Linguistic Distances

John Nerbonne
Alfa-informatica
University of Groningen
j.nerbonne@rug.nl

Erhard Hinrichs
Seminar für Sprachwissenschaft
Universität Tübingen
eh@sfs.uni-tuebingen.de

Abstract

In many theoretical and applied areas of computational linguistics researchers operate with a notion of linguistic distance or, conversely, linguistic similarity, which is the focus of the present workshop. While many CL areas make frequent use of such notions, it has received little focused attention, an honorable exception being Lebart & Rajman (2000). This workshop brings a number of these strands together, highlighting a number of common issues.

1 Introduction

In many theoretical and applied areas of computational linguistics researchers operate with a notion of linguistic distance or, conversely, linguistic similarity, which is the focus of the present workshop. While many CL areas make frequent use of such notions, it has received little focused attention, an honorable exception being Lebart & Rajman (2000).

In information retrieval (IR), also the focus of Lebart & Rajman's work, similarity is at heart of most techniques seeking an optimal match between query and document. Techniques in vector space models operationalize this via (weighted) cosine measures, but older tf/idf models were also arguably aiming at a notion of similarity.

Word sense disambiguation models often work with a notion of similarity among the contexts within which word (senses) appear, and MT identifies candidate lexical translation equivalents via a comparable measure of similarity. Many learning algorithms currently popular in CL, including not only supervised techniques such as memory-based learning (k-nn) and support-vector machines, but also unsupervised techniques such as Kohonen maps and clustering, rely essentially on measures of similarity for their processing.

Notions of similarity are often invoked in linguistic areas such as dialectology, historical linguistics, stylometry, second-language learning (as a measure of learners' proficiency), psycholinguistics (accounting for lexical "neighborhood" effects, where neighborhoods are defined by similarity) and even in theoretical linguistics (novel accounts of the phonological constraints on semitic roots).

This volume reports on a workshop aimed at bringing together researchers employing various measures of linguistic distance or similarity, including novel proposals, especially to demonstrate the importance of the abstract properties of such measures (consistency, validity, stability over corpus size, computability, fidelity to the mathematical distance axioms), but also to exchange information on how to analyze distance information further.

We assume that there is always a "hidden variable" in the similarity relation, so that we should always speak of similarity with respect to some property, and we suspect that there is such a plethora of measures in part because researchers are often inexplicit on this point. It is useful to tease the different notions apart. Finally, it is most intriguing to try to make a start on understanding how some of the different notions might construed as alternative realizations of a single abstract notion.

2 Pronunciation

John Laver, the author of the most widely used textbook in phonetics, claimed that "one of the

most basic concepts in phonetics, and one of the least discussed, is that of **phonetic similarity** [boldface in original, JN & EH]" (Laver, 1994, p. 391), justifying the attention the workshop pays to it. Laver goes on to sketch the work that has been done on phonetic similarity, or, more exactly, phonetic distance, in particular, the empirical derivation of confusion matrices, which indicate the likelihood with which people or speech recognition systems confusion one sound for another. Miller & Nicely (1955) founded this approach with studies of how humans confused some sounds more readily than others. Although "confusability" is a reasonable reflection of phonetic similarity, it is perhaps worth noting that confusion matrices are often asymmetric, suggesting that something more complex is at play. Clark & Yallop (1995, p. 319ff) discuss this line of work further, suggesting more sophisticated analyses which aggregate confusion matrices based on segments.

In addition to the phonetic interest (above), phonologists have likewise shown interest in the question of similarity, especially in recent work. Albright and Hayes (2003) have proposed a model of phonological learning which relies on "minimal generalization". The idea is that children learn e.g. rules of allomorphy on the basis not merely of rules and individual lexical exceptions (the earlier standard wisdom), but rather on the basis of slight but reliable generalizations. An example is the formation of the past tense of verbs ending in [ɪŋ], 'ing' (fling, sing, sting, spring, string) that build past tenses as 'ung' [ʌŋ]. We omit details but note that the "minimal generalization" is minimally DISTANT in pronunciation.

