

Steinsträter, Ch. (1988). *Zur kindlichen Sprachhandlungsfähigkeit: Analyse eines Korpus spontaner Sprechsprache einer Fünfjährigen.* Dortmund: Diss.

Tsohatzidis, S. (ed.) (1994). *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*. London: Routledge.

Wagner, K.R., Steinsträter, Chr. (1987). Wörterbuch der illokutiven Typen zum Korpus Teresa. In: Wagner, K.R. (Hrsg.), *Wortschatz-Erwerb: 59-81*. Bern: Lang.

Zechner, K. (2013). Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains. http://www.cs.cmu.edu/~zechner/sigir01.pdf (15.2.2013)

4.4. R-motifs of speech acts

Problem

State the distribution of length of R-motifs of speech acts in a stage play, set up indicators of the distribution and compare different stage plays.

Procedure

First read the problem description *1.16. Length of R-motifs*, choose a stage play or the transcript of a discussion and take down the corresponding sequence of speech act symbols. Then partition the sequence into R-motifs and state the length of individual motifs. Set up the frequency distribution of lengths, charac-

terize it by means of several indicators and derive a corresponding theoretical model. Interpret the model linguistically.

If you can scale the speech acts in some way, ascribe the individual elements of R-motifs numbers and scrutinize them in two ways:

(1) Compute the mean value of each R-motif and study the resulting time series. Characterize the series in different ways.
(2) For each R-motif state the difference between the greatest and the smallest scaling value, i.e. compute the range of values. Thus each R-motif is characterized by one value. (a) State the distribution of these range values; (b) Set up the time series of these values and evaluate it.
(3) Use all these indicators to characterize a stage play.
(4) Compare several stage plays. You can take also prose or poetry and elaborate on their characteristic R-motifs.
(5) When you have analyzed several texts and obtained the resulting values, join all aspects in a control cycle and add also other properties of texts. I.e. strive for a theoretical approach.

References

Austin, J. L. (1962). *How to Do Things with Words.* Cambridge (Mass.): Cambridge University Press.

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

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

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

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learning and Applications: 637-646.* 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: 81-89.* Wien: Praesens,

Searle, J. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

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

4.5. Scaling speech acts

Problem

Is it possible to find a scaling method for speech acts? (Cf. *Problems in Quantitative Linguistics, Vol. 2: 10.4*) If so, characterize a drama and show the development of degrees in different parts of the drama, e.g. what is characteristic of the catharsis? Compare dramas with comedies and develop a formal description of stage plays.

Procedure

First take into account the fact that speech acts are acts, and acts have properties, and properties can always be scaled. Then consider all kinds of speech acts defined up to now and perform the first tentative scaling. You may define several dimensions and place the acts in at least one of them. If possible, consider the dimensions as parts of a unique space.

For illustration we can take the scaling of children acts performed by Doerge (1975) as follows:

1. labelling
2. repeating
3. answering
4. requesting (action)
5. requesting (answer)
6. calling
7. greeting
8. protesting
9. practising

Having done this conceptual work, analyze a drama, differentiate the individual persons and set up a frequency distribution of their individual speech act inventories. Each person can be characterized by a vector whose elements are weighted averages of individual properties in the given dimension. Thus a text can be characterized by setting up an indicator constructed on the basis of realized vectors of properties of speech acts.

If you succeeded to perform the above operations, show the difference of the indicator: (a) in the development of a given drama, (b) between two dramas, (c) between a drama and a comedy, (d) a drama and a New Year speech of a president.

References

Alston, W.P. (2000). *Illocutionary Acts and Sentence Meaning.* Ithaca: Cornell University Press.

Austin, J.L. (1962). *How to Do Things with Words*. Cambridge (Mass.): Cambridge University Press.

Doerge, F.Ch. (2004). *Illocutionary Acts - Austin's Account and What Searle Made Out of It*. Tübingen: Diss.

Ehrhardt, C., Heringer, H.J. (2011). *Pragmatik.* Paderborn: Fink.

Erler, B. (2010). *The speech act of forbidding and its realizations: A linguistic analysis*. Saarbrücken: VDM Verlag Dr. Müller.

Gaynesford, R.M. de (2009). Illocutionary acts, subordination, and silencing analysis. *Analysis, 69 (3), 488-490.*

Greimann, D., Siegwart, G. (eds.) (2007). *Truth and Speech Acts: Studies in the philosophy of language*. New York: Routledge.

Hindelang, G. (2010). *Einführung in die Sprechakttheorie. Sprechakte, Äußerungsformen, Sprechaktsequenzen*. 5., neu bearbeitete und erweiterte Auflage. Berlin /New York: Mouton de Gruyter.

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

Levinson, S.C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Liedtke, F. (1998). *Grammatik der Illokution: über Sprechhandlungen und ihre Realisierungsformen im Deutschen.* Tübingen: Narr.

Meijers, A.W.M. (1994). *Speech Acts, Communication and Collective Intentionality beyond Searle's Individualism*. Utrecht: A.W.M. Meijers

Searle, J. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

Searle, J.R., Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press.

Staffeldt, S. (2008). *Einführung in die Sprechakttheorie. Ein Leitfaden für den akademischen Unterricht.* Tübingen: Stauffenburg.

Ulkan, M. (1993). *Zur Klassifikation von Sprechakten. Eine grundlagentheoretische Fallstudie*. Tübingen: Niemeyer.

Wunderlich, D. (1976). *Studien zur Sprechakttheorie.* Frankfurt: Suhrkamp.

5. Synergetics

5.1. Word length and polysemy

Problem

In several languages it has been stated that word length and word polysemy are interrelated and display a very conspicuous dependence. Compare the results in Chinese, German, English, Maori and Romanian languages, add the result(s) obtained from your language(s) and (a) state whether the parameters of the function linking these two properties are equal in all languages; (b) if not, investigate the relation between the parameters *a* and *b* of the resulting power function; (c) quantify and measure some further properties of the word in the given languages and find the link between the parameter *b* (exponent) and a third property of word.

Procedure

Departing from the the data from the attached references, compute the variance of the parameters and perform an asymptotic normal test for difference for each pair of languages. The variances of *a* and *b* are either furnished by the applied fitting software or one finds the formulas in text-books on regression analysis. Alternatively, you can test the parallelity of the linearized functions.

If there are differences, state whether *b* depends on *a* and find the form of this dependence. Apply a statistics software package; the relationship need not be preliminarily substantiated theoretically.

Select a third property *X* using Köhler's works or *Problems in Quantitative Linguistics Vol 1 (3rd ed.:141 f.),* and show the interrelation between *X* and length, and *X* and polysemy.

Find the linguistic requirements causing these relationships. Construct a control cycle step by step.

Cf. also *Problems in Quantitative Linguistics Vol. 2: 100.*

References

Breiter, A.M. (1994). Length of Chinese words in relation to their other syntactic features. *JQL 1(3), 224-231.*

Guiter, H. (1974). Les relations (fréquence-longueur-sens) des mots (languages Romanes et Anglais). *Atti del Congresso Internazionale di Linguistica 14/4, 373-381*. Napoli-Macciaroli-Amsterdam: Benjamins.

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

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

Köhler, R. (2002). Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: http://ubt.opus.hbz-nrw.de/volltext/2003/279/

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

Köhler, R. (1999). Der Zusammenhang zwischen Lexemlänge und Polysemie in Maori. In: Genzor, J., Ondrejovič, S. (eds.), *Pange lingua: 27-33.* Bratislava: Veda.

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

Nemcová, E., Rensinghoff, S. (2010). On word length and polysemy in French. *Glottotheory 3(1), 83-88.*

Rothe, U. (1993). Wortlänge und Bedeutungsmenge. Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. *Glottometrika 5, 101-112.* Bochum: Brockmeyer.

Strauss, U., Fan, F., Altmann, G. (2013[3]). *Problems in Quantitative Linguistics Vol. 1.* Lüdenscheid: RAM.

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

5.2. Kelih's Repeat Rate hypothesis

Problem

According to Kelih's (2010) hypothesis,

> *the greater is the first relative frequency (p_1) of the rank-frequency distribution of graphemes, the greater is the repeat rate.*

Since Kelih tested his hypothesis only in Slavic languages;

(1) generalize it to grapheme distributions in other languages;
(2) generalize the hypothesis to other units (phonemes, morphemes, syllables, words,…);
(3) state whether Kelih's hypothesis $RR = 0.2975 p_1^{0.75}$ − for which he obtained $R^2 = 0.86$ − still holds;

(4) compute separately the power function for different language units and state how the parameters change.

Procedure

Collect known rank-frequency distributions of different units from different languages. One can find them on the Internet, in journals devoted to quantitative linguistics and in individual book publications. Collect at least fifty cases. Then compute for each one the repeat rate defined as $RR = \frac{1}{N^2}\sum_{i=1}^{n} f_i^2$, where f_i is the frequency of the unit at rank i, N is the sum of all frequencies and n is the size of the inventory(vocabulary). First compute the above power function for different units and different languages separately. Then unify the results step by step and show the behaviour of *RR* in its relation to the unit, to the language, degree of synthetism, etc.

The relationship cannot be taken for granted. It does not need to hold in short texts. However, we expect its validity because *RR* is a measure of concentration and the first relative frequency is its main component.

Hence find also the boundary conditions which must be inserted in Kelih's formula if an outlier appears in your data.

Show that the relation of entropy to p_1 is just the other way round and find the appropriate formula of this dependence.

References

Kelih, E. (2010). Ein empirischer Regelkreis: Graphemhäufigkeiten in slawischen Sprachen. *Glottotheory 3(2), 23-34*

5.3. Word-length and compositionality

Problem

The hypothesis „*the shorter a word, the more frequently it occurs in compounds*" (Altmann 1988) has been tested in Polish by Hammerl (1990). All parts of speech have been pooled but neither a dependence formula has been derived nor the different propensities of individual parts of speech have been scrutinized. Hammerl generalized this hypothesis by the statement: "*the number of compounds n_L (whose components have length L) is a function of the length L of the components*."

Test the hypothesis, distinguish parts of speech and derive a unique formula of dependence.

Procedure

First clear the concepts: definition of compounds, length of one component (the head) or mean length of all components, etc.

Then use Hammerl's data and find the probability distribution for Polish data or at least the dependence function $n_L = f(L)$. Check your result on data from another language. If you find differences, seek for the cause in the morphological state of the languages.

Distinguish individual word classes and adapt your theoretical result to the new empirical results. Strive for a general theoretical framework in which the boundary conditions are contained in the parameters. Use Altmann's version and perform the investigation in several languages using only nouns as heads.

Find, at least exploratively/inductively, the form(ula) of the dependence and substantiate the parameters by the kind of word class (= different parameters for nominal, verbal, etc. compounds) and by the morphology of language. Do not employ only semantic definitions of compounds but emphasize the morphological form.

References

Altmann, G. (1988). Hypotheses about compounds. In: Hammerl, R. (ed.), *Glottomerika 10: 100-107*. Bochum: Brockmeyer.

Hammerl, R. (1990). Überprüfung einer Hypothese zur Kompositabildung (an polnischem Material). In: Hammerl, R. (ed.), *Glottometrika 12: 73-83*. Bochum: Brockmeyer.

