Linguistic Challenges for Computationalists

John Nerbonne*
Humanities Computing
University of Groningen
NL 9700 AS Groningen, The Netherlands
j.nerbonne@rug.nl

Abstract

Even now techniques are in common use in computational linguistics which could lead to important advances in pure linguistics, especially language acquisition and the study of language variation, if they were applied with intelligence and persistence. Reliable techniques for assaying similarities and differences among linguistic varieties are useful not only in dialectology and sociolinguistics, but could also be valuable in studies of first and second language learning and in the study of language contact. These techniques would be even more valuable if they indicated relative degrees of similarity, but also the source of deviation (contamination). Given the current tendency in linguistics to wish to confront the data of language use more directly, techniques are needed which can handle large amounts of noisy data and extract reliable measures from them. The current focus in Computational Linguistics on useful applications is a very good thing, but some further attention to the linguistic use of computational techniques would be very rewarding.

1 Introduction

The goal of this paper is to urge computational linguists to explore issues in other branches of linguistics more broadly. Computational linguistics (CL) has developed an impressive array of analytical techniques, especially in the past decade and a half, techniques which are capable of assaying linguistic structure of various levels from fairly raw textual data. The goal will be to note how these techniques might be applied to illuminate other issues of broad interest in linguistics.

The thesis my plea is based on—that there are opportunities for computational contributions to "pure" linguistics—is not absolutely new, of course, as many computational linguists have been involved in issues of pure linguistics as well, including especially grammatical theory. And we will naturally attempt to identify such work as we become more concrete (below). We aim to spark discussion by identifying less discussed areas where computational forays appear promising, and in fact, we will not dwell on grammatical theory at all.

It is best to add some *caveats*. First, the sort of appeal we aim at can only be successful if it is sketched with some concrete detail. If we attempted to argue the usefulness of computational techniques to general linguistic theory very abstractly, virtually everyone would react, "Fine, but how can we contribute more concretely?" But we can only provide more concrete detail on a very limited number of subjects. Of course, we are limited by our knowledge of these subjects as well, but the first *caveat* is that this little essay cannot be exhaustive, only suggestive. We should be delighted to hear promptly of several further areas of application for computational techniques we omit here.

Second, the exhortation to explore issues in other branches of linguistics more broadly takes the form of an examination of selected issues in non-computational linguistics together with suggestions on how computational techniques might shed added light on them. Since the survey is to be brief, the suggestions about solutions—or perhaps, merely perspectives—of necessity will also be brief. In particular, they will be no more than *suggestions*, and will make no pretense at demonstrating anything at all.

Third, we might be misconstrued as urging you to ignore useful, money-making applications in favor of dedicating yourselves to the higher goal of collaborating in the search for scientific truth. But both the history of CL and the usual modern attitude of scientists toward applications convinces me that the application-oriented side of CL is very important and eminently worthwhile. Per-

*We are grateful to the Netherlands Organization for Scientific Research, NWO, for support (project "Determinants of Dialect Variation, 360-70-120, P.I. J. Nerbonne). It was stimulating to discuss the general issue of engineering work feeding back into pure science with Stuart Schieber, who organized a course with Michael Collins of MIT at the Linguistics Institute of the Linguistic Society of America at MIT, Summer 2005 contrasting science and engineering in CL.

haps you should indeed turn a deaf ear to the seductions of filthy mammon and consecrate yourself to a life of (pure) science, but this is a matter between you and your clergyman (or analyst). You have not gotten this advice here.

Fourth, and finally, we might be viewed as advocating different sorts of APPLICATIONS, namely the application of techniques from one linguistic subfield (CL) to another (dialectology, etc.). In this sense modern genetics APPLIES techniques from chemistry to biological molecules to determine the physical basis of inheritance, anthropology APPLIES techniques from nuclear chemistry (carbon dating) to date human artefacts, and astronomy APPLIES techniques from optics (glass) and electromagnetism (radio astronomy) to map the heavens. In all of these case is the primary motivation is scientific curiosity, not utilitarian, and this view is indeed parallel to the step advocated here.

