Distances and Trees in Linguistics[1]

William S-Y. Wang[2]

Introduction

This talk is divided in two parts. In Part I, I will discuss the various uses that computers have in linguistics. I will briefly survey how they have been used in the past, and speculate a bit on future applications. In Part II, I will report on some of my own recent research, using computers to study the evolution of languages, especially the languages of China.

Part I

I will begin with some observations on the term “computational linguistics”. In the last decade or two, we see this term more and more in a variety of contexts, in the names of conferences, associations, and even university courses. It is a convenient label, and there is no harm in using it as long as we are clear about what we are labeling.

I think it is important to keep in mind that the term “computational linguistics” is not parallel in content with many other terms in linguistics, such as “anthropological linguistics” or “psycholinguistics”, or “phonology” or “pragmatics”. These terms all refer to content aspects of language; these are, respectively, the role of language in various cultures, the relations between language and psychological mechanisms, how sounds pattern in language, and the use of language in different social contexts. In other words, they all deal with delimited, intrinsic aspects of language on which we are accumulating knowledge.

On the other hand, computational linguistics does not deal with any delimited segment of language *per se*. Rather, computers can be used with profit in all aspects of linguistic scholarship. In fact it is often a measure of the maturity of any aspect of linguistics

[1] Synopsis of paper presented at the 10th Pacific Asia Conference on Language, Information and Computation.

[2] Professor of Language Engineering at the City University of Hong Kong, and Professor of Graduate School of the University of California at Berkeley. E-mail address: WSYW@VIOLET.BERKELEY.EDU,EEWSYW@CITYU.EDU.HK.

to see how much computational power it can harness. The more it can harness this power, the more likely it can achieve results which are statistically significant and of cumulative value.

We use tape recorders in our fieldwork to preserve the speech sounds of various languages. The linguist who uses this tool has a tremendous advantage over others who rely on pencil and paper alone. For one thing, he has accumulated a database which can be analyzed over and over again, by himself later, as well as by others which wish to verify or build upon his work, and thus make the scholarship cumulative. The support of appropriate tools in all sectors of human endeavor, including, of course, linguistic research. There is a Chinese saying from the Lunyu:

论语：工欲善其事，必先利其器

Roughly translated, this means that to do our work well, we must make sure we have the best tools. And computers are the best tools *par excellence* for many intellectual tasks.

I will dwell on these points at some length partly as a reaction against the excesses of abstract theorizing that have pervaded linguistics these past decades. Linguistics has lost too much of its intellectual resources chasing unicorns such as "ideal speaker-hearers" and "homogeneous speech communities." Anytime we study language in the real world, whether synchronically or diachronically, we find the data inevitably very rich and highly complex.

Part II

A language is a collection of elements, which are organized at various levels. In the phonological hierarchy, for instance, we build from distinctive features, to segments, to syllables, and so on. In the grammatical hierarchy, we build from morphemes, to words, and to constructions of various sizes.

A classic set of questions in linguistics has to do with how these elements are related to each other. These questions may be posed syntagmatically, along the axis of time. Thus to the question of how the words of a sequence are related to each other grammatically, we have the ubiquitous constituent structure tree. To take a phrase made famous by the late Professor Y.R.Chao,

無肺病牛

If these four characters are analyzed as (1,(2,3),4), then the phrase means: cows with no lung disease. If they are analyzed as ((1,2),(3,4)), then meaning would be: sick cows without lungs. The difference lies in the scope which the first character modifies, whether the scope is just the 2nd character, or the compound built from characters 2 and 3. The ambiguity derives from whether or not to treat 2 and 3 as a constituent. In such trees, the central idea is that a sequence of elements is a constituent if and only if they derive from a single node. In other contexts, this condition of deriving exhaustively and exclusively from a single node is called *monophyletic*.

Relations among linguistic elements may be studied paradigmatically as well, that is, how they are related to each other within a set. As an example from phonology, we may consider the trees built from distinctive features. These trees are the bases of defining an important concept in phonological theory - the natural class. This concept underlies almost all formulations of rules and discussions of sound change. Again, natural classes are segments which are monophyletic on the distinctive feature tree.

A paradigmatic example from grammar uses semantic features instead of distinctive features. Such was the method used in an early paper by Katz and Fodor, in analyzing the word "bachelor." Similar methods have been used more recently in constructing semantic networks; for example by Winograd.

The earliest use of trees was in diachronic linguistics; it dates back to the middle of the 19th century, with August Schleicher. Here the elements were individual Indo-European languages, which Schleicher arranged in terms of their phylogenetic relations. Charles Darwin had used trees to represent a few years earlier to represent phylogenetic relations among organisms. Since Schleicher was deeply impressed with Darwin's ideas, he might have been influenced by the latter in the use of this method as well.