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

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

5.4. Allomorphic complexity

Problem

Find a numerical expression for the allomorphic complexity of a language. Compare languages and generalize. Find the relations to other properties.

Procedure

Take a random sample of lemmas from a dictionary, or compile a list of the lemma types found in a text. For each lemma state the number of allomorphs of

the corresponding lexemes. You should take into account also the allomorphic variants which occur in spoken language. In any case, distinguish written and spoken language. If you work with a text, you may take into account each word only once or you can count each word as many times as it occurs (a procedure expressing the weight of the word). Determine the empirical distribution of allomorph numbers, i.e. state how many words (morphemes) have 1,2,3,… allomorphs.

Derive a theoretical distribution or function that can capture this phenomenon and can be linguistically well substantiated.

(1) Characterize the language or text by some indicators, e.g. entropy, repeat rate, Ord's criteria (on the basis of the above distribution). E.g. a language whose entropy is zero, does not have allomorphs and tends to analyticity, monosyllabism. Its repeat rate is then 1. A language with great entropy tends to synthetism, it has a number of grammatical categories. Compare inflectional and introflectional languages.
(2) Show in which intervals of the mean and the variance of your individual results languages can be placed. Analyze several languages.
(3) Find links between your indicators of allomorphic complexity and other properties of language, e.g. phoneme inventory, mean word length, etc.
(4) For languages which do not use alphabetic writing systems, examine only the spoken forms. In languages in which the written form strongly differs from the spoken one, analyze both the written and the spoken form and compute the divergence.

References

None.

5.5. Control cycle

Problem

Having read this volume, select from each chapter the pertinent units and their properties, and join them in a control cycle.

Procedure

1. First prepare a list of units and properties taking into account each chapter and consider the properties as vertices of a graph. Then, leaning against the conjectures in the given chapter or your own ones, join the corresponding verti-

ces with edges to represent the linguistic interrelations – without marking the direction of dependence.

2. Begin to theorize:
   (a) Set up hypotheses about the direction of the edges (many of them will be bidirectional) on linguistic reasons.
   (b) Define data and indicators that could be used for testing the given hypothesis.
   (c) Perform measurements in texts, dictionaries, languages.
   (d) Set up intuitively - or using a software - a mathematical hypothesis and test it on your data.
   (e) Use the technique of synergetic linguistics and find the factors ("forces") which may cause the given dependence.
   (f) Derive the hypothesis from the unified theory, i.e. substantiate it both linguistically and mathematically.

3. Perform step 2 for as many edges as possible and construct step by step your own image of the dynamics in language. If necessary, introduce new functions, new factors, new indicators. The capturing of just one of the many existing dependences is better than pure qualitative description or classification.

4. Strive for continuous extending your theory in all kinds of linguistic material. Do not mix texts. Care for boundary conditions and ascribe them to text sorts, authors, languages. If your function seems to deviate strongly for some data, generalize the formula. It is always possible.

References

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

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

Köhler, R. (2002). Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik: http://ubt.opus.hbz-nrw.de/volltext/2003/279/

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). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Language. Word Length Studies and Related Issues: 207-316*. Boston: Kluwer.

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

6. Various Issues

6.1. Scaling dogmatism

Problem

Content analytical methods include the use of the occurrence of selected words in a text as indicators of properties of the text under analysis, the author, the adressee, or the medium. Often, antonyms and other polar expressions are employed such as *good - bad* for purposes of the so-called sentiment analysis. We will focus here on a fuzzy property: dogmatism, and assume for the moment that appropriate indicators include the pairs *must - must not; whisper - shout; white - black; stand - run; consent - reject,* etc. Collect all words whose meaning lies between these extremes and propose a scaling. Then analyse the dogmatism of a text.

Procedure

In cross-linguistic studies, one may find different numbers of words between such extremes (they need not be antonyms) depending on the individual language. Hence, determine a fixed scale for all of them, say beginning from zero for the lowest rank of dogmatism up to the highest value for the maximum dogmatism of an expression. Increase the dogmatism values by 1 in each intermediate step. Finally, divide all values by the maximum rank. In this way one obtains for all phenomena values in the normalized interval <0; 1>.

Now evaluate a text underlining the expressions that can be scaled in this sense. There are also phrases expressing some of these categories without the respective word. Consider them, too, and ascribe them a degree. Then study the obtained sequence in the following way:

(1) Compute the mean dogmatism and its empirical variance for the whole text.
(2) Compute the same for a drama, but consider each person separately.
(3) Compare a press text with a juridical text and a poetic text.
(4) Study the course of dogmatism in the text deployment.
(5) Study the autocorrelation of the dogmatism degrees.
(6) If the course is smooth, capture it using an empirical function.
(7) Search for the dependence of dogmatism on the text sort, text aim, author, etc.
(8) Study the course of dogmatism in the life of a writer and compute the correlation with his age.

References

Christie, R. (1991). Authoritarianism and related constructs. In: J.P. Robinson, P.R.Shaver, L.S. Wrightsman (eds.), *Measures of Personality and Social Psychological Attitudes, Vol. 1*: *501-571.* San Diego, CAL: Academic Press.

Ehrlich, H.J. (1978). Dogmatism. In: H. Londorn, J.E. Exner (eds.), *Dimensions of personality: 129-164.* New York: Wiley.

Vacchiano, R.B. (1977). Dogmatism. In: T. Blass (ed.), *Personality variables in social behaviour: 281-314.* Hillsdale, NJ: Erlbaum.

6.2. Word frequency and initial clusters

Hypothesis

According to G. Fenk-Oczlon (2001: 438) „[..] initial consonant clusters are relatively rare within the class of the most frequent words."

Generalize the hypothesis to a more exact conjecture: *the greater the frequency of words (= the smaller the rank), the smaller the volume of initial consonant clusters.*

Of course, it can be tested only in languages with consonant clusters. Under "volume" is meant the number of consonants in the cluster.

Test the hypothesis.

Procedure

Consider a language with rich consonant clusters. Select the first thousand most frequent words from a frequency dictionary (of word tokens) and establish classes of words beginning with C-, CC-, CCC-, … Then ascribe to each of the words its rank according to its frequency. State whether the ranks (means or sums) are equal in all groups. Apply a non-parametric test such as the U-test or the H-test, etc. If necessary, form a unique group of all words beginning with more than one consonant and test its (mean) ranks against those beginning with exactly one consonant.

If there are initial clusters consisting of five consonants in the language you analyze, propose a function expressing the increase of volume of initial clusters with decreasing frequency (= increasing rank).

References

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

6.3. Frequency and position in text

Hypothesis

Frequent words appear in the text earlier, seldom words later on.
Test the hypothesis.

Procedure

Determine the rank-frequency distribution of the words in a lemmatized text. Annotate the lemmas with their position (or the position of the corresponding sentence) in which it occurred in the text for the first time. If the hypothesis is correct, then there must be a correlation between the frequency (or the rank) and the first occurrence of the lemma. The more frequent a lemma (= the smaller its rank), the earlier it appears in a text.

Study texts of different sorts and state in which sort this trend is best visible. Derive a function expressing this trend - if it exists. If you find a positive answer, explain this phenomenon and insert the forces which are active at its creation. Incorporate them in the formula.

In a second step, omit all synsemantics (function words) and repeat the procedure only with autosemantic lemmas. You may discover very different trends. Is it possible to classify the texts into text sorts with respect to the given trend?

The hypothesis seems to be quite logical, but it must be tested, especially the form of deviations.

References

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

6.4. Hapax legomena and synthetism

Hypothesis

The more synthetic a language is, the greater is the ratio of hapax legomena of word-forms in texts.

The hypothesis follows from the examinations by Tuldava (1995: 115ff) and Popescu, Altmann (2008) and seems to be a natural consequence of synthetism. In highly synthetic languages almost every word has several different word-forms depending on the complexity of the morphological system, and some

forms occur only once. The tail of the word-form rank-frequency distribution is the longer, the more word-forms there are in the text. It has been observed that in highly synthetic languages the respective part of the Zipf-function $f = cr^{-a}$ (where r = rank, f = frequency) is placed below the hapax legomena while in highly analytic ones it lies above the hapax legomena. Test the above hypothesis.

Procedure

Evaluate several texts in your language. Determine the rank-frequency distribution of word-forms for each text separately and fit iteratively the power function $f(r) = cr^{-a}$ to the data, where $f(r)$ is the frequency, r is the rank and a, c are iteratively obtained parameters. Then compute for each text the analytism indicator

$$A = \frac{2^a c}{(1,2744V - 18,6979)^a}$$

(where V is the size of the text vocabulary) proposed on empirical grounds by Popescu, Altmann (2008). Calculate the mean of all A for the analyzed texts. You obtain an indicator A between 0,2 and 5,1 and can find the place of your language among 20 languages analyzed in this way in the above article.

Check the result using the Greenberg-Krupa indices (Greenberg 1960; Krupa 1965) and find the correspondence between them and A. Improve the above indicator or propose new ones. If you have analyzed very "extreme" languages, normalize the indicator A.

References

Crystal, D. (2006). *How language works*. New York: The Overlook Press.

Ernst, P. (2005). *Deutsche Sprachgeschichte*. Wien: Facultas Verlags- und Buchhandels AG.

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

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

Napoli, D.J. (1996). *Linguistics: An Introduction*. New York: Oxford U. Press.

Popescu, I.-I., Altmann, G. (2008). Hapax Legomena and language typology. *Journal of Quantitative Linguistics 15(4), 340-369.*

Tuldava, J. (1995). *Methods in Quantitative Linguistics*. Trier: WVT.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Zipf's law in Italian texts. *Journal of Quantitative Linguistics 16(4), 354-365.*

6.5. Compound degree

Problem

Study the indegrees and outdegrees of compounds consisting of two elements. State the probability distribution of the degrees. Then set up an assortativity hypothesis of the compounds in the given language.

Procedure

Extract the compounds which consist of two elements from a dictionary. It goes without saying that the orthographic conventions concerning compounds vary from language to language, and that compound recognition is not always a straightforward procedure.

Decompose the compounds into head and modifier; the English compound "living room", e.g. consists of the head "room" and the modifier "living". For the present study, it is determind that the head has one outdegree and the modifier one indegree. On this basis, find

(1) The compositionality of the given word, i.e. the number of compounds in which it occurs (as head or modifier). Determine the probability distribution of the variable X = compositionality (number of compounds in which it occurs) and $f(x)$ = number of different words with compositionality x. Find a discrete theoretical distribution capturing your data and substantiate it linguistically.

(2) For each word separately, state the difference between its outdegrees and indegrees, i.e. determine the distribution of the variable D = difference between indegrees and outdegrees of a word, and $f(d)$ = number of words with a given difference d.

(3) Consider a word and state its compositional outdegree. Then compute the mean outdegree of all words with which it forms a compound. If both the outdegree of the word and the mean outdegree of associated words are high, the word displays assortative compounding. If the outdegree is high and the mean outdegree of the associated words is low, the word displays dissortative compounding. If none of these tendencies can be found, the compounding is neutral (cf. Newman 2002, 2003). This is an elementary classification. In books on networks one can find more complex methods.

