Proceedings of the fourth conference of the International Quantitative Linguistics Association

Prague, August 24-26, 2000

Edited by R. H. Baayen

PROGRAM QUALICO 2000, AUGUST 24-26, PRAGUE

AUGUST 24

9.30–10.00	OPENING SESSION
Session 1.1	
10.15–10.35	**Patrick Juola** The rate of language change
10.35–10.55	**Lisa Lena Opas-Hnninen, Pekka Hirvonen, and Fiona Tweedie** Jet set bar or juttutupa?
10.55–11.15	**Wayne Cowart** The Statistical Profile of Coordination
11.15–11.45	BREAK
session 1.2	
11.45–12.05	**Setsuko Wakabayashi, Jun-ya Morishita, and Yasunori Motomura** Contextual influences of top-down inferencing in a language process by EFL listeners
12.05–12.25	**Yuen Wah Grace Tse** The Grammatical Factors Influencing the Choice Between the Use and Omission of the Definite Article Preceding Organization Names: A Statistical Analysis
12.25–12.45	**Jaan Mikk** Prior knowledge of text content and values of text characteristics
12.45–14.15	LUNCH BREAK
session 1.3	
14.15–14.35	**V.A. Dolinsky** Logos and number. The study of associative fields in the study of consciousness
14.35–14.55	**Jan Králík** Remarks on Quantitative Data Understanding
14.55–15.15	**Victor Kromer** Parameter-free model of rank polisemantic distribution
15.15–1545	BREAK

session 1.4

15.45–16.05 **Edda Leopold**
A model-theoretic inspection of the concept of language as system of competing and cooperating forces

16.05–16.25 **A.A.Polikarpov**
Menzerath's law for morphemic structure of words: a hypothesis for the evolutionary mechanism of its arising and its testing

16.25–16.45 **Peter Kunsmann and Johannes Gordesch**
A note on systems theoretical model of usage

16.45–17.05 BREAK

session 1.5

17.05–17.25 **Andy Way**
Applying insights from Quantitative Linguistics to Problems of Machine Translation.

17.25–17.45 **Oliver Cromm**
Big Is Beautiful? What Resources Most Influence the Performance of Translation Extraction from Non-parallel Corpora

17.45–18.05 **Kyo Kageura and Sandra Yamilet Santana**
A comparative observation of English and Spanish Technical Terminology

AUGUST 25

session 2.1

8.40–9.00 **A.A. Polikarpov and D. Khmelev**
Basic assumptions of Sign's life cycle for mathematical modelling of language evolution

9.00-9.20 **Dariusch Bagheri**
Semantic Relations in the Lexicon and their Synergetic Modelling — Towards an Integration of Lexical and cognitive Relations.

9.20–9.40 **Reinhard Köhler and Gabriel Altmann**
Probability Distributions of Syntactic Units and Properties

9.40–10.00 **Peter Meyer**
Qualifying quantities. Reflections on the notion of explanation in quantitative linguistics

10.00–1.15 BREAK

session 2.2

10.15–10.35 **Mirjam Ernestus and Evert Wattel**
Comparing lines which separate clusters in two-dimensional plots

10.35–10.55 **Adam Pawlowski and Maciej Eder**
Quantity or stress? Sequential analysis of Latin prosody

10.55–11.15 **Svitlana Budzhak-Jones**
Quantitative methods, language contact, code-switching, borrowing

11.15–11.35 **Mark Kaunisto**
Relations and proportions in the formation of blend words

11.35–11.45 BREAK

session 2.3

11.45–12.05 **Johan Carlberger and Viggo Kann**
Some applications of a statistical tagger for Swedish

12.05–12.25 **Akira Ushioda**
Word clustering and Part-of-Speech Tagging

12.25–12.45 **Ludmila Uhlířová**
On the so-called language modelling in automatic speech recognition

12.45–14.15 LUNCH BREAK

session 2.4

14.15–14.35 **V.A. Dolinsky and D. Rainova**
Experimental study of semantics of a word

14.35–14.55 **Omar Larouk**
Using presupposition logic in the recognition of the implicit information of the user in information retrieval system

14.55–15.15 **Rychkova Liudmila**
Quantitative Text Investigation Based on Specially Oriented Linguistic Full-text Data Bases

15.15–15.30 BREAK

session 2.5

15.30–15.50 **Yoshio Narisawa**
Co-occurrence of Antonyms

15.50–16.10 **Marc Hug**
Partial Disambiguation of Very Ambiguous Grammatical Words

16.10–16.30 **Jaroslava Hlavacova**
Rarity of words in language corpora

16.30–17.00 BREAK

session 2.6

17.00–17.20 **E.I. Sicilia-Garcia, Ji Ming, and F.J. Smith**
A dynamic model for each significant word

17.20–17.40 **Sibasis Mukherjee**
On sign test

17.40–18.00 **D. Khmelev**
Disputed authorship resolution using relative empirical entropy for Markov chain of letters in a text

AUGUST 26

SESSION 3.1

8.40–9.00 **Patrick Juola**
A linear model of complexity (and its flaws)

9.00–9.20 **Karl-Heinz Best**
Verteilungen sprachlicher Einheiten in Texten und im Sprachsystem

9.20–9.40 **Edda Leopold**
Length-distribution of words with coinciding frequency

9.40–10.00 BREAK

SESSION 3.2

10.00–10.20 **Zahra Mustafa**
Non-courseware factors involved in using multimedia in foreign language instruction

10.20–10.40 **Victor Zakharov**
A range of linguistic tools for the information retrieval system on conservation and preservation

10.40–11.00 **L. Alfonso Urena, Manuel Buenaga, and J. Maria Gomez**
Using and evaluating WSD in information retrieval

11.00–11.20 **Marc Weeber, Rein Vos, and Harald Baayen**
Word association statistics for the lowest-frequency words

11.20–11.40 BREAK

SESSION 3.3

11.40–12.00 **Stefan Th. Gries**
Particle Movement A Multifactorial Analysis of Syntactic Variation

12.00–12.20 **Sheila M. Embleton and Eric S. Wheeler**
Computerized Dialect Atlas of Finnish: Dealing with Ambiguity

12.20–12.40 **Michael Oakes**
Computer Estimation of Vocabulary in a Protolanguage from Word Lists in Four Daughter Languages

12.40–13.00 **Peter Grzybek**
Remarks on the Sentence Length of Proverbs

13.00–13.15 CLOSING REMARKS

13.15–14.30 LUNCH

14.30–18.00 TOUR OF PRAGUE CASTLE

19.00–22.30 FAREWELL DINNER

Patrick Juola
The Rate of Language Change

Language changes over time. Lexemes float in and out of the language, syntactic patterns change, and the latest fashionable accent drives out last week's. Part of the power and role of corpus and quantitative linguistics is to trace these changes, watching rare forms swell and supplant previously common forms. The importance of these studies [e.g. Biber et al., 1998] should not be downplayed. One flaw, however, is that, as in so many other areas of quantitative linguistics, individual studies may lack scope, and results from different studies may be difficult to compare. This problem is exacerbated when one treats linguistic form as the result of competing pressures [as in the theories of Koehler (1987) or MacWhinney & Bates (1989)]. To the problems of determining the delicate balance of forces is added the problem that the overall structure is itself a moving target; does this language have a high rate of lexical change because the panoply of forces underlying the language situation are in an unstable situation — or is it simply that the language itself is changing rapidly due to societal pressures?

Another problem derives from the statistical nature of most corpus processing. The more detailed the facets of language studied, the larger the corpus necessary to acquire a reasonable-sized sample and (therefore) a meaningful number. In the case of lexical frequency, for example, it might require several million or billion words to find and be confident of any measurement of degree of lexical change. This means, paradoxically, it can be harder (and require more data) to measure gross-scale language changes than small-scale ones.

Recent work in language classification (Juola, 1997) suggests that the cross-entropy between two languages can be a meaningful measurement of linguistic difference, even from remarkably small samples. Juola (1997) has shown, for example, that differences in authorship can reliably be inferred from only a few paragraphs. Furthermore, the specific technique yields meaningful difference estimates at many different scales, including stylometry as above, classification of genres and sublanguages (Juola, in preparation), language identification, or typological (sub)family identification (Juola, 1998).

The particular technique employed determines difference by measuring a simple two-sequence variation of "mean match length within a database." (Farach et al., 1995; Wyner, 1996). This is defined as the length $L_n(x, y)$ of sequences $x_1, x_2, \ldots, x_n$, and $y_1, y_2, \ldots$ as the length of the longest substring $x_i, x_{i+1}, \ldots$ that matches a contiguous prefix $y_1, y_2, \ldots$. This value converges in the limit to the value $\frac{\log n}{H}$ as n increases. Using this technique, one can estimate entropy of a sequence or corpus by using a sliding window of n observations and calculating L at each point in the data stream and thus the mean match length $\hat{L}$ and the estimated cross-entropy $\hat{H}$. This yields a number with many of the useful properties of a true distance measure — in particular, the triangle inequality holds, implying that the distances determined from this measure can be used to define a meaningful space and progression of samples within this space. Furthermore, this metric can be shown to be a simple variation on the monkey-at-a-typewriter scenario, and thus sensitive to any and all sorts of language variation, thus leading to an accurate, precise, and useful categorization of the *overall* rate of language change.

We confirm this in a set of pilot work using published (U.S.) English data at yearly intervals over the years 1965–2000. These data are taken from the *National Geographic* magazine, which has been published at monthly intervals since 1888, thus yielding over a century of high-quality, professionally edited text of an approximately uniform genre and content. Individual articles have been extracted and compared using the cross-entropy distance defined above. The results of this experiment show first, that language change can be observed to be a regular and continuous process of evolutionary change (as opposed to random, unrelated, and reversible processes) and second, that the rate is non-uniform.

Further work in this area and using this technique is almost inevitable, extending the time period, genres represented, and languages under study. A more important question to be analyzed is a question of the relationship between the rates of change of individual components and overall change — does rapid lexical replacement stabilize or destabilize morphological and syntactic patterns? Furthermore, what are the historical and cultural conditions that accompany periods of rapid or slow language change? And, from an application standpoint, how can these findings be related to questions of "language policy"?

References

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics : Investigating Language Structure and Use*. Cambridge University Press, Cambridge.

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A., and Ziv, J. (1995). On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the 6th Annual Symposium on Discrete Algorithms (SODA95)*. ACM Press.

Juola, P. (1997). What can we do with small corpora? Document categorization via cross-entropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Edinburgh, UK. Department of Artificial Intelligence, University of Edinburgh.

Juola, P. (1998). Cross-entropy and linguistic typology. In *Proceedings of New Methods in Language Processing 3*, Sydney, Australia.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics*, 14:251–57.

MacWhinney, B. and Bates, E., editors (1989). *The Cross-Linguistic Study of Sentence Processing*. Cambridge University Press, Cambridge.

Wyner, A. J. (1996). Entropy estimation and patterns. In *Proceedings of the 1996 Workshop on Information Theory*.

Lisa Lena Opas-Hänninen, Pekka Hirvonen, & Fiona Tweedie

Jet set bar or juttutupa?

The main consequence of contacts between languages is borrowing, ie."the incorporation of foreign features into a group's native language by speakers of that language" (Thomason and Kaufman 1988: 37). Borrowing begins, and is at its strongest, in the lexicon; with prolonged intensive contacts, morphosyntactic and phonological features will also begin to be borrowed. Language contacts are typically found in border regions and multilingual communities. However, the present dominance of English as a world language, combined with the spreading of Anglo-Saxon popular culture into the daily lives of many peoples across the world, has brought English into contact with most European languages. This paper examines some of the consequences of such contacts.

Among the manifestations of the ever-present, powerful influence of English on other languages are the names of businesses, especially, it seems, of those catering for younger generations or dealing in freetime activities. Both of these criteria are met by restaurant businesses. In the study at hand, we investigated the extent and geographical distribution of the influence of English and other foreign languages on restaurant names in Finland.

Our data was collected in a truly Labovian fashion: from the white and yellow pages of all of Finland's 1999 telephone catalogues, we gathered the names of all restaurants (including bars, cafs, grills and pubs), comprising a total of approximately 5000 businesses. We keyed in all the restaurant names to Excel worksheets, coded them for source language (English, Finnish, Swedish, Russian, French, Italian, Spanish, or other), type of loan (unadapted loanword, adapted loanword, loan shift, loan translation, or hybrid; the classification is essentially that of Lehiste 1988), type of word (place name, person name, noun [including simple, modified and compound nouns], or phrase), type of establishment (restaurant, bar, pub, caf, or grill), region (represented by area codes), name of community, and type of community (ie. whether it was a largish city or part of an outlying area). In the processing of our data, we used a variety of methods for dealing with categorical data, such as chi-squared tests for differences in multinomial populations and tests for 3 and 4-dimensional contingency tables.

In a pilot study we examined the restaurant names in the greater Helsinki area (south and capital region), the Oulu area (north), and the Joensuu area (east). This data comprised approximately 1400 restaurant names. The main results for the languages used in the three regions, given as percentages of the total of the restaurant names, were as follows:

code	ENG	FIN	ITA	SWE	FRE	SPA	RUS	OTH	Region
09	25.87	38.56	7.22	6.29	4.33	2.58	1.24	13.92	(Helsinki)
08	18.47	66.20	3.48	1.39	0.70	1.74	0.70	7.32	(Oulu)
013	17.36	63.64	4.96	1.65	0.83	2.48	1.65	7.44	(Joensuu)

There is a significant difference ($p_i 0.000001$) across the three area codes in the influence of languages other than Finnish. This influence is by far the greatest in the Helsinki region, with the Oulu and Joensuu regions fairly equal. English is by far the most frequent source of loans in all three regions. Helsinki leads in English loans and has the highest frequency of loans from all the other languages as well, except Russian, where the Joensuu region's proximity to the Russian border seems to have given rise to more frequent Russian-influenced names than elsewhere. One finding that does not show in the above table is that the influence of English is much less in the outlying areas of the Oulu and Joensuu regions than in the centers.

In our paper we discuss the total data and address the following questions: To what extent does the influence of English on Finnish business names vary according to region (south vs. north, east vs. west) and according to the status of the community (center/non-center of region)? In other words, does the pattern suggested by the pilot study apply across the whole country? Is there any relationship between this variation and the degree of adaptation of English (-based) business names? In discussing these questions we will also pay some attention to the interaction of English with other languages.

References

Lehiste, Ilse. 1988. Lectures on Language Contact. Cambridge, MA: MIT Press.

Thomason, Sarah & Terrence Kaufman. 1988. Language Contact, Creolization, and Genetic Linguistics. Berkeley: University of California Press.

Wayne Cowart

The Statistical Profile of Coordination

Coordinate structures present one of the classic puzzles in the theory of human linguistic ability. Despite many intriguing proposals, there is as yet nothing like consensus on how syntactic principles and processing systems construct the relationship between each individual conjunct and other constituents outside the coordinate structure.

This paper explores an unusual approach. Relying on English examples, it considers whether the underlying difficulty with syntactic accounts of coordination may be that listeners and readers do not in fact rely upon syntactic/structural analyses to link conjuncts into the sentences that contain them. The investigation adopts the premise that each of the conjuncts and the balance of the sentence containing a coordinate structure are syntactically intergrated up to the point of building links among conjuncts or between conjuncts and the rest of the sentence. In order to assess the role of syntactic or structural modes of analysis the paper contrasts the statistical profile of acceptability ratings obtained with sentences having or lacking coordinate structures.

Putting the matter in slightly different terms, the paper questions whether those relations that link a conjunct into its surrounding sentence are constructed and represented within the same hierarchical framework that's applied in the absence of coordination. Some of the motivation for this line of investigation is apparent in familiar prescriptive errors writers make.

(1) a. The water samples (from the pond) need testing
    b. Jill's ability and her desire to help has led to a career in medicine

Thus, in (1a) extending the subject NP via incorporation of a prepositional phrase or relative clause readily establishes the hierarchical nature of the subject/verb agreement relation. "Need" continues to agree with the plural "samples" whether or not "from the pond" is included. However, there are coordinate structures where a prescriptively irrelevant NP near the agreeing verb can confuse writers. Thus, in (1b) it appears that the second conjunct has somehow overcome whatever control mechanism links the coordinate structure as a whole to the agreement marking on the verb.

Temporarily neglecting possible syntactic accounts, facts of this sort might be explained by holding that relationships between individual conjuncts and constituents outside the coordinate phrase are constructed within a nonhierarchical memory representation in which linear proximity significantly affects the relationships constructed. Such a theory might hold that though each conjunct has internal syntactic structure, interpretive processes are primarily responsible for establishing links between conjuncts and other constituents. An approach of this sort has some attractions: it correctly predicts that there will be no c-command relations among conjuncts; it offers an account of the seemingly extrasyntactic character of the 'coordination of likes' principle by making interpretive processes responsible for coordinate integration; it entails the Coordinate Structure Constraint because it provides for none of the syntactic paraphernalia in the highest node in the coordinate structure that would be needed to implement extraction.

This paper will review a number of experimental results relevant to this proposal, including recent evidence on Binding Theory violations ($p < .001$) that seems to follow from this proposal. The paper draws on reanalyses of some recent published results (Quattlebaum 1994; Sobin 1997) to show that speaker judgments of sentences containing coordinate structures seem to be notably more variable and clouded than are those with otherwise similar sentences lacking any coordinate structure. For example, Sobin (Sobin 1997) had a dozen subjects rate a variety of coordinate and noncoordinate structures. In the four sentence types in (2) (respondents judged three examples of each type) more than 90% of respondents gave the noncoordinate (2a) examples the highest possible rating while only a third of respondents gave the apparently acceptable coordinate (2c) cases similarly high ratings.

(2) a. Three cups are on the table
    b. Two keys is on the desk
    c. A book and a pen are on the desk
    d. A cup and a napkin is on the table

At the other end of the rating scale, only 3% used the lowest rating for (2a), while 17% used that rating for the (2c) cases. Similar contrasts are evident for the unacceptable (2b) and (2d) examples. Nearly 80% give the lowest rating to the noncoordinate (2b) cases, but only 42% used that rating for the coordinate (2d) cases. No respondents used the highest rating for the (2b) cases, but 28% of the responses for the (2d) cases were in that category. For at least these cases, this difference in the spreadness of responses to noncoordinate and coordinate cases in Sobin's data is reliable, chi square=16.1, df=2, $p < .001$ for the (2a) and (2c) acceptable examples, and chi square=21.7, df=2, $p < .001$ for the unacceptable (2b) and (2d) cases. (The six scale values Sobin used, 0-5, were collapsed to three for this analysis on data reconstructed from the appendix to his paper.)

These results show that there is something notably muddled and unclear about speaker judgments with these coordinate structures. However, the principal finding to be reported demonstrates another sort of contrast that comes somewhat nearer the question whether a significantly different organizational approach is applied to coordinate structures.

Examples like (1a) demonstrate the ability of syntactic structure to impose rigid control over what elements within a sentence interact with what other elements in forming agreement relations. However, various investigations over the last decade (see, for example, (Bock 1995)) have shown that, despite the relevant grammatical constraints, speakers can experience a certain amount of interference when an NP directly adjacent to a verb conflicts the with agreement marking indicated by the head of the phrase containing the adjacent NP, as in (3a).

(3) a. The key to the cabinets is lost
    b. The water samples in/and the material from the pond/ponds needs testing

This paper, however, will report evidence bearing on sensitivity to the amount or nature of the material separating conjuncts from an agreeing verb, and will contrast these findings with those for noncoordinate structures. For example, in (3b) the plurality of the distractor "pond/ponds" is prescriptively irrelevant to the agreement marking on "needs", whether the subject NP is coordinate or not. Indeed, in the two variants of this sentence with a non-coordinate subject NP the alternate forms of the distractor have no discernable effect on judged acceptability. However, the distractor does produce a reliable effect (p=.05) when the subject NP is coordinate. Apparently the boundary surrounding the coordinate structure is permeable, allowing "needs" to sometimes spuriously 'agree' with the singular "pond" in a way that does not occur with a non-coordinate subject NP.

The paper argues that these and other effects motivate further exploration of the possibility that linear, non- hierarchical memory structures, in addition to syntactic/structural reprresentations, are involved in the analysis of coordinate structures.

References

Bock, K. (1995). Producing agreement. Current Directions in Psychological Science 4(2): 56-61.

Quattlebaum, J. A. (1994). A Study of Case Assignment in Coordinate Noun Phrases. The Language Quarterly 32(3-4): 131-47.

Sobin, N. (1997). Agreement, default rules, and grammatical viruses. Linguistic Inquiry 28: 318-343.

Setsuko Wakabayashi, Jun-ya Morishita, & Yasunori Motomura

Contextual influences of top-down inferencing in a language process by EFL listeners

Greene (1986) introduces heterarchically controlled processing passes by which information from different types of processing can be pooled before deciding on appropriate representations for linguistic inputs. The basis of language processing here is memory and inputs pass through linguistic processing with the knowledge stored in memory. Linguistic knowledge is explained influential here but how it can be is not clear.

Abe et al. (1994) describe language processing by means of situational meanings utilizing linguistic and general knowledge. They also explain that processing is operated with listeners' cognition as well as feeling and conation. They claim that interpretation is achieved when listeners cognitively form a semantic representation of speech and a speaker's world expressed in a certain situation or context. How how those inputs influence their performance is not discussed.

Sperber and Wilson (1986) and Blakemore (1992) emphasize the important role of relevance to context and meaning in processing.

Wakabayashi (1997) explains language processing (1) by the speaker's internal representations of sound and other stimuli providing contextual information, which process different quantity and quality of information in processing. (2) A feedback link from attempted interpretation by which following interpretations can be influenced, (3) the link dose this by affecting the selection of items in the perceived context derived from internal representation of other stimuli, (4) to yield the currently relevant context operates to selection process of relevant context which, in turn, (5) differentially activates items of stored non-linguistic knowledge (thereby influencing Top-Down inferencing). How language processing utilizes contextual input by feedback link is examined in order to explain how an interpretation is achieved by EFL listeners.

Performances of Japanese listeners in 3 classes (about 30 students per class) of EFL at a university are quantitatively studied for 4 weeks in order to investigate how contextual input operates in language processing.

Analysis of study shows significant roles of contextual information for an interpretation. Language processing is greatly affected by top-down processing.

References

Abe, J., Momouchi, Y., Kaneko, Y. and I, K (1994). Human language information processing in Cognitive Science & Information Processing - 12 (in Japanese). Science Sha, Tokyo.

Blakemore, D. (1992). Understanding utterances. Blackwell, Oxford UK. Greene, J (1986). Language Understanding: A cognitive approach. The Open University Press, Milton Keynes.

Otake, T. (1995). 'Auditory perception' in Y. Otsu (Ed.) Psycholinguistics (in Japanese), Cognitive Psychology 3 (in Japanese). University Tokyo Press,Tokyo.

Sperber, D. and Wilson, D. (1986). Relevance: communication and cognition. Harvard University Press,Cambridge, MA.

Wakabayashi, S. (1997). A model for listening comprehension processing by novice L2 listeners and its implication. In Working Papaers of Information Scienece Centre vol.2, No.1. Himeji Dokkyo University.

Wakabayashi, S. (1998). Contextual effects of video vision of listening comprehension practice. Kenkyukiyou. 12: 206-220. Himeji Dokkyo University, Foreign languages.

Yamadori, A. (1997). Cerebral organization of language genesis: A three-layered structure. Japanese Psychological Review. Vol.40, No.3: 343-355.

Yamadori, A. (1998). Hitoha naze kotoba wo tsukaeruka (in Japanese).. Koudansha Gendaishinsho, Tokyo.

Yuen Wah Grace Tse

The Grammatical Factors Influencing the Choice Between the Use and Omission of the Definite Article Preceding Organization Names: A Statistical Analysis

Since early times, grammarians and linguists such as Sweet (1891), Poutsma (1914), Kruisinga (1932), Curme (1935), Jespersen (1909-1949), Long (1961), Zandvoort and van Ek (1975), and Quirk et al. (1985) have drawn our attention to the intricate problem of the usage of the definite article with English proper names. Although some of the rules provided by these studies seem quite sound as far as they go, they result from subjective impressions based on intuitive examples.

The present study takes up this issue by using a quantitative approach that is able to discern more precise and scientific characteristics of article usage with proper names, and to construct a statistical model in an attempt to explain the difference between what is observed and what is expected (cf. de Haan and van Hout 1986; Leech et al. 1994; McEnery and Wilson 1996).

Based on a sample of authentic newspaper data taken from the British National Corpus, the present study focusses on multi-word organization names (e.g. Lincat Group, the Law Society, the Women's Ministry, British Airways, the Department of Social Security) which take varied forms and occur very frequently in the modern commerical world.

This corpus-based study makes use of a multivariate statistical technique, logistic regression, to estimate the influence of certain grammatical factors (e.g. a premodifying proper noun/common noun/genitive noun/adjective, a head noun, a prepositional phrase as postmodification, etc.) that are hypothesised as influencing the choice between the use and omission of the definite article preceding multi-word organization names. Logistic regression is also able to identify the set of grammatical factors useful in predicting the use or non-use of the definite article.

The present study shows that, for example, if the premodifying item of the organization name is a proper noun, then it is less likely that the definite article should precede the organization name; on the contrary, if the premodifying item of the organization name is a common noun, then it is more likely that the definite article should precede the organization name. By making precise predictive statements about a language phenomenon, this investigation sheds light on the theory of linguistic performance rather than that of linguistic competence.

References

Curme, G. O. (1935). Parts of speech and accidence. D. C. Heath and Company, Boston.

de Haan, P. and van Hout, R. (1986). Statistics and corpus analysis: A loglinear analysis of syntactic constraints on postmodifying clauses. In: Aarts, J. and Meijs, W. (eds.) (1986). Corpus linguistics II: New studies in the analysis and exploitation of computer corpora. Rodopi, Amsterdam.

Jespersen, O. (1909-1949). A modern English grammar on historical principles. Completed and edited by Niels Haislund. Ejnar Munksgaard, Copenhagen.

Kruisinga, E. (1932). A handbook of present-day English. Part II English accidence and syntax 2. Fifth edition. P. Noordhoff, Groningen.

Leech, G., Francis, B. and Xu, X. (1994). The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In: Fuchs, C. and Victorri, B. (eds.) Continuity in linguistic semantics. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Long, R. B. (1961). The sentence and its parts: A grammar of contemporary English. The University of Chicago Press, Chicago and London.

McEnery, T. and Wilson, A. (1996). Corpus linguistics. Edinburgh University Press, Edinburgh.