2 Computational Linguistics

Computational Linguistics (CL) is often characterized as having a theoretical and an application-oriented, or engineering side (Joshi 99; Kay 02). The theoretical side of CL is concerned with processes involving language and their abstract computational characterization, including processes such as analyzing (parsing), and producing (generating) language, but also storing, compressing, indexing, searching, sorting, learning and accessing language. The computational characterization of these processes involves investigating algorithms for their accuracy and time and space requirements, finding appropriate data structures, and naturally testing these ideas, where possible, against concrete implementations.

The application-oriented, or engineering side of the field concerns itself with creating useful computational systems which involve language manipulation in some way, e.g. lexicography tools; speech understanding (in collaboration with speech recognition); machine translation, including translation aids such as translation memories, multilingual alignment, and specialized lexicon construction; speech synthesis, especially intonation; term extraction, information retrieval, document summarization, data (text) mining, and question answering; telephone information systems and natural language interfaces; automatic dictionary and thesaurus access, grammar checking, including spell-checking; document management, authoring (especially in multi-author systems), and conformance to specifications in so-called "controlled language systems"; foreign language aids (such as access to bilingual dictionaries), foreign language tutoring systems, and communication aids (for the handicapped). See Cole et al. (1996) for further discussion of these, and other areas of application for language technology.

We have been overly compulsive about listing the engineering activities not only to remind the reader how extensive these are, but also to emphasize that the breadth of these activities would be unthinkable if it were not for a rich "infrastructure" of language technology tools which the field is constantly creating. For the most part the techniques we urge you to apply more broadly have been developed in order to build better and more varied applications, as this has been the great motor in the recent dynamics of computational linguistics. But some of the techniques have also been useful in theoretical computational linguistics, and the distinction will play no role here. In fact, perhaps the simplest view is to acknowledge that applications and theory make use of common technology, a sort of technical infrastructure, and to emphasize the opportunities this provides.

3 Dialectology

We shall examine dialectology first because it is an area we have directly worked in, and for which we therefore need to rely less on speculation about the potential benefits of a computational approach. Given the greater amount of direct experience with this work, we may use it to distill some of the characteristics we need to seek in other areas in which computational techniques might be promising.

Dialectology studies the patterns of variation in a language and especially its geographic conditioning (Chambers & Trudgill 80). In London people say [wɒtə] for 'water', with a voiceless [t] and no trace of final [r], in New York most people say [wɑɾɻ], with a "tapped" [t], and in Boston [wɑɾə]. These differences are systematic, but not exceptionless, and they appear to involve potentially every level of linguistic structure, pronunciation, morphology, lexicon, syntax, and discourse. Because differences appear to involve exceptions, it is advantageous to process a great deal of mate-

rial and to apply statistical techniques to the analysis. Fortunately, dialectologists have been assiduous in collecting and archiving a great deal of data, especially involving pronunciation and lexical differences.

Once we have agreed that we need to subject a great deal of data to systematic analysis, we have *a fortiori* accepted the need for automating the analysis, and since it is linguistic material, it would be strange if this did not lead us to computational linguistics. In fact EDIT DISTANCE, well-known to computational linguists by its wide variety of applications, may be applied fairly directly to the phonetic transcripts of dialect pronunciations (Nerbonne *et al.* 99). The application of edit distance to pronunciation transcripts yields, for each pair of words, at each pair of field work sites, a numerical characterization of the difference. Because pronunciation differences are characterized numerically, we thereby initiate a numerical analysis of data that dialectologists had normally regarded as categorical—with all the advantages which normally accrue to numerical data analysis.

Nerbonne (2003) discusses at greater length the computational issues in analyzing, presenting and evaluating dialectological analyses, including those which go beyond pronunciation. These issues include the use of lemmatizers or stemmers to clean up word-form data for lexical analysis, raising the edit distance from strings to sets of strings in order treat data collections with alternative forms, and the proper treatment of frequency in detection of linguistic proximity. Opportunities for the application of standard CL techniques in computational linguistics abound. Heeringa (2004) summarizes current thinking on measuring dialectal pronunciation differences, including the thorny issue of evaluating the quality of results. Figure 1 illustrates the results of applying these techniques to Bulgarian data.