(4) Find all kinds of compounding (high-high, high-low, low-high, low-low) and compare the result with that of another language. Set up a classification of compounding in a language. If you consider not only classes but exact numbers, set up a tendency in the given language and compare it with other languages.

(5) Perform all these operations for compounds found in a corpus involving also the occurrence (frequency) of individual compounds.

(6) Compare the frequency of individual words in a corpus with the degree of their compositionality. Can the hypothesis be maintained that the greater the frequency of a word, the greater its compositionality? Test the hypothesis statistically. Use the results obtained in task (1).

Substantiate all hypotheses linguistically.

References

Batagelj, V., Mrvar, A., Zaveršnik, M. (2002). Network analysis of dictionaries. In: T. Erjavec, J. Gros (eds.), *Jezikovne tehnologije/Language Technologies: 143-148.* Ljubljana: *http:/nl.ijs.si/ijst20/zbornik/sdjt02-24batagelj.pdf.*

Ferrer i Cancho, R., Solé, R.V., Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E, 69, 051915 (Santa Fé Institute Working Papers 03-06-092).*

Newman, M.E.J. (2002). Assortative mixing in networks. *Physical Review Letters 89, 280701 (arXiv: cond-mat/0205405).*

Newman, M.E.J. (2003). Mixing patterns in networks. *Physical Review E 67, 026126 (arXiv: cond-mat/0209450).*

Steyvers, M., Tennenbaum, J.B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science 29(1), 41-78.*

Tamaoka, K., Meyer, P., Makioka, Sh., Altmann, G. (2008). On the dynamics of the compounding of Japanese kanji with common and proper nouns. *Journal of Quantitative Linguistics 15(2), 136-153.*

6.6. Diversification theory

Problem

Set up a „teorita“ of diversification in language. Describe the phenomenon, show the places where it can appear, list Köhler's requirements and forces that stimulate its rise, associate it with self-organization and find the mechanisms that control its rise and restriction. Derive the necessary formulas.

Procedure

First read all problems in the first three volumes of "*Problems in Quantitative Linguistics*" concerning diversification. They can be found in Volume I, pp. 96, 114, 121; Volume II, pp. 10, 40, 73, 80, 81, 93, 94 and Volume III, pp.:25-28, 33, 65, and in this volume. Take the results presented there as a starting point. Then

choose those requirements and forces that can be accounted for the individual forms of diversification. Consult Köhler's analysis (2005) and the references therein. Add further phenomena of diversification, classify them into phonetic, morphological, semantic, lexicological, syntactic, dialectal, idiolectal, etc. ones and find the distributions of their frequencies, show the dependence of the parameters on the character of the class, because in each class a different combination of requirements and forces may be active. Construct stepwise a hierarchical model. In the first step show only the directions of dependence and the forces; in the next step add the formulas you obtained; in the third step associate the parameters of the formulas with the forces and the levels in the hierarchy, i.e. interpret them.

Test your theory on data from another language, improve it, if necessary, and at last, set up a model from which all your results can be obtained deductively. Not all of the resulting formulas for individual cases of diversification will be equal but in any case strive for a unified theory.

References

Altmann, G. (1996). Diversification processes of the word. *Glottometrics 15, 112-133.*

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.

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.

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

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

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

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

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

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

For measuring diversity, see http://en.wikipedia.org/wiki/Diversity_index

6.7. Givón's hypothesis

Hypothesis

„*The more important an item is in communication, the more distinct and independent coding expression it receives*" (Givón 1985: 206; cf also Krug 2001: 323).

(1) Quantify each relevant concept in the cited hypothesis.
(2) Test the hypothesis on data from one language

Procedure

First define exactly the three basic concepts: *importance in communication; distinctiveness of coding expression*; *independence of coding expression.*

Since these concepts are properties, propose a method for their quantification. Your measure should meet the requirements of at least an ordinal scale. Choose a language and a linguistic phenomenon and perform the first measurements of these properties. If your quantification is prolific, you should obtain two increasing functions. Begin, for example, with the expressions of courtesy in different languages, especially those that have courtesy levels (Javanese, Japanese, Samoan and other Asian languages), or with already elaborated classifications/ orderings of grammatical phenomena (cf. Krug 2001: 329; Quirk et al. 1985).

Finally, set up a control cycle of the above mentioned properties based on either systems theoretical argumentation or differential equations.

References

Givón, T. (1985). Iconicity, isomorphism and nonarbitrary coding in syntax: iconicity in syntax. In: Haiman, J. (ed.), *Iconicity in Syntax: 187-220.* Amsterdam: Benjamins.

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

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

6.8. Laws in language

Problem

Write a book about the concept of law in linguistics. In order to find a framework, scrutinize also the history of this concept. Take into account the following issues:

1. The origin of the concept
2. Law in sciences today
3. Development of the concept of law in linguistics:
    (a) Neogrammarians,
    (b) Zipf,
    (c) Modern linguistics.
4. Hypothesis:
    (a) Definition,
    (b) Kinds,
    (c) Structure and form,
    (d) Range,
    (e) Systematicity,
    (f) Depth and cognitive status,
    (g) Strength,
    (h) Semantics,
    (i) Inception,
    (j) Abstractness,
    (k) Foundation,
    (l) Testability,
    (m) Function.
5. Hypothesis in linguistics; analyse and criticize several linguistic hypotheses from all points of view mentioned in 4. Do not omit works in Russian.
6. Law:
    (a) Linguistic conceptions
    (b) The way to a law
    (c) Kinds
    (d) Forms
    (e) Contents
    (f) Self-regulation
    (g) The function of law in theories
    (h) The role of the ceteris paribus condition
    (i) The meaning of parameters.
    (j) Explanation
7. History and description of some linguistic laws
8. Conclusions:
    (a) results
    (b) Recommendations
    (c) Further development

References

Armstrong, D. (1983). *What is a law of nature?* Cambridge: Cambridge University Press

Ayer, A.J. (1956). What is a law of nature? *Revue Internationale de Philosophie 10, 144-65.*

Beebee, H. (2000). The non-governing conception of laws of nature. *Philosophy and Phenomenological Research 61, 571-594.*

Bunge, M. (1967). *Scientific Research I-III.* Berlin: Springer.

Bunge, M. (1979-1983). *Treatise on Basic Philosophy IV-VI*. Dordrecht: Reidel.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.

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

Dretske, F. (1977). Laws of Nature. *Philosophy of Science 44, 248-268.*

Foster, J. (2004). *The Divine Lawmaker: Lectures on Induction, Laws of Nature, and the Existence of God*. Oxford: Clarendon Press.

Fraassen, B.v. (1989). *Laws and Symmetry.* Oxford: Clarendon Press.

Feynman, R.P. (1990). *Vom Wesen physikalischer Gesetze*. Piper: München.

Giere, R.N. (1999). *Science Without Laws*. Chicago: University of Chicago Press.

Gigerenzer, G., et al. (eds.) (1999). *Das Reich des Zufalls.* Heidelberg: Spektrum.

Hempel, C.G. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

Kneale, K. (1950). Natural Laws and Contrary-to-Fact Conditionals. *Analysis 10, 121-25.*

Lange, M. (2000). *Natural Laws in Scientific Practice*. Oxford: Oxford University Press.

Mackie, J.L. (1974). *The Cement of the Universe*. Oxford: Oxford University Press.

Popper, K. (1949). A Note on Natural Laws and So-Called Contrary-to-Fact Conditional. *Mind 58, 62-66.*

Salmon, W.C. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Salmon, W.C. (1989). Four decades of scientific explanation. In P. Kitcher and W.C. Salmon, (Eds.), *Scientific Explanation (Minnesota Studies in the History of Science*, vol. 13). Minneapolis: University of Minnesota Press.

Schank, R.C. (1986). *Explanation Patterns: Understanding Mechanically and Creatively.* Hillsdale, NJ: Erlbaum.

Schrödinger, E. (1997). *Was ist ein Naturgesetz? - Beiträge zum naturwissenschaftlichen Weltbild* - 5.Ed. München: Oldenbourg.

Schurz, G. (1995). *Kinds of Unpredictability in Deterministic Systems.* Salzburg: Institut für Philosophie.

Schurz, G. (2001). *Ceteris paribus laws*. PE Preprints 5. Universität Erfurt.

Schurz, G. (2001*). Pietroski and Rey on Ceteris Paribus Laws*. PE Preprints 3. Universität Erfurt.

Schurz, G. (1993). *Scientific Explanation: A Critical Survey*. IPS Preprints 1. Universität Salzburg.

Suppes, P. (ed.) (1977). *The Structure of Scientific Theories*. Urbana: University of Illinois Press.

Tooley, M. (1977). The nature of laws. *Canadian Journal of Philosophy 7, 667-698.*

Vollmer, G. (2000). Was sind und warum gelten Naturgesetze? *Philosophia naturalis, Journal for the Philosophy of Nature 37(2,) 205-239.*

6.9. Morphological complexity of words

Problem

Morphological complexity of a word can be measured in different ways. Collect all known indicators and perform all kinds of measurement. Find the distribution of complexity in texts and in language.

Procedure

As morphology studies the structure of words, morphological complexity, its definition, and its measurement crucially depend on the concept of word which is employed for an investigation. Therefore, this concept has to be determined first, then morphological complexity of a word can be defined. For the current purpose, we will differentiate syntagmatic and paradigmatic complexity.

(1) Syntagmatic complexity is concerned when the definition is based on the number of morphs, i.e. of the elements which are obtained after morphological segmentation of an observed word-form. This definition follows form, not content.

(2) Another way to defines complexity is based on the number of lexical and grammatical meanings expressed by the observed word-form. The analysis of the content side of a word deserves special attention. Avoid prejudiced opinions and premature equatings of phenomena in different languages. In Czech, the conditional auxiliar *by* ("*would"*) expresses also the category person (*bych, bys, by, bychom*), but not all persons. Consider the number of categories and lexical meanings as a random variable representing *paradigmatic complexity*.

Languages with high values of whatever kind of complexity are more synthetic than languages with low complexity values. The computed mean complexity for a given text correlates with the Greenberg-Krupa indicators. Perform the comparison.

Analyze texts of different text-sorts and compute the usual statistical indicators (moments, Ord's criteria), find a universally valid distribution holding true for all languages and characterize a language or a text sort in a language by the parameters of the distribution.

References

Bane, M. (2008). Quantifying and Measuring Morphological Complexity. In: C.B. Chang, H.J. Haynie (eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics: 69-76.* Somerville, MA: Cascadilla Proceedings Project.

Di Sciullo, A.M. (2012). *Perspectives on morphological complexity*. Université du Québec à Montréal. http://www.er.uqam.ca/nobel/asymet/pdf/Perspectives%20on%20morphological%20complexity%20AMDS.pdf

Feldman, L.B., Frost, R., Pnini, T. (1995). Decomposing words into their constituent morphemes: Evidence from English and Hebrew. *Journal of Experimental psychology: Learning, Memory & Cognition 21, 947-960.*

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

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

Lempel, A., Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions in Information Theory 22 , 75-81.*

Melinger, A. (2012). *Morphological Complexity in English Prefixed Words: An Experimental nvestigation.* Diss. Buffalo. http://wings.buffalo.edu/linguistics/people/students/dissertations/melinger_diss.pdf

Moscoso del Prado Martín, F. (2011). *The Mirage of Morphological Complexity.* http://mindmodeling.org/cogsci2011/papers/0836/paper0836.pdf

Schreuder, R., Grendel, M., Poulisse, N., Roelofs, A., Voort, M.v.d. (1990). Lexical processing, morphological complexity and reading. In: D. Balota, G. Flores d'Arcais (eds.), *Comprehension processes in reading: 125-141*. Hillsdale, NJ: Lawrence Erlbaum Associates.