Poutsma, H. (1914). A grammar of late modern English: For the use of Continental, especially Dutch students. Part II The parts of speech. Section I A. P. Noordhoff, Groningen.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). A comprehensive grammar of the English language. Longman, London.

Sweet, H. (1891). A new English grammar: Logical and historical. Part I Introduction, phonology and accidence. The Clarendon Press, Oxford. Zandvoort, R. W. and van Ek, J. A. (1975). A handbook of English grammar. 7th edition. The English Language Book Society and Longman Group Ltd, London.

Jaan Mikk

Prior knowledge of text content and values of text characteristics

It is well known that familiar words are shorter, as a rule. On the other hand, familiar words are more often used in texts on well-known topics. Consequently, the word length is an indicator of text familiarity. A question arises: are there more characteristics that have different values in texts on familiar and unfamiliar topics. The aim of the research was to find out which text characteristics correlate with the level of text familiarity.

To achieve the aim, we have to compare some index of text familiarity with the text characteristics. The index of text familiarity was found by experimental research.

In the experiment 30 passages from popular-scientific texts on physics, chemistry, astronomy and biology were used. The average length of the texts was 166 words. Eight free-response questions to every text passage were composed. To equalise the difficulty of the questions to different texts four of them were composed on one sentence from the text, three of the questions controlled comprehension of some short text passages and one question was composed on the main idea of the text.

The questions to every text were answered by 40 students from 9th and 10th grade in Estonian schools. The students were not familiar with the texts and they did not read the texts before answering the questions. I just asked my testees to answer some questions on some topics. Answers of the testees served as the basis for calculating the level of prior knowledge of the content of the texts.

The texts, to which the questions were composed, were analysed to find word length, sentence length and other characteristics of the texts. Then linear correlation coefficients were computed between the text characteristics and the students' level of prior knowledge. 33 text characteristics statistically significantly related to the level of prior knowledge. The most important of the characteristics are given in Table 1.

The correlation coefficients affirmed our hypothesis. If in the text there are more long words (characteristics No 13 and 122), then students can answer fewer questions on the content of the text.

How did the students know which texts were written in shorter words and answered more questions correctly before reading the text? The mysterious correlation has a simple explanation. Some topics are better known in society and students and authors know the topics better as well. Students answer the questions on a familiar topic better and authors do not need many long words to explain the familiar topic. Therefore, there is a statistically significant correlation between the percentage of long words and the percentage of correctly answered questions before reading the text.

No	Text characteristic	Correlation with prior knowledge
5.	Length of sentence in words	-0.36
9.	Length of independent sentence in letter spaces	-0.50
13.	Length of word in letters	-0.45
17.	Percentage of nouns with the abstract suffix -us	-0.46
20.	Percentage of abstract nouns	-0.43
22.	Average abstractness of nouns repeating in the text	-0.46
23.	Percentage of different unknown words	-0.50
64.	Percentage of embedded phrases	-0.38
115.	Percentage of words in the dictionary of foreign words	-0.48
122.	Percentage of words with ten or more letters	-0.47

Table 1. Correlation coefficients between text characteristics and the level of prior knowledge.

The familiarity of text content is expressed in other text characteristics as well. Let us list them.

- Sentences are longer in texts on complicated content matter (characteristics No 5, 9, and 64). Complicated content makes authors write in long sentences as far as the authors feel the need to explain many details of the unknown content.
- If prior knowledge of readers is lower, then the texts are written in a more abstract manner, as a rule (characteristics No 17, 20, and 22). Text abstractness is an important indicator of the complexity of its content.

- The authors use more unknown words, when they write on a complicated content (characteristics No 23 and 115).

We see that texts on a well-known topic differ from texts on a complicated topic in many characteristics. The validity of the results is confirmed by another analogous piece of research (see Mikk & Elts in Quantitative Linguistics, 1993, vol. 52, pp. 228-238). The characteristics can be used to assess the level of prior knowledge of the topic in society.

The characteristics in Table 1 are often used in readability formulae. The formulae are sometimes seen as measures of linguistic complexity of texts. We see in Table 1 that the characteristics are the indicators of the content complexity of texts as well. Readability formulae measure content complexity of texts by all their predictor variables.

The correlation coefficients in Table 1 can be used to formulate the rules for clear writing. Rewriting texts according to the rules was sometimes successful, sometimes not. The explanation of the conflicting results is as follows. The characteristics in Table 1 are to some extent independent from the content of the text - the correlation coefficients are far from maximal value 1. If the author changes the characteristics skilfully, then he/she succeeds in the higher efficiency of text without changing its content. In other cases, the following of the rules of clear writing necessarily leads to changes in the text content.

V.A. Dolinsky

Logos and number. The structure of associative fields in study of consciousness.

Language corresponds with number which is the basis of the world and universal means of cognition. Number structures the associative fields of linguistic units and assigns the frequency distribution of associated words and, therefore, setsdictionary and textual distributions.

A word turns out to be semantically close to number. (1) In Ancient Rome, as well as in Russia until the XVI century, the letters of the alphabet served as figures. The figures were used for cryptography and cipher. *Cipher* is a form of the word *chittre* (French) which means "digital script" and comes from the Arabian word "sifr" (naught, nonentity). Later on, in Europe, "cipher" turned into "zero". Whether are the units with zero frequency such a cipher to the Zipf distribution? (2) The Greek word *logos* is polymorphous. Its semantic field is translated by the notions *word* and *speech*. The term *word* in its broad meaning turns out to be synonymous to the terms *number* and *calculus*. Hence, the famous beginning of the Gospel of John might aswell contain the following peripheral meaning: "In the beginning was Number, and Number was with God, and the Calculus was God".

Trying to approach to the phenomenon of consciousness, we come across with the fact that we can observe and analyse only language. But language is given to aresearcher not only in observation (text or dictionary), but also in experiment (verbal association process). Belief in "mental unity of mankind" is based on the inexplicableability of people for *understanding*. However, the existence of such ability is not the basis for postulating "logic atoms of sense" (G.W.Leibniz) and searching for "generalcomponents of meaning" or "inherent concepts" or "universal semantic primitives" (A.Wierzbicka). This belief is based on a deep, irrational by nature, level of human mentality, intuitively clear semantic unity of mankind, semantic universe, primordial existence of meanings. Speaking about consciousness and its manifestation inlanguage, we turn to "linguistic feeling", i.e. to intuitive and unconscious comprehension of language. The German word *Sprachgefuhl* means the spontaneous response of a native speaker to a certain linguistic form. Such responses are also word associations of native speakers obtained in course of psycholinguistic experiment.

From formal and logic points of view, referring to the verbal associations means a refusal from several basic laws of logic - law of identity (lex identitatis) and law of the excluded middle (exclusi tertii principium). "Energeia" of natural language (W. von Humboldt) does not allow each significant unit to be always equal to itselfwhich is manifested in semantic polymorphism and live dynamics of spontaneous associations. Simultaneously, there is a wide spectrum of semantic opportunitiesbetween "yes" and "no" in linguistic consciousness. Interpretation of these opportunities is inconceivable without attracting probabilistic models of language and consciousness. For building such a model we should turn to an image of *field* which connects physical reality with semantic one.

A word appears in language from wordless sense. Sense disappears from language through a senseless word. "An association occurs without realising the rules, but in accordance with them and, consequently, with reason, though not as outgoing from reason" (I.Kant). "Language by nature is a pre-reasonable function ... Language is not a cover but rather a pre-prepared way or groove" (E.Sapir). "Giving a meaning, we use linguistic matrixes." (C.G.Jung). Words are the *scoops* of sense. Scooping of the deep and capacious content, being the penetration into sense, is an exit into thespace, vacillating between meanings, which is full of individual associations connected by the psyhological unity of a word. The attempts to give a certain fixed meaning to a word equal to the idea of an average (or key) meaning, and therefore, it inevitably turns out to be shallow, poor and superficial. The closer to a "surface"(high-frequency zone) of associative flow, the poorer its composition; the closer to a "bottom" (low-frequency zone), the more various its composition ($m_F < m_{F-1} < \ldots < m_1$). The different degrees of immersion to an associative flow provide for different "catch" of semantic opportunities. Only reaching the depth of sense, we find out that the most numerous (in a spectrum of distribution) group ($m : m_0 \gg m_1$) turns out to be unmanifested associations ($F = 0$), and senses which did not find linguistic forms (neoyazykovlennye smysly).

What images are recalled by the following set expressions and phrases: "to contain sense", "inexhaustible sense", "sense which is not laying on a surface", "to godeep into sense"? Whether the following expressions (in Russian) are of equal value: "What do you associate the word X with?" and "What meaning do you *put* in the word X?"

Language serves for both disclosing and preserving senses and concealing and losing them. The contradictory "concealing-revealing" (H.G.Gadamer), or "folding-unfolding" (D.Bohm) statuses of a word in respect to meanings determines the quantitative structure of associative fields. *On the one hand*, one or several ultra strong associations play the role of social "guards" and "limiters" of meanings; *on the other hand*, a great number of unique unstable associations take the part of individual "pioneers" and "shakers" of meanings. The first *meaning preserving* tendency is manifested in the homogeneity of distribution of associations, limitation of their assortment, domination of a principle of least effort; it serves for preserving language from natural noise destroying its harmony. The second *meaning revealing* tendency is manifested in the heterogeneity of distribution, expansion of assortment, domination of a principle of greatest effort; it serves for preserving language from deliberate attempts of people to emasculate its sense. The mythological analogues of such two extreme poles in the structure of associative fields are the *Tenpentecostal merging of languages* and the *babel (confusion of languages)*; the mathematical analogues are discreteness and continuity; the socio-cultural analogues are totalitarism and anarchy.

It is possible to conclude: A. The world (of meanings) is set a priori. It isopened for any sense which might find a linguistic form (oyazykovleniye). B. The linguistic worlds differ like the geometrical systems of reference do. C. The dependence, bringing the senses of the world into conformity with the units of languages, can be set by a probabilistic measure ($\partial = \hat{I} \ldots 1$). D. The formal technique of setting this function is a mystery of each language, the "concealing-revealing" mechanism of which is intuitively comprehended. An association serves as a key opening the doors of sense. Without occurring an association nothing finds a meaning; for occurring an association nothing has ameaning. "In the main, life means nothing until there is no thinking human being who could interpret its phenomena. It is necessary to explain to those who do not understand. Only incomprehensible has a meaning" (C.G.Jung).

References

Dolinsky V.A'. (1999). Linguistic Associationism: Problems of Sociolinguistics VI. Language and Contemporary Reality. Sofia.

Jan Králík

Remarks on Quantitative Data Understanding

The enormous number of data gathered and presented by quantitative linguists can be seen in several ways, or, e. g., in several levels. I would suggest three views.

The first type of data would be those ones, which could be confronted with natural probabilistic distributions, as, e. g., Binomial, Gaussian, Poisson, etc.

The second type of data could be confronted with - or expressed by - empirical formulae, as., e. g., Zipf-Mandelbrot, Waring-Herdan etc. Many of such formulae can be defined ad hoc.

The third type of data would be those thousands of coefficients, ratios, indexes, correlations, relative numbers, frequencies etc., which are bound with every quantitative research, and which, in quantitative linguistics, possess statistical stability.

These three types of data are neither observed nor considered separately.

As to the third type: single quantitative data (also in tables) can be investigated and compared in many aspects: the stability of results can be standardly tested, the extent of samples can sharpen the issues, data can be seriated, sometimes trends can be observed and quantified, elegant regression curves can be compiled, multidimensional scaling can be applied etc. However, it seems to me that sometimes data and quantitative characteristics are reached, but still their interpretation can not be fully achieved, as they are isolated items in their essence. Today, the number of measurable quantitative characteristics is so enormous that looking for contexts may bring great complications which may cause any interpretation impossible. Important exceptions are represented by those cases, in which single data can form strings, development in time etc., or multidimensional objects, which can be interpreted clearly with direct linguistic connotations.

The second type of data, empirical formulae, result often from just mentioned exceptions. Formulae can be described by words too. Still, for me, some questions remain: May formulae be tested as to their likeliness to the reality? Are we justified to consider empirical formulae in the same way as probabilistic distributions? What can a mathematical improvement of an empirical formula bring? Is it an algebraic complication only, or is it a real sharpening which can be interpreted as a deeper insight? Some of the empirical formulae could be explained as a result of construction from probabilistic distributions. Such a construction brings a real sense to formulae and it makes formulae part of the nature. However, may we hope that every empirical formula possesses a hidden probabilistic base?

Any mathematical model or probabilistic distribution expresses an ideal situation, which we never face in the real world exactly, not even in physics, nor in biology etc. The more we cannot expect exact likeliness between models and data in linguistics. If a good agreement between a model and data is found, two possible steps remain: first, to improve the suggested model mathematically, and second, to ask what the suggested model means. As mentioned above, mathematical improvement of models can complicate future interpretation. In case of models which are similar to those already known from the nature, we are necessarily faced with the question how to represent the likeliness between the nature and quantitative characteristics of the human language. This could be a philosophical question. But quantitative linguistics gives some explicite proofs for quite sure statements already, as, e. g.: Natural laws are projected into human language - or, say, into human speech - independently of particular language.

Examples will be shown.

References

Altmann, G. (1996) The Nature of Linguistic Units. Journal of Quantitative Linguistics, Vol. 3, No. 1, pp. 1-7.

Altmann, G. (1997). The Art of Quantitative Linguistics. Journal of Quantitative Linguistics, Vol. 4, No. 1-3, pp. 13-22.

Bunge, M. (1967). Scientific Research I. Springer. Berlin.

Victor Kromer

Parameter-free model of rank polisemantic distribution

The purpose of this paper is the creation of a model of rank polysemantic distribution with a minimal number of fitting parameters. In an ideal case a parameter-free description of the dependence on the basis of one or several immediate features of the distribution is possible.

Probabilistic polysemantic distribution is peculiar to a word with a concrete frequency in the frequency list. Let's introduce dependence "usage - polysemanticism" proceeding from the following theoretical assumptions:

1. The lexicon of a concrete language (sublanguage) is reflected in the appropriate explanatory dictionary. A certain constitutive text corpus lays in the basis of the language.
2. Each use of a concrete word in the constitutive text corpus gives a new meaning.
3. A native language speaker does not distinguish separate meanings of the words and arranges them into groups of meanings (so-called "dictionary meanings") according to the psychophysical Weber-Fechner's law.

Let's make an assumption, that words in the constitutive text corpus are distributed in accordance with Zipf's law:

$$F = \frac{K}{i^{\gamma}} \tag{1}$$

where F is the frequency of a word with rank i, K is a constant of proportionality and γ is Zipf's parameter for the distribution. According to Weber-Fechner's law, the spread of the process of arranging word-meanings onto groups of meanings, the number of dictionary meanings will be:

$$m_F = \Psi(F+1) + C \tag{2}$$

where $C = 0.5772\ldots$ is the Euler-Mascheroni constant.

Let's designate the number of words in the language considered as L and require equality of the rarest word in the constitutive text corpus, and accordingly the mathematical expectation of the number of its dictionary meanings, to unity. Let's introduce a normalization condition, that is we require equality of the sum of mathematical expectations of the number of word-meanings on all the ranks according to expression (2) and the total of meanings of all the words in the explanatory dictionary. We obtain a set of equations:

$$\begin{cases} \frac{K}{L^{\gamma}} = 1 \\ \sum_{i=1}^{L} \left[\Psi\left(\frac{K}{i^{\gamma}} + 1\right) + C\right] = M. \end{cases} \tag{3}$$

Solving the set of equations (3), we find values of K and g.

It is defined from experimental data that the probability of a word with a certain frequency to have k dictionary meanings is determined by the formula:

$$p_k = \frac{(m_F - 1)^{k-1}}{m_F^k}, \tag{4}$$

where m_F is the mathematical expectation of the number of dictionary meanings for a word having frequency F in the constitutive text corpus. The number of words in the dictionary with k meanings will be:

$$N_k = \sum_{i=1}^{L} p_k = \sum_{i=1}^{L} \frac{(m_F - 1)^{k-1}}{m_F^k}, \tag{5}$$

where mF is determined according to formula (2), and F according to formula (1) with parameters K and g, defined from the set of equations (3) solution.

The model offered is tested on a series of the explanatory and author's dictionaries. The test was produced by comparison of the empirical and theoretical number of words having a certain degree of polysemy

and evaluation of a goodness-of-fit test "Chi-square". For the evaluation of the criterion the words with a certain degree of polysemy were regarded as one class, and the small classes were integrated up to a size not less than 10 words. The least level of significance P, permitting us to reject the null hypothesis about equality of two compared distributions the theoretical and empirical one was calculated. For the Pushkin's language dictionary $P = 0.21$, and for the Dictionary of Russian in 4 volumes $P = 0.64$. The null hypothesis about the correspondence of empirical and theoretical distributions can be accepted. For Ojegov's Dictionary of Russian $P = 0.03$, and that formally allows us to reject the null hypothesis; however by surveying the empirical data the anomalous part of the polysemantic distribution for the words with polysemy degrees of 8 and 9 (inverse dependence) comes to light, and that can be probably explained by extralinguistic factors. When joining these two classes in one class $P = 0.51$, which again allows us to accept the null hypothesis. For the Dictionary of Contemporary Russian in 17 volumes P is practically equal to zero, that is, the sharp distinction in compared distributions is observed. Accepting that the area of one-meaning words can fall out of the general tendency of the relations between volumes of word-groups of different polysemy degree, this means that it is possible as overfilling of peripheral area at the expense of engaging words concerning the lexical sphere outside the concrete dictionary type, as arbitrary exception of some dictionary words from this area, a conversion is possible of the parameter-free model into the one-parameter model with a fitting parameter $L*$ the number of words in the modified dictionary, at the expense of expansion or compression of the area of one-meaning words of the dictionary. Establishing the parameter $L*$ as 157000 words, and by taking words with a number of meanings more than 14 out of consideration, P is equal to 0.03, and that allows us to accept the null hypothesis about equality of two distributions conditionally. There are about one hundred words with a degree of polysemy more than 14 in the dictionary under consideration, and they, as well as the words with one meaning, also fall out of the general tendency, described by the model offered.

So, the model offered allows us to describe polysemantic rank distributions of the lexicon of explanatory dictionaries on the basis of the observed distribution parameters the total number of words and total number of meanings, and that means the model is a parameter-free one. In case of falling out of the area of one-meaning words from the general tendency the model requires introduction of one fitting parameter the number of words in the modified dictionary.

Edda Leopold

A Model Theoretic Inspection of the Concept of Language as System of Competing and Cooperating Forces.

According to Girgerenzer (1981) model construction is a five valued relation. A researcher S chooses an empirical system E as a model for an object domain D in order to fulfill the purpose Z, and selects a mathematical domain N to formalize the relations in E. In his metamodel he makes a further distinction between the empirical system and the object domain induced by the mathematical domain E_N and D_N and the empirical system E_S and the object domain D_S originally choosen by the researcher. He points out that a discrepancy between E_S and E_N leads to wrong conclusions. An example of this was described by the author (Leopold, 2000).

Model construction in synergetic linguistics goes beyond a mere description of language phenomena. The attempt is to explain the structure of language and language dynamics. In analogy to classical physics relations between several linguistic quantities are formalized by mathematical functions, which map an argument x unequivocally onto a value $y = g(x)$. The relation between x and y, however, is not only considered as a description of the area where most of the data points (x, y) are found but the mathematical function $g(x)$ is considered as an explaination of something that is going on between x and y. It is claimed that x influences y (or vice versa) and $g(x)$ is considered as an manifestation of this influence. Futhermore a central hypothesis of synergetic linguistics is that language can be considered as a self-organizing and self-regulating system, with the consequence that the language system adopts a structure which enables the system to serve as a means for communication.

So we can summarize: A synergetic linguist is a subject S, who uses a model of cooperating and competitive forces E, borrowed from the synergetic sciences, in order to explain the influences of linguistic quantities on each other Z, which can be observed in texts D. To this end he/she uses real valued functions N and various analytical techniques like differential equations and the like.

I will consider the purpose Z of the research program of synergetic linguistics more thoroughly and I will show that different purposes require different mappings between the object domain D and the mathematical domain N in order to avoid misinterpretations of the mathematically induced empirical system E_N

Z_1 description by a numerical equation
Z_2 explanation by an isolated numerical equation
Z_3 construction of a theory consisting of system of equations

I will show that the ambitious claims of synergetic linguistics (Z_3) necessarily require the postulation of the existence of a system of linguistic quantities independent of the observer, and that a more elaborated formalization of the quantities involved is needed.

Purpose Z_1 — description by a numerical equation — means simply that the data can be found in a special region of the domain under consideration. In the case of the relation of polysemy Ψ and frequency F for example one may observe that most of the datapoints (Ψ, F) (i.e. pairs of polysemy and frequency) can be found near the graph of the function

$$\psi = af^b, \tag{1}$$

where ψ and f denote polysemy and frequency respectively and a and b are positive parameters. Or more precisely one may observe that the average polysemy $\bar{\Psi}$ of words ocurring in a text F-times can be described by equation (1). Since the mean is an estimator for the expectation one may infer that the conditioned expectation $\mathbf{E}(\Psi|F = f)$ of Ψ when F is given can be described by (1). According to Altmann (1993) a mere description of the data can not be considered as a law.

It should be noted at this point that other statistical functionals are also admissable in equation (1). So one could for instance consider the standard deviation of L or the value of some other parameter of the distribution of L instead of the mean (see Leopold 1998).

Purpose Z_2 — explanation by an isolated numerical equation — requires a theoretical deduction of the equation tested on the data. The adequate explanation of linguistic phenomena is the functional explanation as pointed out by Köhler (1986,1990). Altmann (1981) has postulated the basic premises of such an explanation. Relations of two variables are often explained with differential equations. So the average polysemy $\bar{\Psi}$ of words with frequency is given by

$$\bar{\Psi} = aF^b, \tag{2}$$

which is explained by the plausible assumption that the "change of polysemy is proportional to the ratio of polysemy and frequency", formally

$$\frac{d\Psi}{F} = b\frac{F}{\Psi} \tag{3}$$

It is important to note that the variable on the right hand side of (2) is a quantity of the empirical system representing observations of entities of the object domain, whereas the variable on the left hand side is a mean value and represents a quantity of the empirical system induced by the mathematical domain. The variable $\bar{\Psi}$ can only be calculated by the operations defined in the mathematical domain. This discrepancy was already discussed elsewhere (Leopold, 2000) and it gives a reason why equation (2) cannot be directly derived from a differential equation which could explain it. However one can introduce an error variable ϵ_f which is added to the solution of the differential equation

$$\Psi = af^b + \epsilon_f. \tag{4}$$

Note that (3) accounts only for the "deterministic" part of (4). The error variable ϵ_f however depends strongly on f (heteroscedacity). In the case of the relationship between length and frequency for instance a hypothesis about ϵ_f can be deduced on the assumption that economic forces in the language system are the more mandatory the more frequent a word is. The presence of heteroscedastic errors requires special statistical methods (see e.g. Gallant 1987). A salient problem however when dealing with quantities depending on frequency is that the prior distribution $\mathbf{P}(F = f)$ of the independent variable is extremely skewed.

For **purpose Z_3 — construction of a theory consisting of system of equations** — the explanatory equations have to be symmetric. This means that both sides of the equation have to be quantities of the same empirical system. To this end equation (4) has to be reformulated. We consider a system of four variables polysemy ψ, polytexty ϕ, frequency f, and length l. One can write down a system of *deterministic* equations explaining different pairs of linguistic quantities, like for instance

$$\begin{aligned} l &= g_l(f) \\ \psi &= g_\psi(l) \\ \phi &= g_\phi(\psi) \\ f &= g_f(\phi), \end{aligned}$$

where the functions $g_\cdot(\cdot)$ explain the relations between the variables.

Since these equations are deterministic their variables cannot represent liguistic quantities themselves. They denote values in an abstract mathematical domain N. One can establish mappings between quantities of the abstract mathematical domain N and their observable counterparts in the empirical Model E_S:

$$\begin{aligned} L &= l + \xi_l \\ \Psi &= \psi + \xi_\psi \\ \Phi &= \phi + \xi_\phi \\ F &= f + \xi_f. \end{aligned}$$

Here ξ_l, ξ_ψ, ξ_ϕ, and ξ_f are random variables, which introduce conditioned probability distributions $\mathbf{P}(L|l)$, $\mathbf{P}(\Psi|\psi)$, $\mathbf{P}(\Phi|\phi)$, and $\mathbf{P}(F|f)$ in the domain of the empirical Model.

References

Altmann, Gabriel (1981): Zur Funktionalanalyse in der Linguistik; in: J. Esser & A. Hubler (Hrsg.): Forms and Functions; Gunter Narr: T"ubingen.

Altmann, Gabriel (1993): Science and Linguistics; in: R. Köhler & B. B. Rieger (eds.): Contributions to quantitative linguistics; Proceedings of the first international conference on quantitative linguistics, Qualico, Trier, 1991; Kluwer: Dordrecht, pp. 3–10.

Gallant, Ronald A. (1987): Nonlinear Statistical Models; Wiley & Sons: New York et al.

Girgenzer, Gerd (1981): Messung und Modellbildung in der Psychologie; Ernst Reinhard: München, Basel.

Köhler, Reinhard (1986): Zur linguistischen Synergetik: Struktur und Dynamik der Lexik; (QL 31); Brockmeyer: Bochum.

Köhler, Reinhard (1990): Synergetik und sprachliche Dynamik; in: Walter Koch (eds.): Nat"urlichkeit der Sprache und Kultur; Brockmeyer: Bochum, pp. 96–111.

Leopold, Edda (1998): Frequency Spectra within Word Length Classes; in: Journal of Quantitative Linguistics 5, (3).

Leopold, Edda (2000): Fractal Structures in Language. The Question of the Embedding Space; in: Reinhard Köhler, Ludmila Uhlírová & Gejza Wimmer (eds.): Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hřebíček; wvt: Trier, pp. 163-176.