Frisch, Pierrehumbert & Broe (2004) have also kindled an interest in segmental similarity among phonologists with their claim that syllables in Semitic languages are constrained to have unlike consonants in syllable onset and coda. Their work has not gone unchallenged (Bailey and Hahn, 2005; Hahn and Bailey, 2005), but it has certainly created further theoretical interest in phonological similarity.

There has been a great deal of attention in psycholinguistics to the the problem of word recognition, and several models appeal explicitly to the "degree of phonetic similarity among the words" (Luce and Pisoni, 1998, p. 1), but most of these models employ relatively simple notions of sequence similarity and/or, e.g., the idea that distance may be operationalized by the number or replacements needed to derive one word from another—ignoring the problem of similarity among words of different lengths (Vitevitch and Luce, 1999). Perhaps more sophisticated computational models of pronunciation distance could play a role in these models in the future.

Kessler (1995) showed how to employ edit distance to operationalize pronunciation difference in order to investigate dialectology more precisely, an idea which, particular, Heeringa (2004) pursued at great length. Kondrak (2002) created a variant of the dynamic programming algorithm used to compute edit distance which he used to identify cognates in historical linguistics. McMahon & McMahon (2005) include investigations of pronunciation similarity in their recent book on phylogenetic techniques in historical linguistics. Several of the contributions to this volume build on these earlier efforts or are relevant to them.

Kondrak and Sherif (this volume) continue the investigation into techniques for identifying cognates, now comparing several techniques which rely solely on parameters set by the researcher to machine learning techniques which automatically optimize those parameters. They show the the machine learning techniques to be superior, in particular, techniques basic on hidden Markov models and dynamic Bayesian nets.

Heeringa et al. (this volume) investigate several extensions of the fundamental edit distance algorithm for use in dialectology, including sensitivity to order and context as well syllabicity constraints, which they argue to be preferable, and length normalization and graded weighting schemes, which they argue against.

Dinu & Dinu (this volume) investigate metrics on string distances which attach more importance to the initial parts of the string. They embed this insight into a scheme in which n-grams are ranked (sorted) by frequency, and the difference in the rankings is used to assay language differences. Their paper proves that difference in rankings is a proper mathematical metric.

Singh (this volume) investigates the technical question of identifying languages and character encoding systems from limited amounts of text. He collects about $1,000$ or so of the most frequent n-grams of various sizes and then classifies next texts based on the similarity between the fre-

quency distributions of the known texts with those of texts to be classified. His empirical results show "mutual cross entropy" to identify similarity most reliably, but there are several close competitors.

3 Syntax

Although there is less interest in similarity at the syntactic level among linguistic theorists, there is still one important areas of theoretical research in which it could play an important role and several interdisciplinary studies in which similarity and/or distant is absolutely crucial. Syntactic TYPOLOGY is an area of linguistic theory which seeks to identify syntactic features which tend to be associated with one another in all languages (Comrie, 1989; Croft, 2001). The fundamental vision is that some sorts of languages may be more similar to one another—typologically—than would first appear.

Further, there are two interdisciplinary linguistic studies in which similarity and/or distance plays a great role, including similarity at the syntactic level (without, however, exclusively focusing on syntax). LANGUAGE CONTACT studies seek to identify the elements of one language which have been adopted in a second in a situation in which two or more languages are used in the same community (Thomason and Kaufmann, 1988; van Coetsem, 1988). Naturally, these may be non-syntactic, but syntactic CONTAMINATION is a central concept which is recognized in contaminated varieties which have become more similar to the languages which are the source of contamination.

Essentially the same phenomena is studied in SECOND-LANGUAGE LEARNING, in which syntactic patterns from a dominant, usually first, language are imposed on a second. Here the focus is on the psychology of the individual language user as opposed to the collective habits of the language community.

Nerbonne and Wiersma (this volume) collect frequency distributions of part-of-speech (POS) trigrams and explore simple measures of distance between these. They approach issues of statistical significance using permutation tests, which requires attention to tricky issues of normalization between the frequency distributions.