It is important to report here, as well, that specialists in dialectology—and not only computational linguists—are enthusiastic about the deployment of computational tools. A common remark by dialectologists is that that the new techniques allow a more comprehensive inclusion of all available data, effectively answering earlier complaints that analyses of dialect areas and/or dialect continua relied too extensively on the analysts' choice of material. William

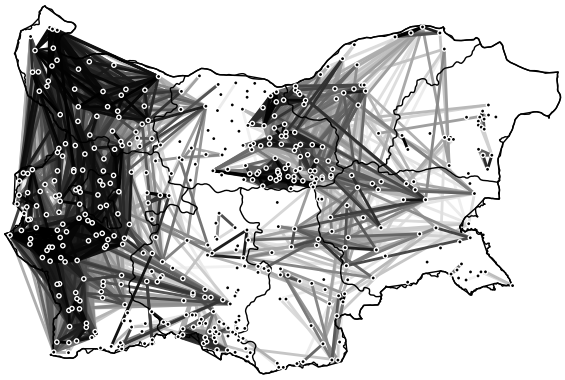

Figure 1: In this line map the average Levenshtein distances between 490 Bulgarian dialects are shown for 36 words. Darker lines join varieties with more similar pronunciations, while lighter lines indicate more dissimilar ones. From collaborative work in progress with Petya Osenova, Bulgarian Academy of Science, and Wilbert Heeringa, Groningen.

Kretzschmar leads the American *Linguistic Atlas Projects* (LAP), and has collaborated in various analyses and workshops (Nerbonne & Kretzschmar 03). He has *inter alia* included a pointer to CL work on the home page of the LAP site he maintains at `http://us.english.uga.edu/`, and he is presently collaborating on a project to publish a second volume of papers focused on computational techniques (Nerbonne & Kretzschmar 06).

Finally, let us note that the computational step may introduce such genuinely novel opportunities that we find ourselves in a position to ask questions which simply lay beyond earlier methodology. Given our numerical perspective on dialect difference, we may e.g. ask, via a regression analysis, how much of the aggregate varietal difference is explained by geography, or whether travel time is a superior characterization of the geography relevant to linguistic variation (Gooskens 04), or whether larger settlements tend to share linguistic variants more than smaller ones—something one might expect if variation diffused via social contact (Heeringa & Nerbonne 02). The introduction of CL techniques enables us to ask more abstract questions in a way we can still link to concrete linguistic analysis.

This work also suggests many related paths of exploration. For example, even if a distance measure allows the mapping of the dialectolog-

ical landscape well, it seems ill-equipped to assay one extreme result of dialect differentiation, i.e. the failure of comprehensibility. The reason for this failure is the fact the comprehensibility is not symmetrical, while linguistic distance by definition is: it may reliably be the case that speakers of one variety understand the speaker of another better than *vice versa*. For example, Dutch speakers find it easier to understand Afrikaans than vice versa (Gooskens & vanBezooijen 06). If this is due to language differences, it calls for the development of an asymmetrical measure of the relative difficulty of mapping from one language to another, or something similar.[1]

The computational work has been successful in dialectology because there were large reservoirs of linguistics data to which analyses could be applied, i.e., dialect atlases, because distinguishing properties resisted simple categorical characterization, and naturally because there were promising computational techniques for getting at the crucial phenomena.

As we turn to other areas, we shall ask ourselves whether we are likely to satisfy these desiderata. When even one is missing, the result can be disappointing. For example, sociolinguistics has largely succeeded dialectology in attracting scholarly interest. The linguistic issues are not wildly different—different social groups use different language varieties, and these may differ in all the ways in which geographical varieties do (pronunciation, lexicon, etc.). It would be straightforward and interesting to apply the techniques sketch above to linguistic varieties associated with different social groups. But there is no tradition in sociolinguistics like that of the dialect atlas, i.e. collecting speech samples from a large set of sociolects. So the opportunity does not present itself.