$y = g(x)$ / description of data / process of self-organization

A. Polikarpov

Menzerath's Law for Morphemic Structures of Words: A Hypothesis for the Evolutionary Mechanism of its Arising and its Testing

1 Aim of the paper

In the most general formulating Menzerath's Law sounds likefollows: the longer some"construct" (the whole) the shorter should be its "components" (parts) [Altmann, 1980; Altmann, Schwibbe, 1989]. In its historically initial form [Menzerath, 1954] it described the reverse proportional dependence of the average length of syllables in words on length of words (measured by number of contained in them syllables). Later on this law was expanded for describing regularities of various units on various levels of language organization (syntactic, textual, etc.) and even for describing other semiotic, bilogic, etc. phenomena). Nevertheless, it wasn't theoretically founded and even wasn'tempirically studied on the basic sign level of any national variety of Human Language organisation, level of morphemic units. Units of any other language levels (beginning from a word level) are formed mainly by combination of these basic units into more complex ones. That's why quantitative-structural regularities for sign units of any other upper lying levels of language sytem can't be independent on regularities happening on the basic, morphemic level, can't be properly understood without theoretical and empirical study of regularities on it. In this paper there is present (1) a hypothesis for understanding the evolutionary mechanism responsible for the Law arising and (2) some data for testing the hypothesis. For reaching the second goal the database "Chronological Morphemic and Word-Formational Dictionary of Russian Language" (CMWDRL) containing on the whole more than 180,000 words prepared at the Laboratory for General and Computer Lexicology and Lexicography at Moscow State Lomonosov University [Polikarpov, Bogdanov, Kryukova, 1998] is used. In this paper those results are present which concern analysis of only root and affixally derived Russian words (more than 50,000 different words).

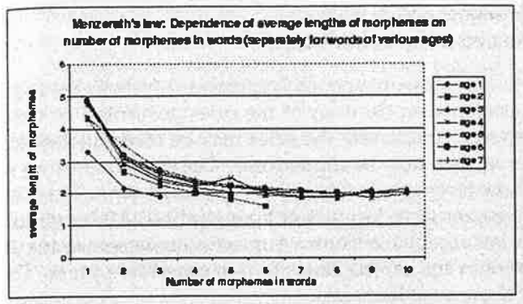
2 An hypothesis

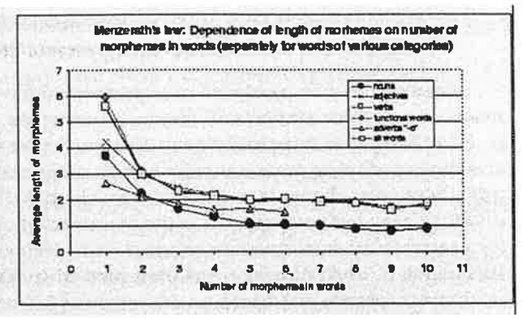
According to the model of sign life cycle [Polikarpov, 1993] it is natural to expect that the most probable (statistically dominant) direction for the categorial development within the nest of derivationally connected words will be the movement from some relatively concrete, objectively oriented categorial semantics of each word-base towards its derivatives of more abstract and subjectively oriented parts of speech categories. So, there should be a tendency to begin a word-formational tree mainly from nouns, to continue it with adjectives, verbs, adverbs, pronouns, etc., and to end it with words of pure syntactic (functional) quality like conjunctions and prepositions. This direction of the categorial development most basically is predetermined by the fundamental fact of the inescapable development of any word's integral lexical semantics during speech acts mainly into the direction of its greater abstractness. More abstract lexical semantics seeks corresponding more abstract categorial form (which is more organic to it) and finds it in acts of word-formation, producing further derivatives from previously derived words.

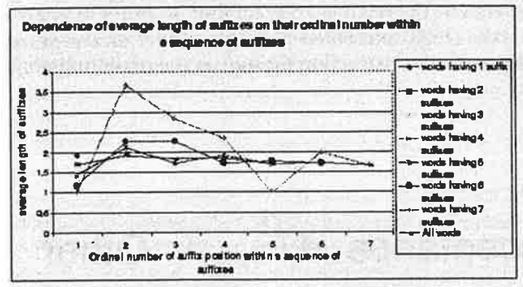
One of the most remarkable consequences of the mentioned process is probabilistic categorial, age, frequency and length ordering of morphemes within a wordform. It means that derivation affixes which are more distant to their root should be proportionally more grammatical, more frequent, and, finally, shorter than less distant ones. It is clear that just this phenomenon leads to the inescapable gradual diminishing of the average length of affixes and, correspondingly, of morphemes on the whole within longer wordforms, i.e. it leads to the existence of the law under study.

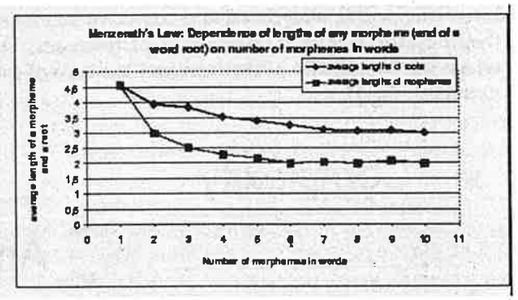
3 Data and results

Below there is present some data on this point showing the validity of the derived general regularity and its significant variation for morphemic structures of words of various ages, categorial form, for roots of words as opposed to all morphemes together, for suffixes in different positions within a word (figures 1–4).









By use of this data Menzerath's Law in this study is certainly corroborated for morphemes of words of any part-of-speech category and words of any age. Words of different ages and categorial form demonstrate not only the astonishing analogy in following the law, but also the significant differences. For instance, words of the same length proportionally to the decline of their age (7 grades of ages - from the most ancient words of Indo-European and older origin, to gradually younger and younger words up to the youngest words of the 20th century) are built with the use of gradually longer morphemes. This, presumably, demonstrates that, on the average, younger and less grammatical words are built by younger morphemes.

A phenomenon of the average morphemes' length reverse proportional dependence on the ordinal number of morphemes' position within a word (i.e., the more distant a suffix from a root, the smaller is the length of a suffix), in our opinion, is of even more fundamental importance for the theory than a Menzerathian itself. As a matter of fact, Menzerath's law itself is a derivative of the more basic positional dependence of morphemes' length on place and, therefore, function of morphemes in a word.

4 Oscillation phenomenon

While empirical data obtained from CMWDRL for the chains of suffixes of Russian words in general is in line with just proclaimed dependence there is also observed some slighter tendency for oscillations, i.e. rhythmically repeating and gradually diminishing "plus" and "minus" weak deviations from the main dependence stream. Some evolutionary explanation for it also will be present at the Conference. Figures 1-4.

Menzerath's law: Dependence of average lengths of morphemes on number of morphemes in words (separately for words of various ages)

Peter Kunsmann & Johannes Gordesch

A Note on a Systems Theoretical Model of Usage

Language may be conceived of as an autopoietic (i. e. self-organizing, self-regulatory) system. Speakers making use of this system will choose a particular utterance on the basis of the rules governing its units (i. e. de Saussures linguistic signs). In a particular speech community the rules may be conventionalized and thus specifying grammatically correct utterances or they may be stigmatized. The social relevance of utterances, then, depends on the extent to which their use is accepted without such stigmatization. The term divided usage has been applied in the literature when two or more forms have been studied which compete for grammatical correctness in contrast with deviant but acceptable forms. A questionnaire containing 48 such items of divided usage was presented to speakers in the United States in two successive years. The relative acceptance was based on the return of 300 of these questionnaires.

What motivates a speaker to choose utterances deviating from the norms set by the speech community? Linguistic, social, situational and cognitive factors determine the relative acceptability. In order to analyse the linguistic factors, Kunsmann, Gordesch and Dretzke (1998) presented a simple model of acceptance where the complexity of the item and the type of grammatical construction figured as the main influential variables (Fig.1).

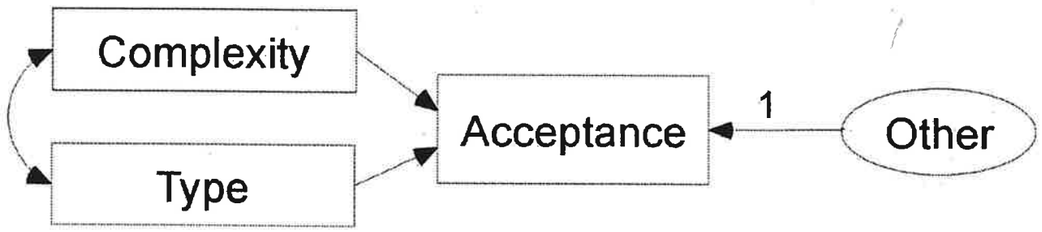


Figure 1. Linguistic Model of Acceptance

1	at least 1 point for each sentence
2	disruption of sentence intonation (1 pt)
3	syllable length of the divided usage form (1pt for each syllable)
4	phonological distribution (C-clusters) (1 pt)
5	the divided usage form constitutes a frozen phrase (1 pt)
6	syntactic complexity of the sentence (1 pt for each syntactic clause)
7	complexity of the divided usage form (1 pt for each modifier, morphological or syntactic unit)
8	application of some movement rule in the sentence (1 pt for each)
9	the divided usage form lacks a case assigner (1 pt)
10	in coordinate structures: inclusion of a lexical noun phrase (1 pt)
11	semantic complexity of matrix sentence (1 pt for weak semantic content)
12	reduction of 1 pt when the divided usage form is socially marked in the speech community

Figure 2: Complexity: criteria for assigning numerical values

1	case assignment to object
2	deletion of -ly for adverb
3	split infinitive
4	agreement (singular/plural)

Figure 3: Grammatical category types

Complexity was defined by a weighted set of criteria and a complexity index computed. The items listed on the questionnaire were grouped into fifteen grammatical types. The relationship between the complexity index and the grammatical types was found to be such that a high complexity index increased the acceptability for some categories while it decreased for others.

Measuring linguistic complexity is still a difficult task. Therefore, for the present analysis a slightly revised set of criteria will be used (Fig.2). The grammatical category types used are listed (Fig.3).

Mathematical models using techniques related to nonlinear theory of dynamic systems are developed in some detail and applied to empirical data.

References

Kunsmann, P.; Gordesch, J.; Dretzke, B.(1998). Native Speakers Reactions to Modern English Usage. Journal of Quantitative Linguistics, Vol. 5, No. 3, pp. 214-223.

Andy Way

Applying Insights from Quantitative Linguistics to Problems of Machine Translation

Machine Translation LFG-DOT (Way, 1999) has been developed as a novel model for Machine Translation (MT) based on DataOriented Parsing (DOP: Bod, 1998) allied to the syntactic representations of Lexical Functional Grammar (LFG: Kaplan & Bresnan, 1982). We shall demonstrate that LFG-DOT can cope with translational phenomena which prove problematic for many other systems.

DOP has produced interesting results for a range of NLP problems. DOP language models consider past experiences of language to be significant in both perception and production. DOP prefers performance models over competence grammars: models based on large collections of previously occurring fragments of language are preferred to abstract grammar rules. New language fragments are handled with respect to existing fragments from the corpus, which are combined using statistical techniques to determine the most probable analysis for the new fragment.

DOP has been used already as a basis for MT-Data-Oriented Translation (DOT: Poutsma, 1998). DOP models typically use surface PS-trees as the chosen representation for strings. The DOT translation model relates tree-fragments between two (or more) languages with an accompanying probability, linking source target translations at all possible nodes in accordance with the principle of Compositionality of Meaning. Once the most likely parse of the source language sentence has been produced, the tree structure of the target is assembled, and the string is (trivially) derived.

DOT is an interesting model, yet it is not guaranteed to produce the correct translation when this is noncompositional and considerably less probable than the default, compositional alternative.

One linguistic phenomenon which has received a good deal of treatment in the MT literature is *Headswitching* (Sadler et al., 1989, 1990; Sadler & Thompson, 1991). These involve cases where what is a dependant of node X in a source language becomes the governor of X in the target. Examples include:

(1) EN: John has just gone → FR: Jean vient de partir.
EN: John likes to swim → NL: Jan zwemt graag.
→ DE: Johannes schwimmt gerne.

When faced with such data (as well as others such as Relation Changing verbs, e.g. EN: *like* → FR: *plaire*), DOT's model of adherence to left-most substitution in the target given *a priori* left-most substitution in the source is too strictly linked to the linear order of words. Given this, as soon as this deviates to any significant degree between languages, DOT has a significant bias in favour of the incorrect translation (assuming the corpus to be representative). Even if the correct, non-compositional translation is achievable in such circumstances, there are likely to be many other wrong alternatives with higher probabilities. In this case, the correct translation will be dismissed, unless all possible translations are inspected manually.

LFG, however, can capture and provide representations of linguistic phenomena other than those occurring at surface structure. Given this, the functional structures of LFG have been harnessed to the techniques of DOP to create a new model, LFG-DOP (Bod & Kaplan, 1998). LFG-DOP permits (via the *Discard* operator) the relaxation of certain constraints on LFG representations, thereby creating generalised fragments against which new input can be compared, and the best analysis constructed.

Way (1999) proposes that LFG-DOP has the potential to be used as the basis for an innovative MT model, LFG-DOT. Being able to link exactly those source-target elements which are translations of each other using LFG's τ-equations, LFG-DOT overcomes the problems specific to the DOT system. Furthermore, LFG-DOT promises to improve upon the LFG-MT model (Kaplan et al., 1989), which has been shown to have serious difficulties with Headswitching data (Sadler et al., 1989, 1990; Sadler & Thompson, 1991). Furthermore, LFG-DOP's *Discard* function enables both unseen and ill-formed input to be handled with relative ease, thereby increasing robustness compared to LFG-MT models. Finally, DOP's statistical model also gives a "level of correctness" figure to alternative translations. This is useful in cases where the default translation in LFG-MT (and in many other systems) cannot be suppressed when the specific translation is required.

The problematic translation data will be presented and discussed, and results will be provided showing how DOT, LFG-MT and LFG-DOT attempt to cope with such phenomena, with differing degrees of success.

References

Bod, R. (1998): Beyond Grammar: An Experience-Based Theory of Language, CSLI Publications, Stanford, California.

Bod, R. and R. Kaplan (1998): "A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis", in COLING: Proceedings of the 17th International Conference on Computational Linguistics & 36th Conference of the Association for Computational Linguistics, Montreal, Canada, 1:145-151.

Grishman, R. and M. Kosaka, (1992): "Combining Rationalist and Empiricist Approaches to MT", in Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, pp. 263-274.

Kaplan, R. and J. Bresnan, (1982): "Lexical Functional Grammar: A Formal System for Grammatical for Grammatical Representation", in J. Bresnan (ed.) The Mental Representation of Grammatical Relations, MIT Press, Cambridge, Mass., pp.173-281.

Kaplan, R., K. Netter, J. Wedekind and A. Zaenen (1989): "Translation by Structural Correspondences", in Fourth Conference of the European Chapter of the Association for Computational Linguistics, Manchester, pp.272-281.

Poutsma, A. (1998): "Data-Oriented Translation", in Ninth Conference of Computational Linguistics In the Netherlands, Leuven, Belgium.

Sadler, L., I. Crookston, D. Arnold and A. Way (1990) 'LFG and Translation', in Third Conference on Theoretical and Methodological Issues in MT , University of Texas, Austin, pp.121-130.

Sadler, L., I. Crookston and A. Way (1989) 'Co-description, projection, and 'difficult' translation', Working Papers in Language Processing 8, Department of Language and Linguistics, University of Essex, Colchester.

Sadler, L., and H. Thompson (1991): "Structural Non-correspondence in Translation", in Fifth European Conference on Computational Linguistics, Berlin, Germany, pp.293-298.

Way, A. (1999): "A hybrid architechture for robust MT using LFG-DOP", Journal of Experimental and Theoretical Artificial Intelligence 11:441-471.

Oliver Cromm

Big Is Beautiful? What Resources Most Influence the Performance of Translation Extraction from Non-parallel Corpora

Our Aim: Example-Based Machine Translation (EBMT) of Compounds. EBMT has been applied mainly to sentences (as in TM systems), and in research also to word groups like idioms. We think that it is promising to apply it to compounds, which often can be inferred from their constituents.

In order to study Japanese-to-German EBMT of Compound Nouns, a database of example translations had to be assembled. Other sources being inadequate, we turned to extraction from non-parallel text. The results are quite useful, making us optimistic for the final example-based translation as well. Nevertheless, the pitfalls of our experiments underscore the need for better electronic dictionaries.

Compound Nouns

Both Japanese and German show a lot of nominal compound expressions, and they often constitute important technical terms. Our preliminary study showed that many Japanese noun+noun compounds are translated not into German Compound Nouns, but into adjective+noun or noun+noun(genitive); thus, in our experiment, we considered these three types as possible target expressions.

Non-parallel Corpora

Extraction of translations from non-parallel corpora using collocational information (e.g. Rapp 99) still fails with low-frequency expressions. As for compounds, we may attack them from the homeground of constituent translations, given a dictionary of base words.

The Experiment

Corpora:

1. Japanese Parliamentary Record (1.6 MB);
2. German Newspaper (900 MB)

Dictionary: jddict, about 10000 entries

1. Extract noun-noun sequences relying on the Japanese tagger
2. Look up the constituents in the dictionary
3. Combine any feasible translations to form German nominal expressions
4. Try to verify these in the German corpus

Results And Considerations How To Improve Them

Due to the small dictionary, nearly 2/3 of the extracted expressions could not be treated at all. Of the others, a sample was taken to assess quality, and checked by several bilinguals. For the expressions not succesfully translated, we tried to judge the reasons.

For roughly 1/4 of all sample expressions, our method appears not applicable. For about 2/3 of all expressions within reach of our method, an acceptable translation was found, giving a recall of 45% within the sample (16% overall) and "precision" of 85%.

In comparison to similar studies (e.g. Tanaka 99), we use a rather small dictionary, but rather large monolingual (German) corpus. It could be expected that dictionary insufficiency is a bigger problem here than corpus insufficiency.

Extrapolations from our sample indeed suggest that a doubling of our corpus size would raise the overall recall by some 2% only, while a dictionary of about 40000 entries could be expected to double it with more than 15% increase of recall, thus raising the number of possibly useful examples to more than 3000 (up from now about 1500) translated Japanese expressions from just 1.6 MB Japanese text.

Conclusions

Despite claims to the contrary, finding compound translations from translations of their constituents is a feasible method, even for languages far from each other. Statistical methods on large corpora can give good suggestions for translations, which should then be verified by humans or from parallel text. Though automated processes on monolingual resources can help multilingual NLP, at least good dictionaries are an essential, and for many language pairs, more lacking than bigger corpora (cf. Grefenstette 1999). Dictionaries have to be assembled by hand first, but would then prove valuable in many applications.

References

Grefenstette, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. ASLIB '99 Translating and the Computer 21, London, UK

Rapp, R. (1999). Korpusbasierte Lexikonerstellung mittels nicht-paralleler Texte. GLDV '99 Frankfurt, Germany

Tanaka, T. and Matsuo, Y. (1999). Extraction of translation equivalents from non-parallel corpora. Proceedings of TMI-99, Chester, England

Kyo Kageura and Sandra Yamilet Santana

A Comparative Observation of English and Spanich Technical Terminology

1 Introduction

The quantitative nature of English and Spanish parallel terminologies (Kotani & Kori 1990) is observed, in terms of heads (defined as the deepest head) and modifiers (all others), focusing on content-bearing elements (functional elements are removed, and sex and number variations are normalised manually), following Kageura et. al (1999), which observed overall tendencies of English and Spanish terminologies. The basic quantities of the data are shown in the table below; N indicates the number of running constituent elements and V the number of different elements. The number of terms are 30056, and average length in terms of the number of elements are 1.86 in English and 1.95 in Spanish. The ratio of head elements in English (53.76%) is bigger than that in Spanish (51.31%).

	All		H		M		M_H		M_M	
	Eng	Spn	Eng	Spn	Eng	Spn	Eng	Spn	Eng	Spn
N	55948	58619	30080	30080	25868	28539	16290	18668	9578	9871
V	10122	8476	7059	5782	5685	5162	2622	2468	3063	2694
N/V	5.53	6.92	4.26	5.20	4.55	5.53	6.21	7.56	3.13	3.66

2 Viewpoints of Analyses

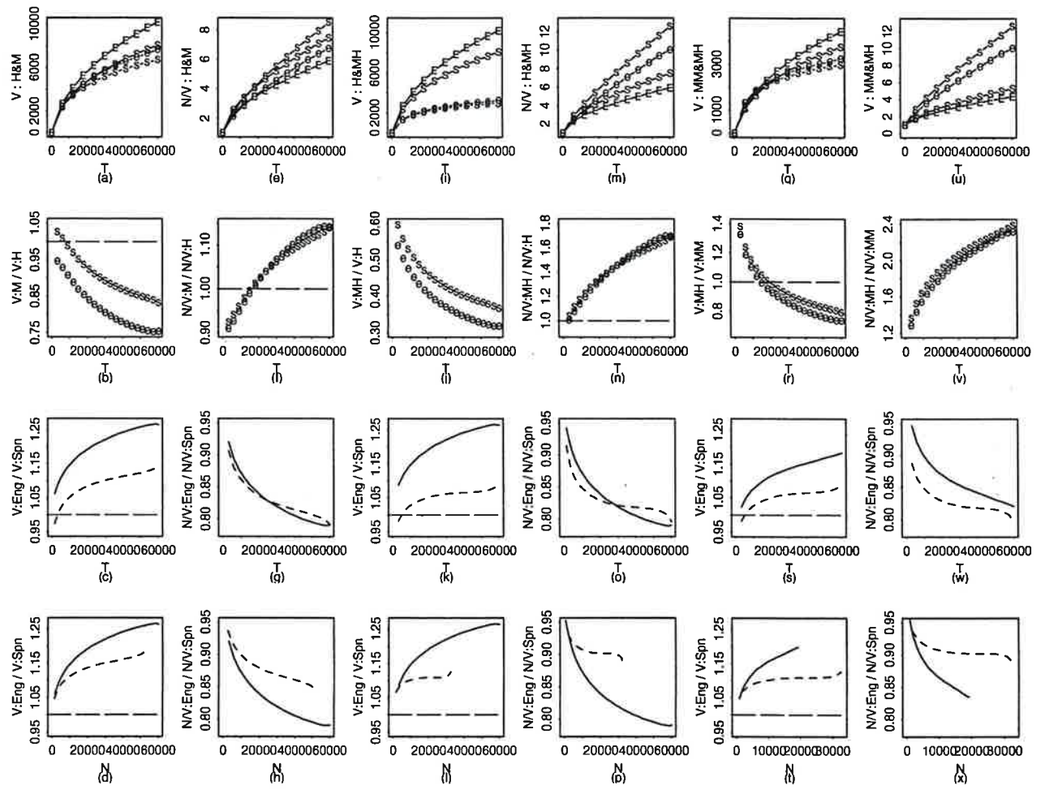
Assuming that the use of elements in heads and modifiers are separately determined, or independent, within a terminology, the number of constituent elements (V) as well as the average use of an element (N/V) were observed, from the viewpoint of (1) head elements vs. modifier elements and to (2) elements used in heads vs. elements only used in modifiers. We thus observed characteristics of 'positions' as well as of 'elements'. For succinctness, let H refer to the set (and distribution) of head elements, M modifier elements, M_H elements of H in modifiers, and M_M elements occurring only in modifiers.[1] We trace the transitions of these points of observation up to twice the original data size using binomial inter- and extrapolation (Baayen to appear; Good & Toulmin 1956). The average length of a term and the token ratio of heads and modifiers are assumed to be constant irrespective of sample size. Taking advantage of this, we give interpretation from the point of view of terminology (i.e. based on T), although the quantitative model is constructed of the relation between V and N [2].

3 Observations

The figure illustrates patterns of the above observation points. The left two columns show relations between H and M , the center two between H and M_H, the right two between M_H and M_M. Panels in the top row show transitions of V and N/V, those in the second low show intra-lingual ratio in English and Spanish, and those in the bottom two show inter-lingual ratios of the observation points with respect to T and N. 's' and 'e' indicates Spanish and English respectively, and capital letters as well as solid lines indicate H (left two and center two colums) and M_M (right two).

3.1 Head (H)/Modifier (M)

From panels (a) and (d), we can observe that the same type of difference between H and M is observed in both languages, i.e., the groth of V is slower and the average use is higher in M . Both languages consume more element types in head than in modifiers. The second row shows that, within that language-independent tendency, the difference between H and M is (or become) smaller in Spanish than in English.[3] The ratios of average frequencies of H and M show similar pattern in Spanish and English, which indicates that the difference of the patterns of use of elements between H and M are parallel in English and Spanish.



The inter-lingual difference becomes bigger in H than in M, implying that the two language differ in their formation of conceptual nuclei in terminology construction.

3.2. Head (H) / Head Elements in Modifier (M_H)

What was said in 3.1 can be said about the relation between H and M_H as well. The difference between H and M_H observed in V and N/V is, or becomes, smaller in Spanish; and H shows larger difference between the two languages compared to M_H.

3.3 Non-head Elements (M_M)/(M_H) Head Elements in Modifiers

The relations between M_H and M_M are similar to those between M_H and H. Particularly notable here is large V of M_M whose token number is small, which result in its low average frequency. The difference between M_H and M_M with respect to V becomes smaller in Spanish as terminology grows, while the difference with respect to N_V is generally bigger in Spanish. The difference between Spanish and English is smaller in M_H than in M_M with respect to V and also to N/V in terms of N.

4 Conclusions

The same patterns of differences between H and M, H and M_H and M_M and M_H were observed in English and Spanish. In general, the differences between these sets or distributions of elements are smaller in Spanish than in English. Inter-language distance is bigger in H than in M, and in M_M than in M_H. From the bottom two rows, we can also see that English uses more element types while Spanish tend to use the same elements more frequently, irrespective of the positions or set of elements. Our observations here were rather intuitive, and more rigid analyses and description, both statistically and with respect to the theoretical background of the study of terminology, will be carried out, following the preliminary result we reported here.

Notes

1. Elements in H have terminologically a 'special' status, which is related to the distinction of terminology construction and term formation (Sager 1990).
2. The problem related to this is complex, related to the model of term formation and terminological growth.
3. Though not obvious from here as we do not observe them here directly with respect to N, this difference is not attributed to the different ratio of token frequencies. This applies to most discussions below.

Acknowledgements

This work is a part of the project "A study on ubiquitous information systems for utilization of highly distributed information resources," funded by the Japan Society for the Promotion of Science. The first author would like to thank Dr. Harald Baayen of the Max Planck Institute for Psycholinguistics, whose programs we modified and used here.

References

Baayen, R. H. (to appear). Word frequency distributions. Dordrecht, Kluwer.

Good, I. J. and Toulmin, G. H. (1956). "The number of new species, and the increase in population coverage, when a sample is increased," Biometrika. 43(1), p. 45–63.

Kageura, K. et. al (1999). "A quantitative morphological structure of English and Spanish technical terminology," 6th Int'l Conference on Social Communication. p. 1277–1285.

Kotani, T. and Kori, A. (1990). Dictionary of technical terms. Tokyo, Kenkyusha.