Homola & Kuboň (this volume) join Nerbonne and Wiersma in advocating a surface-oriented measure of syntactic difference, but base their measure on dependency trees rather than POS tags, a more abstract level of analysis. From there they propose an analogue to edit distance to gauge the degree of difference. The difference between two tree is the sum of the costs of the tree-editing operations needed to obtain one tree from another (Noetzel and Selkow, 1999).

Emms (this volume) concentrates on applications of the notion 'tree similarity' in particular in order to identify text which is syntactically similar to questions and which may therefore be expected to constitute an answer to the question. He is able to show that the tree-distance measure outperforms sequence distance measures, at least if lexical information is also emphasized.

Kübler (this volume) uses the similarity measure in memory-based learning to parse. This is a surprising approach, since memory-based techniques are normally used in classification tasks where the target is one of a small number of potential classifications. In parsing, the targets may be arbitrarily complex, so a key step is select an initial structure in a memory-based way, and then to adapt it further. In this paper Kübler first applies chunking to the sentence to be parsed and selects an initial parse based on chunk similarity.

4 Semantics

While similarity as such has not been a prominent term in theoretical and computational research on natural language semantics, the study of LEXICAL SEMANTICS, which attempts to identify regularities of and systematic relations among word meanings, is more often than not predicated on an implicit notion of 'semantic similarity'. Research on the lexical semantics of verbs tries to identify verb classes whose members exhibit similar syntactic and semantic behavior. In logic-based theories of word meaning (e.g., Vendler (1967) and Dowty (1979)), verb classes are identified by similarity patterns of inference, while Levin's (1993) study of English verb classes demonstrates that similarities of word meanings for verbs can be gleaned from their syntactic behavior, in particular from their ability or inability to participate in diatheses, i.e. patterns of argument alternations.

With the increasing availability of large electronic corpora, recent computational research on word meaning has focused on capturing the notion of 'context similarity' of words. Such studies follow the empiricist approach to word meaning summarized best in the famous dictum of the British

linguist J.R. Firth: "You shall know a word by the company it keeps." (Firth, 1957, p. 11) Context similarity has been used as a means of extracting collocations from corpora, e.g. by Church & Hanks (1990) and by Dunning (1993), of identifying word senses, e.g. by Yarowski (1995) and by Schütze (1998), of clustering verb classes, e.g. by Schulte im Walde (2003), and of inducing selectional restrictions of verbs, e.g. by Resnik (1993), by Abe & Li (1996), by Rooth et al. (1999) and by Wagner (2004).

A third approach to lexical semantics, developed by linguists and by cognitive psychologists, primarily relies on the intuition of lexicographers for capturing word meanings, but is also informed by corpus evidence for determining word usage and word senses. This type of approach has led to two highly valued semantic resources: the Princeton WordNet (Fellbaum, 1998) and the Berkeley Framenet (Baker et al., 1998). While originally developed for English, both approaches have been successfully generalized to other languages.

The three approaches to word meaning discussed above try to capture different aspects of the notion of semantic similarity, all of which are highly relevant for current and future research in computational linguistics. In fact, the five papers that discuss issues of semantic similarity in the present volume build on insights from these three frameworks or address open research questions posed by these frameworks. Zesch and Gurevych (this volume) discuss how measures of semantic similarity—and more generally: semantic relatedness—can be obtained by similarity judgments of informants who are presented with word pairs and who, for each pair, are asked to rate the degree of semantic relatedness on a predefined scale. Such similarity judgments can provide important empirical evidence for taxonomic models of word meanings such as wordnets, which thus far rely mostly on expert knowledge of lexicographers. To this end, Zesch and Gurevych propose a corpus-based system that supports fast development of relevant data sets for large subject domains.

St-Jacques and Barrière (this volume) review and contrast different philosophical and psychological models for capturing the notion of semantic similarity and different mathematical models for measuring semantic distance. They draw attention to the fact that, depending on which underlying models are in use, different notions of semantic similarity emerge and conjecture that different similarity metrics may be needed for different NLP tasks. Dagan (this volume) also explores the idea that different notions of semantic similarity are needed when dealing with semantic disambiguation and language modeling tasks on the one hand and with applications such as information extraction, summarization, and information retrieval on the other hand.