4 Diachronic Linguistics

Diachronic linguistics investigates how languages change, and, most spectacularly, how a single language many evolve into many related ones. It regularly attracts a good deal of scholarly attention (Gray & Atkinson 03; Eska & Ringe 04) as computational biologists have applied their techniques for tracking genetic evolution to linguistic data. Although the scholarship is at times forbidding in its expectations about philological expertise, the problem appears to allow neat enough formulations so that one may be optimistic about computational investigations.

Essentially, we are given a set of cognate words in several putatively related languages, and we construct hypotheses about the most recent common ancestor—the protolanguage—as well as a simple set of sound changes leading from the protolanguage to the individual descendants. For example, we note that the word for father has an initial /f/ in Germanic (English *father*), /p/ in Romance, Greek and Indic (French *père*, Greek *patera*, and Hindi *pitā*), and no initial consonant in some Celtic languages (Irish *athair*). This suggests that we postulate a /p/ in the protolanguage and changes from /p/ to /f/ for Germanic and /p/ to $\emptyset$ for the relevant Celtic varieties. But we gain confidence in these postulates only when the same rules are shown to operate on other forms, i.e. when the correspondences recur (as the p/f/$\emptyset$ definitely does). It is surprising that CL should turn over to the biologists such a well-structured problem in linguistic computation.[2]

Our community has contributed to this area, especially Brett Kessler, who investigated how to test when sound correspondences exceed chance levels (Kessler 01), and Grzegorz Kondrak, who modified the edit distance algorithm mentioned above, in order to identify cognates, align them, and on that basis postulate recurrent sound combinations (Kondrak 02). But these studies deserve follow-ups, tests on new data, and extensions to other problems. Among many remaining problems we note that it would be valuable to detect borrowed words, which should not figure in cognate lists, but which suggest interesting influence; to operationalize the notion of semantic relatedness relevant to cognate recognition; to quantify how regular sound change is; or to investigate the level of morphology, which is regarded as especially probative in historical reconstruction. But we emphasize that there are likely to be interesting opportunities for contributions with respect to detail as well, perhaps in the construction of instruments to examine data more insightfully, to measure hypothesized aspects, or to quantify the empirical base on which historical

[1] Nathan Vaillette, University of Massachusetts has explored this problem using relative entropy in unpublished work.

[2] See also Benedetto et al. (2002) for attempt to reconstruct linguistic history using relative entropy, but especially Goodman (2002) for criticism of Benedetto.

hypotheses are made.

5 Language Acquisition

Studies of children's acquisition of language are interesting to all sorts of inquiries because language is a defining characteristic of us as humans. They occupy an important position in linguistics due to the linguistic argument that innate, specifically linguistic mechanisms must be postulated to account for acquisition (Pinker 94, Chap. 9). The innate organizing principles of language are postulated to be part of human genetic constitution, and therefore the source of universal properties which all languages share. At the same time psychologists have shown that some acquisition is mediated by sensitivity to statistical trends in data (Saffran *et al.* 99). And children naturally need minimally to learn which of all the languages they are genetically predisposed toward is the one in use locally. Finally, CL has explored machine learning techniques extensively over the past decade (Manning & Schütze 99). Surely CL is positioned to contribute crucially to this scientific discussion with interesting implemented models of specific phenomena, and in particular with models aimed at broader coverage or so one would think.

On the other hand, machine learning techniques do not translate to computational models of acquisition very directly, at least not as normally used by CL, namely to optimize performance on technical tasks that may have no interesting parallel in a child's acquisition of language, e.g. the task of recognizing named entities, persons, places and organizations. In addition, even idealized simulations of acquisition might wish to impose restrictions on the sort of mechanisms to be used, e.g. that they may apply incrementally, and on the input data, e.g. that it reflect children's experience.

Fortunately, these differences in tasks, mechanisms and input data may be overcome, and CL has not been inactive in examining language acquisition. Brent (1997) is an early collection of articles on computational approaches to language acquisition, including especially Brent's own work applying minimal description length to the problem of segmenting the speech stream into words, and using only phonotactic and distributional information (Brent 99b; Brent 99a). There have been a number of other studies focusing on phonotactics (Nerbonne & Stoianov 04; Nerbonne & Konstantopoulos 04), the acquisition of morphophonemic rules (Gildea & Jurafsky 96; Albright & Hayes 03), morphology (Goldsmith 01), and syntax (Niyogi & Berwick 96). Albright and Hayes's work is especially worth recommending to a CL audience as it is clear and explicit about linguistic concerns in modeling acquisition computationally.