Sager, J. C. (1990). A practical course in terminology processing. Amsterdam, John Benjamins.

A.A.Polikarpov & D.V.Khmelev

Basic Assumptions about a Sign's Life Cycle for Mathematical Modelling of Language System Evolution

1. Aim of the paper. Modelling of language evolution should be based on some assumptions concerning its micro-level, i.e. level of its micro-units' development. A sign (morphemic, lexemic and phraseologic) is an elementary (micro-) unit on some certain level of language organisation. A sign's polysemy evolution in time is the most fundamental ontological fact. That is why it has become the starting point in the building of the mathematical model for the development in time and synchronic correlation of the whole system of a sign's features. Correspondingly, it can lead to building a theory of the organisation and historical development of language systems as a whole.

2. Basic assumptions.

(1) A sign's polysemy development is a branching process of generating new meanings from previously acquired (and, correspondingly, losing some previously generated) ones.

(2) According to the increase of the ordinal number i of meaning's generation within a sign there should proportionally grow the average degree of meaning's abstractness A_i (or, in other words, decrease the average degree of meaning's filling by some number of semantic components B_i). This means that $A_i = 1/B_i$.

(3) The more abstract, on the average, the meanings of some generation of a sign are, the greater stability L_i (length of life) specific to each of them.

(4) The more abstract, on the average, each meaning of some generation of a sign, the lower generating activity G_i (number of meanings of the next generation produced from a meaning in its life) specific to each of them is.

(5) The more abstract meanings of some generation, the greater sense volume V_i (number of senses covered by each of them) that is specific, on the average, to each of them.

(6) The greater sense volume of meanings of some generation, the higher, on the average, the frequency of use U_i for each of them is.

These assumptions provide us with the ability to draw some useful conclusions for modelling of some other functional dependences for any language sign, as well as for ensembles of them, i.e. for a language system as a whole.

3. Dying out, giving birth, and preserving of sign's meanings.

Consequence 1. From the fact of a finite number of features in any sign's meaning it follows that maximal possible number of generations of meanings in a sign can not exceed some n.

Consequence 2. From assumptions (1)-(4) it follows that $L_1 \le L_2 \le \cdots \le L_n$ and $G_1 \ge G_2 \ge \ldots G_{n-1}$.

We shall consider evolution of a sign in continuous time. Let $\gamma_i = 1/L_i$ and $\beta_i = G_i/L_i$. Clearly, γ_i is a decay rate of meanings of generation i in a sign and β_i is a rate for generating new meanings (meanings of the next generation $i+1$) by each meaning of a generation i. Let us assume that during small intervals of time Δt every meaning of a generation i independently of all other sign meanings does the following:

1) dies with probability $\gamma_i \Delta t + o(\Delta t)$,

2) if $1 \le i \le n-1$ then it generates a meaning of the next generation $i+1$ with probability $\beta_i \Delta t + o(\Delta t)$.

Otherwise a meaning just preserves itself, continues its existence (with probability $1-(\gamma_i+\beta_i)\Delta t+o(\Delta t)$).

It is easy to prove that within the model activity of a meaning belonging to a generation i, i.e. average number of meanings of a generation $i+1$ produced by a meaning of a generation i, equals G_i.

4. Drawing a curve for a sign's polysemy dynamics. It is much more difficult to check a conclusion that a polysemy curve of a typical sign should have only one global maximum. Assumption (7) reduces the model to the branching process with the definite number of types of particles (see [1]). Denote [1] by P_σ^i the probability of the fact that a meaning of a generation i will generate for the time t meanings determined by the vector $\sigma = (\sigma_1, \ldots, \sigma_n)$: σ_1 meanings of a generation 1, ..., σ_n meanings of a generation n. Define generating functions

$$F^i(s) = \sum_{\sigma \ge 0} P_\sigma^i(t) s^\sigma,$$

where $s^\sigma = s_1^{\sigma_1} \times \cdots \times s_n^{\sigma_n}$. Also define a vector generating function $F(t,s) = (F^1(t,s), \cdots, F^n(t,s))^T$. It follows from [1, p.119, theorem 3] and from assumption (7) that $F(t,s)$ satisfies the system of equations

$$\partial F(t,s)/\partial t = f(F(t,s)) \tag{1}$$

with the initial conditions $F(0,s) = s$. Here $f(s) = (f^1(s), \ldots, f^n(s))$ where $f_i(s) = \gamma_i - (\gamma_i + \beta_i)s_i + \beta_i s_i s_{i+1}$ (we put $\beta_n = 0$). It is impossible to find an explicit solution of the system (1) for all initial conditions and all values of parameters. Nevertheless, behaviour of the average number of sign meanings $M(t)$ at the moment t is described by the system of linear differential equations. It is possible to obtain the following formula for $M(t) = L_1 p_1(t) + G_1 L_2 p_2(t) + \ldots + G_1 \ldots G_{n-1} L_n p_n(t)$ where $p_i(t)$ for $t \geq 0$ is a density for the sum of i exponentially distributed independent random variables of means $L_1, \ldots, L_i$.

Theorem 1. Under assumptions (1)–(7) we have only two qualitatively different kinds of behaviour for $M(t)$ when $t \geq 0$:

1. If $G_1 > 1$ then there exists a unique maximum at $t^* > 0$: $M(t^*) > M(t)$ for all $t \in [0, \infty] \setminus \{t^*\}$. Also $M'(t) \geq 0$ for all $t \in [0, t^*]$ and $M'(t) \leq 0$ for all $t \in [t^*, \infty]$.
2. If $G_1 < 1$ then $M'(t) \leq 0$ for all $t \geq 0$ and $M(t)$ reaches its global maximum at $t^* = 0$: $M(0) = 1$.

5. Regularities for the dynamics of a sign's sense volume, frequency of use, length, etc. These regularities are deduced on the basis of conclusions made earlier on the polysemy quantitative dynamics and some other "qualitative" assumptions (see "Basic assumptions").

6. Arriving at the conclusion on general shape of momentary word polysemy, frequency of use, length, etc. distributions in language. For deducing the general shape of momentary distribution in language for these features we assumed that some independent source generates signs for a language with some constant rate β_0. It is possible to show that in time this system arrives at some stationary state and to find its numerical characteristics.

Further details on this point, analytical deriving of other dependences, as well as presenting of some empirical data for testing the deduced form of polysemy distributions of lexemic signs (words) in various types of dictionaries of languages of various types — Russian, English, Chinese, Vietnamese, Mongolian, Hungarian, Estonian, Turkmen, Turkic, Tartar, Azerbaijan, etc. — will be made in the extended version of this paper.

References

Sevast'janov B.A. (1976) Vetvyashchiesya protsessy. (Russian) [Branching processes] Izdat. "Nauka", Moscow.

Dariusch Bagheri

Semantic Relations in the Lexicon and their Synergetic Modelling — Towards an Integration of Lexical and cognitive Relations.

The lexicon is considered as a list of items each of which is associated with a defintion representing the main aspect of its meaning. The combination of an item and its meaning is called a lexeme. The definition of a lexeme usually contains wordforms of other lexemes thereby forming the lexical relations between lexemes.

There are many different kinds of relations, e.g, hyperonomy-hyponymy, part-whole, cause-effect etc. This paper restricts its investigation to the well-known hyperonym respectively hyponomy relation. First an overview of the approach in the quantitative paradigm is given. Then the method of investigation commonly applied is discussed and criticized. The questions raised lead into the theoretical background. The distinction between 'logical' concepts, pschological concepts and lexemes is considered as crucial to give a satisfactory explanation for the way the (cognitive) hyperonym-hyponym relation is reflected in the lexicon. Finally an attempt is made to sketch a synergetic model of this relation and its parameters.

References

Altmann, Gabriel & Kind, B. (1983) "Ein semantisches Gesetz". Glottometrika 5, S. 1-13.

Hammerl, Rolf (1987) "Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen". Glottometrika 8. S. 113-129.

Hammerl, Rolf (1988) "Neue Modelltheoretische Untersuchungen im Zusammenhang mit dem Martingesetz der Abstraktionsebenen". Glottometrika 9. S. 105-120.

Hammerl, Rolf (1989) "Neue Perspektiven der sprachlichen Synergetik: Begriffsstrukturen - kognitive Gesetze". Glottometrika 10. S. 129-140.

Hammerl, Rolf (1991) "Methodologische und methodische Probleme der Erstellung von Definitionsfolgen und Lexemnetzen". In: Sambor, J. & Hammerl, R. (Hrsg.) Definitionsfolgen und Lexemnetze. Ldenscheid: RAM-Verlag.

Kay, Paul (1971) "Taxonomy and Semantic Contrast". In: Language, Vol. 4, Number 4: 866-887.

Martin, Robert (1974) "Syntaxe de la Dfinition lexicographique: tude quantitative des Dfinissants dans le 'Dictionnaire fondamental de la Langue Franaise'". In: David, J. & Martin, R. Statistique et Linguistique. Paris: Klincksieck. S. 61-71.

Rosch, Eleanor, Mervis, Carolyn B., Gray, Wayne D., Johnson, David M & Boyes-Braem, Penny (1976) "Basic Objects in Natural Categories". In: Cognitive Psychology 8: 382-439.

Sambor, Jadwiga (1991) "Definitionsfolgen und Suche nach semantischen Indefinibilien". In: Sambor, J. & Hammerl, R. (Hrsg.) Definitionsfolgen und Lexemnetze. Ldenscheid: RAM-Verlag.

Schierholz, Stefan J. (1991) Lexikologische Analysen zur Abstraktheit, Hufigkeit und Polysemie deutscher Substantive. Tbingen: Max Niemeyer Verlag. (Diss.)

Skorochod'ko, Eduard F. (1981) Semantische Relationen in der Lexik und in Texten. Bochum: Brockmeyer.

Smith, Edward E. & Medin, Douglas L. (1981) Categories and Concepts. Cambridge (Massachusetts): Harvard Univ. Press.

Reinhard Koehler & Gabriel Altmann
Probability Distributions of Syntactic Units and Properties

In (Köhler 1999), an attempt has been made to set up a basic functional-analytic model of a syntactic subsystem in the framework of synergetic linguistics. In this paper, functional dependencies among selected properties, viz. frequency, complexity, length, depth of embedding, and information, and the quantities polyfunctionality and synfunctionality are postulated, derived, and empirically tested on data of the Susanne corpus (Sampson 1995). The analysis of the probability distributions of the quantities under consideration was postponed and will be tackled in the present study. It will be shown that the properties of syntactic constructions are lawfully distributed according to only a few distributions which belong to a common family of probability distributions, and that hypotheses can be set up from which the corresponding distributions can be derived, thus explaining the empirical findings. The empirical database is extended by another language in form of the German Negra-Korpus (Brants 1999:102). The empirical tests yield results which are compatible with the hypotheses.

References

Brants, Thorsten (1999). Tagging and Parsing with Cascaded Markov Models. Automation of Corpus Annotation. Saarbrücken: Universität des Saarlands.

Khler, Reinhard (1999). Syntactic structures: Properties and Interrelations. In: Journal of Quantitative Linguistics 6/1, 46-57.

Sampson, G. (1995). English for the Computer. Oxford.

Peter Meyer
Qualifying Quantities Reflections on the Notion of Explanation in Quantitative Linguistics

According to a widespread conception, quantitative linguistics should in the long run explain empirical quantitative findings (such as Zipfs Law) by deriving them from highly general statistical 'laws' that are assumed to be part of a forthcoming general theory of human language (cf. Best (1999) for a summary of possible theoretical positions). This view is rejected here since it implicitly takes the 'qualitative' linguistic units and properties it deals with as unexplained explainers and since the quantitative notions employed in formulating the supposed explanatory laws are necessarily circular or even meaningless. An alternative position is sketched that takes the conceptual dependency of quantitative results on some prior 'qualitative' description into account.

1 Explanatory Schemes in Quantitative Linguistics: An Assessment

For the purposes of this paper I will briefly discuss the theoretical treatment of word length distribution in German texts as presented in Altmann/Best (1996). The fact that the negative binomial distribution can be fitted well to the empirical distribution of word length (measured as number of syllables) in a wide variety of German texts is explained by stipulating an underlying self-regulative mechanism that consists in mutual influence between neighboring word length classes. In accordance with the general framework proposed in Wimmer et al. (1994), it is assumed that this influence implies a proportionality relation between neighboring classes:

$$P_x = g(x)P_{x-1}, \tag{1}$$

where $g(x)$ is a proportionality function that, for the German texts in question, is assumed to be representable as

$$g(x) = \frac{a+bx}{cx}. \tag{2}$$

Here, a is taken to represent something like the length-invariant part of the German lexicon, whereas b is an author-specific modification factor (the author chooses to employ shorter or longer words, according to stylistic and other needs) and c is supposed to stand for the communicative interests of potential text recipients (such as minimizing the effort of decoding a given message).

Mathematically deriving an observed distribution from assumed underlying stochastic regularities does not, however, in itself constitute an explanation of the phenomena in question. In our sample case, equations (1) and (2) are just a mathematically equivalent reformulation of the original observation, viz. the negative binomial distribution. But even if derivability were unidirectional, we would have to find independent justification for the explanans in order to obtain explanatory value. No such justification seems available for the time being, however, except for some rather impressionistic translations of the supposedly explaining laws into intuitions about how languages, understood as systems, or their users are organized. The parameters a, b, c do not represent anything one could actually measure or provide an empirically testable interpretation for. The interpretations we endow them with are grounded on seeing what happens if we increase or decrease them and giving them a somehow fitting interpretation accordingly; or on correlating observed parameter values with general descriptions of text type, style etc. There is, obviously, no empirically based way of deciding which of several competing interpretations of the parameters is the right one, or which of several competing derivations of an observed stochastic phenomenon is to be preferred.

It is no solution to stick, in our explanations, to empirically interpretable parameters, as suggested in Meyer (1997), where word length distribution in some Eskimo texts was derived from directly interpretable assumptions about distribution of syllable numbers in morphemes and morpheme numbers in words. Though the explanans is empirically falsifiable in this case, it stands itself in need of an explanation. Thus, we still do not leave the realm of description in this case; we just enlarge it.

The only option that seems left to us is to stop searching for ever more fundamental laws as soon as a sufficient number of cross-linguistically valid statistical generalizations with some plausible though necessarily vague pre-theoretical interpretation has been found. Explanation would then reside in assuming that the general laws found form part of the underlying mechanisms of language itself, or are somehow wired in in the heads of language users.

Such an account may be accused of reductionism or even preformationism in a fashion comparable to explaining principles of Universal Grammar by simply equating them with some assumed though unknown genetically encoded faculty of the mind (cf. Stewart/Cohen 1997). It is implausible to assume that, for instance, some mechanism of least effort should actually be on duty in the heads of speakers. Rather, it makes sense to interpret Zipf's Law as some emergent property of a chaotic process involving many speakers, their macrophysical interactions and their internal constitution. Cf. Mandelbrot (1961), where the language of a monkey hammering randomly on the keys of a typewriter is shown to follow Zipf's Law. Generally speaking, it is false to assume that if certain stochastic observations are describable as a solution to an optimization problem then these observations must also be accounted for by stipulating an underlying mechanism that actually embodies the optimization strategy. See Mandelbrot (1983) for illustrating examples.

2 Reconsidering the Role of Qualitative Concepts in Quantitative Linguistics

I propose that a deeper reason for the dilemma sketched above can be found in the fact that standard explanatory approaches in Quantitative Linguistics simply take the existence of some system of linguistic units with certain postulated stochastic properties for granted. An alternative, though speculative, view is outlined below; the full version of the paper will discuss a specific sample problem, viz. segmentation of sentences into words.

1. Quantitative linguistics depends with conceptual necessity on the discrete classificatory notions of some sort of traditional qualitative linguistics (cf. Itkonen 1983). In measuring properties of linguistic units, we already presuppose a theoretical framework or at least an implicit justificatory practice that individuates and classifies these units and their properties in the first place by assigning them a certain role within empirical statements in a comprehensive linguistic description. This holds regardless of whether some of the descriptive notions implied are themselves fuzzy or vague.

2. Standard qualitative analyses of specific linguistic phenomena (e.g., segmentations of utterances into words and of words into morphemes) can be considered successful applications of normative notional schemes. Highly formalized examples of such schemes can be found in contemporary generative grammar (minimalist program, HPSG). Note that there might be different, mutually incompatible ways of applying a general descriptive scheme to a set of linguistic data, that is, different solutions to a general description problem. In addition, compliance with a descriptive scheme is certainly a matter of degree.

3. A central explanatory task of a general theory of human language is, then, to explain why certain description schemes have proven to be well applicable across very different languages, whereas other, theoretically possible ones, are not suited for linguistic descriptions. It seems plausible to assume that certain very general cognitive restrictions are responsible for this.

4. However, although grammatical analysis is nomologically dependent on the behavior and cognitive states of the language-using individuals, it is not descriptively reducible to (physicalistic or naturalistic) statements about individual or collective behavior and/or the internal constitution of individuals. For a discussion of some of the philosophical and science-theoretical assumptions alluded to here cf. Brandom (1994) and Stephan (1999).

5. An appropriate metaphor for a principled explanatory approach to qualitative linguistics might then roughly look as follows: The good applicability of certain descriptive schemes, as opposed to others, is an emergent consequence of chaotic invisible hand processes that are constrained by internal conditions of the individual symbol-processors in such a way as to produce attractors that can be mapped well onto certain highly general formal description schemes.

6. Whether some such explanatory account will actually be possible is itself an open empirical question. But if it were so, it could also be a sound basis for quantitative explanation provided that chaotic communicative processes with the right sort of attractors, when described within an admissible description scheme, indeed show just the stochastic regularities quantitative linguistics observes. This has been the very point of my conjecture, viz. that true explanation in quantitative and qualitative linguistics, respectively, cannot be achieved separately.

References

Altmann, G. and Best, K-H. (1996) Zur Länge der Wörter in deutschen Texten, Glottometrika 15, 166–180.
Best, K-H. (1999) Quantitative Linguistik: Entwicklung, Stand und Perspektive, Göttinger Beiträge zur Sprachwissenschaft 2, 7–23.
Brandom, R. B. (1994) Making It Explicit. Reasoning, Representing, and Discursive Commitment, Cambridge, Mass./London.
Cohen, J. and Stewart, I. (1997) Figments of Reality. The Evolution of the Curious Mind, Cambridge.
Itkonen, E. (1983) Causality in Linguistic Theory. A Critical Investigation into the Philosophical and Methodological Foundations of 'Non-Autonomous' Linguistics, London/Canberra.

Mandelbrot, B. B. (1961) On the Theory of Word Frequencies and on Related Markovian Models of Discourse in: Jakobson R. (ed.), Structures of Language and Its Mathematical Aspects, New York.

Mandelbrot, B. B. (1983) The Fractal Geometry of Nature, updated and augmented edition, New York.

Meyer, P. (1997) Word Length Distribution in Inuktitut Narratives: Empirical and Theoretical Findings, Journal of Quantitative Linguistics 4, 143–155.

Stephan, A. (1999) Emergenz. Von der Unvorhersagbarkeit zur Selbstorganisation, Dresden/München.

Wimmer, G., Köhler, R., Grotjahn, R. and Altmann, G. (1994) Towards a Theory of Word Length Distribution, Journal of Quantitative Linguistics 1, 98–106.

Mirjam Ernestus & Evert Wattel

Comparing lines which separate clusters in two-dimensional plots

The perceptual classification of alveolar stops in Dutch as [t] or [d] is strongly correlated to the closure and burst durations of the stops: the shorter the stop, the greater its probability to be classified as [t]. In addition, there is an influence of the height of the preceding vowel: stops following high vowels must be longer before they are classified as voiceless than stops following low vowels (Ernestus 2000).

There are several methods to determine the correlation between the classification of the stops and their closure and burst duration (e.g. discriminant analysis, CART). If we wish to determine the influence the height of the preceding vowel on this correlation, we must compare the correlation holding for stops following high vowels with the one holding for stops following low vowels. That is, we must compare the line separating [t]s and [d]s in a plot which represents the closure and burst duration of the stops and contains only stops following high vowels with the separation line in a same type of plot which contains only stops following low vowels.

There is no technique available which can compare separation lines in different plots. That is why we developed a method which separates clusters in two-dimensional plots and compares the separation lines in the different plots.

The calculation of the separation lines is based on the idea that every separation line is assiociated with an error, and that this error is smaller if the elements of the clusters are positioned on the correct side of the line and further from it. The optimal separation line is the line with the smallest error. This method provides us with optimal separation lines which separate clusters at least as well as Discriminant Analysis and CART (Breiman et al. 1984).

In order to compare lines in different plots, the characteristics of each line are computed as many times as there are elements in the plot, with every computation being based on all elements minus one, and every element being removed from the data set once (n-fold validation). The differences between the characteristics of all lines calculated in this way for one plot indicate the probability that the position and slope of the optimal line change when a new stop is incorparated into the plot, and therefore the standard error of the separation line. With the standard errors of the lines which have to be compared being available, the comparison of the lines is possible.

References

Breiman L., J.H. Friedman, R.A. Olshen & C.J. Stone (1984), Classification and regression trees. New York: Chapman & Hall.

Ernestus M.T.C.(2000), Voice Assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface. Utrecht: LOT.

Adam Pawlowski & Maciej Eder

Quantity or stress? Sequential analysis of Latin prosody

An overview of a number of encyclopaedias and academic handbooks in linguistics indicates that the most frequently quoted example of a language with the quantity-based prosody is Latin. This kind of prosody is commonly contrasted with the tonality systems found in Asian languages as well as with the dynamic expiratory accent (stress) typical of the most of Indo-European languages. The function of quantity in Latin becomes prominent especially in texts that comply with specifically Latin norms of versification. Acquaintance with Latin versification allows us to represent a text (nowadays inescapably written) as a sequence of metrical feet composed of long and short syllables.

Clear as the picture seems to be, it remains slightly ambiguous. One can't help asking why the vast majority of Indo-European languages (especially the Romance languages derived from Latin) exhibit stress in the absence of quantity. It is also difficult to accept the hypothesis that the prominent role that stress unquestionably played in vulgar Latin was due exclusively to foreign influences and had no foundation in some intrinsic features of classical Latin. Students of the phonetic development of preclassical Latin have noted that vowels in certain positions did not undergo any modification, which can be taken to indicate that they were dynamically stressed. As Safarewicz argues: ?Ten znany nam akcent aciski uwzgld-nia, jak wida, dwa czynniki: miejsce w wyrazie (poda na sylab przedostatni albo trzeci od koca) i iloczas sylab (bo wybr sylaby drugiej czy trzeciej od koca zalea od iloczasu sylaby przedostatniej)." (Safarewicz 1988, 521). The fact that classical Latin versification did not depend on word stress but only on the length of syllables is usually explained as a conse-quence of Latin's return at some stage to tonality (ibid.).

Based on the arguments given above, we put forward a hypothesis that classical Latin displayed quantity as well as dynamic accent and that both could determine the textual rhythm. Subsequently, we wanted to find out what the relationship was between these two kinds of linear arrangement. Taking for granted the universality and cultural versatility of quantity-based versification, one could expect that sequences of long and short syllables show a high degree of orderedness (especially in artistic texts from the classical period).

To test the hypothesis as outlined, the texts studied were segmented into, respectively, stressed / unstressed and long / short syllables. The quantification procedure consisted in assigning value 1 to stressed or long syllables and value 0 to short or unstressed ones. It was also assumed, as it is generally the case, that Latin words were stressed paroxitonically if the penultimate syllable was long and proparoxitonically if it was short. In this way, we produced parallel representations of Latin texts which were subsequently subjected to sequential statis-tical treatment. This allowed us to precisely determine the degree of the linear arrangement of syllables in each case and compare the corresponding representations of samples.

The quantification is exemplified by the treatment of the following distich (Verg. Aen. IX 93-94):

> Filius huic contra, torquet qui sidera mundi:
> O genetrix , quo fata vocas, aut quid petis istis?

which can be represented as a sequence of metrical feet:

```
d s s s d s

d d s s d c
```

or long and short syllables:

```
-   - - - - - - -   - -

-   -   - - - - -   -
```

or else stressed and unstressed syllables:

```
- - - - - - - - - - - - - -

- - - - - - - - - - - - - - -
```

Linear models of autoregression and/or moving-average stochastic processes, based on the procedures of the ARIMA method, were estimated for the numerical series under research. The percentage of the original variance explained by the estimated models for the correspon-ding samples was then compared (the greater this value, the more regular the arrangement of the series). So far, this method has proved to be an efficient tool of description and analysis of textual rhythm. It allowed the distinction of so called functional styles (Pawlowski 1997), types of verse (ibid.) and accentual systems (Pawlowski 2000).

The result of our research was quite surprising. The percentage of the original variance explained by the estimated models turned out to be significantly greater for the series coded with regard to stress than for the series coded with regard to quantity. Translated into linguis-tic terms and generalised, this result means that Latin texts read in a 'contemporary manner', thus without quantity but with stress, should sound much more rhythmical than they do when we include in reading quantity or both quantity and stress together. One concludes that stress, based on the so called accentual isochrony, played in Latin a similar role as in the contempo-rary Indo-European languages.

References:

Box G., Jenkins G. (1976), Time series analysis: forecasting and control. San Francisco etc.: Holden-Day.

Crystal D. (1997), The Cambridge Encyclopdia of Language. Cambridge: CUP.

Gottman J. (1984), Time-series analysis: a comprehensive introduction for social scientists. Cambridge etc.: Cambridge University Press.

Pawlowski A. (1997), Time-Series Analysis in Linguistics. Application of the ARIMA Method to Some Cases of Spoken Polish. In: Journal of Quantitative Linguistics 4/1-3, 1997, 203-221.

Pawlowski A. (2000), Analyse quantitative compare de la prosodie des langues á accent fixe et á accent libre. In: M. Rajman, J.C. Chappelier (red.), JADT 2000, Actes des 5es journées internationales d'analyse statistique des données textuelles. Lausanne: EPFL, 2000, 531-534.

Safarewicz J. (1988), Jêzyki italskie (Eng. Italic languages). In: Bednarczuk L. (ed.) (1988), Jêzyki indoeuropejskie (Indo-European languages). Warszawa: PWN.

Whiteley P. (1980), Time Series Analysis. In: Quality and Quantity 14, 225-247.

Svitlana Budzhak-Jones

Fine Tuning Statistical Procedures in Language-Contact Research: Evidence from Ukrainian-English Bilingualism

Recently, variationists have successfully applied statistical procedures in language-contact research to disambiguate the language membership of items occurring in bilingual discourse (e.g., Poplack & Meechan 1995). We argue, however, that while quantitative methods are an indispensable tool for determining the status of lone items, they must be used with greater discretion in order to produce further results.