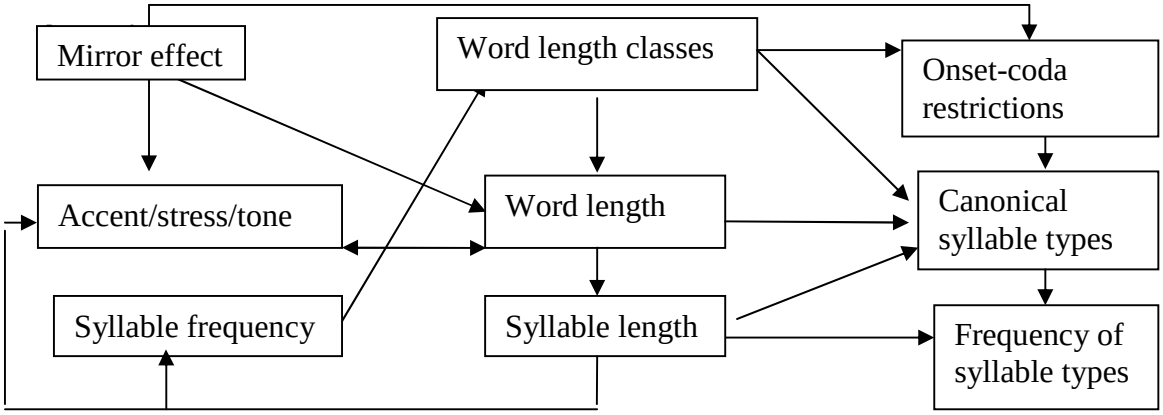
Siegel, J. (2004). Morphological simplicity in pidgins and creoles. *Journal of Pidgin and Creole Languages 19, 139-162.*

Stockall, L. (2012). A single route, full decomposition model of morphological complexity: MEG evidence. http://web.mit.edu/marantz/Public/Recent/StockallMarantz.pdf

6.10. Syllable length

Problem

E. Kelih (2012: 150) presented a control cycle of syllable length displayed in the figure below:



Test these hypotheses.

Procedure

Testing these hypotheses is possible only on data from a set of several languages. You should acquire as many data as available. The quantities mentioned in the following list have to be interpreted as mean values in all cases.

(1) The longer the word, the shorter are its syllables.
(2) The structure of syllables in short words is more complex than in long words. Measure the complexity in individual word length classes.
(3) The number of canonical syllable types (V, CV, CVC,…) depends on the word length class.
(4) For the rank-frequency distribution of individual syllables and canonical syllable types Zipf's law holds in some of its forms.
(5) The longer (= more complex) the syllable, the smaller is its frequency.
(6) The longer the syllable *type*, the smaller is its frequency.
(7) The onset does not have as many restrictions as the coda, in other words, the onset is on the average more complex than the coda (if there are consonant clusters in the language).
(8) The more symmetric is the syllable structure concerning onset and coda (= mirror effect), the smaller is the number of syllable types.
(9) The more suprasegmental properties a language possesses, the shorter is the word length on the average (Kempgen 1990: 119).
(10) The greater is the phoneme inventory, the shorter are the syllables.

Select some of the hypotheses, obtain data and propose functions capturing the given dependences. Start with Köhler's (1986, 2005) systems theoretical approach or derive the hypotheses by means of differential equations. Publish at least the results of your counts and measurements.

Enlarge the control cycle by adding further properties which are linked with at least one of the properties presented in the above control cycle. Approach a theory step by step.

References

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

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

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

Köhler, R. (2005). Synergetische Linguistik. 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).

6.11. Nominal compounding tendencies in German

Problem

Lewizkij and Matskulyak (2009) subdivided nouns into 34 semantic groups and analysed their tendencies to form compounds with words of a number of parts of speech. Test whether nouns in other languages display the same tendency to form compounds with other nouns, adjectives and verbs. The data are ordered in decreasing order of the number of nominal compounds.

Noun group	**N+N**	**A+N**	**V+N**
Objects and instrument	805	45	148
Person	689	73	45
Space and place	564	52	64
Properites of objects	483	36	36
Acts and behaviour	421	27	24
Buildings	404	26	51
Activity	320	34	37
Language and speech	291	25	30
Collective names of humans, organisations	288	30	22
Numbers, measurement units	248	26	10
Somatisms	234	43	26
Plants	209	28	31

Stuffs and materials	196	28	39
Time, age	192	11	29
Existence	181	21	7
Abstract notions	177	17	21
Motion	164	14	17
Science, culture, traditions	141	18	12
State and his attributes	136	18	2
Mental sphere	136	12	19
Eating and drinking	129	10	12
Spiritual sphere	122	13	21
Human attributes	120	13	18
Plays, shows	113	9	6
Physical properties	95	7	12
Terms	92	8	5
Documents, money	88	6	6
Natural phenomena and states	85	7	9
Animals	69	13	15
Properties of the Man	64	11	13
Physiological sphere	61	3	3
Possessorial sphere	60	8	2
Perception	24	8	3
Proper names	5	3	0

The above classification of nouns is one of the many possible ones. A different way of forming the classes and using different texts would probably yield other results.

Perform similar investigations also for other languages on data from different texts.

Procedure

The homogeneity of the columns cannot be tested by means of the chi-square test because there are too few data in some of the cells. You could either pool some similar classes so that all numbers in the tables exceed the value 5 – or you apply the information statistic 2I instead of the chi-square test . Apply also a rank test, e.g. Kendall's W-test, too, in any case because the numbers in the table depend on individual texts.

In order to test the adequacy of the above classification, find a theoretical rank-order distribution expressing the propensity of nouns to interact with dif-

ferent semantic classes. (Order separately the A+N and the V+N classes.) Perform this investigation for another language and compare the results. If you do not find an adequate distribution, perform a different classification of nouns and repeat the complete procedure.

For the three rank order distributions compute Ord's criteria *I* and *S*. Compare the results from German with those in other languages.

Extend the investigation and find also compound nouns consisting of noun + adverb.

References

Booij, G. (2007). *The Grammar of Words: An Introduction to Morphology* (2nd edition). Oxford: Oxford University Press.

Dunabeitia, J.A., Perea, M., Carreiras, M. (2007). The role of the frequency of constituents in compound words: Evidence from Basque and Spanish. *Psychonomic Bulletin & Review, 14(6), 1171-1176.*

Frisson, S., Niswander-Klement, E., Pollatsek, A. (2008). The role of semantic transparency in the processing of English compound words. *British Journal of Psychology, 99(1), 87-107.*

Haspelmath, M. (2002). *Understanding Morphology*. London: Arnold.

Lewizkij, V., Matskulyak, Y. (2009). Semantische Kombinierbarkeit von Komponenten in der Struktur der deutschen Komposita. *Glottometrics 19, 11-42.*

Libben, G., Jarema, G. (eds.), (2006) *The Representation and Processing of Compound Words*. New York: Oxford University Press.

Pollatsek, A., Hyona, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes, 20(1), 261-290.*

6.12. Quantification exercise

M.A.K. Halliday (2004) classified categorically some adjectives in three different ways: classes, subclasses and subsubclasses. Find a generic term for each set of classes and replace the categorical classification by an ordinal one. (The "type" can be omitted.) Then normalize the classification in order to get a scale in the interval <0; 1>. Analyze texts and characterize them numerically according to the adjectives occurring in them (Halliday 2004, p. 317).

Table 6(7) Adjectives frequently occurring as post-Deictic

type	sub-type		examples
expansion	elaborating	identity	identical, same; different ('non-identical'), other [note a + other, another]; respective
		exemplification	certain, particular, given; various, different ('various'), odd; famous, well-known, infamous, notorious; special
	extending		complete, entire, whole
	enhancing	space-time	above, aforementioned, earlier, preceding; subsequent, future
		comparison	similar, different ('non-similar'), comparable
projection	modality: modalization	probability	certain, possible, probable
		usuality	customary, habitual, normal, ordinary, typical, usual, regular
	modality: modulation	obligation	necessary, required
		readiness	intended, desired
	report	locution	alleged, so-called, self-styled
		idea	hypothetical, purported, expected, evident, obvious

Procedure

Ignore the "type" because it contains only two classes. Consider instead the second column and find a generic term for all these classes. Include in your considerations also psycholinguistic aspects. Decide which of the classes expresses the lowest degree of this generic property. Ascribe it degree 1. Then order the other classes accordingly and ascribe them the subsequent degrees. Dividing all classes by the highest degree yields values in the interval <0;1>.

Evaluate professional texts and find the frequency of individual relative degrees. You have now a continuous distribution - you can also consider it discrete if you do not normalize the degrees - and should find a theoretical function. Begin with the beta-distribution/function and characterize the text in terms of the resulting parameters. Compare texts.

Consider the third column and perform all operations you made for the second column. You obtain more degrees and the texts will be characterized by slightly different parameters.

If the beta-distribution is not adequate, find another one and derive it formally justifying it by means of linguistic arguments.

Find a better classification of adjectives or of another word class and perform the described steps with this one.

References

Halliday, M.A.K. (2004). *Introduction to functional grammar* (third edition, revised by Ch.M.I.M. Matthiessen). London: Arnold.

Farsi, A.A. (1968). Classification of adjectives. *Language Learning 18, 45-60.*

Levi, J. N. (1973). Where do all those other adjectives come from? In: C. Corum, T. C. Smith-Stark, A. Weiser (eds.), *You Take the High Node and I'll Take the Low Node. Papers from the 9th Regional Meeting of the Chicago Linguistic Society, April 13-15: 332-345.* Chicago: Chicago Linguistic Society.

Ney, J.W. (1982). The order of adjectives and adverbs in English. *Forum Linguisticum 6, 217-257.*

Rijkhoff, J. (2002). *The Noun Phrase*. (Oxford Studies in Typology and Theoretical Linguistics). Oxford: Oxford University Press.

6.13. Corpus linguistics and theory

Problem

The problem can be appropriately presented by means of a quotation from Halliday (2006: 130)

"At a recent conference devoted to modern developments in corpus studies, I was struck by the way that a number of speakers, at the conference were setting up an opposition between "corpus linguists" and "theoretical linguists" - not a conflict, I mean, but a distinction, as if these were members of two distinct species. I commented on this at the time, saying that I found it strange because corpus linguistics seemed to me to be, potentially at least a highly theoretical pursuit. Work based on corpus studies has already begun to modify our thinking about lexis, about patterns in the vocabulary of languages; and it is now beginning to impact on our ideas about grammar. In my view, this impact is likely to be entirely beneficial. Corpus linguistics brings a powerful new resource into our theoretical investigations of language."