Most relevant to the sort of CL contribution I have in mind is the series of workshops organized by William Gregory Sakas of CUNY, *Psycho-Computational Models of Human Language Acquisition*. The first took place in 2004 in Geneva in coordination with COLING and the second in 2005 in Ann Arbor in coordination with the ACL's special interest group on natural language learning (`http://www.colag.cs.hunter.cuny.edu/psychocomp/`). It is clear from the proceedings of these workshops that new syntheses of linguistic, psychological and computational perspectives enjoy a good deal of interest (Yang 04).

It is also clear that there is an enormous interest in further questions about segmentation, alignment, constituency, local and long-distance relations, modification, and ill-formed input in addition to the usual questions about the generality of solutions wit respect to various language types.

Finally, it is worth emphasizing here more than elsewhere that contributions need not take the form of simulations of human learning (even if this is the case for most of the studies cited). There is great potential interest in characterizing easy vs. difficult material, in what happens when second and third languages are learned (contamination), and in how languages are lost. In addition to simulation, we should also be thinking of how to operationalize measures of language proficiency that could use speech as directly as possible. At the moment, extremely crude measures such as mean length of utterance (MLU) and type/token ratio enjoy great popularity, but one suspects that this is due more to their ease of computation than to their reflection of linguistic sophistication. Ideally we should like to automate our detection of the mastery of various linguistic structures, rules and exceptions. That is clearly a long way off in its full generality, but perhaps realizable in some instances with standard techniques.

6 Language Contact

Language contact study is an active branch of linguistics focused on recognizing and analyzing the ways in which languages borrow from one another (Thomason & Kaufmann 88; van Coetsem 88). It is growing in popularity, perhaps due to increases in mobility and the realization that multilingual speakers often, albeit unconsciously, impose the structures of one language on another. Mufwene (2001) urges us to view extreme contact effects such as koinéization, creolization and pidginization as various degrees to which language mixtures may develop (instead of as the results of very different processes, as earlier scholarship had held). Language contact study is, moreover, linked to second-language acquisition in an obvious way: if second-language speakers habitually impose elements of their native language onto another, then those element are good candidates for long-term borrowing whenever these languages are in contact.

It might seem as if we could use the same tools for the study of contact effects that we developed for dialectology. After all, if one variety of a language adopts elements of another, it should become more similar. Indeed given the sort of data in dialect atlases, one can perform these analyses and determine the convergence of some varieties toward a putative source of contamination, at least the convergence with respect to other varieties (Heeringa *et al.* 00; Gooskens & Heeringa 04). Furthermore, one could examine the role of geography in this convergence.

But language contact data collections are not usually designed as dialect atlases, with a number of distinct collection sites, and a controlled set of linguistic variables to be assayed. Recently, we obtained data of a rather different sort, and set ourselves the task of developing computational tools for its analysis.[3] Watson collected recordings of Finnish emigrants to Australia in the mid 1990's (Watson 96), and this group could be divided into adult emigrants and child emigrants, using puberty (16 years old) as the dividing line. The challenge was the development of a technique to determine whether there were significant changes in the syntax of the two groups.

[3]What follows is an informal synopsis of work in progress being conducted with Wybo Wiersma of Groningen and Lisa Lena Opas-Hänninen, Timo Lauttamus and Pekka Hirvonen of Oulu University.

Following an obvious tack from CL, we settled on using n-grams of part-of-speech tags (POS tags) assigned by the TnT tagger (Brants 00) as a probe to determine syntactic similarity. In order not to be swamped by fine distinctions we used trigrams of a small tag set (50 tags). Up to this point we were rediscovering an idea others had introduced (Aarts & Granger 98). To compare one corpus with another, we measured the difference in the two vectors of trigram frequencies using cosine (*inter alia*). To determine whether the difference is statistically significant, we applied permutation-based statistics, roughly resampling the union of the two data sets (using some complicated normalizations) and checking the degree of difference. A difference is significant at the level $p < p'$ iff it is among the most extreme p' fraction of the resampled data.