Glottochronology was a method used in diachronic linguistics. It was inspired by carbon-14 dating, especially as it was developed in archeology. The limitation of the original formulation was that the method was applicable to languages only one pair at a time, thus missing out information which bear upon relations of groups of three or larger.

Given this diversity of uses to which trees have been put, it is necessary to define these various trees in terms of the following four [4] properties.

We begin by noting that a tree is a graph consisting of nodes, branches, and tips. Tips are distinct from nodes in that they are each connected to a single branch. We may [1] distinguish an unrooted tree from a rooted one. The branches in an unrooted tree have no directionality. Trees which represent semantic networks typically have no directionality.

We may [2] distinguish trees in which each of the nodes, *not tips*, is connected to exactly three branches. These we call binary trees, as opposed to nonbinary trees. The exception that needs to be made is that the root of a binary tree is connected to exactly two branches. A tree displaying phonological features, where each distinctive feature is either + or -, e.g., voicing or nasality, would be a rooted binary tree.

We may [3] distinguish trees the lengths of whose branches are significant, as opposed to those whose branch lengths are not significant. In a tree designed to show the syntactic constituents of a sentence, the lengths of the branches have no meaning *per se*.

Lastly, among trees which are rooted and whose branch lengths are significant, we may [4] further distinguish the trees whose tips are all equidistant from the root, from the trees whose tips may not be all equidistant from the root. In a tree showing the historical derivation of a family of languages, we may wish to require that all the languages have evolved exactly the same amount from their common ancestor. Alternatively, we may wish to make no such assumption, and the branch lengths reflect the differences, if any, in the amounts to which each language has changed from the common ancestor.

I hope to provide examples for all the properties discussed above, and point up the strengths and weaknesses of each type of tree. Syntactic trees, for instance, have difficulty in displaying discontinuous constituents. Phylogenetic trees are an effective representation of common inheritance; but do not portray easily results due to horizontal transmission.

I will also discuss the magnitude of the computation in working with trees. As the number of tips increases, the number of trees grow extremely fast, so the task of evaluating all trees exhaustively is highly demanding of computing power. I will discuss the standard methods of tree construction, as well report on some on-going work using matrix solutions. Lastly, I will present some linguistic results in the use of trees in analyzing the historical relations among the major Chinese dialects.

References

Fitch, W.M. and E.Margoliash. 1967. Construction of phylogenetic trees. *Science* 155.279-284.

Felsenstein, J. PHYLIP: Phylogeny Inference Package. Department of Genetics, University of Washington.

Greenberg, J.H. 1957. *Essays in Linguistics*. University of Chicago Press.

Halle, M. 1959. *The Sound Pattern of Russian*. The Hague: Mouton.

Hennig, W. 1966. *Phylogenetic Systematics*. University of Illinois Press.

Hoenigswald, H. and L.Wiener, eds. 1987. *Biological Metaphor and Cladistic Classification*. University of Pennsylvania Press.

Katz, J. and J. Fodor. 1964. The Structure of a Semantic Theory. In Fodor, J. and J. Katz, eds. 1964. *The Structure of Language: Readings in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice Hall.

Meyers, R. and W.S-Y. Wang. 1963. *Tree representations in linguistics*. Ohio State University: Project on Linguistic Analysis Report #3.

Qiao, S.Z. and W.S-Y. Wang. Matrix solutions for additive trees. In preparation.

Saitou, N. and M.Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol.Biol.Evol.* 4.406-425.

Sneath, P.H.A. and R.R.Sokal. 1973. *Numerical Taxonomy*. New York: W.H.Freeman.

Wang, W.S-Y. 1988. Diannao zai yuyanxue li de yunyong. (The use of computers in linguistics.) *Proceedings of ROC Computational Linguistics Workshop I*, pp.257-287. Taiwan. [电脑在语言学里的运用]

Wang,W.S-Y. and Z.W.Shen. 沈钟伟 1992. Fangyan guanxide jiliang biaoshu. (Quantitative description of dialect relationship) *Zhongguo Yuwen* 227.81-92. [方言关系的计量表达]

Wang, W.S-Y. Glottochronology, Lexicostatistics, and other Numerical Methods. Pp.1445-1450. *Encyclopedia of Language and Linguistics*. Pergamon Press.

Winograd, T. 1986. “Computer Software for Working in Language”. In Wang, W.S-Y. ed. 1986. *Language, Writing and the Computer*. New York: W.H. Freeman, pp.61-72.

Xu, Tongqiang. 徐通锵 1991. 历史语言学。北京：商务印书馆。[Historical Linguistics.]

Zhong, Yang et al. 钟扬 等 1994. 分支分类的理论与方法。北京：科学出版社。[Theory and Method in Classification.]