Analyze this quotation. Does corpus linguistics have the above described status?

Procedure

First read several works on the concept *theory* as discussed in the philosophy of science in general, state what it is, what are its components, how does it develop, what does it need for its establishment.

Then state whether corpus linguistics contributed to the establishment of laws – the fundamental, indispensable components of any theory - and not only to placing at our disposal data for testing the statements of a theory.

Ask whether the so-called "theory of grammar" is a real theory and not only a description of grammatical rules. Does grammar contain laws?

Does corpus linguistics contain hypotheses that could be derived mathematically and substantiated linguistically? Do the statements of corpus lin-

guistics hold true for all languages at all times? What is theoretical in corpus linguistics?

Are "many" data necessary and sufficient for the establishment of a theory?

Did Einstein have many data when he established the relativity theory?

The opinions on the Internet differ. There are even journals emphasizing the theoretical aspect of corpus linguistics (e.g. Corpus Linguistics and Linguistic Theory) but in linguistics, "theory" usually means not purely empirical research. Look at the problem from the point of view of the philosophy of science and state which parts or directions of linguistics have a theory. List all "theoretical" issues in corpus linguistics, i.e. hypotheses, laws, explanations, etc.

References

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus Linguistics, Investigating Language Structure and Use*. Cambridge: Cambridge UP.

Facchinetti, R. (2007). *Theoretical Description and Practical Applications of Linguistic Corpora*. Verona: QuiEdit.

Halliday, M.A.K. (2006). *Computational and Quantitative Studies*. (Volume 6 in the Collected Works of M.A.K. Halliday, edited by Jonathan J. Webster). London-New York: Continuum.

McCarthy, D., Sampson G. (2005). *Corpus Linguistics: Readings in a Widening Discipline*. London-New York: Continuum.

Svartvik, J. (ed.) (1992). *Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82)*. Berlin: Mouton de Gruyter.

Teubert, W., Čermáková, A. (2007). *Corpus Linguistics*. London-New York: Continuum.

6.14. Small inventories

Hypothesis

If the inventory of a kind of linguistic entities is small, then the ranked frequencies of the ordered classes follow a regular probability distribution or a regular ranking series.

Define restricted inventories, e.g. phonemes, letters, parts-of-speech, affixes, speech acts, etc. State their frequencies in individual texts, rank them in decreasing order and find a sequence or a function capturing this course.

Procedure

First define the entities of an inventory. Then evaluate a not too short text and state the frequencies of the elements of the given inventory. Order them in decreasing order and ascribe them ranks. Show that the ranked frequencies abide by one of the following sequences (all of them can be considered continuous functions like (1), or discrete sequences like (2), or distributions like (3)).

Zipf-Alekseev:

$$(1) \qquad y_r = ar^{b+c\log r}, \qquad \text{r} = 1,2,3,\ldots$$

Altmann:

$$(2) \qquad y_r = \frac{\binom{b+r}{r-1}}{\binom{a+r}{r-1}} y_1, \quad r = 1,2,3,...$$

where y_1 is the frequency of the first class;

Negative hypergeometric distribution

$$(3) \qquad P_r = \frac{\binom{-M}{r-1}\binom{-K+M}{n-r+1}}{\binom{-K}{n}}, \quad r = 1,2,3,...,n+1, \;\; K \geq M \geq 0, \; n \in \mathbf{N}.$$

You can use a software or proceed as follows: Estimate the parameters from the first ranked classes. Apply at least one of the above sequences and test the difference between observed and computed frequencies (chi-square test or determination coefficient). Apply the best formula to several texts in order to corroborate the result.

Then derive the above formulas either from difference or differential equations. Interpret the basic equation linguistically, i.e. ascribe the parameters some forces which are active in speaking/writing. That means, substantiate the hypothesis. If the above formulas are not adequate, derive a further formula (use a software and find a good fit).

Generalize or specialize your finding showing that one of the formulas holds true for phonemes, the other for word classes, etc.

If you applied the above hypotheses to different inventories in a language, show the differences between them, e.g. depending on the levels, extent, etc.

Compare languages. Are the parameters linked to some other linguistic property, e.g. synthetism? Compare text-sorts. Do they differ? Both kinds of comparison can be performed without having a model, quite empirically.

(See also the problem *4.1. Speech act distribution*)

References

Altmann, G. (1991). Word class diversification of Arabic verbal roots. In: Rothe, U. (ed.), *Diversification processes in language: grammar: 57-59*. Hagen: Rottmann.

Altmann, G. (1993). Phoneme counts. *Glottometrika 14, 55-70.*

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

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

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

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

Best, K.H. (2001). *Quantitative Linguistik. Eine Annäherung*. Göttingen: Peust & Gutschmidt.

Hammerl, R. (1989). Untersuchungen zur Verteilung der Wortarten im Text. *Glottometrika 11, 142-156.*

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

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

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

Lauter, J. (1966). *Untersuchungen zur Sprache von Kants "Kritik der reinen Vernunft"*. Köln: Westdeutscher Verlag.

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

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

Uhlířová, L. (2000). On language modelling in automatic speech recognition. *J. of Quantitative Linguistics 7, 209-216.*

Wimmer, G., Altmann, G. (2001b). Some statistical investigations concerning word classes. *Glottometrics 1, 109-123.*

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

Ziegler, A. (1998b). Word class frequencies in Brazilian-Portuguese texts. *J. 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 Ludek Hřebíček: 295-312.* Trier: WVT.

6.15. Borrowing

Problem

Borrowing is a very regular process abiding by the Piotrowski-law. Collect all published results, both theoretical derivations and practical fittings, show the differences, boundary conditions and deviations from the predictions according to this law. Show as many phenomena abiding by this law as possible.

Procedure

Take the newest literature and trace down the history of the problem. Find new vistas and substantiate the necessity of the hypothesized development. Decide whether the Piotrowski law is a result of self-regulation or self-organization. What can impede its effect?

Describe the history of this problem beginning with the works by Graudina (1964) and Piotrovskaja, Piotrovskij (1974).

References

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung: 59-90.* Göttingen: Herodot.

Altmann, G., Buttlar, H.v., Rott, W., Strauss, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical Linguistics: 104-115.* Bochum: Brockmeyer.

Beöthy, E., Altmann, G. (1982). Das Piotrowski-Gesetz und der Lehnwortschatz. *Zeitschrift für Sprachwissenschaft 1, 171-182.*

Best, K.-H. (2001). Wo kommen die deutschen Fremdwörter her? *Göttinger Beiträge zur Sprachwissenschaft 5, 7-20.*

Best, K.-H. (2003). Anglizismen – quantitativ. *Göttinger Beiträge zur Sprachwissenschaft 8, 7-23.*

Best, K.-H. (2003a). Slawische Entlehnungen im Deutschen. In: Kempgen, S. (ed.), *Rusistika – Slavistika – Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag: 464-473.* München: Sagner.

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

Best, K.-H. (2005). Diversifikation der Fremd- und Lehnwörter im Türkischen. *Archiv orientální 73, 291-298.*

Best, K.-H. (2005). Ein Modell für das etymologische Spektrum des Wortschatzes. *Naukovyj Visnik Černivec'kogo Universytetu: Hermans'ka filolohija 266, 11-21*.

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

Best, K.-H. (2009). Zur Entwicklung der Entlehnungen aus dem Japanischen ins Deutsche. *Glottometrics 19, 80-84.*

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

Graudina, L.N. (1964), Razvitie nulevoj formy roditel'nogo množestvennogo u suščestvitel'nych - jedinic izmerenija. In: *Razvitie grammatiki i leksiki sovremennogo russkogo jazyka: 210-221*. Moskva, Nauka.

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

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

Leopold, E. (2005). Das Piotrowski-Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 627-633.* Berlin-New York: de Gruyter.

Müller-Hasemann, W. (1983). Das Eindringen englischer Wörter ins Deutsche ab 1945. In: Best, K.-H., Kohlhase, J. (ed..), *Exakte Sprachwandelforschung: 143-160.* Göttingen: Herodot.

Piotrovskaja, A.A., Piotrovskij, R.G. (1974). Matematičeskie modeli diachronii i tekstoobrazovanija. In: *Statistika reči i avtomatičeskij analiz teksta; 361-400*. Leningrad: Nauka.

Sağlam, M.Y. (2004). Lehnwörter im Türkischen. Eine lexikologische Wortschatzuntersuchung. *Muttersprache 114, 115-122.*

Stuhlpfarrer, M. (2010). Anglizismen im Russischen. *Glottotheory 3(1),97-109.*

6.16. Frequency and irregular verbs

Problem

It is known that from time to time an irregular verb looses its irregularity and enters into the class of regular ones. This depends, usually, on the frequency of the given verb. Hence, one can set up the hypothesis that entities behaving ir-

regularly are more frequent than the regular ones of the same category. Choose a number of entities and test the hypothesis.

Bybee and Hopper (2001: 1) state two hypotheses belonging to this problem: “irregular morphological formations with high frequency are less likely to regularize“ and “high frequency forms resist analogical change, such as regularization of irregulars, because their frequency makes them easy to access whole and there is no need to re-form them by regular rule.” (2001: 17)

Procedure

Study the regular and irregular verbs in one of Indo-European languages. Evaluate a long text or a corpus and let enumerate all verbs with their frequencies. Write a lexeme as a set of its forms, e.g. *take* = {*take, takes, took, taken*} and sum up the frequencies of all its word-forms. Then set up the rank-frequency distribution of all verbs (lexemes) and test the following hypotheses:

1. The greater the allomorphic cardinality of the lexeme (= number of different word-forms), the lower is its rank.
2. Irregular verbs are more frequent on the average than regular ones.
3. Consider only the irregular verbs and test whether the shorter the lemma, the more frequent is the verb (or vice versa).
4. Study the change from irregularity to regularity historically. If you find such a process, set up the stochastic death-process concerning the transition of irregular verbs into the class of regular ones. A corresponding test is not easy since one cannot find reliable data in all cases.

Define a function for Hypothesis 1 and 3. For Hypothesis 2, perform a statistical test.

References

Bybee, J., Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In: Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic structure: 1-24.* Amsterdam-Philadelphia: Benjamins.

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

Hare, M.L., Ford, M., Marslen-Wilson, W.D. (2001). Ambiguity and frequency effects in irregular verb inflection. In: Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic structure: 181-200.* Amsterdam-Philadelphia: Benjamins.

6.17. Ord's plane

Problem

Compute the curves or areas in the <I,S> coordinate system for the following distributions:

(a) Neyman A
(b) Hyperpoisson
(c) Hyperpascal
(d) Ferreri-Poisson
(e) Waring
(f) Geometric

All these distributions can be found in Wimmer, Altmann (1999).

Procedure

First derive the first raw moment, then the second and third central moments of the given distributions. Set up the indicators

$$I = \mu_2/\mu_1'$$

$$S = \mu_3/\mu_2$$

and draw the graph of <I, S> in Cartesian coordinates. If it is a function, find S = f(I); if it is an area, find the boundaries of the area.