Because the technique is still under development, we cannot yet report much more. The differences are indeed statistically significant, which, in itself, is not surprising. The corpora are quite raw, however, so that the differences we are finding to-date are dominated by hesitation noises and errors in tagging. The promise is in the technique. If we have succeeded in developing an automated measure of syntactic difference, we have opportunities for application to a host of further questions about syntactic differences, e.g., about where these differences are detectable, and where not; about the time course of contamination effects (do second-language learners keep improving, or is there a ceiling effect?); and about the role of the source language in the degree of contamination. Some crucial computational questions would remain, however, concerning detecting the source of contamination.

7 Other Areas

As noted in the introduction, this brief survey has tried to develop a few ideas in order to convince you that there are promising lines of inquiry for computationalists who would seek to contribute to a broader range of linguistic subfields. We suspect that there are many other areas, as well.

We have deliberately omitted grammatical theory from the list of potential near-term adopters of computational techniques. There are two reasons for eschewing a sub-focus on grammar here, the first being the fact that the potential relevance of computational work to grammatical the-

ory has been recognized for a long time, as grammar has been cited since the earliest days of CL as a likely beneficiary of closer engagement (Kay 02). But second, even as computational grammar studies uncover new means of contributing to the study of pure grammar (van Noord 04), it seems to be a minority of grammarians who recognize the value of computational work. Many researchers have explored this avenue, but the situation has stabilized to one in which computational work is pursued vigorously by small specialized groups (Head-Driven Phrase Structure Grammar and Lexical-Functional Grammar), and largely ignored by most non-mainstream grammarians. We deplore this situation as do others (Pollard 93), but it unfortunately appears to be quite stable.

In addition to the areas discussed above, it is easily imaginable that CL techniques could play an interesting role in a number of other linguistic subareas. As databases of linguistic typology become more detailed and more comprehensive, they should become attractive targets for data-mining techniques (`http://www-uilots.let.uu.nl/td/`). Psycholinguistic studies of processing are promising because they provide a good deal of empirical data. We shall be content with a single example. Moscoso del Prado Martin (2003) reviews a large number of studies relating the difficulty of processing complex word forms, i.e., those involving inflectional and/or derivational structure to the "family size" of a word form, i.e. how many other word forms are related to it. He is able to show that a simple characterization of family size and frequency due to information theory correlates highly with processing difficulty.

8 Conclusions

We have urged computational linguists to consider how much they might contribute to curiosity-driven research into language, i.e. linguistic theory, focusing on examples in dialectology, diachronic linguistics, language acquisition and language contact. We have suggested that there are many avenues to pursue for those with a broader interest in language, and also that the tools and training one receives in developing language technology will be of direct use. We have not suggested that contributions in pure science are any easier or harder to make, and the experience has been general that the dynamics involved in pursuing non-applied goals are every bit as demanding, and every bit at provocative: a successful effort invariably suggests new questions and new avenues to explore.

We have been careful to avoid deprecating application-research and, at the risk of repetition, restate that the development of useful applications is a most valuable aspect of current CL. We encourage colleagues to think of both channels of activity rather than to force a choice of one over the other.

If we are right that most of the interesting techniques for exploring issues in non-computational linguistics have arisen through the development of techniques for engineering activities, then we may have another case where applied science furthers the progress of pure science (Burke 85). In making this remark, we are reneging on the promise in Section 2 not to concern ourselves with whether a particular technique originated in theoretical vs. applied CL, but given the preponderance of applied work in CL, it would be surprising if it were not true in many instance that techniques from engineering were being conscripted for work in theory.

The use of a stemmer to extract lexical differences from lists of word forms in dialectology (Nerbonne & Kleiweg 03) is an example of the sort of contribution where a technique developed only for application purposes could be put to a purely scientific use, that of detecting lexical overlap across a dialect continuum. The Porter stemmer which was used for this purpose is not to be confused with a genuine lemmatizer, which is interesting both linguistically and practically. But it usually reduces word forms to the same stem when they in fact are elements of the same inflectional paradigm. It was developed for use in information retrieval (Porter 80), not for the purpose of exploring linguistic structure or its processing, but its use in dialectology has no ambitions toward practical application.