According to the principle of accountable reporting, one of the postulates of variationist studies, the researcher should incorporate all instances of a studied phenomenon in one analysis, irrespective of whether they occurred or could have occurred (Labov 1966). Code-switching and borrowing, however, pose a major problem for this requirement since they often result in identical surface realizations. By combining all other-language tokens together, a researcher may be analyzing a mixed corpus containing two different phenomena. Moreover, if tokens representing one linguistic phenomenon considerably outnumber the tokens of the other phenomenon in the same corpus, then the statistical procedure applied to such a mixed corpus will mask the effects of the lesser represented phenomenon. This may result in a less accurate determination of the constituents in a mixed corpus. In order to overcome this problem we suggest that the researcher should carefully consider all indications of language membership (either linguistic or extra-linguistic) and separate mixed tokens into two groups: those which are most likely to be code-switched versus those which are most likely to be borrowed. Only after such a detailed triage of the mixed tokens and no other indications of language membership, should quantitative methods be applied.

1. Pisla do duze, duze takoji exclusivnoji highschool v Filadelfiji.
   Went-F to very very such-F.Gen exclusive-F.Gen highschool- in Philadelphia-F.Loc
   [She] went to such a very, very exclusive highschool in Philadelphia (20/173)

For this project we examined all other-language incorporations extracted from 36 hours of natural performance data of 25 Ukrainian-English bilinguals. All mixed tokens (like, e.g., exclusivnoji and highschool in 1) were analyzed with respect to morphological marking, agreement, syntactic position and flagging relevant to each part of speech in both languages in contact. These results were then systematically compared to their counterparts in each of the unmixed corpora. In doing so we tested the following hypothesis: If the patterns established by English-origin nouns are similar to those established by unmixed Ukrainian counterparts, they are produced by the same grammar, and hence are borrowed, irrespective of the nouns etymology. If English-origin nouns ascertain the same patterns as their English counterparts, while at the same time differing from those of their Ukrainian counterparts, they are produced by English, and hence, are code-switched.

Quantitative analysis of all mixed forms considered together did not reveal unambiguous distributional and variable patterns in favor of either language. When separated into most likely candidates for code-switching and most likely candidates for borrowing status, these groups produced distinct patterns in each part of speech. English-origin tokens with overt Ukrainian morphology irrespective of the part of speech reflected the distribution and conditioning of the inflected Ukrainian counterparts. The results of bare forms depended on the part of speech. Uninflected verbs, participles and adjectives which usually disallow null-marking in Ukrainian, patterned similar to their English counterparts. Uninflected nouns and adverbs showed mixed results, since both of them allow null-marking in Ukrainian as well as English. Hence, quantitative methods can and should be used in language-contact research, although with greater discretion.

References:

Labov, W. 1966. The Social Stratification of English in New York City. Washington, D.C.: Center for Applied Linguistics.

Poplack, S. and M. Meechan. 1995. Patterns of language mixture: Nominal structure in Wolof-French and Fongbe-French bilingual discourse. In L. Milroy & P. Muysken (eds.), One Speaker, Two Languages. Cambridge, UK: Cambridge University Press. 199-232.

Mark Kaunisto

Relations and Proportions in the Formation of Blend Words

The paper I propose outlines earlier studies on blending as a word-formational process, and introduces my current work on the topic, including a theory on the structuring of blend words. Blending has attracted lexicographers' interest in the 20th century a great deal, which can be partly explained by the fact that the use of blending in forming new words seems to be increasing. Blending, or contamination or amalgamation (sometimes used in slightly differing senses), as the process has also been called, is a phenomenon in which two words are combined so as to form a new word, joining together orthographic, phonemic items of both source words, e.g. brunch from breakfast and lunch. The more intricate ways of forming blend words, the practicalities in blending, have also been a topic of discussion among linguists. The usual comment on blending, however, is that it follows no clear-cut rules (e.g. Bauer 1983:235).

Recently the question has also been looked into from the viewpoint of Optimality Theory (especially with Spanish blendings, e.g. by Piñeros (1998)). Bergström (1906) pointed out in his thesis the possibility of investigating "the quantity of the contribution of each element in each different case" (p. 46). Unfortunately he does not specify exactly what he had in mind. It is evident that he did not think much of this kind of approach, saying that he preferred not to consider the structuring elements of blend words "as a relation or proportion almost in a mathematical sense and way" (p. 16), later adding that based on his material he did not see enough reason to present any such rules (p. 46). He does, however, present a table of blend words according to the stress patterns of the source words and the resulting blend word. But the question of mathematical relations and proportions, mentioned by Bergström but not commented on since by linguists, is, in fact, one that seems to deserve further investigation, especially in the light of some other relevant aspects at the stage of forming new blend words. Considering the structure of blends, one starting point could be the question of what is a well-formed blend word. It may be argued that the deletion of any items from the source words presents a certain amount of 'danger' or 'threat' as to the understandability of the final blend word. Ideal blends then would naturally be ones where the ending of the first source word and the beginning of the second one overlap, resulting in a way in no deletion at all. Examples of this kind of 'ideal' blending (Cannon (1987:144) calls this the "traditional" kind of blending) include shamateur, sexploitation, slanguage, filmania, netiquette, and palimony. But this type of blending covers only a fraction of all blend words, as there are also blends where elements of either one or both of the source words has undergone deletion. In fanzine (from fan and magazine), magazine has had elements deleted from it, whereas fan has remained intact. In tangemon (from tangerine and lemon), on the other hand, both source words have been shortened. One possible criterion in blending that seems justifiable from a cognitive point of view is the tendency of language towards economy. It can be argued that economy in language generally results in shorter forms, but it could also be seen as having an effect on where and how deletion actually takes place. In blending two words together one might assume that there would be a natural tendency to preserve as much from the shorter source word as possible, and thus to minimize the loss of information on the source form that would be under a greater 'threat'. It could be proposed then that as a result of such a tendency, the relation between the part of the shorter form represented in the final word and the entire source word is greater than (or equals) that between the part of the longer word in the final word and the entire longer source word, i.e. a greater percentage of the originally shorter source word is retained in the blend.

One question arising from this theorizing, of course, is what do we exactly refer to by 'length' - orthographical or phonological length? Although it has been argued that letters do not have a significant weight in word-formation in general, the argument here could be tested with the numbers of orthographical units. In the case of phonemes, we would first have to solve a number of problems dealing with questions such as what counts as one phoneme (problematic cases being affricates and diphthongs) and what to do with complexities caused by regional differences in pronunciation in such calculations. If we consider then that two words, X and Y, are blended to form a third word, Z, and that X is represented in Z by A (a part of X, that is) and Y is represented in Z by B, we can present the hypothesis in the following axiom:

if $x > y$, then $a : x < b : y$

where

x = the number of letters/phonemes in X
y = the number of letters/phonemes in Y
a = the number of letters/phonemes in A
b = the number of letters/phonemes in B

Thus for brunch we could present the following:

X = breakfast, Y = lunch;
A = br, B = unch;
x (breakfast) = 9, y (lunch) = 5,
$a = 2, b = 4$;

in which case the final blend does conform with the hypothesized axiom, as $a : x = 0.22$ is smaller than $b : y = 0.8$.

For tangemon, if X is tangerine and Y is lemon, the corresponding breakdown would be $x = 9, y = 5, a = 5, b = 4$; and again the axiom holds, as $a : x = 0.56$ is smaller than $b : y = 0.8$. A notable thing here is the fact that the e in the middle of the blend is counted twice, due to the overlap of the source words in that element.

Other blend words that conform with this axiom include, for example, motel, dumbfound, smog, smaze, kissagram, plumcot, heliport, wargasm, and Eurasia (plus the ideal blends with which $a : x = b : y = 1$) There are some blends, however, that seem to go against this theory, e.g. Oxbridge, infotainment, and zebrule. Some blends are problematic to analyse, as some endings may already be regarded as suffixes, e.g. -zine in fanzine, videozine, and letterzine, making it also difficult to collect a good, representative pool of blend words for in-depth examination from the viewpoint of this theory.

It appears that the tendency suggested might only be one possible factor explaining why some blendings have been formed as they have. But the extent to which such words seem to conform with the idea is noteworthy, and adds to our knowledge on the structuring of blend words. In addition, the observations made have other possible implications: for example, as regards the question of the well-formedness of blend words, the conformity with the proposed theory might in future be looked into from the viewpoint of the words' lifespan, i.e. it could be argued that the incoherence of blendings and the theory might be one reason why they have become short-lived, a characteristic which is rather common among blend words.

Johan Carlberger & Viggo Kann

Some applications of a statistical tagger for Swedish

We will briefly describe a part-of-speech (POS) tagger for Swedish and discuss some applications: rule-based and probabilistic grammar checking, word prediction and keyword extraction.

In POS tagging of a text, each word and punctuation mark in the text is assigned a morphosyntactic tag. We have designed and implemented a tagger based on a second order Hidden Markov Model (Charniak et al., 1993). Given a sequence of words $w_{1..n}$, the model finds the most probable sequence of tags $t_{1..n}$ according to the equation:

$$T(w_{1..n}) = \arg\max_{t_{1..n}} \prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1}) P(w_i|t_i). \quad (1)$$

Estimations of the two probabilities in this equation are based on the interpolation of relative counts of sequences of 1, 2 and 3 tags and word-tag pairs extracted from a large tagged corpus.

For unknown words, we use a statistical morphological analysis adequate for Swedish and other moderately inflecting languages. This analysis is based on relative counts of observed tags for word types ending with the same 1 to 5 letters. This captures both inflections (tense *-ade* in *hmtade* (fetched)) and derivations (nounification *-ning* in *hmtning* (pick-up)).

We also perform an analysis that finds the last word form of compounds, which are common in Swedish. The possible tags of the last word form indicate possible tags (and probability estimation) for an unknown compound word. These two analyses are heuristically combined to get estimations of $P(w_i|t_i)$, which enables unknown words to work in the model. This method combines morphological information for unknown words with contextual information of surrounding words, and resulted in a tagger that tags 98 % of known and 93 % of unknown words correctly (Carlberger and Kann, 1999).

Grammar checking

The tagger was developed to be a part of a program for grammar checking of Swedish text. The text is first tagged, and then checked for grammatical errors by so called error rules. An agreement error detecting rule could for example state that if a determiner is followed by any number of adjectives and then a noun where the noun disagrees in gender, then the determiner should be corrected, and the correct inflection should be suggested. The grammar checker is efficient and has good recall and precision (Domeij et al., 2000). At `http://www.nada.kth.se/theory/projects/granska/` there is a web version of the tagger and grammar checker.

By letting the grammar checker use a syntactically disambiguated instead of an undisambiguated text, error rule writing becomes easier and tests showed an increase in both recall and precision. The drawback is when incorrect tagging causes false alarms, but such can be avoided to a great extent by writing "tagging correction rules" and by other means.

It is inefficient to try each error rule at each position in the text. We therefore perform a statistical optimization, where each rule is analyzed in advance. For each position in the rule the possible matching words and taggings (tag pairs) are computed. Then, using statistics on word and tag pair relative frequencies, the position of the rule that is least probable to match a Swedish text is determined. This means that this rule is checked by the matcher *only* at the positions in the text where the words or tag pairs of this least probable position in the rule occur.

Probabilistic grammar and spell checking

Using error rules in grammar checking makes it hard to cover all possible combinations of ungrammatical constructions. This fact led us to investigate a method for finding errors with a probabilistic approach.

A suspicious sentence is identified using some measure related to the probability given by Equation (1). Changes are made to the sentence in order to make it less suspicious, and if a good enough alternative sentence is found, it is suggested as an alternative to the original sentence. This method may capture not only errors that spell checkers findm, but also errors where the misspelled word happens to be a word in

them word list, or when adjacent are words interchanged, or when a is omitted, or when a an extra word has slipped into the sentence by mistake.

The simple algorithm we devised successfully identified and corrected errors of these types in very short sentences, but whether this method can successfully treat more complicated sentences remains to be investigated. Anyhow, the approach seems promising, at least for identifying suspicious sentences.

Keyword extraction

A problem with indexing of documents and keyword extraction is that words appear in different forms and as parts of compound words. In a typical web search engine, if one uses the noun *lagar* (laws) as a keyword, one will not find documents containing the word *lag* (law) or *grundlagar* (constitutional laws). A solution to this problem could be to index documents with the base forms of occurring words and use the base forms of keywords as actual keywords.

This approach will improve coverage, but precision can be further improved. Not knowing the syntactic function of an occurrence of *lagar* in a text, the possible base forms are *lag* (law) as well as the verb *laga* (mend). To improve precision, a better solution is to tag the texts and only use base forms implied by the tagging.

With the compound analysis algorithm included in our tagger, compound words can also be mapped to their correct base forms as well as to the base forms of their constituents. For example, *grundlagar* is mapped to *grund*, *lag* and *grundlag*.

Word prediction

Word prediction is the problem of guessing which word will appear after a given sequence of words. For example, the word following *The boy will not* is more likely *be* than *been*. A word prediction algorithm ranks the words in a lexicon according to an estimated probability of each word to appear after a given sequence of words.

As the tagging algorithm estimates probabilities of sequences of words, it can be used as for word prediction. Lexicon words are ranked by the probabilities given by tagging the given sequence of words with each lexicon word inserted at the end. However, this is an inefficient solution, since there are as many different sequences to examine as there are lexicon words.

In most applications, only the most probable words are of interest, and there can be restrictions on what words are selectable. A more efficient approach would therefore be to only investigate words that are among the most probable ones. Our solution involves two simple steps: First, categorize words according to restriction criterion and possible tag. Second, select the most common in each category and group these words by restriction criterion, and use only selected words in the algorithm. This solution effectively reduces the number of words to consider. Since $P(w|t)$ is the only term of Equation (1) that depends on w, this method gives (almost) the same result as the naive approach.

The algorithm was improved by including $P(w_i|w_{i-1})$ in the tagging equation, by promoting previously occurring words in the text and by adjusting the probability estimations for unknown words for more suitable rank among known words. An implementation used for speeding up typing showed a saving of almost 50 % of keystrokes.

References

Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993). Equations for part-of-speech tagging. In 11th national Conf. Aritificial Intelligence, pages 784–789.

Carlberger, J. & Kann, V. (1999). Implementing an Efficient Part-Of-Speech Tagger. In Softw. Pract. Exper. 29(9), pages 815–832.

Domeij, R., Knutsson, O., Carlberger, J., & Kann, V. (2000). Granska – and efficient hybrid system for Swedish grammar checking. To appear in Proc. Nodalida-99.

Akira Ushioda:
Word Clustering and Part-Of-Speech Tagging

This paper presents automatic construction of word clusters from plain texts and application of the clusters to corpus-based natural language processing for the purpose of alleviating the data sparseness problem.

Plain texts of 50 million words are extracted from Wall Street Journal articles and used to construct word clusters with mutual information clustering algorithm (Brown et al. 1992, Ushioda 1996). The vocabulary to be clustered is the set of 70,000 most frequently occurring words in the WSJ texts and the number of classes is 500. A series of experiments is conducted to examine the usefulness of the clusters obtained for improving Part-Of-Speech (POS) tagging accuracy of Hidden Markov Model (HMM)-based POS taggers. The base form of the HMM tagger is constructed following Church's trigram-based POS tagger (Church 1988). In this model the joint probability of the tag sequence $T = t_1 t_2 ... t_n$ occurring with a given word sequence $W = w_1 w_2 ... w_n$ is given by:

$$Pr(T|W)Pr(W) = Pr(t_1)Pr(t_2|t_1)\prod_{i=3}^{n} Pr(t_i|t_{i-1}, t_{i-2})\prod_{j=1}^{n} Pr(w_j|t_j)$$

In this formula, $Pr(t_i|t_{i-1}, t_{i-2})$ is a state transition probability which, in this model, refers to a probability that a word with tag t_i appears after a tag sequence of $t_{i-1}t_{i-2}$ is observed. $Pr(w_j|t_j)$ is an emission probability which refers to a probability that a word with tag t_j turns out to be w_j. With this model, the tagging accuracy of unknown words is much smaller than the total tagging accuracy because the emission probability $Pr(w_j|t_j)$ for unknown words cannot be estimated from the training set.

We then implemented the HMM tagging model which incorporates class information for tagging unknown words. In this model, class emission probabilities $Pr(C_j|t_j)$ are used instead of $Pr(w_j|t_j)$ when the word is unknown.

Figure 0.1 shows the POS Tagging accuracy of unknown words with different clusters and with varied training text size. The x-axis is the average mutual information (AMI) of the entire WSJ POS-tagged corpus with respect to the clusters, and the y-axis is the tagging accuracy of unknown words. Beselines show tagging accuracy without using class information. Tagging accuracy of unknown words is considerably increased by using class information, and the better the cluster (with respect to AMI) we use, the better the result we obtain. For example, in the case of T4 (245,000 words training data) with CL7, the accuracy is increased from 43.5 % to 77.8 % with class information. In terms of tagging error rate, the error rate is less than half (56.5 → 22.2: 39 %). Figure 0.2 shows the overall tagging error rate with varied training text size.

Class-based HMM taggers are also compared with other major POS tagging methods in the case of small training data (1000 sentences). Table 1 compares the class-based HMM tagging model (HMM-CLASS) with Weischedel's (1993) HMM-based tagger (HMM-MORPH), Brill's (1993) transformation-based tagger and Ratnaparkhi's (1998) Maximun Entropy tagger (ME). Although ME performs slightly better than HMM-CLASS for unknown words, HMM-CLASS performs best in total word accuracy.

	Unknown Word Accuracy	Total Word Accuracy
(a) HMM	49.74	86.55
(b) HMM-MORPH	71.70	91.00
(c) Transformation-Based	79.52	90.82
(d) Maxmum Entropy (ME)	80.54	91.67
(e) HMM-CLASS	79.01	91.98
(f) HMM-CLASS + DTree	83.48	92.80
(g) HMM-CLASS + ME	85.54	93.17

Table 1: Comparison of POS Tagging Accuracies

One distinctive feature of HMM-CLASS is that it does not use morphological (spelling) features of words at all. In all the other models in the table, morphological features of words play a central role for tagging

unknown words. In order to incorporate morphological features into HMM-CLASS, the original training set is split into 90 % training data and 10 % held-out data, and HMM-CLASS tagging with 10-fold cross-validation is conducted. The parameters tuned in the cross-validation are those of a learner which learns error patterns of tagging unknown words using HMM-CLASS. Morphological features of unknown words, as well as the tagging results of the base tagger (HMM-CLASS), are used as features of the learner. We tested two types of learner, a decision tree (HMM-CLASS+DTree) and a Maximum Entropy model (HMM-CLASS+ME). HMM-CLASS+ME performs best and its accuracy for unknown words is by 6.5 percentage points higher than HMM-CLASS. In terms of an error rate of unknown words, the error of HMM-CLASS+ME is 25.7 % smaller than that of ME, and 29.4 % smaller than that of Brill's transformation-based tagger. This clearly shows that a considerable benefit is obtained by combining distributional characteristics of words (expressed in the form of clusters) and morphological features of words.

References

Brill, E. (1993) A Corpus-Based Approach To Language Learning. Doctoral dissertation, University of Pennsylvania.

Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R. (1992) "Class-Based n-gram Models of Natural Language". *Computational Linguistics*, Vol. 18, No 4, pp. 467–479.

Ratnaparkhi, A. (1998) *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Doctoral dissertation, University of Pennsylvania.

Ushioda, A. (1996) "Hierarchical Clustering of Words". *Proceedings of The 16th International Conference on Computational Linguistics*.

Ludmila Uhlírová

Title: On the so-called language modelling in automatic speech recognition

The topic of the paper belongs to the multidisciplinary field of man-machine communication, in which

1. the computer is not (only) the medium, but the other party in the dialogue;
2. the dialogue is spoken; therefore, the computer must recognise speech (voice), and synthesise (produce) a spoken answer.

To recognise speech, firstly it is necessary to segment a continuous acoustic signal (more exactly: spectral information) into discrete parts (segments, units, chunks, ...), and to decode (identify, label, ...) them in phonetic terms. During the last three decades, several mathematical methods of the acoustic signal analysis have been developed. One of them, widely used over the world and corroborated also by Czech research teams, is based on the so-called Hidden Markov Models. HMM is a statistical recogniser which uses an explicit mathematical model of speech to calculate the probability that a speech signal corresponds to a particular sequence of words. In principle, the method of HMM is applied both to the recognition of isolated words and continuous speech. However, if the task is to recognise continuous speech, additional supporting information is searched for - information about combinatoric rules of words. The combinatorics of words (or, sometimes also of otherwise defined units) is considered to be a linguistic feature of speech (in contrast to purely acoustic, phonetic ones), and the model of combinatoric properties of word n-tuples is called language model. The term language model has been coined in speech recognition literature just in this sense, and it is also in this sense that it is used in the present paper. The crucial question of language modelling, to put it very briefly, is the following: What is the probability that a word which is expected to occur with a certain probability in a spoken text will be followed by a concrete word or by a sequence of n words? (So far $n = 1, 2, 3$, but rarely > 3.) In language models, probabilities of all possible combinations, usually called n-grams, are estimated, i. e. they are approximated by relative frequencies from a big corpus of training texts. The problem is that with the increasing size of the training corpus the number of possible n-grams increases astronomically and, therefore, it is difficult, even impossible, to seize it in a reasonable way; moreover, as many researchers report, the experience made so far has shown that the vast majority of theoretically possible word sequences do not occur in any tested text at all, whereas it is quite probable that a word sequence which does not occur in the training corpus still will be attested, and then it must be handled in a special way. In connection with the problem of the increasing dictionary size two questions arise: (1) How to reduce a huge number of n-grams and how to make modelling more effective? (2) Recent results reported in literature have shown that speech recognition procedures are successful and their further elaboration promising if the analysed language is English. However, what will happen if the same procedures are applied to data from a language with rich morphology and free word order?

Quantitative linguistics is expected to offer linguistic arguments which might help to further develop mathematical procedures and to apply them to data from various languages. In the present paper, we deal with advantages and limitations of (a) adding more abstract information, namely part-of-speech tags, into the training dictionary of words, and (b) adding information about subword phonetic units - syllables. The analysed language is Czech.

References

Jelinek, F. (1999). Statistical methods for speech recognition. IMT Press, Cambridge, Mass.

Psutka, J. (1995): Komunikace s potaem mlkuvenou e. Academia, Praha.

Hanl, V. (1999). Theory of structured cogitation in speech recognition. Eurospeech 99, Budapest, September 5-9. Conference proceedings on CD-ROM.

Brookes, M. (2000). Speech processing.

`Http://www.ee.ic.ac.uk/hp/staff/dmb/courses/speech/speech.htm`

V.A. Dolinsky and D. Rainova
Experimental study of semantics of a word.

An object of quantitative analysis of the data of psycholinguistic experiment isto elicit from native speakers' memory the knowledge of semantics of lexical units as of how they are presented in psychological space of speakers before and irrespectiveof introspection of a linguist. E. Sapir noted that "associative connection is the essence of a language". F. de Saussure pointed to the distinctive features ofassociative relations in language such as the uncertainty of the order and quantitative openness. The quantitative research of semantic fields discovers a specific picture of the world inherent in people as a language collective. This picture reveals that way by which each language dismembers reality, calling and attributing values to such reality. "Each language draws its own boundary-lines in the amorphous mass ofthoughts, arranges the centres of gravity in different places and lays various emphasises on them" (L. Hjelmslev).Words, being units of a language system, are brought to conformity with the meanings - symbolic correlates in extralinguistic reality. Various meanings or lexico-semantic variants of a word form corresponding, inherent in them only, links of probabilistic nature, a cognitive map of which is located on different planesprojecting semantic reality.

The research of the cognitive structure of language is based on a fragment oftwo experimental databases received from the answers of testees (students of the Moscow Institutes of Higher Education and native Russian speakers) on the word-stimulus BELYI (white).

Experiment 1 (directed).

Testees were offered to fill in a word-buildingmatrix (motivation family of words) with the basis BEL- by words known to them (nouns, adjectives and adverbs). The number of testees was 210. It was required toexpand the matrix by all units of internal lexicon known to them, motivated by an internal form of the word-stimulus. The questionnaires included only a scheme ofsemantic-morphological motivation graph with basic exhibits (20 units). Testees expanded the family of words up to 108 units.

1.	pro -	bel		8.		bel	- i - la
2.		bel	- ok	9.		bel	- izn - a
3.		bel	- j - o		...		
4.	na -	bel	- o	105.	zhemchuzhno -	bel	- yi
5.	do -	bel	- a	106.	serebristo -	bel	- yi
6.		bel	- j - mo	107.	cherno -	bel	- yi
7.	po -	bel	- k - a	108.	snezhno -	bel	- yi

Experiment 2 (free).

Testees were offered to give the first and sole response tothe word-stimulus BELYI. The number of testees was 1010. The received responses were grouped in two tables. One of them separately classified the syntagmaticresponses (nouns) as follows:

151	sneg	16	klyk	8	platok	5	kon'
47	tsvet	14	sharf	7	lebed'	5	krolik
23	svet	12	aist	7	medved'	5	parus
21	dom	12	prostynya	6	pudel'	5	pustota
18	list	11	shar	6	khalat	5	smert'
17	flag	10	zayats	6	khleb	5	stena ...

The paradigmatic responses (adjectives with semantics of colour signs) are separately presented as follows:

175	chernyi	7	ceryi	1	oranzhevyi
18	krasnyi	1	golyboyi	1	fioletovyi
9	siniyi	1	zheltyi	...	
7	zelenyi	1	lilovyi		

The data of the free association experiment and directed experiment for fillingin word-building matrixes (motivation families of words) can be compared. Despite of differences connected with various levels of an internal lexicon presented in eachof the experiments, in both cases we deal with the projections in the sole word of extralinguistic meanings having an ontological reality of the semantic field, andleaving traces on all levels of the language consciousness of speakers.

As a result of the quantitative analysis of the materials of both experiments,we succeeded in finding dictionary lexico-semantic variants of a word, which were not psyhologically adequate to the consciousness of native speakers of modernRussian, but also in revealing a meaning of the lexeme BELYI (white), not presented in modern explanatory dictionaries as follows:

A. clean, pure;
B. empty;
C. symbolical, spiritual; those relating to the life-death archetype.

A. "chistyi". Experiment 1: belovoy, nabelo, po-belomu. Experiment 2: chistota,chistyi, gryaznyil, nevinnost', zapachkat', postel'.
B. "pustoy". Experiment 1: probel. Experiment 2: pustota, pustoy, bestsvetny, nikto,pustynya.
C. "simvolichesky, dukhovny; those relating to the life-death archetype". Experiment1: belokroviye, bel'mo, belena, belok. Experiment 2: smert', mertvets, bol'nitsa, bolezn', krug, kvadrat, simvol, savan, nebo, ogon', traur, Bog, chyort, shaman,grustny, angel, mag, chudny, radost', sovershenstvo, tishina, kholod.

References

Dolinsky V.A'., Rainova D. (1998). Experimental Study of Verbal Associations basedon the Directed Experiment. Methodology of Mathematical Modelling. VI.

Sofia. Nalimov V.V. (2000). Scattering ideas. In a way and at the cross-roads. "Progress -tradition", Moscow.

Omar Larouk

Using presupposition logic in the recognition of the implicit information of the user in Information Retrieval System

The project deals with automatic text processing and computer aided information search. The Documentation Information System (DIS) is based on a linguistic model for the recognition of a written text to determine the reference function (Noun Phrase). This paper describes the effectiveness of other approaches to ease access at the user/system interface. We propose to "optimise" Information search through text processing and query formulation in natural language using the contribution of "presupposition logic". We are looking at implicit textual information in the question of man/machine interfaces (aspects of the conversation). The SII group has designed a model for the processing of information stock documents based on a logico-semantic approach for knowledge representation and information retrieval.

1. Introduction

The aim of an IRS (Information Retrieval System) is to facilitate access to users who do not know what type of information they are looking for. This information can be found in a Knowledge-Based System (KBS : Database, Data Banks, etc...) wherein lies the importance between indexing and interrogation. The aim of the SII group is to design an automatic Documentation Information System (DIS) based on the information stock documents to be processed before indexing and the user queries [Salton and Mc Gill,1]. The documents are processed by indexing in order to correlate the information content of these documents. This content is necessary to determine "descriptors". The users queries are often combinations of descriptors together with connector logic. A classical documentary system compares the query descriptors with the document descriptors in order to retrieve the required documents [Croft and al., 8].

Documentary automization is blocked by the problem of indexing and interrogation due to the constant updating of the input information which remains static once in the data base. Therefore the idea was put forward to use natural language as textual data for automatic indexing because language integrates the implicit information in the questions.

Texts contain referential descriptions and the demand for information concerns these references. To summarize , information search in a knowledge based system can have the following double structuring mechanism : this structuring can be carried out using "full" information such as (Noun Phrases, sentences) [Sparck Jones, 10] or properties extracted from the text in the form of simple or complex predicates.

We propose to "optimise" Information search through text processing and query formulation in natural language using the contribution of "presupposition logic". We are looking at implicit textual information in the question in man/machine interfaces (aspects of the conversation).

2. Perspectives of this approach

The representation of a document should form an integral part of the stock information documents. On interrogation, the document must be retrieved if it is relevant. In the same way, the modelling of the stock information documents should be structured in the form of a KBS taking into account the formulation of the query in order to use the presupposition information.

One of the problems of the Man/Machine interfaces (aspects of the conversation) is the following :

The user uses natural language, whereas the system uses formal language. With this in mind it is necessary to express what is implicit. The opposition presupposed/posed is of use for optimising the Man/Information System interface by calculating the hidden semantic information.

The logico-semantic approach contributes in the following way :

- Indexing must be designed in such a way so as the documents can be found in the corpus, while reducing the noise. Silence can be considered as unacceptable whereas a little noise is possible. From which came the idea of a system, which would extract all the NP of a text, so reducing the silence. The contribution of the predicate approach in an interrogation allows the dichotomy NP/VP. We can provide the focus (NP), and we expect to find the topic (VP). Therefore, we need two database (one of NP and one of VP). The NP which would not be relevant for indexing purposes would still be useful in a database system to ease information retrieval.
- The idea is to use presuppositions deduced from the query to search for terms (descriptors) relevant to the query. We believe it is necessary for the morpho- syntactic analysis to follow the same procedure as that for a written text in order to facilitate the search for relevant documents.

3. Conclusion

The fundamental characteristics of language in the case of Man/Information System interface are related to methods of automatic language processing. Therefore it is necessary to have a better understanding of the natural language before automatic processing.

Certainly, system structuring should be related to the "utility" of the application, but the development of a global model requires much forethought concerning the design. The whole problem of performance is to optimise the existing links between the configuration of the system and the user needs, and better understand the relationships between the notion of presupposition and queries in natural language.

The aim of an Information Retrieval System for the user is to retrieve information stored in a data bank, but also to interact, hence the necessity to take this final objective into account i.e. consulting the base using a query in natural language. However, the processing of this query comes within the scope of automatic language processing. Natural Language possesses a "natural logic" which links it to presupposition logic.

References

Salton G, Mc Gill M.J; Introduction to modern information retrieval; Mc Graw-Hill; 1983; New York.

Larouk O; WebData Compression for Information Collection : Organisation, Navigation and Filtering with linguistic relationships of inclusion , Second International Conference on Information Fusion Organis par International Society of Data Fusion and Data Mining, July 6-8, 1999, Sunnyvale, CA, USA. (in CD-Rom).

Larouk O; Modeling users needs : Schema of interrogation and Filtering the answers from the WEB in mode Co-operation International Conference on Knwoledge Organisation ISKO Lille, publi in Advances in Knwoledge Organisation, August 24-30th, 1998, pp. 105-115.

Larouk O; "Application of Non-Classical Logics to Optimize Textual Knowledge Representation in an Information Retrieval System (IRS)" in HEURISTICS : THE JOURNAL of Knowledge Engineering; Volume 6, Number 1 Spring 1993; Gaithersburg, MD- USA; pp. 24-37.

Croft W.B, Lucia T.J, Cohen P.R; "Retrieving Documents by Plausible Inference : A Preliminary Study";In Y. Chiaramella (Ed.); ACM Conference on Research and Development in Information Retrieval; Grenoble; June 13-15; 1988; pp. 481- 494.

[9] Van Rijsbergen C.J; "A non-classical logic for information retrieval"; The Computer Journal; Vol. 29, n6; 1986; pp. 482-485.

Sparck Jones K; "User models, discours models, and some others"; in Y. Chiaramella (d.); ACM/SIGIR: 11 th Conference on Research & Development in Information Retrieval; Grenoble; 1988; pp. 13-29.

Hintikka J; The Semantics of Questions and the Questions of Semantics; North-Holland; Amsterdam; 1976.

Rychkova Liudmila

Quantitative Text Investigation Based on Specially Oriented Linguistic Full-text Data Bases

It is normally considered that quantitative text investigation at the level of lexies [1] is the most efficient. Numerous works by Yuhan Tuldava [2], for example, are devoted to quantitative systemic research of lexics. The research of dependence between the sentence length and the nature of the text [3] in fact, also reflects the quantitative characteristics of lexical units combination in the text, considered at the very exterior level. Only the development of styleometry and the diagnostics being its part actualized quantitative investigations at the level of syntax: it is considered that namely "the syntactic structure makes the author's manner of the thought development more explicitly seen" [4].

The compound quantitative text investigation is possible only with the support on specially oriented linguistic full-text data bases. Such data bases are formed from linguistically annotated language corpora. The directions of linguistic annotation are reflected through the definite indexation system admitted for a definite corpus and are defined by the users' needs in a data base. The latter allows to define the objects typology and the complex of attributes-qualities, characterizing each of object types. Thus, each of exact objects is made explicit in a data base by means of composite labels application which contain syntactic and semantic information.

It is evident that effectuation of research in the framework of the bank of such linguistic data bases will allow to consider objectively the genre, style and other text peculiarities making up the initial corpus.

Grounded on the fact of national text corpora creation one can consider justified to provide quantitative investigations: a) based on the most general semantic-syntactical text structure; b) of text and linguistic objects qualities splitting according to their reference to a definite text type or to a definite text component part; c) of multiword lexical units in the aspect of their typology, their composite elements particularities, word order, "breaks" of linea structure.

References

Martchuk, Yn. N. (1980) On translation modelling (Voprosy informatsionnoy teorii i praktiki.) – M, VINT, N43, p.83. [1]

Tuldava, Yuhan (1987) Problems and methods of quantitative - systemic lexies research. – Tallinn: Valgus. [2]

Lesskis, G.A. (1963) On the dependence of sentence length and the nature of the text (Voprosy yazikoznaniya), N3, p. 92–112. [3]

Martynienko, G.Ya (1996) The Complexity of syntactic structures and style diagnostics (Prikladnaya Linguistika). - St.Peterburg: Izd-vo S.Peterb. Un-ta, p. 435–436. [4]

Yoshio Narisawa

Co-occurrence of Antonyms — Research Based on English Corpus

This paper extends Charles and Miller's (1989) co-occurrence hypothesis, which states that antonymous adjectives co-occur in the same sentence with frequencies far greater than predicted by chance. This paper also extends Fellbaums co-occurrence and antonymy, which extend Charles and Miller's hypothesis to nouns, verbs, showing that semantically opposed concepts co-occur in the same sentence with higher frequencies than predicted by chance. In searching the COBUILD CD-ROM for intrasentential as well as intersentencial co-occurrences of semantically opposed verbs, nouns, adjectives, adverbs and prepositions, we found that words expressing antonymous concepts co-occur with higher-than-chance frequencies in the identical context. As Farghal (1995: 20) rightly puts it, 'The linguists have been preoccupied with the structural systems of language — phonology and syntax. This emanates from the fact that they are most susceptible to scientific analysis. The fact remains, however, that without grammar very little can be conveyed, without lexis nothing can be conveyed'. It is only recently that lexis has been a full candidate for serious language research to the linguists.

Computer-assisted research of a large body of data has put the renewed light to the discussion of lexical relations of English language, e.g., synonymy, antonymy, hyponyms, etc. and may be capable of revealing various features of English language which early linguists were not aware of. Many of the linguists suffered from the fact that analysis lacked any formal basis, while the corpora that were used were commonly rather small, privately owned collections of the data, accessible only to few people. Consequently, much of the research that was done then proves impossible to verify, is inconsistent, and because of the fact that it was carried out on a relatively small scale, it is hardly conclusive about anything.

In our data from COBUILD CD-ROM we found the following co-occurrences of antonyms with neighbouring sentences. They show that antonyms do not always co-occur within a sentence. It is of interest to point out that they occur in a shared context common to a pair of the antonyms.

Unlike Christine Fellbaums findings for adjectives, nouns and verbs they can co-occur in neighbouring sentences. The co-occurrence of antonyms cannot be confined to the same sentence. Examples are:

> (1) Now there is relative peace in Uganda. The physical wounds may have healed, but the psychological scars of violence have been slow to fade, creating obstacles to people trying to rebuild their communities in . . .

Because the searches focus on the co-occurrences of words, counts are based on words including some variants (parent, parents, child, children). By taking word frequency, the expected number and chance probability of co-occurring words were obtained; based on the statistics, the expected rate and the chance probability of occurrence were far much smaller than the actual occurrence. The figure that we obtained reflects very strong co-occurrence of antonyms in a context. Each word (w1 or w2) and the number of its occurrence (n_1 or n_2) in COBULD CD-ROM were searched. The expected number of w1 and w2 co-occurring in a shared context consisting of 50 words is:

$$\text{Expected number of co-occurrence} = \frac{n_1 * n_2}{100,000} \tag{1}$$

Next the ratio, r, of actual and expected co-occurrence is listed; this figure shows clearly in each case that antonyms co-occur far more frequently than statistically expected:

$$r = \frac{\text{actual co-occurrences}}{\text{expected co-occurrences}} \tag{2}$$

Finally, the chance probability of finding w1 and w2 in any context is calculated in the corpus:

$$\text{chance probability} = \frac{n_1 * n_2}{100,000^2} \tag{3}$$

Evidence from Cobuild CD-ROM shows that antonymous words with any word class in the same context co-occur with much higher-than-chance frequencies. The data examined in this paper strengthen Christiane Fellbaums co-occurrecence hypothesis of antonyms (1995) and extend it beyond the word classes of

adjectives, nouns and verbs, showing that semantically opposed concepts of all the other word classes like pronouns, prepositions and adverbs co-occurr The result of this study rules out Fellbaums intrasentential co-occurrence of semantically opposed words.

Marc Hug

Partial Disambiguation of Very Ambiguous Grammatical Words

In textual databases like FRANTEXT, the user can list occurrences of grammatical elements such as conjunctions, prepositions etc, and even punctuation marks. In other databases, on the contrary, as the CD-ROM from the paper <Le Monde>, Paris, this is impossible, because such units have not been coded. In both cases, however, the linguist is often frustrated because he/she cannot select a particular kind of use of such units, which usually are very ambiguous. The problems involved with these ambiguities reflect in the fact that the "categorized database" within FRANTEXT excludes from categorization some of the most frequent and most ambiguous French words, as <que> or <si>. Furthermore, no distinction has been made between relative and interrogative use of words like <qui>, <que>, <quand>, <o>, <lequel>. Similar problems arise when one deals with English words like <which> or <that>, or German words like <wer>, <was>, <der> etc. Starting from the case of French words <qui> and <que>, and specially from their interrogative use, I shall try to examine what proportion of their occurrences might be disambiguated by means of a micro-syntax (examination of two or three adjacent words). Of course, these proportions are very different when comparing different words and different uses. The goal of the research is to enhance performances of categorizing programs.

Jaroslava Hlavacova

Rarity of words in language corpora

Rarity is the measure which describes how even or uneven is distribution of words in a language corpus. Its value express the difference between the frequency of a word in the corpus and in the language. The rarity together with the frequency in the corpus, yields more reliable information about the frequency of the word in the language than the frequency in the corpus alone.

The term rarity was introduced last year at the conference Text, Speech, Dialogue 99 (Hlavacova & Rychly, 1999). In this contribution some new aspects of the measure are presented. All examples are taken from the Czech National Corpus (CNC) with 100 mil. words.

Definition

Let x be a word in its broadest sense (it can be a word form, a lemma, a tag or whatever sequence of letters). $f(x)$ is the frequency of the word in the corpus. In other words it is number of positions where the word occurs in the corpus, where positions are unique numbers representing the order of words within the whole corpus. N is number of words in the corpus (position of the last word of the corpus). Let us for every word x divide positions of the whole corpus into $f(x)$ intervals, each of them having (approximately) the same length. In general the i-th interval is $< [(i-1)N/f(x)]+1, [iN/f(x)] >$ for all $i = 1, f(x)$. Note: In fact the intervals have exactly the same length only when the number N is divisible by $f(x)$. Then $N/f(x)$ is an integer and $[N/f(x)] = N/f(x)$. Otherwise, the lengths of intervals differ, but by 1 at most. When working with tens or hundreds of millions of words, this difference won't influence the results. Reduced frequency: $r(x) = \sum(Fx(i))$ for $i = 1, \ldots N$, where $Fx(i) = 1$, if the word x occurs in the i-th interval, $Fx(i) = 0$, otherwise. Rarity $R(x)$ is the quotient of the normal and reduced frequencies: $R(x) = f(x)/r(x)$.

Properties

$R(x) >= 1$ for all words x. If a word is distributed entirely evenly in the whole corpus, its reduced frequency is equal to its frequency and $R(x) = 1$. There is no lemma in the CNC with $f(x) > 8$ and $R(x) = 1$. On the other hand, if all the occurrences of a word fall into one interval, its reduced frequency is 1 and $R(x) = f(x)$. If the frequency of such a word is high, so is its rarity. We can say, that the higher rarity of a word, the more uneven its distribution in the corpus, which implies that the word is probably rare (not frequent) in the language. This can distinguish the rare words in the language which have only accidentally high frequency in the corpus (a scientific term used many times in one article, unusual proper name in one novel) from those that are really common. Unfortunately, the opposite is not true. Low rarity tells nothing about the word, because it can be caused by low frequency as well as by very even distribution within the corpus. The following table proves that by an example of real data taken from CNC.

word - lemma	English	frequency	reduced frequency	rarity
a	and	2809239	1767491	1.59
naskakovat	to jump into	138	87	1.59
znepjemnit	make unpleasant	89	56	1.59
lehkovn	light-minded	51	32	1.59
mastika	ointment	27	17	1.59

The rarity itself cannot serve as a measure of word commonness. It is necessary to look also at the frequency. The rarity 1.59 from the table corresponds with the rarity of a randomly distributed word. The probability that such a word occurs in the i-th interval follows the binomial distribution, but for high frequencies it can be approximated by the Poissons one. Then the rarity of a randomly distributed word would be $e/(e-1)$, which is 1.59. Let us now have a look at the most frequent words - lemmas from the CNC:

Their rarity is very close, rather slightly higher than, the rarity of the hypothetical randomly distributed word.

word - lemma	English	frequency	reduced frequency	rarity
a	and	2809239	1767491	1.59
v	in / at	2665425	1601007	1.66
bt	to be	1689681	937477	1.80
na	at / in	1632182	983832	1.66
z	from / of	862104	503969	1.71

Possible use of the rarity

The concept of rarity was developed in order to help with automatic decisions concerning the commonness of a word in a language. It can be used for instance as a tool for selection words into a dictionary. It could also serve as a correction for making frequency dictionaries.

References

Hlavacova, J., Rychly, P. (1999). Dispersion of Words in a Language Corpus. Proceedings Text, Speech and Dialogue, Springer Verlag Berlin Heidelberg.

E. I. Sicilia-Garcia, Ji Ming, & F. J. Smith

A Dynamic Language Model for each Significant Word

We present a new statistical language model based on two new concepts. A collection of language models are generated, one for each content word. The language models for the content words in the cache are then selected and combined using a new method based on a logical union of conjunctions. First results are encouraging.

Statistical language models rely on the assumption that the future use of a language will follow similar linguistic patterns to those used in the past. The most prevalent statistical language models are n-grams [1], which used word phrases (n-grams) to represent the language. Different techniques have been developed to obtain improvements in the statistical language model performance. Among them, the cache language models based upon the observation that whenever a particular word type appears within a text or conversation, it has an increased chance of reappearing within the following few sentences [2]. Also, it is likely that a human needs only a few content words, usually less than a single sentence, to recognise accurately the specific domain of a whole sentence and therefore (also) the domain of the following sentence or utterance. A human is therefore able to anticipate the likely words, phrases and structures peculiar to the domain. The use of individual language models is an attempt to mimic this ability of humans.

To build these models, first a relatively small corpus is compiled for each significant word by collecting together all the contexts of the word (i.e. all the sentences containing the word or all the paragraphs containing the word). Then an n-gram language model is generated from this word corpus. With modern PCs it is just possible to do this even for a language with 1000,000 word types. A conventional model developed from a global corpus is also created and combined with the individual word models for words in the cache.

Different ways of combining these probabilities have been studied. Preliminary forms were based on linear models but the improvement was insignificant compared with the computational time. Lately, a more sophisticated method was used, based on an exponential decay in probability with distance from the significant word along with a cut off. However, in this paper we present a completely new approach to combine models, the Probabilistic-Union model[3] previously applied in noisy speech recognition.

The union model is based on the logical principle of a disjunction of conjunctions and implemented as a probabilistic sum of products. It is best explained with an example of 4 models with probabilities P_1 to P_4. There are then 4 union models of order 1 to 4:

$$P^{(1)}_{Union}(w) = \Psi_1(P_1 \cdot P_2 \cdot P_3 \cdot P_4) \quad (1)$$

$$P^{(2)}_{Union}(w) = \Psi_2(P_1P_2P_3 \oplus P_1P_2P_4 \oplus P_1P_3P_4 \oplus P_2P_3P_4) \quad (2)$$

$$P^{(3)}_{Union}(w) = \Psi_3(P_1P_2 \oplus P_1P_3 \oplus P_1P_4 \oplus P_2P_3 \oplus P_2P_4 \oplus P_3P_4) \quad (3)$$

$$P^{(4)}_{Union}(w) = \Psi_4(P_1 \oplus P_2 \oplus P_3 \oplus P_4) \quad (4)$$

where $P^{(k)}_{Union}(w) = P^{(k)}_{Union}(w|w_1^n)$ is the union conditional probability of order k. $P_i = P_i(w|w_1^n)$ is the conditional probability for the significant word w_i and Psi_k is a normalizing constant. The symbol $\oplus$ is a probabilistic sum, i.e. its equivalent for 1 and 2 is:

$$P_{1\text{and}2} = P_1 \oplus P_2 = P_1 + P_2 - P_1P_2. \quad (5)$$

Equation (1) is a product and is the form commonly used in signal processing. The equation (4) is a sum which can be generalized to a weighted sum commonly used in language processing.. The other two are the interesting new models.

The combination of the global language model with the probabilistic-union model is defined as follows:

$$P(w|w_1^n) = \alpha P_{Global}(w|w_1^n) + (1-\alpha)P_{Union}(w|w_1^n). \quad (6)$$

One advantage of using this method is to reduce the effect of a zero or very low probability in any of the word models which reduces the overall probability. Another advantage of the Union model is that contributions from two or more words reinforce one another when included in the form of a product. Experiments

have been carried out with a 10 million word subset of the Wall Street Journal corpus and we have obtained an improvement of 17% in perplexity with the exponential decay model and , so far, 20% with the union model of order 5 using the 6 nearest words from the cache. Further improvement is expected.

References

Jelinek F., Merialdo B., Roukos S., Strauss M. (1991). A Dynamic Language Model For Speech Recognition. Proceedings of the Speech and Natural Language DARPA Workshop. pp. 293-295. [1]

Kuhn R., DeMori R. (1992). Corrections to A Cache Based Natural Language Model for Speech Reproduction. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 12(6), pp. 691-692. [2]

Ming J., Stewart D., Hanna P., Smith F. J. (1999). A Probabilistic Union Model for Partial and Temporal Corruption of Speech. Automatic Speech Recognition and Understanding Workshop, ASRU 99, Vol. 1, pp 43-46, Keystone, Colorado, USA, December. [3]

Sibasis Mukherjee

On The Sign Test

In many kinds of investigation of the interest to a linguist , only an ordinal Level can be achieved while conducting various non-parametric tests. Let us take an example , the case where informants are asked to rate two sentences on a scale of acceptability, or politeness , or some similar variable . Such variable (Like coherence) cannot really be measured in units with equal interests, and we cannot attach mush interest to the magnitude of differences between ratings; we could however, claim that a sentences rated say 5 on an acceptability scale had been rated as more acceptable than a sentence given a rating on 3.

To carry out the sign test,we first record the sign of difference for each pair of scores , subtracting consistently . Tied scores are dropped from the analysis and the number of pairs (N) reduced accordingly.We now find the number of pairs with the less frequent sign , and say it is x . As an example of situation where the sign test will be appropriate,consider the case,where a group of people has been asked rate a sentence on scale of acceptability from 0 (totally unaccepted) to 5 (totally accepted) for (i) informal spoken Bengali and (ii) formal written Bengali. The investigator predicts that the sentence will be judged as more acceptable in the informal spoken.form than in the formal written formal Bengali. Please refer to the following two table:

No of Samples	Informal Spoken	Formal Written
1	3	2
2	4	2
3	1	0
4	5	4
5	5	5
6	4	5
7	2	0
8	3	5
9	3	4
10	4	2
11	2	3
12	4	3
13	2	1
14	3	2
15	2	1

Table 1: Acceptability ratings for a sentence in informal spoken and formal written Bengali

Now, it seems from table no 2 that we have 4 negative and 10 positive differences . Therefore x=4 and N=14 . The critical value of x at the 5 percent level for N=14 in a directional test is 4 . Since the calculated value of x is equal to the critical value , we can reject the null hypothesis at the 5 percent level , and conclude that there is significant difference between the two sets of scores , which is clearly in the predicted direction.

The samples were taken from a suburban town of Calcutta ,India with equal number of men and women and the present paper envisages to compute the probability of obtaining any particular degree of deviation of the rate of acceptability from the equality under the null hypothesis.

Informal Spoken	Formal Written	Sign of(Informal-Formal)
3	2	+
4	2	+
1	0	+
5	4	+
5	5	0
4	5	-
2	0	+
3	5	-
3	4	-
4	2	+
2	3	-
4	3	+
2	1	+
3	2	+
2	1	+

Table 2: Sign differences in acceptability Scores.

D.V. Khmelev
Disputed Authorship Resolution Using Relative Empirical Entropy For Markov Chain of Letters in a Text

A new statistical method in the analysis of literary style for disputed authorship resolution is considered here. It was tested in the following experiment.

We take a training set of 304 text samples (novels, stories and short stories) of 82 different authors in Russian. The total size of training text samples for each author exceeds 100,000 symbols. We also take one control text sample per each author. The size of the control sample exceeds 100,000. The total volume of text samples is about 120Mb.

Each control text is considered to be anonymous. Our method determines the true author for 69 control texts. In 3 (resp. 2, 1) cases the true author is second (resp. third, fourth) in the list of pretenders to its own text. Note that the analysis of frequencies of isolated letters guesses the true author for just 2 control texts.

Let us describe the method in detail. Consider a sequence of letter of text as a Markov chain. The matrices of transition frequencies of letters pairs are calculated over all texts for each author. Therefore we know (approximately) the probability of transition from one letter to another for each author. The author of the control text is guessed by the principle of maximal likelihood, i.e., for each matrix we calculate the probability of anonymous text and we choose the author with the maximal corresponding probability. The chosen author is supposed to be the true author.

Suppose we take the logarithm of each probability, change a sign, and divide it by the length of control text; then each of the numbers obtained is called the *relative empirical entropy*. Relative empirical entropy is more convenient for computing than actual probabilities. Besides, the chosen author (who often happens to be a true author) has the minimal relative entropy.

