The universality of simple distributional methods: Identifying syntactic categories in Mandarin Chinese.

Martin Redington, Nick Chater, Chu-Ren Huang,
Li-Ping Chang, Steve Finch, & Keh-jiann Chen*

Introduction

The problem of acquiring language is a difficult and complex one. The traditional assumption is that infants acquire language by virtue of innate knowledge, which reduces the problem to one of tuning this innate knowledge of language in general, to the specific characteristics of the language spoken in the infant's early environment (*e.g.* Chomsky, 1980). Nevertheless, without empirical assessment of potential learning mechanisms and sources of information, it may be premature to decide which aspects of language, if any, require drawing on innate knowledge. This paper considers a simple distributional learning mechanism, which does not draw on explicit prior knowledge. This method has previously been shown to be informative about the syntactic category membership of individual words in English, French, and German. We ask whether it can provide similar constraints in Chinese.

Acquiring Syntactic Categories

In acquiring syntactic categories (such as *noun, verb, etc.*), language learners face two inter-related problems. They must discover the set of syntactic categories in the language, and also identify the syntactic category membership of individual words.

It seems both plausible and likely that human infants possess some innate constraints on the identity of the syntactic categories[1] and on the form of the rules governing how these categories can be combined into sentences (the Universal Grammar).

However, the extent to which these constraints can aid the identification of the syntactic categories of individual words (which cannot be known innately) is unclear. It may well be that semantic or pragmatic information (for instance, knowing that *ball* refers to a round object [see Pinker, 1984], or that *want ball* is likely to result in an adult providing a round object [see Snow, 1988]) can provide both clues that *ball* is a noun, and

*Redington & Chater: Department of Experimental Psychology, University of Oxford, South Parks Rd., Oxford, UK. OX1 3UD. E-mail: [fmr|nick]@psy.ox.ac.uk. Huang: Institute of History and Philology, Academia Sinica, Taipei, Taiwan. E-mail: hschuren@ccvax.sinica.edu.tw. Chang & Chen: Institute of Information Science, Academia Sinica, Taipei, Taiwan. E-mail: [lpchang|kchen]@iis.sinica.edu.tw. Finch: Human Communications Research Centre, 1–4 Buccleuch Place, Edinburgh, UK. EH8 9LW. E-mail: steve@cogsci.ed.ac.uk.

[1] For example, it has been proposed that infants might possess innate 'action' and 'object' categories, corresponding to nouns and verbs, (Pinker, 1984).

a way to relate this category information to innate grammatical rules. But given that relatively little is known about how infants represent the world, and which features of the environment are most salient (or define 'objects' and 'actions'), empirical assessment of the role and potential contribution of such mechanisms is difficult.

A different potential source of information about syntactic category membership is distributional constraints—simple relationships within the linguistic input. Proposals that distributional mechanisms might contribute to the language learning process (notably by Maratsos, 1979, 1988) have been heavily criticised (see in particular Pinker, 1984), but these criticisms generally reveal a deep misunderstanding of the scope, role, ambitions, and mechanisms of distributional analysis. These mechanisms are not proposed as a solution to the problem of language learning in general. They are proposed as potential sources of information, relevant to particular aspects of language acquisition; it is not claimed that they are the only source of information, or the most informative ones.

However, as potential constraints on particular aspects of language, distributional mechanisms have a big advantage: it is possible to assess, and empirically quantify, their informativeness. If, as their critics suggest, such mechanisms are worthless, this should soon become apparent. However, a positive assessment establishes the feasibility of the proposal that distributional information may actually contribute[2] to particular aspects of child language acquisition.

A simple distributional mechanism to provide constraints on syntactic category membership

The method used here has been rediscovered several times (Rosenfeld *et al.*, 1968; Kiss, 1973), and was first described in its present form by Finch & Chater (1991). It is based on the notion of the *replacement* test, from theoretical linguistics. If word A can be replaced by word B, throughout a corpus, without loss of syntactic well-formedness, then they share the same syntactic category. This notion of replacement can be generalised to words which share a common *distribution*. Thus words which are similarly distributed (*i.e.* tend to appear in the same linguistic contexts) will share the same syntactic category.

Given a representation of the distributions of contexts in which each word appears, then words of a similar syntactic nature should possess similar representations. For these purposes, 'context' can be defined as the words immediately preceding and succeeding the target word.