This would appear to bea genuine case of an engineering technique serving a purpose in curiosity-driven research. To the extent CL is involved in other pure science (beyond CL proper), this sort of cross-fertilization must be standard. Only time will tell whether it will remain true of future computational forays into pure linguistics.

References

(Aarts & Granger 98) Jan Aarts and Sylviane Granger. Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In Sylviane Granger, editor, *Learner English on Computer*, pages 132–141. Longman, London, 1998.

(Albright & Hayes 03) Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90:119–161, 2003.

(Benedetto *et al.* 02) Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language trees and zipping. *Physical Review Letters*, 88(4):048702, 2002.

(Brants 00) Thorsten Brants. TnT — a statistical part of speech tagger. In *6th Applied Natural Language Processing Conference*, pages 224–231, Seattle, 2000. ACL.

(Brent 97) Michael Brent, editor. *Computational Approaches to Language Acquisition*. MIT Press, Cambridge, 1997.

(Brent 99a) Michael Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning Journal*, 34:71–106, 1999.

(Brent 99b) Michael Brent. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science*, 3:294–301, 1999.

(Burke 85) James Burke. *The Day the Universe Changed*. Little, Brown & Co., Boston, 1985.

(Chambers & Trudgill 80) J.K. Chambers and Peter Trudgill. *Dialectology*. Cambridge University Press, Cambridge, 1998, [1980].

(Cole *et al.* 96) Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue. *Survey of the State of the Art in Human Language Technology*. National Science Foundation and European Commission, `www.cse.ogi.edu/CSLU/HLTsurvey/`, 1996.

(Eska & Ringe 04) Joseph F. Eska and Don Ringe. Recent work in computational linguistic phylogeny. *Language*, 80(3):569–582, 2004.

(Gildea & Jurafsky 96) Daniel Gildea and Daniel Jurafsky. Learning bias and phonological rule induction. *Computational Linguistics*, 22(4):497–530, 1996.

(Goldsmith 01) John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

(Goodman 02) Joshua Goodman. Extended comment on 'Language trees and zipping'. *Condensed Matter Archive*, Feb. 21, 2002. `arXiv:cond-mat/0202383`.

(Gooskens & Heeringa 04) Charlotte Gooskens and Wilbert Heeringa. The position of Frisian in the Germanic language area. In Dicky Gilbers, Maartje Schreuder, and Nienke Knevel, editors, *On the Boundaries of Phonology and Phonetics*, pages 61–88. CLCG, Groningen, 2004.

(Gooskens & vanBezooijen 06) Charlotte Gooskens and Renée van Bezooijen. Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical? *Literary and Linguistic Computing*, 21, 2006. accepted, 7/2005.

(Gooskens 04) Charlotte Gooskens. Norwegian dialect distances geographically explained. In Britt-Louise Gunnarson, Lena Bergström, Gerd Eklund, Staffan Fridella, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundgren, and Mats Thelander, editors, *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2, June 12-14, 2003*, pages 195–206. Uppsala University, Uppsala, Sweden, 2004.

(Gray & Atkinson 03) Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(27 Nov.):435–439, 2003.

(Heeringa & Nerbonne 02) Wilbert Heeringa and John Nerbonne. Dialect areas and dialect continua. *Language Variation and Change*, 13:375–400, 2002.

(Heeringa 04) Wilbert Heeringa. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Unpublished PhD thesis, Rijksuniversiteit Groningen, 2004.

(Heeringa *et al.* 00) Wilbert Heeringa, John Nerbonne, Hermann Niebaum, and Rogier Nieuweboer. Measuring Dutch-German contact in and around Bentheim. In Dicky Gilbers, John Nerbonne, and Jos Schaeken, editors, *Languages in Contact*, pages 145–156. Rodopi, Amsterdam-Atlanta, 2000.

(Joshi 99) Aravind K. Joshi. Computational linguistics. In Robert A. Wilson and Frank C. Keil, editors, *The MIT Encyclopedia of the Cognitive Sciences*, pages 162–164. MIT Press, Cambridge, MA, 1999.