References

Ord, J.K. (1972). *Families of frequency distributions.* London: Griffin.
Wimmer, G. Altmann, G. (1999). *Handbook of discrete univariate probability distributions.* Essen: Stamm.
Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM. (esp. 111-112).
Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics 3.* Lüdensheid: RAM. (passim)

6.18. Block distribution of modal expressions
(Frumkina's Law)

Hypothesis

The block distributions of modal expressions are characteristic of text-sorts.

Test the hypothesis.

Procedure

Modal expressions such as modal verbs, paraphrases, lexicalised modality etc. (cf. the problem *2.18. Modality Marking* in this volume) should be typical of text-sorts because of the different functions of the texts and hence different prominence of speech acts (cf. the *Problems* I: 58; II: 118-127; III: 108, 128, 135, 155, 156) and other discourse elements. This fact should find a conspicuous reflex in the specific shape of their distribution.

A simple way of testing this hypothesis is the fitting of the negative hypergeometric distribution to corresponding data (cf. the problems connected with the Frumkina Law, Problems I: 117; II: 51, 54, 71, 129). This distribution can be written as

$$P(X = x) = \frac{\binom{-M}{x}\binom{-K+M}{n-x}}{\binom{-K}{n}}, \quad x = 0,1,...,n$$

or as

$$P(X = x) = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}, \quad x = 0,1,2,...,n$$

The parameters *M*, *K*, and *n* must be estimated from the data. Partition the texts under consideration into blocks of equal lengths and determine the number of blocks with 0,1,…,*n* occurrences of modal elements, separately for each set of equivalents of “must”, “may”, “can” etc. in the language of the texts. The appropriate block size must be determined empirically. The optimal block size is found when the negative hypergeometric distribution yields the best results.

As each set of modal expressions yields an individual set of parameters, an obvious method for text comparison and text categorisation is forming a vector of the parameter values or even of frequencies, applying one of the common distance measures such as the Euclidian distance or the angle between the vectors, and using one of the usual classification procedures.

Having solved the problem for modal expression in several texts, generalize the problem:

(1) Find those entities whose block-distribution is characteristic of a special text-sort, i.e. perform counts for different entities and apply the above formula.
(2) State which block length is optimal for individual entities, or, specify the optimal block length of entities in specific text-sorts.
(3) Derive the recurrence function of the above formula (= difference equation) and interpret the parameters linguistically. To this end employ Köhler's (2005) "requirements" and insert the supposed ones for different entities.

References

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

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

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

Best, K.-H. (2005). Sprachliche Einheiten in Textblöcken. *Glottometrics 9, 1-12.*

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

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

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

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

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

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

Suhren, S. (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen „De lütte Prinz"*. Staatsexamensarbeit, Göttingen.

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

Zwirner, E., Ezawa, K. (eds.) (1966, 1968, 1969). *Phonometrie, Erster-Dritter Teil*. Basel-New York: Karger.

Zwirner, E., Zwirner, K. (1935). Lauthäufigkeit und Zufallsgesetz. *Forschungen und Fortschritte 11, Nr. 4: 43-45*. (Also in: Zwirner & Ezawa (eds.), Dritter Teil: 55-59.).

Zwirner, E., Zwirner, K. (1938). Lauthäufigkeit und Sprachvergleichung. *Monatsschrift für höhere Schulen 37: 246-253*. (Also in: Zwirner & Ezawa (eds.), Third Part, 68-74.)

6.19. Sonority sequences

Problem

The individual sounds of a word have a certain degree of sonority. Describe the course of sonority in words taken from a dictionary and find typical word structures, course tendencies, test hypotheses and make general statements.

Procedure

Refer to Restle & Vennemann's (2001) sonority scale (cf. also Vennemann's 1982, 1988; Murray, Vennemann 1983) given as

1. Voiceless plosives
2. Voiced plosives
3. Voiceless fricatives
4. Voiced fricatives
5. Nasals
6. Lateral liquids (/l/)
7. Central liquids (/r/)
8. High vowels
9. Mid vowels
10. Low vowels

Ascribe each sound in a selected word its sonority degree. Do this with all words in a dictionary or with a sample from it. Then count the frequencies of individual types. Set up a rank-frequency sequence and find a function capturing it. Distinguish words of different length.

Study the properties of the sonority sequences in this rank-frequency sequence. Set up some general hypotheses. Do short words display different sonority structures than longer ones? Study the beginning and the end of the words.

Perform the same analysis for individual words in a text taking into account all affixes. When you evaluate poetic texts show whether they differ in

mean sonority or in its course from scientific texts. Set up hypotheses, develop necessary tests and test your hypotheses statistically.

Kelih (2012) collected the objections against sonority. Do not try to use sonority as a criterion of syllable separation - the discussion is more than 100 years old. Instead, find a possibility of its empirical measurement, even if it is composed of different properties. That is, find a finer scaling method than that given above. Discover regularities and derive them theoretically – if you succeed, you are entering a research phase at the beginning of theory construction.

Perform a sonority analysis in a family of cognate languages. Compare cognate words and state the differences, divergences, directions of development. Consider the longest form of the word as basis and scale the missing sounds in related languages with zero.

References

Blevins, J. (1995). The syllable in phonological theory. In: Goldsmith, J. (ed.), *The Handbook of Phonological Theory: 206-244*. Oxford: Blackwell.

Cetnarowska, B., Żygis, M. (2004). Syllabification across a prefix stem boundary in Polish. The role of semantic compositionality. *Zeitschrift für Slawistik 49, 42-60.*

Clements, G.N. (1990). The role of the sonority cycle in core syllabification. In: Kingston, J., Beckman, M.E. (eds.), *Papers in laboratory phonology I. Between the grammar and physics of speech: 283-333.* Cambridge: Cambridge University Press.

Donegan, P.J. (1978). *On the natural phonology of vowels*. Ohio: Univ.-Diss.

Foley, J. (1970). Phonological distinctive features. *Folia Linguistica 4, 87-92.*

Foley, J. (1977). *Foundations of theoretical phonology*. Cambridge: Cambridge University Press.

Kelih, E. (2012). *Die Silbe in slawischen Sprachen.* München-Berlin-Washington: Sagner.

Meinschaefer, J. (2003). *Sonorität. Sprachstruktur und Sprachverstehen.* Tübingen: Narr.

Murray, R.W., Vennemann, T. (1983). Sound change and syllabic structure in Germanic philology. *Language 59, 514-528.*

Ohala, J.J. (1992). Alternatives to the sonrity hierarchy for explaining segmental sequential constraints. In: Zilokowski, M., Bose, M., Deaton, K. (eds.), *Papers from the rgional meting of the Chicago Linguistic Society 1990, Vol. 2. The parasession on the syllable in phonetics and phonology: 319-353.* Chicago: Chicago Linguistic Society.

Restle, D., Vennemann, T. (2001). Silbenstruktur. In: Haspelmath, M. et al. (eds.), *Language typology and language universals: 1310-1336.* Berlin-New York: de Gruyter.

Vennemann, T. (1982). Zur Silbenstruktur der deutschen Standardsprache. In: Vennemann, T. (ed.), *Silben, Segmente, Akzente. Referate zur Wort-, Satz-*

und Versphonologie anlässlich der vierten Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Köln, 2.-4. März 1982: 261-305. Tübingen: Niemeyer.

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

Wiese, R. (1988*). Silbische und lexikalische Phonologie*. Tübingen: Niemeyer.

6.20. Verb classes

Problem

Classify verbs of a language according to the biological development of Man (*nutrition,…, perception…, motion,…, feeling,…, intellect* and all intermediate states). Ascribe the groups some ordinal (evolutionary) numbers. Then analyze texts and show the distribution of these classes in texts. Can one set up a hypothesis concerning text-sorts?

The results depend on the classes you estalished. You can perform a scaling with respect to "intensity" or "bodily effort" or evolutionary chronology, etc.

A very thorough analysis of German verbs sorted in dozens of classes can be found in Trier (1933), Dornseif (1934), Mater (1966), Ballmer, Brennenstuhl (1986), etc. Restrict the number of classes to some fewer, more general ones, otherwise no text will represent a sufficiently large sample. In any case begin with the formulation of a hypothesis and establish the classes adequately.

Procedure

Select a classification (of any sort) of verbs form the rich existing literature. Introduce a scale according to a property you have chosen. Biological development is one of the possibilities. E.g. *eat, breath* are "lower (earlier) acts" than e.g. *intend, ponder, say*. Or *intend* is a "higher effort" than *jump*. Do not take verbal phrases into account, reduce them to simple verbs. Each class can be separately scaled.

Then analyze several texts and show the differences in proportions of these classes. Set up confidence intervals for text sorts. Show differences between texts. Analyze the work of an individual writer in the course of years. If you have scaled the verbs according to some property, search for a regular change.

References

Ballmer, T.T. (1982). *Biological foundations of linguistic communication*. Amsterdam: Benjamins.

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

Dornseiff, F. (1934). *Der deutsche Wortschatz nach Sachgruppen*. Berlin: de Gruyter (1965).

Kipper, K., Korhonen, A., Ryant, N., Palmer, M. (2008). A Large-Scale Classification of English Verbs. *Language Resources and Evaluation Journal 42, 21-40.*

Korhonen, A., Briscoe, T. (2004). Extended Lexical-Semantic Classification of English Verbs. *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*. Boston, MA.

Levin, B. (1993). *English Verb Classes and Alternations*, Chicago: University of Chicago Press.

Mater, E. (1966). *Deutsche Verben*. Leipzig: VEB Bibliographisches Institut.

Trier, J. (1933). *Der deutsche Wortschatz im Sinnbezirk des Verstandes.* Heidelberg: Winter.

Author Index

Abelin, A. 61
Acha, J. 14
Agricola, C. 65, 89
Agricola, E. 65, 89
Aikhenvald, A.Y. 46
Alston, W.P. 100
Altmann, E.G. 82
Altmann, G. 6, 10-12, 18, 23, 25, 28, 32, 34-38, 40, 41, 44, 46-50, 54, 55, 58, 59, 63-67, 70, 71, 75, 78-80, 83, 85, 86, 91-96, 98, 99, 101, 103, 105, 107, 110, 111, 113, 114, 121, 127-130, 132, 134
Altmann, V. 67
Alvarez-Lacalle, E. 82
Andreev, S. 71
Andreevskaja, A.A. 38
Anis, A.A. 91
Arduini, S. 32
Armstrong, D. 116
Asher, R.E. 58
Au, T.K. 13
Austin, J.L. 94-96, 99, 101
Ayer, A.J. 117
Baayen, H. 75
Bailey, R.W. 76
Ballmer, Th.T. 66, 68, 95, 137, 138
Balota, D. 119
Bane, M. 119
Bansal, A.R. 91
Barabási, A.-L. 82
Batagelj, V, 112
Beaugrand, R. de 84
Beckman, M.E. 136
Beebee, H. 117
Bekkerman, R. 18
Bektaev, K.B. 80, 134
Beliankou, A. 22, 24, 25, 35, 96
Bennett, C.H. 73
Beöthy, E. 47, 129
Bergem, W. 51
Best, K.-H. 26-28, 31, 32, 40, 41, 58, 59, 71, 78, 80, 114, 128-130, 134
Biber, D. 126
Bin Ma 73
Black, M. 52
Blakemore, D. 58
Blass, T. 109
Blevins, J. 136
Blum, L. 51
Blumenberg, H. 51
Boehnke, K. 28
Bohunická, A. 51
Bolinger, D. 60, 61
Booij, G. 123
Boroda, M.G. 75
Bortz, J. 28
Boschtan, A. 26-28, 31, 32, 71
Bose, M. 136
Brainerd, B. 41, 80, 129, 134
Breiter, A.M. 102
Brennenstuhl, W. 95, 137, 138
Briscoe, T. 138
Brown, C.H. 62
Brown, D. 131
Brown, R. 13
Brownell, H. 14
Brunet, E. 73, 74
Bucková, M. 22, 35, 56, 68, 87, 96, 97, 99
Buiati, M. 88
Bulitta, E. 44, 65, 89
Bulitta, H. 44, 65, 89
Bunge, M. 117
Burdinski, V. 80, 134
Burghardt, H. 35, 68, 87, 96, 98, 99
Burwick, F. 13
Busemann, A. 67, 68
Buttlar, H.v. 129
Bybee, J. 3, 8, 58, 109, 115, 131
Cai, Z. 84
Calbert, J.P. 58

Carlson, M. 76
Carreiras, M. 123
Carrel, P. 84
Carroll, J.B. 75, 76, 78
Cartwright, N. 117
Cazeaux, C. 52
Čech, R. 40, 41, 78, 92, 132, 134
Čermáková, A. 126
Cetnarowska, B. 136
Chamoli, A. 91
Chang, C.B. 119
Chung, C.B. 58
Christie, R. 109
Clements, G.N. 136
Cohen, P.R. 88
Cole, P. 20
Collins, M. 18
Condoravdi, C. 58
Conrad, S. 126
Contini-Morava, E. 46
Corbett, G.G. 131
Corral, A. 82
Corrigan, R. 13
Corum, C. 125
Cover, J.A. 117
Cover, P.M. 73
Craig, C.G. 46
Cruse, D.A. 44, 65, 89
Crystal, D. 111
Curd, M. 117
Czernin, F.J. 51
d'Arcais, G.F. 119
Damerau, F.J. 18
Davidson, D. 52
de Grada, E. 14
Deaton, K. 136
Decker, R. 35. 68, 87, 96, 98, 99
Demuth, K. 46
Derrida, J. 53
Dezsö, Z. 83
Di Sciulo, A.M. 119
Diaz-Guilera, A. 82
Dimri, V.P. 91
Dirvens, R. 53
Dixon, R.M.W. 46
Doerge, F.Ch. 100, 101
Doležel, L. 76
Dömötör, Z. 48
Donegan, P.J. 136
Dornseiff, F. 46, 137, 138
Dorow, B. 87
Drebet, V.V. 36-38, 44, 46
Dressler, W. 84
Dretske, F. 117
Dugast, D. 74
Dunabeitia, J.A. 123
Eckmann, J.-P. 82
Eder, T. 51
Ehrhardt, C. 101
Ehrlich, H.J. 109
Eickmeyer, H.-J. 58
Ejiri, K. 74
Elam, K. 76
El-Yaniv, R. 18
Erler, B. 101
Ernst, P. 111
Exner, J.E. 109
Ezawa, K. 134, 135
Fachinetti, R. 126
Falkum, I.L. 36
Fan, F. 49, 65, 66, 74, 80, 93, 96, 99, 103, 132, 134
Farsi, A.A. 124
Fass, D. 53
Feder, J. 91
Feldman, L.B. 119
Felici, A. 58
Fenk, A. 19
Fenk-Oczlon, G. 7, 8, 19, 109
Fenz, E. 60, 61
Fergusson, F. 76
Ferrer-i-Cancho, R. 50, 82, 113
Feynman, R.P. 117
Firth, J.R. 60, 61
Fish, D. 13
Flammini, A. 83
Foley, J. 136
Ford, M. 131

Försterling, F. 15
Foster, J. 117
Fox, B. 20
Fraassen, B.v. 117
Frawby, W. 58
French, P.L.61
Freytag, G. 76
Frisson, S. 123
Frost, R. 119
Frumkina, R.M. 78, 80, 134
Garthwaite, G.H. 83
Gaynesford, R.M. de 101
Genzor, J. 22, 35, 38, 52, 68, 87, 96, 97, 99, 103
Gibbons, J.D. 28, 63, 78
Giere, R.N. 117
Gigerenzer, G. 117
Girdenienė, S. 14
Girke, W. 14
Givón, T. 1-9, 11, 12, 16, 17, 20, 21, 27-29, 32-34, 115
Goikoetxea, E. 14
Gottmann, J.M. 41
Graesser, A.C. 84
Graudina, L.N. 14, 130
Greenbaum, S. 115
Greenberg, J.H. 111, 118, 119
Greene, S.B. 14
Gregory, E.D. 88
Greiman, D. 101
Grendel, M. 119
Grigolili, P. 88
Gross, B. 84
Grzybek, P. 23, 35, 57, 68, 78, 87, 88, 93, 96, 98, 99, 107
Guiraud, H. 74
Guiter, H. 102
Günderson, K. 97.
Gunkel, L. 20, 21, 27-29, 33
Hahn, E. 14
Haiman, J. 8, 115
Halliday, M.A.K. 9, 16, 21, 27-29, 32, 84, 123-126
Hammerl, R. 85, 105, 138
Hare, M.L. 131
Harizanov, V. 58
Harsh, P.W. 76
Hasan, R. 9, 16, 84
Haspelmath, M. 133, 136
Haverkamp, A. 51
Haynie, H.J. 119
Hecker, U.v. 15
Hempel, C.G. 117
Henschelmann, K. 14
Herdan, G. 74
Heringer, H.J. 101
Hermodsson, L. 14
Hindelang, G. 101
Hintikka, J. 51
Hinton, I. 60, 61
Hippisley, A. 131
Hoey, M. 16
Hoffmann, L. 21
Holdcroft, D. 97
Holman, E.W. 62
Hopper, P. 2, 3, 8, 109, 115, 131
Householder, F. 60, 61
Hřebíček, L. 70, 75, 91-93, 110
Hudson, R. 128
Hundsnurscher, F. 65
Hurst, H.E. 91
Hyönä, J. 85
Jäger, S. 80, 134
Jarema, G. 123
Jastrembski, J.E. 37
Jayaram, B.D. 93
Jeffries, L. 65
Job, M. 65
Johnson, M. 53.
Johnson, N.L. 47
Jones, M.N. 18
Jones, S. 65
Joshi, A. 84
Judt, B. 128
Kaahinen, J. 85
Kang, Ch. 14
Kantemir, S. 62, 64, 65
Kasof, J. 14

Katz, S.M. 82
Kaufmann, S. 58, 85
Kelih, E. 22, 23, 25, 35, 40, 41, 57, 59, 68, 87, 88, 96, 98, 99, 103, 104, 114, 119-121, 136
Kelley, H.H. 14
Kempgen, S. 120, 121, 129
Kijko, S.V. 36-38, 44, 46
Kim, K.-O. 61
Kingston, J. 136
Kipper, K. 138
Kitscher, P. 117
Kittay, E.F. 65, 89
Kneale, K. 117
Koch, W. 68, 75
Köhler, R. 6, 8, 18, 21, 23-28, 32, 34-37, 44, 46, 47, 49, 50, 54-58, 64, 65, 68, 70, 74, 75, 78, 80, 85, 87, 88, 91-99, 101-103, 105, 107, 113, 114, 121, 128-130, 134
Kohlhase, J. 129, 130
Kokol, P. 88
Korhonen, A. 138
Kornai, A. 75
Körner, H. 130
Kotz, S. 47
Kövecses, Z. 53
Kratzer, A. 58
Krug, M.G. 2, 3, 115
Krupa, V. 51, 93, 111, 118, 119
Kunz, M. 82
Küper, Ch. 14
Kurz, G. 51
LaFrance, M. 14
Lakoff, G. 53
Lange, M. 117
Laufer, B. 75
Laufer, J. 47, 58, 114
Leacock, C. 37
Lee, J.Y. 14
Leech, G. 115
Lehrer, A.J. 65, 89
Lehrer, K. 89
Lempel, A. 119
Leopold, E. 130
Levi, J.N. 125
Levickij, V.V. 36-38, 43, 44, 46, 60, 61, 67, 68, 123
Levin, B. 138
Levin, S.R. 51
Levinson, S.C. 101
Libben, G. 123
Liedtke, F. 101
Lienert, G.A. 28
Lin, Sh.-K. 73
Liu, H. 82
Lloyd, E.H. 91
Londorn, H. 109
Lowerse, M.M. 84
Lučak, M. 67, 68
Lukjanenkov, K.F. 134
Lutzeier, P.-R. 65
Lyons, J. 44
Mackie, J.L. 117
Mačutek, J. 22, 23, 35-37, 40, 41, 57, 68, 70, 71, 78, 86-88, 91, 93, 96, 98, 99
Mahlmann, R. 51
Makioka, Sh. 113
Malkiel, Y. 60, 61
Mandelbrot, B. 90, 91
Manetti, L. 14
Marriott, P. 131
Marslen-Wilson, W.D. 131
Martináková-Rendeková, Z. 75
Marx, F. 51
Mater, E. 137, 138
Matskulyak, Y. 121, 123
McArthur, L.Z. 14
McCarthy, D. 126
McKoon, G. 14
McNamara, D.S. 84
Meijers, A.W.M. 101
Meinschaefer, J. 136
Melinger, A. 119
Menzer, F. 83
Mettinger, A. 65, 86

Mewhort, D.J.K. 18
Meyer, P. 113
Michajlov, M.N. 14
Mikros, G. 22, 35, 96
Ming Li 73
Mislovičová, S. 22, 35, 51, 57, 96, 98, 99
Mizutani, S. 128
Mološnaja, T.N. 134
Montemurro, M.A. 82
Morgan, J. 20
Morris, J. 85
Moscoso del Prado Martin, F. 119
Moses, E. 82
Motter, A.E. 82
Mrvar, A. 113
Müller, B. 19
Müller, W. 65, 89
Müller-Hasemann, W. 130
Murphy, M.L. 44, 65, 89
Murray, R.W. 135, 136
Murthy, K.P.N. 88
Nadarejšvili, I.Š. 75
Napoli, D.J. 111
Nation, I. 75
Naumann, S. 21-25, 34, 35, 57, 68, 87, 96-99
Nemcová, E. 47, 50, 58, 103, 114
Newman, M.E.J. 112, 113
Newman, S.S. 61
Ney, J.W. 125
Nichols, J. 60, 61
Niswander-Klement, E. 123
Nordquist, R. 51, 53
Novikova, N.L. 14
Nuzban, O. 62, 84, 65
Oates, T. 88
Obradović, I. 25, 26, 96
Ohala, J.J. 60, 61, 136
Oliveira, J.G. 83
Ondrejovič, S. 38, 75, 93, 103
Ord, J.K. 11, 41, 132
Orgoňová, O. 51
Orlov, J. 75
Ortony, A. 51
Pagliuca, W. 3, 58
Palatella, L. 88
Palmer, F.R. 58
Palmer, M. 58, 138
Paluoš, M. 88
Pascual, G. 14
Paškovskij, V.E. 80
Patzke, U. 14
Pavelka, J. 51
Perea, M. 123
Perkins, R. 3, 58
Peterfalvi, J.M. 60, 61
Peters, E.E. 91
Pickering, M.J. 15
Pierrehumbert, J.B. 83
Piotrowski, R.G. 6, 28, 32, 34, 44, 46, 47, 49, 50, 54, 55, 58, 64, 70, 78, 80, 91, 93, 94, 103, 105, 107, 114, 121, 130, 134
Piotrovskaja, A.A. 80, 130
Pnini, T. 119
Polikarpov, A.A. 37
Politi, M. 83
Pollatsek, A. 123
Pompe, B. 88
Popescu, I.-I. 18, 23, 35, 40, 41, 47, 49, 59, 64, 70, 71, 75, 78, 79, 86, 91, 93-95, 110, 111, 114
Popper, K. 117
Pörings, R. 53
Portner, P. 59
Poulisse, N. 119
Preisach, C. 35, 68, 87, 96, 98, 99
Punter, D. 53
Pustet, R. 93
Quirk, R. 115
Rabiner, L.R. 18
Rádl, Z. 82
Ratcliff, R. 14
Rehm, R. 76
Rensinghoff, S. 103
Reppen, R. 126
Restle, D. 135, 136

Reza, F.M. 73
Richards, I.A. 53
Ricœr, P. 51, 53
Rieger, B. 74, 92
Rieser, H. 58
Rijkhoff, J. 20, 27,-29, 32, 125
Riška, A. 48
Robinson, J.P. 109
Roeck A. de 83
Roelcke, Th. 32, 33
Roelofs, A. 119
Rothe, U. 7, 25, 47, 50, 59, 95, 98, 102, 114, 128
Rott, W. 129
Roy, A.K. 41
Rudolph, E. 14, 15
Rudolph, U. 15
Ryant, N. 138
Sacks, S. 51
Saeed, J.I. 59
Sağlam, M.Y. 130
Salmon, W.C. 117
Salonen, J. 85
Sampson, G. 126
Sanada, H. 23, 35, 50, 68, 88, 114
Sanford, A.J. 15
Sapir, E. 61
Sarkar, A. 83
Scalas, F. 83
Schank, R.C. 117
Schierholz, S.J. 37
Schmidt, P. 83
Schmidthauser, B. 15
Schmidt-Thieme, L. 35, 68, 87, 96, 98, 99
Schmill, M.D. 88
Schreuder, R. 119
Schrödinger, E. 117
Schroeder, M. 41, 91
Schurz, G. 117
Schwarz-Friesel, M. 51
Schweers, A. 128
Searle, J. 94-96, 98, 88, 101
Senft, H. 75
Serrano, M.A- 83
Shaver, P.R. 109
Shibatani, M. 7
Sichel, H. 75
Siegel, J. 119
Siegwart, G. 101
Skinner, B.F. 12
Skirl, H. 51
Smith, A.E. 74
Smith-Stark, T.C. 125
Solé, R.V. 113
Spoelders, M. 85
Srebrjanskaja, I.I. 80
Staffeldt, S. 101
Stechow, A.v. 58
Steinsträter, Ch. 98
Stewart, A.J. 15
Steyvers, M. 113
Stockall, L. 119
Stopelli, P. 44
Strauss, U. 65, 66, 80, 93, 96, 99, 103, 129, 132, 134
Strecker, B. 21
Stroyny, K. 15
Stuhlpfarrer, M. 130
Suhren, S. 80, 134
Suppes, P. 118
Svartvik, J. 115, 126
Sweetser, E.E. 59
Tamaoka, K. 113
Taxidou, O. 76
Tennenbaum, J.B. 113
Teubert, W. 126
Thomas, J.A. 73
Timberlake, A. 58
Tishby, N. 18
Togia, P. 65
Tooley, M. 118
Traugott, E. 2,3
Trier, J. 137, 138
Tsohatzidis, S. 98
Tuldava, J, 38, 110, 111
Tuzzi, A. 64, 111, 114
Tweedie, F. 75

Uhlířová, L. 80, 128, 129, 134
Ulkan, M. 101
Underhill, P. 53
Vacchiano, R.B. 109.
Vanderveken, D. 101
Van der Velde, M. 15
Vater, H. 58
Vázquez, A. 83
Venäläinen, P. 85
Vennemann, T. 135-137
Vidya, M.N. 93
Vollmer, G. 118
Voort, M.v.d. 119
Waddell, C. 71
Wagner, K.R. 98
Weinstein, S. 84
Weiser, A. 125
Weiss, D. 15
Wenhrynowytsch, A.A. 43, 44, 46
Wescott, R.W. 60, 62
Weydt, H. 14
Wichmann, S. 62
Wiese, R. 137
Wimmer, G. 31, 32, 36-38, 44, 50, 54, 55, 63, 64, 70, 75, 78, 80, 83, 93, 107, 128, 129, 132, 134
Wimmerová, S. 75, 80, 91, 93
Winter, Y. 18
Wrightsman, L.S. 109
Wunderlich, D. 59, 101
Yde, Ph. 85
Zanette, D.H. 82
Zaveršnik, M. 113
Zechner, K. 98
Zhu, J. 128
Ziegler, A. 10, 11, 78, 93, 129
Zifonun, G. 20, 21, 27-29, 32
Zilokowski, M. 136
Zipf, G.K. 43, 103, 107, 111, 116
Ziv, J. 119
Zörnig, P. 9, 11, 12, 23, 35, 78, 83, 86
Zwirner, E. 134, 135
Zwirner, K. 135
Żygis, M. 136

Subject register

Activity 60, 88
adjective 19, 26, 28, 36-39, 46, 48, 62-65, 81, 88, 121, 123, 124
adnominal modifier 26-35, 70
adverb12, 19, 48, 81, 84, 124, 125
alliteration 69, 81
analytism 111
anaphoric distance 9,10
angle 42
antonymy 62-65, 88, 89
arc 41, 85, 86
assonance 69, 81
assortativity 112
asymmetry 22, 32
autocorrelation 13, 22
auxiliary 19, 88, 89, 118
Behaghel-Hawkins Law 8
Beta function 124
bigram 17, 18, 41, 86
binomial d. 48, 63
birth-and-death process 37, 49
block distribution 132-134
B-motif 66, 67
Boroda's F-motif 95
borrowing 129
Carroll's vector 75
cataphoric persistence 11, 12
causality 12, 13
chaotic series80, 91
clause union 1, 2
cluster 60, 109
coding expression 115
coherence 8, 9, 15, 16
cohesion 8, 9, 15, 16, 32, 83, 84
co-lexicalization 3, 4
complementation scale 2-4
complexity, allomorphic 105, 106
- adnominal modifiers 31,32
- global 25
- local 25
- morphological 118
- syntactic 25, 26

compositionality 104-106
compound 105, 106, 112, 113, 121, 122
conjunction 15, 16
content word 19, 82
control cycle 55, 69, 78, 108, 107
corpus linguistics 126-126
descriptivity 60
distance 13, 80-83, 133
diversification 46, 47, 57, 113, 114
D-motif 23
dogmatism 60, 108
dominance 76
efficiency 48
emotionality 60
entropy 18, 22, 41, 59, 71, 72, 86, 87, 94, 106
entropy, discriminative 48
event integration 1, 2
evolution 70-72, 76
excess 18, 32
Ferreri-Poisson d. 132
finiteness 16, 17
Fourier analysis 13, 76
fractal 41, 42
frequency 109, 110, 130, 131
Frumkina's law 78, 79, 132, 133
function word 119, 38, 40, 110
geometric d. 132
Givón's hypothesis 3-5, 7
gradual transition 3
grammaticality 2
grapheme 103
Greenberg-Krupa index 111, 118
hapax legomena 110, 111
Hausdorff-Besicovitch dimension 90
Hidden Markov chain 22
homogeneity 10, 28, 30, 38, 89, 122

hreb 92, 93
Hurst exponent 13, 22, 41,77,89,90, 97
Hyperpascal d. 132
Hyperpoisson d. 132
interjection 19, 61, 81
inventory 126, 127
Kendall's W 10, 58, 63, 122
language
- Arawak 16
- Atabashkan 16
- Austronesian 60
- Carib 16
- Chinese 102
- English 2, 17, 25, 48, 59, 63, 102
- German 19, 24,-26, 28, 36, 37, 43, 66, 89, 102, 121-123
- Indo-European 63, 131
- Indonesian 60
- Iroquois 16
- Japanese 48, 115
- Javanese 115
- Maori 102
- Nimboran 48
- Papuan 16
- Polish 105, 106
- Quechuan 16
- Romanian 102
- Samoan 115
- Slovak 48
- Sundanese 60, 61
- Tibeto-Burmanian 16
- Turkic 16
- Ugro-Finnish 48
- Uto-Aztecan 16

law 115, 116, 125
length 54, 56
L-motif 22, 35, 57, 68, 87, 95
location system 48
Lyapunov coefficient 22, 77
Markov chain 22, 41
metaphoricality 50-56
Minkowski sausage 41
mixed NB distribution 24
modal verb 133
modality 57-60
negative binomial d. 24, 46, 47, 97
Neyman A d. 132
n-gram 18, 86, 87
negative hypergeometric d. 79, 127, 133
noun 44, 45
noun phrase 19, 20
numeral 19
Ord's criterion 18, 22, 32, 33, 41, 71, 86, 87, 94, 106, 118, 123, 132
ornamentality 29, 75, 77
parts-of-speech 17, 18, 64, 70
peakedness 22
persistent 90
phonestheme 60, 61
Piotrowski-law 129
Poisson process 47
polysemy 36-42, 49, 102
polytextuality 21, 95
position 110
postposition 19, 48, 49, 81
prefix 46, 47
preposition 48, 49
pronoun 19, 81
property 2, 3, 31, 33, 34, 45, 65, 69-71, 76-78, 89, 95, 102, 108, 124, 137
radian 40, 42
range 34, 35, 90, 91, 99, 116
remote referent 7
repeat rate 18, 22, 59, 86, 94, 103, 104, 106
repetition 21, 22, 80
rhyme 70, 71
R-motif 13, 21, 22, 98, 99
roughness 18
scaling 33, 34, 100, 108
semantic bond 1
semantic integration 1
sequence 86

similarity 3, 60
skewness 20, 82
sonority 135, 136
sound symbolism 60, 61
speech act 94, 95, 97-99, 100
speech act distribution 94
speech act motif 95, 96, 98, 99
stage play 75-77
style 69-71
submorpheme 60
subordinating morpheme 4
Swadesh list 37, 38
syllable 119, 120
symmetry, directive 48
symmetry, planar 48
synergetics 102-107
synonymy 43, 49
synthetism 106, 110, 111
thematic concentration 75
thematic continuity 8,9
time series 76
transitivity, directive 48
TTR 73, 74
valency 33, 75, 81, 91
variance 28, 32
verb 137, 138
verb, irregular 130, 131
verb-adjective ratio 75
vocabulary richness 75
voice 5
Waring d. 132
Wavelets 76
word length 102, 104-106
Zipf function 120
Zipf-Alekseev function 56, 94, 127