Here, the most frequent words (*e.g.* the top 1,000) are chosen as target words, and the very high frequency words (*e.g.* the top 150) are chosen as context words. The target words will generally comprise the bulk of any corpus (*e.g.* for the Chinese corpus, the most frequent 1,000 words make up 67% of the total), whilst the top 150 words will tend to be closed-class words, which tend to occur in stereotypical relationships to open-class words (for instance, in English, determiners are always followed by adjectives or nouns).

In a single pass through the corpus, the dependencies between the target and context words are recorded by incrementing the value of a cell indexed by the appropriate tar-

[2] We stress that this finding says nothing whatsoever about the feasibility or otherwise of other potential sources, or whether distributional sources (or any other sources) actually *do* play a role in human language learning.

get and context word, in a contingency table corresponding to the appropriate context position; last but one word, previous word, next word, or next but one word. Once this process is complete, for each target word, there is a row in each contingency table forming a 150-dimensional vector representing the observed distribution of each of the 150 context words in that position. These vectors can be strung together to form a 600-dimensional vector, representing the distribution of local contexts within which each focus word appeared[3].

Standard hierarchical clustering techniques can be applied to these vectors, producing a hierarchically structured representation of the similarity of distributions of contexts. This classification provides a straightforward constraint on the syntactic category of individual words: words which are close together in the space of possible distributions of context (and thus in the hierarchical classification, or dendrogram) should belong to the same syntactic category, and words which are distant in this space (or in the dendrogram) should belong to different syntactic categories.

Results from English Text and Transcribed Speech

Previously, this method has been applied to very large corpora of English text, taken from USENET newsgroups, and the *Wall Street Journal.* The resulting dendrograms reveal correspondences with syntactic structure at many levels. Large branches of the dendrogram correspond to gross syntactic categories, with nouns and verbs being particularly obvious and coherent, and smaller categories such as adjectives, adverbs, pronouns, conjunctions also showing up clearly. Finer divisions are also apparent, with divisions between plural and singular nouns, different verb cases (the present participle shows up particularly well), and possessive pronouns and determiners. At a very fine level of detail, semantic relationships become apparent. For example, within large coherent noun clusters there are sub-clusters of 'food', 'computer', and 'organisation' related nouns within large noun clusters.

However, the relationship between syntactic structure and that of the dendrogram is not perfect, and whilst many divisions have a straightforward syntactic interpretation, many do not. A second problem is that the dendrogram is a hierarchical structure, and gives no clues, apart from the intuition of the observer, as to where to draw divisions between syntactic categories. Nevertheless, empirical analysis shows that discrete categories formed by 'cutting' the dendrogram at any level share a high degree of mutual information (and reliably more than would be expected by chance alone) with the classification of each word according to it most common syntactic category[4].

Similar results have been obtained for transcribed English speech (Redington, Chater & Finch, 1993), taken from the CHILDES corpus (MacWhinney & Snow, 1985). This analysis was performed on 2.5 million words of adult speech, recorded in domestic North American settings. This provided a better approximation (than written text) to the language to which children are actually exposed.

[3] For a more detailed description of this particular algorithm, see Finch, 1993. Detailed descriptions and comparisons of similar methods are also given in Grünwald, 1994.

[4] The method described here averages contexts over all occurrences of a word, and therefore only picks up the most common syntactic reading (although with English corpora ambiguous noun-verbs, such as *fire*, form their own clusters). There are various possible solutions to this problem (see for instance Redington *et al.*, 1993; Brill & Marcus, 1992).

Analysis of a Mandarin Chinese Corpus

An important feature of any theory of language acquisition is that it should apply universally across languages. Whilst different potential sources of information may have different values for different languages, and exploited to differing extents by learners, a theory which relies on the idiosyncrasies of, for instance, English, to account for some part of the language acquisition process is of limited value. Given that some critics (again, Pinker, 1984, being the clearest example) argue that distributional methods are uninformative about syntactic categories in English (which is untrue), it is important, given that it is possible, to demonstrate their value across languages.

The corpus for the current analysis consisted of approximately 1.15 million words of written Mandarin Chinese. This was from a variety of sources, of recent origin (within 3 years). A description of the ongoing Mandarin Chinese corpus project is provided in Huang & Chen (1992). The original corpus was in character format (equivalent to an English corpus with all of the spaces removed), and this was segmented into words automatically (again, see Huang & Chen, 1992 for details of the segmentation). This segmented corpus is comparable in nature, for the purposes of this analysis, to the Wall Street Journal or USENET corpora: whilst analysis of transcribed speech (and ideally child-directed speech) would have been preferable, in the absence of such corpora, written texts are the best available approximation to language learners' input.

As described above, the 1,000 most frequent words in the corpus were used as the target words, with the 150 most frequent words used as context. The target words occurred at least 100 times in the corpus, and the target and context words comprised 50 and 67% of the corpus respectively. The analysis was performed in exactly the same manner as described in Finch & Chater (1991), with the Spearman Recurrent Correlation Coefficient (ρ) being used as the metric of similarity between context vectors, and average link clustering (Sokal & Sneath, 1963) being used to produce the dendrogram given the table of similarities between target words.

Results

In order to assess the quality of the clustering, a canonical classification of the 1,000 target words was performed. This was based on the categories in use in the ongoing corpus project (Huang and Chen, 1992). This set of 46 categories is very fine-grained compared to the usual conception of syntactic categories within the developmental literature, and a coarser superset of these categories (see Table 1) was used in assessing the informativeness of the analysis.

Figure 1 shows the overall structure of the dendrogram resulting from the analysis, and two small subclusters of the dendrogram. The former has been cut at the point where the similarity between clusters (ρ) is 0.17, chosen by hand for the relatively obvious concordance between these clusters and the canonical classification. Only clusters containing more than 10 items are shown, and these have been labelled by hand with their dominant syntactic category (a minimum of 60% [and generally more than 80%] of the items within each cluster conform to this label). Some degree of syntactic structure is also apparent within the many smaller clusters not shown here. However, qualitatively the clustering is relatively poor compared to results with English, with relatively little detail at the medium level concerning cases. This is most likely due to the relatively

Category	Example	Number
Noun	*sir, time, problem*	356
Verb	*call, bring, listen to*	277
Preposition	*be at, from, to*	32
Adjective	*international*	8
Adverb	*completely, just, again*	137
Pronoun	*it, oneself, who, they*	22
Coordinate conjunction	*as well as, and , or*	6
Subordinate conjunction	*and yet, because of, moreover*	34
Determiner	*some, every, many, all*	27
Particle		7
Postpositional constituent	*and so on, even, within*	31
Unclassified		53
negligible categories		10

Table 1: The 11 syntactic categories, with English translations of example target words in the Mandarin Chinese analysis. The number of target words in each category is also listed. 63 target items (approximately 6%) had no canonical classification (punctuation marks, words whose syntactic usage had not been analysed within the corpus project), or were in categories consisting of fewer than 5 target items (*e.g.* relative and aspect markers). These were excluded from the analysis of the goodness of clustering.

small size of the corpus, compared to analyses of 40 million word corpora. Even so, the relationship between the dendrogram's structure and the canonical classification is readily apparent to the naked eye. It is important not to read too much into the organisation of the high level clusters shown, which is relatively variable across languages and corpora. The relevant feature here is that the clusters in the dendrogram mirror to some extent the syntactic relationships between the Chinese words.

The small sample clusters have been chosen by hand as an example of appropriate clustering (there are many clusters with no clear canonical interpretation). As in previous analyses, at a high level of similarity, some semantic influences on the pattern of clustering were observed. For instance, *like* and *love*; *mother* and *father*; *year* and *night* and *afternoon* were clustered together. Again, this effect was reduced due to the small corpus size.

To provide a more quantitative assessment of the 'goodness of clustering', we calculated the mutual information[5] between the dendrogram classification, and the canonical classification, at various levels of similarity. For purposes of comparison, we also calculated the amount of mutual information that the classifications would be expected to share if items were allocated to the dendrogram categories at random, whilst the number and size of categories were held constant. Figure 2 clearly shows that at all interesting levels of the dendrogram (where the number of dendrogram clusters is significantly lower than the number of items, and significantly greater than one), the dendrogram classification is much more informative about the canonical classification than one would expect by chance alone; for instance, when $\rho = 0.15$, the dendrogram classification shares 70% of the information in the canonical classification, whereas the *highest* value (over 1,000 simulations) was 38%. Thus in a very precise info-theoretic sense, distributional analysis conveys a significant amount of information concerning the syntactic categories of the individual Chinese words.

[5] The mutual information, M, describes the extent to which the classification of an item according

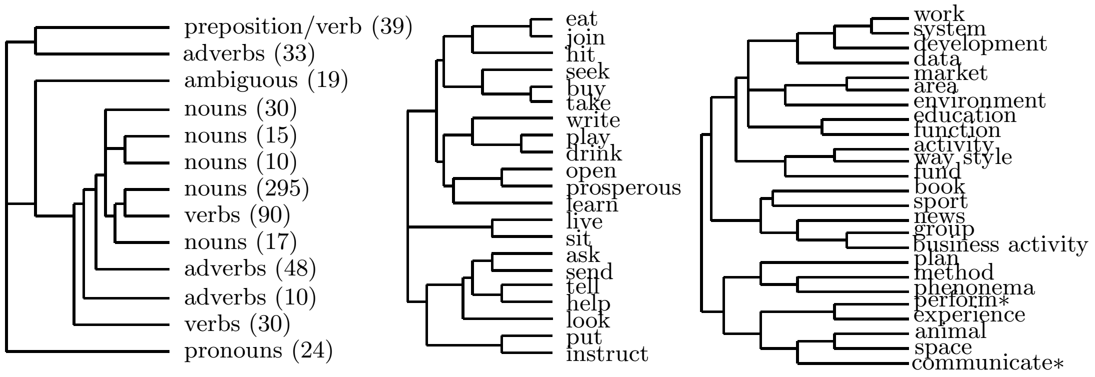

Figure 1: The overall structure of the dendrogram (left, with the number of items in each cluster in parentheses). Syntactic structure was also apparent in the many smaller clusters which, for reasons of clarity, are not shown. Centre and right are subclusters of the noun and verb clusters. These have been chosen as examples of 'good' clustering, but are not exceptional of the clusters shown in the annotated dendrogram. The 2 items marked with an asterisk ($*$) are inappropriately clustered, being verbs under the canonical classification.

Discussion

Implications for Early Category Acquisition

It is important to realise that we do not propose that human infants utilise this particular algorithm, or generally propose the mechanism outlined here as a *model* of the identification of words' syntactic categories, or of processes contributing to this feat. This algorithm is only one of myriad possible ways to exploit the structure of natural language. However, the fact that this method does provide information and constraints about the identity of words' syntactic categories, *across a variety of languages*, proves the feasibility and utility of this particular source of information (that is, distributional information), and makes real the possibility that it may be one of the sources exploited by real language learners.

The Nature of the Language Acquisition Device

A second consideration concerns the *empiricist* nature of such mechanisms. Statistical learning mechanisms are traditionally seen as being as opposition to approaches to language acquisition which stress the role of innate knowledge.

Let us suppose that human infants do possess distributional learning mechanisms, and exploit them, in order to acquire language. How can this be squared with nativist arguments, such as the poverty of the stimulus (that rapid, consistently successful language acquisition must imply that the learning problem that infants face is massively

to the dendrogram reduces any uncertainty as to its canonical classification. $M_{ij} = I_i + I_j - I_{ij}$, where $I_i = -\sum_i p(i) \log_2 p(i)$, ($I_i$ is the amount of information in the canonical classification, and $p(i)$ is the probability of an item being a member of category i, and similarly for I_j) and I_{ij}, the joint information in both classifications, $= -\sum_{ij} p(ij) \log_2 p(ij)$, where p_{ij} is the probability that an item is a member of canonical category i, and dendrogram cluster j.

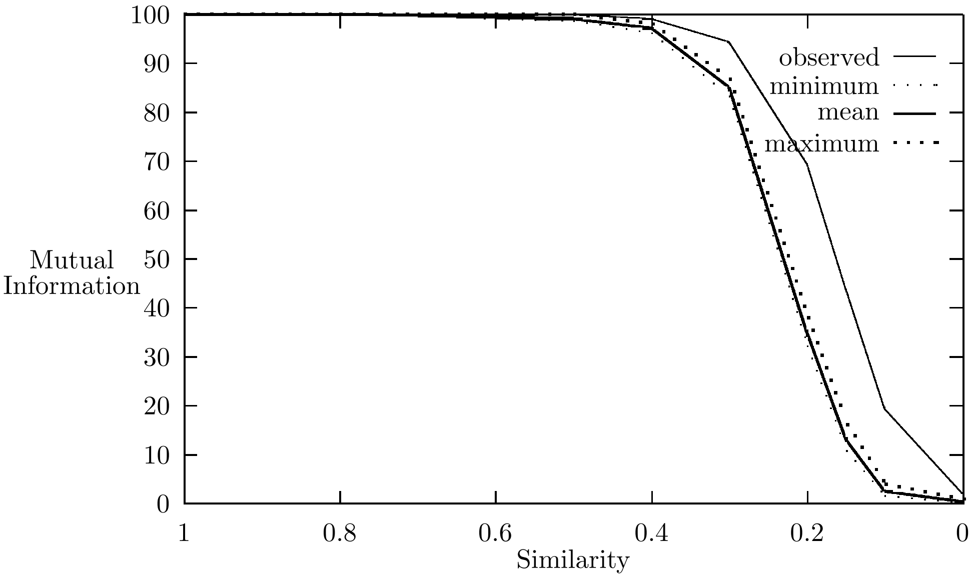

Figure 2: The mutual information (as a percentage of the information in the canonical classification) between the categories obtained from the dendrogram at a range of levels of similarity between clusters (ρ), and the canonical classification. Also shown are the minimum, mean, and maximum values of 1,000 Montecarlo simulations, where items were allocated to categories at random, with the number and size of categories matched to those at the appropriate level of similarity in the dendrogram.

simplified by the possession of innate knowledge of language)? If some aspect of language can be gleaned from (or at least constrained by) the early linguistic environment, in an efficient manner, then evolution may have equipped infants with the apparatus to effect this, rather than encoding the end result directly.

To possess and successfully utilise a distributional learning mechanism, whether specific to language or not, is to possess knowledge of language in general, in a sense that is not generally implied by Chomsky's (1980) Language Acquisition Device (LAD). However, if infants can exploit distributional information in language using 'general' learning mechanisms, then the LAD need only concern itself with how to integrate the information provided with specialise innate linguistic knowledge. If specialised language learning mechanisms are required, then these effectively constitute part of the LAD, but in the sense of *knowing how* to acquire language; what cues to look for, rather than innate grammatical knowledge. In either case, the potential contribution of distributional information may have a profound effect on the nature of the Language Acquisition Device.

Before empirically investigating whether distributional learning mechanisms actually do contribute to human language learning, we must ask whether they are feasible in principle. Can aspects of language be efficiently constrained by distributional information? In the case of words' syntactic categories, across a variety of languages, the answer is yes.

References

Brill, E. & Marcus, M. (1992). Tagging an unfamiliar text with minimal human supervision. *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language*, American Association for Artificial Intelligence.

Chomsky, N. (1980). *Rules and Representations.* Cambridge, Mass: MIT press.

Finch, S. (1993). *Finding Structure in Language.* Ph. D. Thesis, Centre for Cognitive Science, University of Edinburgh.

Finch, S. P. & Chater, N. (1991). A hybrid approach to the automatic learning of linguistic categories. *AISB Quarterly*, *78*, 16–24.

Grünwald, P. (1994). *Automatic grammar induction using the* MDL *principle.* Master's thesis, University of Amsterdam, Amsterdam, 1994.

Huang, C. & Chen, K. (1992). A Chinese corpus for linguistic research. *Proceedings of COLING-92.* Nantes, France. 1214–1217.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, *7*, 1–41.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271–295.

Maratsos, M. (1979). How to get from words to sentences. In D. Aaronson & R. Rieber (Eds.), *Perspectives in Psycholinguistics.* Hillsdale, NJ: LEA.

Maratsos, M. (1988). The acquisition of formal word classes. In Y. Levy, I.M. Schlesinger & M.D.S. Braine (Eds.), *Categories and Processes in Language Acquisition.* Hillsdale, NJ: LEA.

Ninio, A. & Snow, C. E. (1988). Language acquisition through language use: The functional sources of children's early utterances. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine (Eds.), *Categories and Processes in Language Acquisition.* Hillsdale, NJ: LEA.

Pinker, S. (1984). *Language Learnability and Language Development.* Cambridge, Mass: Harvard University Press.

Radford, A. (1988). *Transformational Grammar*, 2nd Edition. Cambridge: Cambridge University Press.

Redington, M., Chater, N., & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: LEA. pp 848–853.

Rosenfeld, A., Huang, H. K. & Schneider, V. B. (1969). An application of cluster detection to text and picture processing. *IEEE Transactions on Information Theory*, *15*, 672–681.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* San Francisco: W. H. Freeman.