(Kay 02) Martin Kay. Introduction. In Ruslan Mitkov, editor, *Handbook of Computational Linguistics*, pages xvii–xx. Oxford University Press, Oxford, 2002.

(Kessler 01) Brett Kessler. *The Significance of Word Lists*. CSLI Press, Stanford, 2001.

(Kondrak 02) Grzegorz Kondrak. *Algorithms for Language Reconstruction*. Unpublished PhD thesis, University of Toronto, 2002.

(Manning & Schütze 99) Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, 1999.

(Moscoso del Prado Martín 03) Fermin Moscoso del Prado Martín. *Paradigmatic Structures in Morphological Processing: Computational and Cross-Linguistic Experiemental Studies*. Unpublished PhD thesis, Radboud University Nijmegen, 2003.

(Mufwene 01) Salikoko Mufwene. *The Ecology of Language Evolution*. Cambridge University Press, Cambridge, 2001.

(Nerbonne & Kleiweg 03) John Nerbonne and Peter Kleiweg. Lexical variation in LAMSAS. *Computers and the Humanities*, 37(3):339–357, 2003. Special Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr.

(Nerbonne & Konstantopoulos 04) John Nerbonne and Stasinos Konstantopoulos. Phonotactics in inductive logic programming. In Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining. Proceedings of the International IIS: IIPWM '04 Conference held in Zakopane, Poland*, Advances in Soft Computing, pages 493–502, Berlin, 2004. Springer.

(Nerbonne & Kretzschmar 03) John Nerbonne and William Kretzschmar, editors. *Computational Methods in Dialectometry*, volume 37 (3). 2003. Special Iss. of *Computers and the Humanities*.

(Nerbonne & Kretzschmar 06) John Nerbonne and William Kretzschmar, editors. *Progress in Dialectometry*, volume 21. 2006. Special Issue of *Literary and Linguistic Computing*, accepted to appear in 2006.

(Nerbonne & Stoianov 04) John Nerbonne and Ivilin Stoianov. Learning phonotactics with simple processors. In Dicky Gilbers, Maartje Schreuder, and Nienke Knevel, editors, *On the Boundaries of Phonology and Phonetics*, pages 89–121. CLCG, Groningen, 2004.

(Nerbonne 03) John Nerbonne. Linguistic variation and computation. In *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*, volume 10, pages 3–10, 2003.

(Nerbonne *et al.* 99) John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. Edit distance and dialect proximity. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.*, pages v–xv. CSLI, Stanford, CA, 1999.

(Niyogi & Berwick 96) Partha Niyogi and Robert C. Berwick. A language learning model for finite parmeter spaces. *Cognition*, 61:161–193, 1996.

(Pinker 94) Steven Pinker. *The Language Instinct*. W. Morrow and Co., New York, 1994.

(Pollard 93) Carl Pollard. On formal grammars and empirical linguistics. In Andreas Kathol and M. Bernstein, editors, *ESCOL '93: Proc. of the 10th Eastern States Conference on Linguistics*, Columbus, 1993. Ohio State University.

(Porter 80) Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

(Saffran *et al.* 99) Jenny R. Saffran, E.K. Johnson, Richard N. Aslin, and Elissa L. Newport. Statistical learning of tonal sequences by human infants and adults. *Cognition*, 70:27–52, 1999.

(Thomason & Kaufmann 88) Sarah Thomason and Terrence Kaufmann. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley, 1988.

(van Coetsem 88) Frans van Coetsem. *Loan Phonology and the Two Transfer Types in Language Contact*. Publications in Language Sciences. Foris Publications, Dordrecht, 1988.

(van Noord 04) Gertjan van Noord. Error mining for wide-coverage grammar engineering. In *Proc. of 42nd Meeting of the Association for Computational Linguistics*, pages 446–453. ACL, Barcelona, 2004.

(Watson 96) Greg Watson. The Finnish-Australian English corpus. *ICAME Journal: Computers in English Linguistics*, 20:41–70, 1996.

(Yang 04) Charles Yang. Universal grammar, statistics or both? *Trends in Cognitive Science*, 8(10):451–456, 2004.