Now we shall give a list of 82 authors included to the experiment. There are a lot of well-known Russian writers of the XIX and XX centuries among them: K. Bulychev, O. Avramenko, A. Bol'nykh, A. Volkov, G. Glazov, M. i S. Djachenko, A. Etoev, A. Kabakov, V. Kaplan, S. Kazmenko, V. Klimov, I. Krashevskij, I. Kublickaja, L. Kudrjavcev, A. Kurkov, Ju. Latynina, A. Lazarevich, A. Lazarchuk, S. Lem, N. Leonov, S. Loginov, E. Lukin, V. Chernjak, A.P. Chekhov, I. Khmelevskaja, L. and E. Lukiny, S. Luk'janenko, N.Ju. Markina, M. Naumova, S. Pavlov, B. Rajjnov, N. Rerikh, N. Romaneckijj, A. Romashov, V. Rybakov, K. Serafimov, I. Sergievskaja, S. Scheglov, A. Schegolev, V. Shinkarev, K. Sitnikov, S. Snegov, A. Stepanov, A. Stoljarov, R. Svetlov, A. Sviridov, E. Til'man, D. Truskinovskaja, A. Tjurin, V. Jugov, A. Molchanov, F.M. Dostoevskij, N.V. Gogol', D. Kharms, A. Zhitinskij, E. Khaeckaja, V. Khlumov, V. Kunin, A. Melikhov, V. Nabokov, JU. Nikitin, V. Segal', V. Jan, A. Tolstoj, I. Efremov, E. Fedorov, O. Grinevskij, N. Gumilev, L.N. Tolstoj, V. Mikhajlov, Ju. Nesterenko, A.S. Pushkin, L. Reznik, M.E. Saltykov-HHedrin, V. Shukshin, S.M. Solov'ev, A. Kac, E. Kozlovskij, S. Esenin, A. Strugackij, A. and B. Strugackie and B. Strugackij.

Note that transition frequencies of letter pairs are frequent, easily quantifiable and relatively immune from conscious control. The author of this study performed the same computations on English, French, and German texts and received similar results. Its description needs a separate paper and is beyond the scope of this short thesis. Clearly, further studies in this direction may be the best approach to "stylometry's 'holy grail', the fully automated identifier" (Holmes, 1998).

References

Holmes, D.I. (1998) The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13: 111–17.

Patrick Juola
A Linear Model of Complexity (And Its Flaws)

A key question underlying any investigation of complexity is that of the model under which complexity is measured. For instance, the description of a sphere as an approximation by planes (as is commonly done in computer graphics packages) is long, involved, and 'complex;, while a similar description as a quadratic surface (an equation involving a term like x^2) can be written in a single line. This suggests, among other things, that spheres and similar curves make more extensive use of quadratic than planar components, or, more simply, that a sphere is quadratic. In general, an appropriate model results in a simple(r), shorter, and more plausible description of the events of interest.

(Juola, 1998) argued for Kolmogorov complexity (K-complexity) as a basis for a psychologically valid measure of complexity. Here we develop a model of linguistic complexity based on a different measurement than K-complexity and with a strongly different psycholinguistic base. By comparing the complexity measurements taken in the different frameworks, one implicitly compares the underlying psycholinguistic bases. In particular, the model applied here (linear complexity) ignores most aspects of long-term memory, and therefore quantitatively confirms and measures the role that long-term memory in particular may play in human language.

Previous work (Chater and Hahn, 1997; Juola, 1997) has shown that the size of compressed text samples, an estimate of the Kolmogorov complexity (Li and Vitányi, 1997) of those samples, can be a useful and psycholinguistically valid measure of the comparative complexities of language. The K-complexity of a sample of text, or any other sequence, is defined as the size of the smallest computer program that will output a given sequence, or less formally can be seen as the amount of knowledge required to (re)produce a given sequence. A more complex sequence will require more knowledge to recreate, and hence a larger (compressed) file. Linguistic regularities act to reduce the apparent complexity. By subjecting the text to various kinds of distortion, one observes the role played by the various distorted linguistic forms. For example, a language where syntax (in particular) is crucial to the prediction and generation of subsequent text (such as English), is more strongly affected by word-order distortion than a language where other levels, such as morphology, carry more of the load. K-complexity is, in a formal sense, a maximally general complexity measure that can capture any form of regularity in order to reduce the size of the final program. Other definitions exist where the sort of reconstruction machinery available is limited to a particular architecture or kind of information. If this limited architecture were found to be well-suited to the description of linguistic data, this would be strong evidence about the nature of the cognitive architecture available in the human head. Conversely, if the limitation were to result in poor or needlessly complex models, then the sort of information and processing excluded by the limitation would be shown to be of importance.

Standard compression programs such as *gzip* and *Stuffit* tend to use variations of Lempel-Ziv (Lempel and Ziv, 1976) complexity. This technique describes future text as a concatenation of strings seen in past text: the (novel) string "New York" is composed of the (previously seen) strings "New " and "York". At this point, the string "New York" is also available (in a sort of long-term memory) to facilitate the description of other (novel) forms. A string is thus modelled as a sequence of increasingly larger (novel) substrings, described in terms of previous elements of the description, and selected from an increasingly large "vocabulary" of available strings.

By contrast, we choose to study an alternate definition : *linear* complexity (Massey, 1969; Schneier, 1996). The architecture is a *linear feedback shift register* (LFSR) composed of an ordered set of registers and a (linear) feedback function. The register set acts as a queue, where the past few elements of the sequence line politely up in order, while the feedback function predicts the next single text element and adds it to the end of the queue (dropping the element at the head, of course). The linear complexity of a given sequence is defined as the size of the *smallest* LFSR generating a given sequence, and can be efficiently determined by a simple computer program (Massey, 1969).

We have, then, at least two competing definitions of the "complexity" of a given language sample. Is there reason to believe they could differ? Not only is the answer "yes," but there are interesting *cognitive* reasons behind the answer. LZ-complexity is the complexity of a sequence given free access to an essentially unlimited "vocabulary" of pseudo-lexemes in long-term memory. Linear complexity, by contrast, focuses attention on the last few elements of the sequence of interest. It thus makes no use of, and no provision for, long-term memory. If, as one expects, long-term storage is an important part of human language

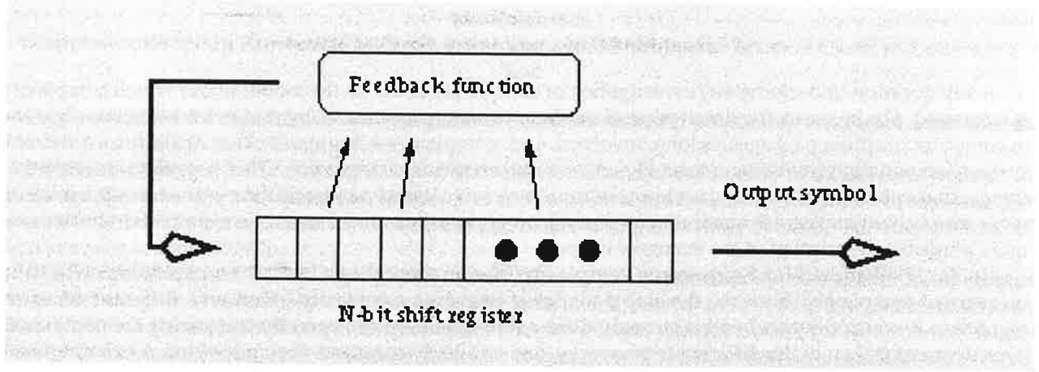


Figure 1: Sample Linear-Feedback Shift Register (LFSR)

processing, then (just as spheres are badly described as collections of planes), it should be very badly modelled via LFSRs. It is therefore unsurprising that the linear models of human language show much less match with our intuitions than LZ-models by the same tests used earlier (Juola, 1998) to demonstrate the well-foundedness of modelling linguistic complexity.

What, then, is the purpose of presenting an avowedly bad model of human language? Primarily this : only by contrast with a bad model can the fitness of any particular model be shown. In particular, a model limited to ignore particular sorts of regularity and cognitive machinery, in this case, long-term memory, produces worse results than a model taking that into account, thus numerically and quantitatively demonstrating the importance of that particular regularity. A similar but more detailed examination could shed light on additional aspects of the architecture of human language. We thus have a new method for empirically investigating the architecture and processes underlying human language processing.

References

Chater, N. and Hahn, U. (1997). Representational distortion, similarity, and the Universal Law of Generalization. In Proceedings of the Interdisciplinary Workshop on Similarity and Categorization (SimCat 97), pages 31–36, University of Edinburgh.

Juola, P. (1997). A numerical analysis of cultural context in translation. In Proceedings of the Second European Conference on Cognitive Science, pages 207–210, Manchester, UK.

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

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. IEEE Transactions on Information Theory, IT-22(1):75–81.

Li, M. and Vitányi, P. (1997). An Introduction to Kolmogorov Complexity and Its Applications. Graduate Texts in Computer Science. Springer, New York, 2nd edition.

Massey, J. L. (1969). Shift-register synthesis and BCH decoding. IEEE Transactions in Information Theory, IT-15(1):122–7.

Schneier, B. (1996). Applied Cryptography, Second Edition : Protocols, Algorithms and Source Code in C. John Wiley and Sons, Inc., New York.

Karl-Heinz Best

Verteilungen sprachlicher Einheiten in Texten und im Sprachsystem

1 Summary

This paper presents the investigations of the 'Göttingen Project on Quantitative Linguistics' which aims at discovering the laws controlling the frequency distributions of different kinds of linguistic units in texts and lexica. Up to now, about 40 languages have been investigated with promissing results.

2 Vorbemerkung

In diesem Vortrag wird — anschließend an Best (1998) — über die neueren Forschungsergebnisse des Göttinger Projekts berichtet. Dieses Projekt widmet sich der Frage, ob und wie weit die theoretisch begründete Hypothese, dass Wörter unterschiedlicher Länge sich entsprechend dem Ansatz $P_x = g(x)P_{x-1}$ mit verschiedenen Spezifizierungen für $g(x)$ in Texten verteilen, sich a) für Wortlängen immer wieder (auch bei weiteren Daten) mit Erfolg prüfen und b) auf andere Einheiten und Klassen von Einheiten übertragen lässt.

3 Theoretische Grundlagen

Die erwähnte Theorie wurde als Theorie der Wortlängenverteilung zuerst von S.G.Čebanov (1947) formuliert, der annahm, dass die Poisson-Verteilung hierfür ein gutes Modell sein müsste. Wenig später kam W. Fucks (1955) — offenbar ohne Kenntnis der Arbeit von Čebanov — zum gleichen Ergebnis. In späteren Arbeiten konnte gezeigt werden, dass es sich bei diesem Vorschlag lediglich um einen Spezialfall eines noch allgemeineren Gesetzes handelt, wie auch schon von Fucks in Betracht gezogen worden war. Den gegenwärtigen Stand der Theorie stellen nach wie vor die Arbeiten von Wimmer u.a. (1994) sowie Wimmer & Altmann (1996) dar. Dieser Ansatz geht davon aus, dass zwischen der Länge x und der Länge $x-1$ von Wörtern in Texten eine variable Proportion bestehen sollte, was zu der bereits genannten Gesetzeshypothese $P_x = g(x)P_{x-1}$ geführt hat, wobei $g(x)$ je nach Sprache, Autor, Textsorte etc. unterschiedliche Formen annehmen kann. Mit $g(x) = a/x$ kommt man nach Umformungen zur Poisson-Verteilung, dem ältesten Gesetzesvorschlag; $g(x) = a/b + x$ modelliert die sog. Zipfschen Kräfte (Unifikations- und Diversifikationskraft) und führt zur Hyperpoisson-Verteilung, die für viele Verteilungen das Grundmodell zu sein scheint.

4 Ergebnisse

Anknüpfend an Best (1998) kann festgestellt werden:

1. Verteilung von Wortlängen. Dies ist das mit Abstand am besten erforschte Phänomen. Verglichen mit Best (1998) sind als Sprachen Arabisch, Bulgarisch und (schottisches) Gälisch hinzugekommen; außerdem wurden erstmals Daten zu einigen deutschen Dialekten erhoben: Pfälzisch, Schweizerdeutsch. Zu vielen Sprachen, die schon vorher berücksichtigt worden waren, sind weitere Daten erhoben worden. Die Ergebnisse bestätigen wiederum, dass die o.a. Gesetzeshypothese die sprachlichen Phänomene zutreffend erfasst. Probleme mit einem chinesischen Datensatz konnten behoben werden. Neu ist die Erkenntnis, dass die Wortlängen in slavischen Sprachen offenbar meist ebenfalls der Hyperpoisson-Verteilung folgen, wenn man nur die nullsilbigen Wörter unberücksichtigt lässt. Der Wortstatus der Nullsilbigen ist ja mindestens problematisch. Erste Untersuchungen zu Wortlängen in Lexika zeigen, dass auch in alphabetischen und in Frequenzwörterbüchern im Prinzip die gleichen Verteilungsgesetze gelten (Best 1999, Best [Hrsg.] 1999).

2. Verteilung anderer Einheiten: Satz-, Silben-, Morph-, clause-, Satzgliedlängen, Längen rhythmischer Einheiten und Satzgliedtiefen Es wurden über die in Best (1998) erwähnten Untersuchungen hinaus weitere empirische Daten erarbeitet und überprüft. Die Ergebnisse zeigen, dass die Theorie der Verteilung von Wortlängen offenbar sinngemäßauf beliebige andere Einheiten übertragen werden kann. Es gibt bisher keinen einzigen Fall, keinen einzigen Datensatz, der auf etwas Anderes schließen lässt.

3. Verteilungen von Wortarten, Satzgliedern und Satzgliedfunktionen Bisher waren nur Wortartverteilungen — in geringem Umfang — auf ihre Verteilungen in Texten hin untersucht und als Rangordnungen behandelt worden. Weitere Bearbeitungen von Wortarten sowie erste Versuche zu Satzgliedern und Funktionen von Satzgliedern deuten darauf hin, dass diese den gleichen Verteilungsgesetzen gehorchen wie Wortlängen.

5 Perspektiven

Nur für die Verteilung von Wortlängen in Texten ist eine einigermaßen zufriedenstellende Datenbasis erreicht; wünschenswert ist jedoch eine Einbeziehung weiterer Sprachen. Für alle andern Einheiten liegen bisher einige Erkenntnisse zu den sprachlichen Verhältnissen im Deutschen vor; andere Sprachen konnten nur in geringem Maße berücksichtigt werden. Hier ist dringend eine Verbreiterung der Datenbasis in jeder Hinsicht geboten. Das bedeutet, dass in der Syntax verstärkt Sätze, Teilsätze (clauses) und Satzglieder (phrases) untersucht werden sollten, in der Morphologie die Morphe, in der Phonetik die Silben, etc. Dies ist eine wesentliche Perspektive für die weitere Arbeit des Göttinger Projekts. In diesem Zusammenhang ist noch auf folgenden Aspekt hinzuweisen: Die Länge der sprachlichen Einheiten ist nur ein mögliches Maßfür ihre Komplexität; man kann diese auch noch anders messen, z.B. durch ihre 'Tiefe'. So kann man die Tiefe eines Satzgliedes danach bemessen, wie viele hierarchisch geordnete Attribute es enthält. Das nominale Satzglied 'Das Haus' enthält kein Attribut und hätte damit die Tiefe 1; 'das Haus meines Vaters' mit 1 Attribut hat die Tiefe 2, 'das Haus des Bruders meines Vaters' die Tiefe 3, usw. Nun kann man untersuchen, wie viele Satzglieder der verschiedenen Tiefen in einem Text vorkommen und prüfen, ob auch diese den vorgeschlagenen Verteilungsgesetzen folgen. Ein allererster Versuch an einem deutschen Gedicht hat dies bestätigt. Man kann also prüfen, ob die bekannten Verteilungsgesetze nur für die Längen der Einheiten oder für beliebige Komplexitätsmaße gelten. Abschließend sei eine Spekulation erlaubt: Für die Textstrukturierung zeichnen sich womöglich 3 durchgehende Gesetzmäßigkeiten ab. Wenn man einmal die Häufigkeitsverteilungen von Klassen von Einheiten gleicher Art als 'horizontale' Strukturierung betrachtet, so scheinen diese alle dem Verteilungsgesetz $P_x = g(x)P_{x-1}$ zu gehorchen, dessen erste Form als Čebanov-Fucks-Gesetz bezeichnet wurde. Die spezifische Form für $g(x)$ kann verschieden ausfallen, die grundlegende Gesetzmäßigkeit ist aber immer die Gleiche. Möglicherweise können Verteilungen von Längenklassen sprachlicher Einheiten (z.B. Wörter verschiedener Silbenzahl) und Klassen von Kategorien derselben Einheiten (also z.B. Wortarten) als prinzipiell gleich strukturiert betrachtet werden. Betrachtet man nicht Klassen von Einheiten einer Sprachebene, sondern die einzelnen Einheiten, so folgen sie offenbar der Zipf-Mandelbrot-Verteilung (Knüppel 1997, Uhlířová 1995). Statt innerhalb einer Sprachebene zu bleiben, kann man auch das Zusammenspiel über die Sprachebenen hinweg betrachten, also die Abhängigkeit der Größe eines Konstrukts zur Größe seiner Konstituenten, allgemein als Menzerath-Altmann-Gesetz bekannt (Altmann & Schwibbe 1989; Hřebíček 1997). Diese 'vertikale' Strukturierung hat offenbar ebenfalls einen sehr allgemeinen Charakter, wie Köhlers 'linguistische Synergetik' zeigt, der ja wesentlich die gleiche Gesetzmäßigkeit zugrunde liegt (Köhler 1986). Es wird eine Aufgabe für die Zukunft sein, diese Spekulation zu überprüfen und entweder zu bestärken oder zu falsifizieren.

Literatur

Altmann, Gabriel und Schwibbe, Michael H. 1989. Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Hildesheim: Olms.

Best, Karl-Heinz. 1998. Results and perspectives of the Göttingen project on quantitative Linguistics. Journal of Quantitative Linguistics 5: 155–162.

Best, Karl-Heinz. 1999. Quantitative Linguistik: Entwicklung, Stand und Perspektive. Göttinger Beiträge zur Sprachwissenschaft 2: 7–23.

Best, Karl-Heinz. 1999. (Hrsg.) Häufigkeitsverteilungen in Texten. Trier: Wissenschaftlicher Verlag Trier (im Druck).

Čbanov, Sergej Grigorévič. 1947. O podčinenii rečevych ukladov indoevropejskoj gruppy zakonu Puassona. Doklady Akademii Nauk SSSR. Tom 55/2: 103–106.

Fucks, Wilhelm. 1955. Theorie der Wortbildung. Mathematisch-Physikalische Semesterberichte. Bd. 4. 195–212.

Hřebíček, L. 1997. Lectures on Text Theory. Prag: Academy of Sciences of the Czech Republic, Oriental Institute.

Knüppel, Anke. 1997. Untersuchungen zum Zipf-Mandelbrot-Gesetz im Deutschen. Staatsexamensarbeit, Göttingen. Gekürzt in: Best, Karl-Heinz (Hrsg.), Häufigkeitsverteilungen in Texten. Trier: Wissenschaftlicher Verlag Trier (im Druck).

Köhler, Reinhard. 1986. Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Bochum: Brockmeyer

Uhlířová, Ludmila. 1995. On the Generality of Statistical Laws and Individuality of Texts. A Case of Syllables, Word Forms, their Length and Frequencies. Journal of Quantitative Linguistics 2: 238–247.

Wimmer, Gejza und Altmann, Gabriel. 1996. The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (Hg.), Glottometrika 15. Trier: Wissenschaftlicher Verlag Trier. 112–133.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, und Altmann, Gabriel. 1994. Towards a Theory of Word Length Distribution. Journal of Quantitative Linguistics 1: 98–106.

Für weitere Informationen: Eine Bibliographische bersicht zu den Forschungen wird im Internet unter der folgenden Adresse fortlaufend aktualisiert: http://www.gwdg.de/ kbest/projekt.htm.

Edda Leopold

Length-Distribution of Words with Coinciding Frequency

It is a well known fact that frequent words on average are shorter than rare words (see e.g. Zipf 1932; Guiter 1974). For the time being, however, the exact form of the respective regression function is still an open question (see e.g. Köhler 1986; Hammerl 1990; Zörnig et al. 1990). Therefore attempts have been made to examine the two dimensional probability distribution (joint distribution) of pairs of frequency and length

$$\mathbf{P}(F = f, L = l).$$

Although no satisfying regression function between length and frequency has been found yet, a lot is known about the respective joint distribution. The projection onto the frequency axis

$$\mathbf{P}(F = f) \quad = \quad \sum_l \mathbf{P}(F = f, L = l)$$

is Zipf's law in the spectral version. The projection onto the length axis

$$\mathbf{P}(L = l) \quad = \quad \sum_f \mathbf{P}(F = f, L = l)$$

has been explored by Wimmer et al. (1994) and others and follows a Poisson-like distribution. The conditioned probability distribution

$$\mathbf{P}(F = f|L = l) \quad = \quad \frac{\mathbf{P}(F = f, L = l)}{\mathbf{P}(L = l)}$$

of frequency when a fixed length is given has been studied by Leopold (1998). She has examined frequency spectra of words with coinciding length in Dutch, English, Finnish, German, and Polish. She has found that Zipf's law is still valid when words of equal length are considered, but parameters are different from the unconditioned case and vary with word length. Her results have been corroborated for German by Knüppel (1997).

In this contribution we approach the relation between length and frequency of words from the other side. We consider the conditional probability distribution of word length for fixed word frequency

$$\mathbf{P}(L = l|F = f) \quad = \quad \frac{\mathbf{P}(F = f, L = l)}{\mathbf{P}(F = f)}$$

A salient result is that variance of length changes considerably with frequency. This is partly due to Zipf's law, but it can also be explained by the fact that the requirement of production effort is the larger the more frequent a word is. We present material from different languages and discuss how the findings can be explained within a system theoretic framework.

References

Guiter, Henry (1974): Les rélations fréquence - longueur - sens des mots (langues romanes et anglais); in: XIV congresso internazionale di linguistica filologia romanza; Napoli, 15. - 20. April 1974; 373 - 381.

Hammerl, Rolf (1990): L"ange – Frequenz, L"ange – Rangnummer: "Uberpr"ufung von zwei lexikalischen Modellen; in: Rolf Hammerl (ed.): Glottometrika 12; (QL 45); Brockmeyer: Bochum, S. 1–24.

Hammerl, Rolf (1991): Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells; wvt: Trier.

Knüppel, Anke (1997) Untersuchungen zum Zipf-Mandelbrot-Gesetz im Deutschen. Staatsexamensarbeit, Goettingen; in: Karl-Heinz Best (ed.), Häufigkeitsverteilungen in Texten. Trier: wvt (in press)

Köhler, Reinhard (1986): Zur linguistischen Synergetik: Struktur und Dynamik der Lexik; (QL 31); Brockmeyer: Bochum.

Leopold, Edda (1998): Frequency Spectra within Word Length Classes; in: Journal of Quantitative Linguistics, Vol. 5, pp. 224–231.

Zipf, George K. (1932): Selected Studies of the Principle of Relative Frequency in Language; Havard Univ. Press: Cambridge/Mass.1968.

Zörnig, Peter & Köhler, Reinhard & Brinkmöller, R. (1990): Differential Equation Models for the Oscilation of the Word Length as a Function of Frequency; in: Rolf Hammerl (eds.): Glottometrika 12; (QL 45); Bochum: Brockmeyer, pp. 25–40.

Wimmer, Gejza & Köhler, Reinhard & Grotjahn, Rüdiger & Altmann, Gabriel (1994): Towards a theory of word length distribution; in: Journal of Quantitative Linguistics, Vol. 1, pp. 98–106.

Zahra Mustafa

Non-courseware factors involved in using multimedia in foreign language instruction

The aim of this paper is to investigate the factors involved in applying multimedia in teaching English as a foreign language. In particular, it deals with the non-courseware factors affecting the use of multimedia in improving the pronunciation and the oral communication skills of students majoring in English for specific purposes at Jordan University of Science and Technology. The factors under investigation are the following: previous instruction on pronunciation using the conventional classroom teaching method, familiarity with multimedia, English proficiency, academic achievement, free time using multimedia, attitude, gender, and social status. The results show that there are variations in the effect of these factors on using multimedia to improve the students' oral skills.

Victor Zakharov

A range of linguistic tools for the information retrieval system on conservation and preservation

Some years ago the Russian Academy of Sciences Library (the Russian abbreviation is BAN) started the development of an information system in the field of conservation, restoration and preservation of library collections (here in our paper we will use for this subject area an abbreviation 'C&R'). This area is a new trend in science, culture and library practice. It appeared on the crossroad of library science, chemistry, physics, microbiology, entomology, paper production, leather production etc.

The information system on C&R of library materials in BAN is a system that includes three subsystems:

- A document bibliographic and full-text information retrieval system on C&R
- A factual database on physical conditions of books and environment in book stocks
- An expert system which should help specialists in library conservation and restoration.

The aim of this paper is to present a range of linguistic tools developed for the system, which includes:

- Frequency dictionaries of lexical units for the subject area
- Frequency dictionaries of word combinations for the subject area
- Dictionary of stop words
- Classification scheme (taxonomy)
- Bilingual information retrieval thesaurus
- Algorithms of generating frequency dictionaries
- Algorithms of automatic classification
- Algorithms of automatic indexing.

The dictionaries used for classification and indexing are produced using automatic and manual procedures. Also the special software was developed to create and use these tools. Linguistic support is provided also by glossaries of nomenclature elements, reference books of possible attribute values for factual databases and by the list of typical queries.

First of all the generalised classification scheme (taxonomy) for C&R was developed which was divided into 13 main divisions on the first level. The peculiarity of C&R is that it has an origin in different "pure" and applied sciences. Thus, the C&R vocabulary belongs to different subject areas. So a few frequency dictionaries were created which correspond to main topic divisions. The volume of elaborated texts was about 3 Megabytes. The analysis of lexical filling of divisions possessed us to correct the proper divisions and to determine links between separate divisions. Additionally we produced "contrast" frequency dictionaries for other subject areas in order to compare them with the C&R ones. The algorithm of generating frequency dictionaries is undependable on languages which texts belong to.

Each frequency dictionary consists of two parts:

1. Lexical entries - word usages with their frequencies in texts
2. Words which form entries' context in analysed texts.

Dictionaries have indices of absolute and relative frequency for each entry. These lists of word usage were automatically normalised, that is, the different forms of the same word were brought to its normal morphological form. The algorithm uses lists of inflexions which create paradigms that include both word-changing and word-building ones. For example, one of such paradigms for English includes endings '-s', '-ed', '-ing'. The other list of inflexions consists of '-al', '-ful', '-ive', '-less', '-like', '-ant', '-ent', '-en', '-some' and represents a word-building paradigm. Lists of inflexions exist separately of computer procedures. This allows to use the same procedure for different languages and to vary the procedure for the same language

(to use it consequently with different paradigms). Besides, there was developed another algorithm of morphological normalisation in search procedures based on word compression. It was used for lexical units' identification, too. The programme generates automatically a code for each lexical usage in text. The code is generated from word form by compression, which meets two conditions: 1) each lexical unit receives a different code; 2) each usage of the same lexical unit receives the same code. In general, the essence of the algorithm lies in the following: 1) the beginning of a word usage is included into the code; 2) then only consonants are included in accordance with the definite rules; 3) the length of such a compressed code is fixed. The end values of coding parameters, such as length of the word beginning, the table of consonants with their characteristics, the length of resulting code, were established experimentally.

The total volume of all frequency dictionaries is about 20,000 entries in different languages - mainly in Russian (65%) and in English (30%). Then the program of combining the separate dictionaries into one common dictionary for C&R subject area as a whole was created. It summarises frequencies of lexical units, which are the same, and the frequencies of word combinations from context. The generated in such a way combined dictionary for C&R includes, among full-meaning words, also link-words, words with common scientific meaning, names, abbreviations etc. So, the algorithm of singling out of the words with common meaning was elaborated. It chooses high-frequency parts according to pragmatically stated thresholds in two dictionaries: a C&R dictionary and a "contrast" one. The units that are represented in both sets form so called stop-word list. The combined dictionary for C&R has total frequencies for the whole area and separate frequencies for each subdomain with subdomain codes and could be used for different purposes, first of all for automatic attribution of full texts. Then the frequency dictionary was also processed by experts and transformed into thesaurus for information retrieval. The thesaurus is intended for automatic indexing in information retrieval system. It consists of 3 parts: (1) an ordered lexical and semantic index of keywords and descriptors; (2) a list of identifiers, i.e. names of unique objects; (3) hierarchy index. Relation types are "broader", "narrower", "synonyms". The main features of the thesaurus are English language entries and broad interpretation of synonymy. The thesaurus is designed in accordance with Russian standards 7.25–80 (monolingual thesaurus) and 7.24–90 (multilingual thesaurus). At present it includes about 2 and half thousand entries grouped in 900 descriptors. The thesaurus is now being used in experimental operation of the bibliographical database.

In the conclusion I would like to underline that we were using the system and developing different linguistic tools almost simultaneously, and thus the system started to provide information virtually from the "zero level". Two processes - the process of developing of means for linguistic support and the proper system operation - are iterative and mutually interdependent.

6 Introduction

Information access methods must be improved to overcome the information overload that most professionals face nowadays. Information Access tasks, like Information Retrieval, help the users to access a great amount of text they find in the Internet and their organizations. About 90% of the information in corporations exists in the form of text (Oracle, 97).

The task of WSD is the identification of the correct sense of a word in a particular context. Improvement in the accuracy of identifying the correct word sense will result in better for many natural language processing tasks (Kilgarriff, 1997) (i. e. in MT (Dagan et al., 1991; Chang et al., 1991), accent restoration (Yarowsky, 1994), or information extraction (Kilgarriff, 1997), etc.). WSD is especially interesting in classification tasks like IR (Voorhees, 1994, Sanderson, 1996), CLIR (Grefenstette, 1998), and TC (Buenaga et al., 1997; Urena et al., 1998b).

In this work, we study the way in which WSD could contribute to improve the effectiveness of these systems, specially IR systems. Information retrieval is concerned with the identification of documents in a collection that are relevant to a given information need, usually represented as a query containing terms, which are supposed to be a good description of what the user is looking for. IR systems may improve their effectiveness (i.e. increasing the number of relevant documents retrieved) by using a process query expansion, which automatically adds new terms to the original query posed by a user. In this paper, we develop a method of WSD based on the integration of linguistic resources (text corpora: SEMCOR and LDB: WORDNET) through the VSM. This disambiguator has been directly evaluated obtaining a high precision in the resolution of lexical ambiguity. Morever, we have realized an indirect evaluation, that is, we have applied the WSD intermediate task to IR task. In particular, we have applied this to the query expansion process, improving the effectiveness of retrieval (precision and recall) considerably.

7 Task description

The basic idea in this approach to disambiguation is that a set of manually disambiguated words can be used to predict the sense of new terms. Usually, a representation of word senses is obtained from a training phase, and after that, new terms are compared to word senses making use of a similarity function. In the VSM (Salton & McGill, 1983), each word sense and a word is represented by a weight vector, where each component represents the importance of some term in the word sense or expression. Basically, using the frequency of the terms that appears in a CW (Diaz et al., 1998; Urena et al., 1998) around the word, that is, using its context. We also constructed weight vectors for each word to disambiguate. The similarity between the word and each sense is obtained with the cosine distance, and assigned to the word sense when the similarity value is high.

Training algorithms provide a way to calculate the weight vectors for the word senses. We have selected the Rocchio and the Widrow-Hoff algorithms to compute the term weights for classes in our approach. The first one is an algorithm traditionally used for Relevance Feedback in IR. The second one comes from Machine Learning.

This approach integrates a set of linguistic resources as knowledge sources (Diaz et al., 1998): the training collection SEMCOR (Miller,et al., 1993) and the lexical database WORDNET. The information of SEMCOR has been combined with lexicalsemantic relations from WORDNET (synonymy relation, meronymy, etc.). This combination is performed by the using initial weights for senses word (Urena et al, 1998).

8 Evaluation and conclusions

Evaluation in text classification operations exhibits great heterogeneity. Several metrics and test collections have been used for different approaches or works. The VSM promotes evaluation based on recall and precision.

For the evaluation of our experiments of WSD applied to IR, we have used the TREC (Text REtrieval Conference) (Harman, 1993) test collection, one of the most used in IR and the Smart system (Buckley, 1985).

We have developed a series of experiments to evaluate our approach to WSD and its application to IR tasks. The results of these experiments show, firstly, that the integration of resources is a very effective approach to WSD (92% precision —table 1). On the other hand, the disambiguation of queries (based on the query expansion process through the LDB) in an IR system improves the effectiveness of retrieval (figure 1). This way, we have realised automatically the total process of query expansion in IR, applying WSD.

References

C. Buckley , Implementation of the SMART Information Retrieval System , Cornell University , 85-686 , 1985.

Buenaga, and J.M. Gómez, and B. Díaz , Using WORDNET to Complement Training Information in Text Categorization , Proceedings of Second International Conference on Recent Advances in Natural Language Processing (RANLP), 1997.

J.S. Chang and J.N. Chen, H.H. Sheng and J.S. Ker , Combining Machine Readable Lexical Resources and Bilingual Corpora for Broad Word Sense Disambiguation , Proceedings of the Second Conference of the Association for Machine Translation, 1996.

I. Dagan and A. Itai and U. Schwall , Two Languages Are More Informative than One , Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), 1991.

A. Díaz and M. Buenaga and L.A. Ureña and M. García , Integrating Linguistic Resources in an Uniform Way for Text Classification Tasks, Proceedings of the First International Conference on Language Resources and Evaluation, 1998.

G. Grefenstette, Cross-Language Information Retrieval, PUBLISHER Ed. by G. Grefenstette Kluwer Academic Publishers, 1998.

D. Harman, The First TExt Retrieval Conference (TREC-1) , Information Processing and Management, 29.4, 1993.

D. Harman , Overview of the Forth Text Retrieval Conference (TREC-4), Proceedings of the Fourth Text Retrieval Conference, 1996.

A. Kilgarriff , What is Word Sense Disambiguation Good for?, Proceedings of Natural Language Processing Pacific Rim Symposium, 1997.

G. Miller and C. Leacock and T. Randee and R. Bunker , A Semantic Concordance, Proceedings of the 3rd DARPA Workshop on Human Language Technology, 1993.

Oracle, Managing Text with Oracle8(TM) ConText Cartridge, An Oracle Technical White Paper, 1997.

G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

M. Sanderson, Word sense disambiguation and information retrieval, Department of Computing Science, University of Glasgow, 1996.

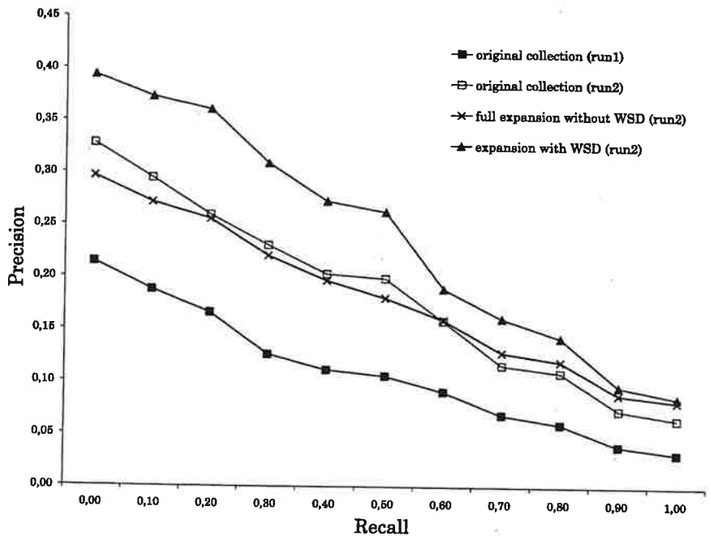
L.A. Ureña and M. Buenaga and M. García and J.M. Gómez , Integrating and evaluating WSD in the adaptation of a lexical database in text categorization task , Proceedings of the First Workshop on Text, Speech, Dialogue –TSD'98– , 1998.

E.M. Voorhees, Query expansion using lexical-semantic relations, Proceedings of the ACM SIGIR, 1994.

D. Yarowsky, Decision List for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French , Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94), 1994.

	Training+WordNet		Baseline
	Rocchio	Widrow-Hoff	
Precision Average	0.905	0.920	0.706

Precision of WSD



Effectiveness for full and WSD expansion with relevance feedback

Marc Weeber, Rein Vos, & Harald Baayen
Word Association Statistics for the Lowest-Frequency Words

It is common practice in word association applications such as information retrieval and collocation extraction systems to discard words with a frequency less than five a priori. This implies that information provided by over 60% of all word types is lost. The arguments for this frequency threshold are two-fold. First, the lowest-frequency words have no practical use in word association applications and second, the statistics to compute the association are problematic. We will show that, for two totally different language data sets and word association tasks, the lowest-frequency words are just as meaningful to these tasks as the words that have a higher frequency. We have therefore investigated the behavior of different word association statistics when applied to the lowest-frequency words.

We used the standard technique of defining a *window* around a seed term and selected those words as potentially relevant terms that appeared more often in these windows than expected under chance conditions. The recent literature has seen some discussion of the appropriate statistical methods for analyzing the 2×2 contingency tables that contain the counts of how a word is distributed inside and outside the windows. The current standard seems to be Mutual Information (Church & Hanks, 1990), however, Dunning (1993) has called attention to the log-likelihood ratio, G^2, as appropriate for the analysis of these tables, especially, when such contingency tables concern very low-frequency words. Kageura (1999) provides additional experimental evidence for this claim. Pedersen et al. (1996) follow-up Dunning's suggestion that Fisher's exact test might be even more appropriate for such contingency tables.

Our primary research objective is to investigate whether there is an optimal window size for two different word association tasks: the extraction of side-effect terms from an English medical abstract corpus and the extraction of Dutch verb-particle combinations from a newspaper corpus. We will show that indeed there seems to be an optimal window size for both G^2 and Fisher's exact test. However, a recurrent pattern of local optima calls this conclusion into question. Upon closer inspection, this recurrent pattern appears at fixed ratios of the number of words inside the window to the number of words outside the window (*complement*). We will relate the recurrent patterns of local optima at fixed window-complement ratios (W/C-ratios) to the distributions of the lowest-frequency words over window and complement.

With these observations, we propose to adjust the word association statistical methods of G^2 and Fisher's exact test by a) excluding the hapax legomena, words occurring only once, a priori, b) considering only the most extreme distributions (i.e., window–complement = 2–0, 3–0, 4–0), and c) adjusting significance levels for the optimal window size.

References

Church, K. W. & Hanks, P. 1990. Word association norms, mutual information, and lexicography. Computational Linguistics, 16 (1), 22–29.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19 (1), 61–74.

Kageura, K. 1999. Bigram statistics revisited: A comparative examination of some statistical measures in morphological analysis of Japanese kanji sequences. Journal of Quantitative Linguistics, 6 (2), 149–166.

Pedersen, T., Kayaalp, M. & Bruce, R. 1996. Significant lexical relationships. In Proceedings of the 13th national conference on artificial intelligence (455–460). AAAI Press / The MIT Press.

Stefan Th. Gries

Particle Movement: A Multifactorial Analysis of Syntactic Variation

A notoriously difficult problem for syntactic research is the existence of so-called allo-sentences, i.e. closely related syntactic variants with truth-conditionally equivalent meanings. Examples in English include the well-known word order alternations Particle Movement, Dative Movement and Preposition Stranding in (1), (2) and (3) respectively.

(1) a. John picked up the book.
    b. John picked the book up.
(2) a. John gave the book to Bill.
    b. John gave Bill the book.
(3) a. Who did you see Bill with?
    b. With whom did you see Bill?

The question arises in what respect the syntactic variants differ from one another, given that they seem to communicate the same truth-conditional meaning. Moreover, what functions do the variants serve and how can we explain the co-existence of two so similar constructions? Finally, which variables govern these alternations? This study investigates these questions for the first of the above-mentioned word order alternation of transitive phrasal verbs, namely the continuous construction in (1a) and the discontinuous construction in (1b).

In the last 80 years, Particle Movement has been extensively studied in a wide variety of approaches from different theoretical perspectives, and about 20 independent variables from different levels of linguistic analysis have been argued to influence speakers' choices for one of the two constructions (cf. Gries 1999:109ff. for a concise review):

- phonological variables such as the stress pattern of the verb phrase (cf., e.g., Van Dongen 1919:352);
- morphosyntactic variables such as the length, the complexity and the determiner of the direct object NP (cf, e.g., Fraser 1966:46; Chen 1986:84);
- semantic variables such as the idiomaticity of the meaning of the VP (cf. Fraser 1974:573; Bolinger 1971:112f., 121ff.);
- pragmatic variables such as givenness of the referent of the direct object NP (cf. Chen 1986:82);
- other variables such as directional adverbials following the transitive phrasal verb construction (cf. Fraser 1974:571f.).

However, in spite of the numerous variables, the analyses suffer from several drawbacks, both specific and general in nature. Most importantly, all previous analyses are only monofactorial in the sense that every single variable is investigated in isolation. For instance, it was argued that, all other things being equal, complex direct object NPs yield a preference of the continuous construction (cf. (4)); also, it was suggested that following directional adverbials (such as directional PPs) yield a preference for the discontinuous construction (cf. (5)).

(4) a. John picked up the large brown book of his father.
    b. ??John picked the large brown book of his father up.
(5) a. ?John picked up the book from the table.
    b. John picked the book up from the table.

But what is problematic about this approach? Is this not an example of one of the most traditional and well-established methods in linguistics, namely the minimal-pair test? The problem lies in the fact that examples such as (5) have been used to support the claim that the following directional PP is more natural in the discontinuous construction. However, the example does not warrant this claim at all: the preference for the discontinuous construction in (5) might as well derive from the fact that short and simple direct

objects already favour the discontinuous construction, as do definite determiners and literal VP meanings. Therefore, the intuitive/introspective acceptability judgements is (5) need not result from the variable that the example is purported to support. If we generalise from (5) to similar analyses we find that, given the complexity of 20 or so interacting variables, we cannot rely on monofactorial analyses in order to describe Particle Movement adequately. Additionally, not a single monofactorial analysis from the literature enables us to make any claims as to the strength of the variables on the choice of construction. What is more, the monofactorial analyses make it impossible to arrive at a cognitively realistic account of Particle Movement since, for native speakers, all of the variables are given at the same time rather than in isolation (which is what monofactorial analyses imply). Lastly, these methods do not allow the analyst to predict speakers' choices of constructions in order to rigorously test our knowledge of the alternation. In other words, more complex techniques are needed that tackle these problems.

This study will show how all of these weaknesses can be overcome by using (multiple) contingency tables, correlational measures and discriminant analysis on a set of corpus data from the British National Corpus (403 sentences with 20 clauses context). I will show

- how the degree of importance of every single variable can be determined;
- how a hypothesis related to utterance processing subsumes all of the variables and can be supported in a multifactorial context;
- how the choices of construction by native speakers in discourse can be very accurately predicted (prediction accuracy ¿ 80%);
- how prototypical cases can be objectively determined.

In sum, this study shows that the analysis of syntactic variation in cognitive/functional linguistics can benefit from multidimensional/multifactorial analyses in very much the same way that the analysis of register variation and the English genitive has profited from Biber's (1988, 1995) work or from Leech, Francis and Xu's (1994) multifactorial study.

References

Biber, Douglas. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. Dimensions of Register Variation. Cambridge: Cambridge University Press.

Bolinger, Dwight. 1971. The Phrasal Verb in English. Cambridge Harvard University Press.

Chen, Ping. 1986. Discourse and Particle Movement in English Studies in Language 10:79-95.

Fraser, Bruce. 1966. Some Remarks on the Verb-Particle Construction in English. In: Dinneen Francis P. Problems in Semantics, History of Linguistics, Linguistics and English. Washington, DC: Georgetown University Press, p. 45-61.

Fraser, Bruce. 1974. The Phrasal Verb in English. By Dwight Bolinger Language 50:568-575.

Leech, Geoffrey, B. Francis and X. Xu. 1994. The Use of Computer Corpora in the Textual Demonstrability of Gradience in Linguistic Categories. In: Fuchs, C. and B. Vitorri (eds.). Continuity in Linguistic Semantics. Amsterdam, Philadelphia: John Benjamins.

Van Dongen, W. A. 1919. He puts on his hat and He puts his hat on. Neophilologus 4:322-353.

Sheila M. Embleton and Eric S. Wheeler

Computerized Dialect Atlas of Finnish: Dealing with Ambiguity

To enable quantitative studies of large volumes of data, it is often appropriate to create machine-readable forms of existing printed works. We have undertaken such a project (Embleton and Wheeler 1997) for Finnish using an important, but out-of-print, dialect atlas (Kettunen 1940), and have reached a stage where the primary data entry has been completed. Next, we need to confirm the accuracy of the data entry in a way that is both efficient for us and still convincing to a potential user of the data or other outside party.

We describe our testing protocol, testing tools and the practical concerns of selecting appropriate sample sizes for statistically-based tests.

A critical issue, however, is the inherent ambiguity in the data itself. Because the original dialect atlas used typographic conventions for marking dialect areas, the delineation of these areas has a different precision than the digital form.

For example, Village A may be on the edge of an area marked with X's, and on the edge of an area marked by O's, but not definitely inside or outside either or both areas. For the atlas reader, the marginal relationship of the village to each of the two dialect features is obvious. However, in digitalizing the map (with the categories we have chosen), it is necessary to assign the village to "X" or "not X", and to "O" or "not O".

We outline our approach to resolving these issues for Finnish. However, we note that the problem is much more general, and needs to be considered in the design of any such conversion of data for quantitative study.

References

Embleton, Sheila M. and Eric S. Wheeler. (1997). Finnish Dialect Atlas for Quantitative Studies. JQL 1997 4.1-3, 99–102.

Kettunen, L. (1940). Suomen murrekartasto [The Dialect atlas of Finland]. Helsinki: Suomalaisen kirjallisuuden seura.

Michael P. Oakes

Computer Estimation of Vocabulary in a Protolanguage from Word Lists in Four Daughter Languages.

Words in genetically related languages which appear to be derived from a common original form are said to be cognate with each other, and both are reflexes of the same form in the protolanguage or common ancestor language. Even if no written records of the protolanguage remain, it is possible to estimate what some of the words in that language might have been, by comparison of their reflexes in the more recent daughter languages. This method of protolanguage reconstruction is called the "Comparative Method", and is described by Crowley (1992, Chapter 5). Although long practised by human linguists, the Comparative Method is extremely time consuming, and only a few parts of the process have previously been automated (Frantz 1970, Damerau 1975, Guy 1994, Lowe & Mazaudon 1994). This paper describes a program which attempts to replicate the methodology described by Crowley almost in its entirety.

Word lists for each of four daughter languages are obtained, such as those given by Nothofer (1975) for four Malayo-Polynesian languages. The first task for the computer is then to select which words are cognate across all four daughter languages. The method used here is that of McEnery & Oakes (1996), where empirical data is used to determine the number of transformations (insertions, deletions or substitutions) or Levenshtein distance between a pair of words above which the word pair is probably not cognate.

The technique of dynamic programming is then used to align all pairs of cognate words at the character level in a pairwise comparison of each of the four daughter language word lists. For example, in comparing the Malay and Tagalog cognate words "telur" and "itlog" (meaning "egg"), the following sound correspondences were observed: NULL → i, t → t, e → NULL, l → l, u → o, and r → g. The sound changes discovered in this way are collated, and each sound change occurring a threshold number of times is deemed to be regular.

By comparing the tables of regular sound changes between each pair of languages, it is possible to identify instances where a sound change is regular across all four languages, such as a "t" in Tongan remaining as "t" in Samoan and Rarotongan, but being replaced by a "k" in Hawaiian, as in the word forms "tapu" and "kapu" (meaning "forbidden"). The lists of sound correspondences which hold true across all four languages are then used to derive the corrresponding sounds in the protolanguage using four principles given by Crowley (Chapter 5, 1992), as follows:

1. Proposed reconstructions must involve sound changes that are plausible. A number of types of sound change commonly occur throughout the world (Crowley, Chapter 2, 1992) such as "lenition", where a "stronger" version of a sound such as "b" becomes replaced by a weaker sound (phoneme) such as "v" (e.g. the Malay "batu" and the Fijian "vatu", meaning "stone").

2. Look for sound correspondences that involve phonetically similar sounds. A table classifying phonemes according to manner of production, position of production and vocalisation or otherwise forms the basis of a matrix of similarity between the phonemes.

3. Reconstructions should involve as few changes as possible between the protolanguage and the daughter langauges. Taking the above example where the corresponding sounds in the daughter languages are "t", "t", "t" and "k", the corresponding sound in the protolanguage would probably be "t", since that would only involve one change with respect to the daughter language.

4. A phoneme should not be proposed for the protolanguage if it is not found in any of the daughter languages.

These four principles are incorporated in a module which estimates the total degree of change (cost) which would have occurred between a proposed phoneme for the protolanguage with each of the corresponding phonemes in the daughter language. A cost of 1 or 2 is given for a sound change according to whether it is favoured by principles (1) and (2) or not. If no sound change has occurred at all between the protolangauge and a daughter langauge, the cost is 0.

At present, the program is no substitute for the experience of the human comparative linguist. In particular, the two other principles given by Crowley have not been implemented, namely those regarding the maintenance of balanced phonological systems and conditioned (such as position dependent) sound changes.

In a separate program, the vocabulary lists for each daughter language are read in again, and each phoneme in each daughter language word is substituted for the corresponding phoneme in the protolanguage if this is known. If these substitutions generate the same sequence of protolangauge phonemes for all

four daughter languages, it can be assumed that the original word in the protolanguage has been plausibly reconstructed.

References.

Crowley, T (1992). An Introduction to Historical Linguistics. Oxford University Press: Oxford.

Nothofer, B (1975). Reconstruction of Proto-Malayo-Javanic. Martinus Nijhoff: 'S-Gravenhage.

Guy, J B M (1994). An Algorithm for Identifying Cognates in Bilingual Word Lists and its Applicability to Machine Translation. Journal of Quantitative Linguistics 1(1), pp 35-42.

Lowe, J B and Mazaudon, M (1994). The Reconstruction Engine: A Computer Implementation of the Comparative Method. Computational Linguistics 20, pp 381-417.

Damerau, F J (1975). Mechanization of Cognate Recognition in Comparative Linguistics. Linguistics 148, pps 5-29.

Frantz, D G (1970). A PL/1 Program to Assist the Comparative Linguist. Communications of the ACM 13(6), pp 353-356.

McEnery A M and Oakes M P (1996). Sentence and Word Alignment in the CRATER project. In Thomas, J and Short, M (eds), Using Corpora for Language Research. Longman: London.

Peter Grzybek

Remarks on the Sentence Length of Proverbs

This paper is related to a research project on quantitative aspects of proverbs. The issue to be discussed here focuses primarily on the question of sentence length and sentence length distribution of proverbs. Proverbs are an interesting object for linguistic studies, since they represent a specific text type on the sentence level, which displays particular features due to its stereotypical form. In traditional paremiology (proverb scholarship), there are quite a number of studies on the sentence length of proverbs from various languages; all these studies have calculated sentence length on the basis of the number of words per sentence. Although most of these studies have been restricted to rather simple statistical procedures such as the calculation of the mean length of proverbial sentences (often not even standard deviations along with mean lengths), partly far-reaching conclusions have been drawn on rather weak evidence, which represent the paremiological state of the art. In this paper, an initial report is given on the status of paremiological studies on proverbs sentence length. In an attempt to relate the question of proverbs sentence length to approaches from quantitative linguistics, theoretical models of sentence length distributions will be discussed, which have been empirically derived and/or theoretically postulated. However, these models have been brought forth with regard to sentences as integral elements of texts, which is not the case with proverbs. Therefore, it is interesting to see how these models fit for the theoretical modeling of proverbs' sentence length. On the basis of more than a dozen of analyses from proverbs of different languages (including Croatian, Estonian, German, Hungarian, Russian, Slovenian, Turkish, and others) it will be shown by way of empirical evidence that there is a common model which fits for proverbs from all these languages. This model is different from the models hypothetically postulated for sentence length distributions in 'ordinary' texts from these languages. Various possible reasons for this divergence will be discussed, such as, e.g.: in case of a proverb collection, we are concerned with heterogeneous data, since a proverb corpus represents which uses to be called a 'quasi text'; proverbs display specific linguistic structures, due to economic processes in the shaping of stereotypical texts; proverbs represent a possibility to study linguistic material up to the sentence level, without the interference of text parameters. a proverb corpus represents 'paradigmatic' linguistic data, similar to a lexicon, which is not subject to syntagmatic processes.

Authors by first author, address of first author, e-mail addresses of all authors, and page number.

Setsuko Wakabayashi, Jun-ya Morishita, Yasunori Motomura Himeji Dokkyo University, 7-2-1 Kamiohno, Himeji 670-8524, Japan. *setsuko@himeji-du.ac.jp* 13

Andy Way, School of Computer Applications, Dublin City University, Dublin 9, Ireland. *away@compapp.dcu.ie* 30

Marc Weeber, Rein Vos, Harald Baayen Dept. Social Pharmacy and Pharmacoepidemiology, A. Deusinglaan 2, 9713 AW Groningen. *marc@farm.rug.nl, rein.vos@zw.unimaas.nl, baayen@mpi.nl* 83

Victor Zakharov, Burenina St., 1-2-158, 195253 Saint-Petersburg, Russia. *vz@laz.usr.pu.ru* 78