Dridan and Bond (this volume) and Hachey (this volume) both consider semantic similarity from an application-oriented perspective. Dridan and Bond employ the framework of robust minimal recursion semantics in order to obtain a more adequate measure of sentence similarity than can be obtained by word-overlap metrics for bag-of-words representations of sentences. They show that such a more fine-grained measure, which is based on compact representations of predicate-logic, yields better performance for paraphrase detection as well as for sentence selection in question-answering tasks than simple word-overlap metrics. Hachey considers an automatic content extraction (ACE) task, a particular subtask of information extraction. He demonstrates that representations based on term co-occurrence outperform representations based on term-by-document matrices for the task of identifying relationships between named objects in texts.

Acknowledgments

We are indebted to our program committee and to the incidental reviewers named in the organizational section of the book, and to others who remain anonymous. We thank Peter Kleiweg for managing the production of the book and Therese Leinonen for discussions about phonetic similarity. We are indebted to the Netherlands Organization for Scientific Research (NWO), grant 200-02100, for cooperation between the Center for Language and Cognition, Groningen, and the *Seminar für Sprachwissenschaft*, Tübingen, for support of the work which is reported on here. We are also indebted to the Volkswagen Stiftung for their support of a joint project "Measuring Linguistic Unity and Diversity in Europe" that is carried out in cooperation with the Bulgarian Academy of Science, Sofia. The work reported here is directly related to the research objectives of this project.

References

Naoki Abe and Hang Li. 1996. Learning word association norms using tree cut pair models. In *Proceedings of 13th International Conference on Machine Learning*.

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

Todd M. Bailey and Ulrike Hahn. 2005. Phoneme Similarity and Confusability. *Journal of Memory and Language*, 52(3):339–362.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90, San Francisco, California. Morgan Kaufmann Publishers.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

John Clark and Colin Yallop. 1995. *An Introduction to Phonetics and Phonology*. Blackwell, Oxford.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology:Syntax and Morphology*. Oxford, Basil Blackwell.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

David Dowty. 1979. *Word Meaning and Montague Grammar*. Reidel, Dordrecht.

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

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

J. R. Firth. 1957. A synopsis of linguistic theory. *Oxford: Philological Society*. Reprinted in F. Palmer (ed.)(1968). Studies in Linguistic Analysis 1930-1955. Selected Papers of J.R. Firth., Harlow: Longman.

Stefan A. Frisch, Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity Avoidance and the OCP. *Natural Language & Linguistic Theory*, 22(1):179–228.

Ulrike Hahn and Todd M. Bailey. 2005. What Makes Words Sound Similar? *Cognition*, 97(3):227–267.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, Rijksuniversiteit Groningen.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Proc. of the European ACL*, pages 60–67, Dublin.

Grzegorz Kondrak. 2002. *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto.

John Laver. 1994. *Principles of Phonetics*. Cambridge Univeristy Press, Cambridge.

Ludovic Lebart and Martin Rajman. 2000. Computing similarity. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 477–505. Dekker, Basel.

Beth Levin. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago and London.

Paul A. Luce and David B. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1–36.

April McMahon and Robert McMahon. 2005. *Language Classification by the Numbers*. Oxford University Press, Oxford.

George A. Miller and Patricia E. Nicely. 1955. An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America*, 27:338–352.

Andrew S. Noetzel and Stanley M. Selkow. 1999. An analysis of the general tree-editing problem. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 237–252. CSLI, Stanford. [1]1983.

Philip Stuart Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing an semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland.

Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Published as AIMS Report 9(2).

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

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

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

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

Michael S. Vitevitch and Paul A. Luce. 1999. Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, 40(3):374–408.

Andreas Wagner. 2004. *Learning Thematic Role Relations for Lexical Semantic Nets*. Ph.D. thesis, Universität Tübingen.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA.