A Stochastic Finite-State Word-Segmentation Algorithm for Chinese

Richard Sproat*
Bell Laboratories

Chilin Shih†
Bell Laboratories

William Gale‡
Bell Laboratories

Nancy Chang§
University of Cambridge

The initial stage of text analysis for any NLP task usually involves the tokenization of the input into words. For languages like English one can assume, to a first approximation, that word boundaries are given by whitespace or punctuation. In various Asian languages, including Chinese, on the other hand, whitespace is never used to delimit words, so one must resort to lexical information to "reconstruct" the word-boundary information. In this paper we present a stochastic finite-state model wherein the basic workhorse is the weighted finite-state transducer. *The model segments Chinese text into dictionary entries and words derived by various productive lexical processes, and—since the primary intended application of this model is to text-to-speech synthesis—provides pronunciations for these words. We evaluate the system's performance by comparing its segmentation "judgments" with the judgments of a pool of human segmenters, and the system is shown to perform quite well.*

1. The Problem

Any NLP application that presumes as input unrestricted text requires an initial phase of text analysis; such applications involve problems as diverse as machine translation, information retrieval, and text-to-speech synthesis (TTS). An initial step of any text-analysis task is the tokenization of the input into words. For a language like English, this problem is generally regarded as trivial since words are delimited in English text by whitespace or marks of punctuation. Thus in an English sentence such as *I'm going to show up at the ACL* one would reasonably conjecture that there are eight words separated by seven spaces. A moment's reflection will reveal that things are not quite that simple. There are clearly eight **orthographic** words in the example given, but if one were doing syntactic analysis one would probably want to consider *I'm* to consist of two **syntactic** words, namely *I* and *am*. If one is interested in translation, one would probably want to consider *show up* as a single **dictionary** word since its semantic interpretation is not trivially derivable from the meanings of *show* and *up*. And if one is interested in TTS, one would probably consider the single orthographic word *ACL* to consist of three **phonological** words—/e^{j} s'i ɛl/—corresponding to the pronunciation of each of the letters in the acronym. Space- or punctuation-delimited

* 700 Mountain Avenue, 2d-451, Murray Hill, NJ 07974, USA. E-mail: rws@bell-labs.com
† 700 Mountain Avenue, 2d-451, Murray Hill, NJ 07974, USA. E-mail: cls@bell-labs.com
‡ 600 Mountain Avenue, 2c-278, Murray Hill, NJ 07974, USA. E-mail: gale@research.att.com
§ Cambridge, UK. E-mail: nc201@eng.cam.ac.uk

(a)

日文章魚怎麼說?
'How do you say octopus in Japanese?'

(b)

Plausible Segmentation

日文	章魚	怎麼	說
ri4-wen2	*zhang1-yu2*	*zen3-me0*	*shuo1*
'Japanese'	'octopus'	'how'	'say'

(c)

Implausible Segmentation

日	文章	魚	怎麼	說
ri4	*wen2-zhang1*	*yu2*	*zen3-me0*	*shuo1*
'Japan'	'essay'	'fish'	'how'	'say'

Figure 1
A Chinese sentence in (a) illustrating the lack of word boundaries. In (b) is a plausible segmentation for this sentence; in (c) is an implausible segmentation.

orthographic words are thus only a starting point for further analysis and can only be regarded as a useful hint at the desired division of the sentence into words.

Whether a language even has orthographic words is largely dependent on the writing system used to represent the language (rather than the language itself); the notion "orthographic word" is not universal. Most languages that use Roman, Greek, Cyrillic, Armenian, or Semitic scripts, and many that use Indian-derived scripts, mark orthographic word boundaries; however, languages written in a Chinese-derived writing system, including Chinese and Japanese, as well as Indian-derived writing systems of languages like Thai, do not delimit orthographic words.[1]

Put another way, written Chinese simply lacks orthographic words. In Chinese text, individual characters of the script, to which we shall refer by their traditional name of *hanzi*,[2] are written one after another with no intervening spaces; a Chinese sentence is shown in Figure 1.[3] Partly as a result of this, the notion "word" has never played a role in Chinese philological tradition, and the idea that Chinese lacks anything analogous to words in European languages has been prevalent among Western sinologists; see DeFrancis (1984). Twentieth-century linguistic work on Chinese (Chao 1968; Li and Thompson 1981; Tang 1988, 1989, inter alia) has revealed the incorrectness of this traditional view. All notions of word, with the exception of the orthographic word, are as relevant in Chinese as they are in English, and just as is the case in other languages, a word in Chinese may correspond to one or more symbols in the orthog-

1 For a related approach to the problem of word-segmention in Japanese, see Nagata (1994), inter alia.
2 Chinese 漢字 *han4zi4* 'Chinese character'; this is the same word as Japanese *kanji*.
3 Throughout this paper we shall give Chinese examples in traditional orthography, followed immediately by a Romanization into the **pinyin** transliteration scheme; numerals following each pinyin syllable represent tones. Examples will usually be accompanied by a translation, plus a morpheme-by-morpheme gloss given in parentheses whenever the translation does not adequately serve this purpose. In the pinyin transliterations a dash (-) separates syllables that may be considered part of the same phonological word; spaces are used to separate plausible phonological words; and a plus sign (+) is used, where relevant, to indicate morpheme boundaries of interest.

raphy: 人 *ren2* 'person' is a fairly uncontroversial case of a monographemic word, and 中國 *zhong1-guo2* (middle country) 'China' a fairly uncontroversial case of a digraphemic word. The relevance of the distinction between, say, phonological words and, say, dictionary words is shown by an example like 中華人民共和國 *zhong1-hua2 ren2-min2 gong4-he2-guo2* (China people republic) 'People's Republic of China.' Arguably this consists of about three phonological words. On the other hand, in a translation system one probably wants to treat this string as a single dictionary word since it has a conventional and somewhat unpredictable translation into English.

Thus, if one wants to *segment* words—for any purpose—from Chinese sentences, one faces a more difficult task than one does in English since one cannot use spacing as a guide. For example, suppose one is building a TTS system for Mandarin Chinese. For that application, at a minimum, one would want to know the phonological word boundaries. Now, for this application one might be tempted to simply bypass the segmentation problem and pronounce the text character-by-character. However, there are several reasons why this approach will not in general work:

1. Many hanzi have more than one pronunciation, where the correct pronunciation depends upon word affiliation: 的 is pronounced *de0* when it is a prenominal modification marker, but *di4* in the word 目的 *mu4-di4* 'goal'; 乾 is normally *gan1* 'dry,' but *qian2* in a person's given name.

2. Some phonological rules depend upon correct word segmentation, including Third Tone Sandhi (Shih 1986), which changes a 3 (low) tone into a 2 (rising) tone before another 3 tone: 小老鼠 *xiao3 [lao3 shu3]* 'little rat,' becomes *xiao3 [lao**2**-shu3]*, rather than *xiao**2** [lao**2**-shu3]*, because the rule first applies within the word *lao3-shu3* 'rat,' blocking its phrasal application.

3. In various dialects of Mandarin certain phonetic rules apply at the word level. For example, in Northern dialects (such as Beijing), a full tone (1, 2, 3, or 4) is changed to a neutral tone (0) in the final syllable of many words: 冬瓜 *dong1-gua1* 'winter melon' is often pronounced *dong1-**gua0***. The high 1 tone of 瓜 would not normally neutralize in this fashion if it were functioning as a word on its own.

4. TTS systems in general need to do more than simply compute the pronunciations of individual words; they also need to compute intonational phrase boundaries in long utterances and assign relative prominence to words in those utterances. It has been shown for English (Wang and Hirschberg 1992; Hirschberg 1993; Sproat 1994, inter alia) that grammatical part of speech provides useful information for these tasks. Given that part-of-speech labels are properties of words rather than morphemes, it follows that one cannot do part-of-speech assignment without having access to word-boundary information. Making the reasonable assumption that similar information is relevant for solving these problems in Chinese, it follows that a prerequisite for intonation-boundary assignment and prominence assignment is word segmentation.

The points enumerated above are particularly related to TTS, but analogous arguments can easily be given for other applications; see for example Wu and Tseng's (1993) discussion of the role of segmentation in information retrieval. There are thus some very good reasons why segmentation into words is an important task.

A minimal requirement for building a Chinese word segmenter is obviously a dictionary; furthermore, as has been argued persuasively by Fung and Wu (1994), one will perform much better at segmenting text by using a dictionary constructed with text of the same genre as the text to be segmented. For novel texts, no lexicon that consists simply of a list of word entries will ever be entirely satisfactory, since the list will inevitably omit many constructions that should be considered words. Among these are words derived by various productive processes, including:

1. Morphologically derived words such as 學生們 *xue2-sheng1*+**men0** (student+**plural**) 'students,' which is derived by the affixation of the plural affix 們 *men0* to the noun 學生 *xue2-sheng1*.
2. Personal names such as 周恩來 *zhou1-en1-lai2* 'Zhou Enlai.' Of course, we can expect famous names like Zhou Enlai's to be in many dictionaries, but names such as 石基琳 *shi2-ji1-lin2*, the name of the second author of this paper, will not be found in any dictionary.
3. Transliterated foreign names such as 馬來西亞 *ma3-lai2-xi1-ya3* 'Malaysia.' Again, famous place names will most likely be found in the dictionary, but less well-known names, such as 布朗士維克 *bu4-lang3-shi4-wei2-ke4* 'Brunswick' (as in the New Jersey town name 'New Brunswick') will not generally be found.

In this paper we present a stochastic finite-state model for segmenting Chinese text into words, both words found in a (static) lexicon as well as words derived via the above-mentioned productive processes. The segmenter handles the grouping of hanzi into words and outputs word pronunciations, with default pronunciations for hanzi it cannot group; we focus here primarily on the system's ability to segment text appropriately (rather than on its pronunciation abilities). The model incorporates various recent techniques for incorporating and manipulating linguistic knowledge using finite-state transducers. It also incorporates the Good-Turing method (Baayen 1989; Church and Gale 1991) in estimating the likelihoods of previously unseen constructions, including morphological derivatives and personal names. We will evaluate various *specific* aspects of the segmentation, as well as the *overall* segmentation performance. This latter evaluation compares the performance of the system with that of several human judges since, as we shall show, even people do not agree on a single correct way to segment a text.

Finally, this effort is part of a much larger program that we are undertaking to develop stochastic finite-state methods for text analysis with applications to TTS and other areas; in the final section of this paper we will briefly discuss this larger program so as to situate the work discussed here in a broader context.

2. A Brief Introduction to the Chinese Writing System

Most readers will undoubtedly be at least somewhat familiar with the nature of the Chinese writing system, but there are enough common misunderstandings that it is as well to spend a few paragraphs on properties of the Chinese script that will be relevant to topics discussed in this paper.

The first point we need to address is what type of linguistic object a hanzi represents. Much confusion has been sown about Chinese writing by the use of the term *ideograph*, suggesting that hanzi somehow directly represent ideas. The most accurate characterization of Chinese writing is that it is **morphosyllabic** (DeFrancis 1984): each

hanzi represents one *morpheme* lexically and semantically, and one *syllable* phonologically. Thus in a two-hanzi word like 中國 *zhong1-guo2* (middle country) 'China' there are two syllables, and at the same time two morphemes. Of course, since the number of attested (phonemic) Mandarin syllables (roughly 1400, including tonal distinctions) is far smaller than the number of morphemes, it follows that a given syllable could in principle be written with any of several different hanzi, depending upon which morpheme is intended: the syllable *zhong1* could be 中 'middle,' 鐘 'clock,' 終 'end,' or 忠 'loyal.' A morpheme, on the other hand, usually corresponds to a unique hanzi, though there are a few cases where variant forms are found. Finally, quite a few hanzi are homographs, meaning that they may be pronounced in several different ways, and in extreme cases apparently represent different morphemes: The prenominal modification marker 的 *de0* is presumably a different morpheme from the second morpheme of 目的 *mu4-di4*, even though they are written the same way.[4]

The second point, which will be relevant in the discussion of personal names in Section 4.4, relates to the internal structure of hanzi. Following the system devised under the Qing emperor Kang Xi, hanzi have traditionally been classified according to a set of approximately 200 semantic **radicals**; members of a radical class share a particular structural component, and often also share a common meaning (hence the term 'semantic'). For example, hanzi containing the INSECT radical 虫 tend to denote insects and other crawling animals; examples include 蛙 *wa1* 'frog,' 蜂 *feng1* 'wasp,' and 蛇 *she2* 'snake.' Similarly, hanzi sharing the GHOST radical 鬼 tend to denote spirits and demons, such as 鬼 *gui3* 'ghost' itself, 魔 *mo2* 'demon,' and 魘 *yan3* 'nightmare.' While the semantic aspect of radicals is by no means completely predictive, the semantic homogeneity of many classes is quite striking: for example 254 out of the 263 examples (97%) of the INSECT class listed by Wieger (1965, 773–76) denote crawling or invertebrate animals; similarly 21 out of the 22 examples (95%) of the GHOST class (page 808) denote ghosts or spirits. As we shall argue, the semantic class affiliation of a hanzi constitutes useful information in predicting its properties.

3. Previous Work

There is a sizable literature on Chinese word segmentation: recent reviews include Wang, Su, and Mo (1990) and Wu and Tseng (1993). Roughly speaking, previous work can be divided into three categories, namely purely statistical approaches, purely lexical rule-based approaches, and approaches that combine lexical information with statistical information. The present proposal falls into the last group.

Purely statistical approaches have not been very popular, and so far as we are aware earlier work by Sproat and Shih (1990) is the only published instance of such an approach. In that work, mutual information was used to decide whether to group adjacent hanzi into two-hanzi words. Mutual information was shown to be useful in the segmentation task given that one does not have a dictionary. A related point is that mutual information is helpful in augmenting existing electronic dictionaries, (cf.

4 To be sure, it is not always true that a hanzi represents a syllable or that it represents a morpheme. For example, in Northern Mandarin dialects there is a morpheme *-r* that attaches mostly to nouns, and which is phonologically incorporated into the syllable to which it attaches: thus *men2+r* (door+R) 'door' is realized as *mer2*. This is orthographically represented as 兒 so that 'door' would be 門兒, and in this case the hanzi 兒 does not represent a syllable. Similarly, there is no compelling evidence that either of the syllables of 檳榔 *bin1-lang2* 'betelnut' represents a morpheme, since neither can occur in any context without the other: more likely 檳榔 *bin1-lang2* is a disyllabic morpheme. (See Sproat and Shih 1995.) However, the characterization given in the main body of the text is correct sufficiently often to be useful.

Church and Hanks [1989]), and we have used lists of character pairs ranked by mutual information to expand our own dictionary.

Nonstochastic lexical-knowledge-based approaches have been much more numerous. Two issues distinguish the various proposals. The first concerns how to deal with ambiguities in segmentation. The second concerns the methods used (if any) to extend the lexicon beyond the static list of entries provided by the machine-readable dictionary upon which it is based. The most popular approach to dealing with segmentation ambiguities is the **maximum matching** method, possibly augmented with further heuristics. This method, one instance of which we term the "greedy algorithm" in our evaluation of our own system in Section 5, involves starting at the beginning (or end) of the sentence, finding the longest word starting (ending) at that point, and then repeating the process starting at the next (previous) hanzi until the end (beginning) of the sentence is reached. Papers that use this method or minor variants thereof include Liang (1986), Li et al. (1991), Gu and Mao (1994), and Nie, Jin, and Hannan (1994). The simplest version of the maximum matching algorithm effectively deals with ambiguity by ignoring it, since the method is guaranteed to produce only one segmentation. Methods that allow multiple segmentations must provide criteria for choosing the best segmentation. Some approaches depend upon some form of constraint satisfaction based on syntactic or semantic features (e.g., Yeh and Lee [1991], which uses a unification-based approach). Others depend upon various lexical heuristics: for example Chen and Liu (1992) attempt to balance the length of words in a three-word window, favoring segmentations that give approximately equal length for each word. Methods for expanding the dictionary include, of course, morphological rules, rules for segmenting personal names, as well as numeral sequences, expressions for dates, and so forth (Chen and Liu 1992; Wang, Li, and Chang 1992; Chang and Chen 1993; Nie, Jin, and Hannan 1994).

Lexical-knowledge-based approaches that include statistical information generally presume that one starts with all possible segmentations of a sentence, and picks the best segmentation from the set of possible segmentations using a probabilistic or cost-based scoring mechanism. Approaches differ in the algorithms used for scoring and selecting the best path, as well as in the amount of contextual information used in the scoring process. The simplest approach involves scoring the various analyses by costs based on word frequency, and picking the lowest cost path; variants of this approach have been described in Chang, Chen, and Chen (1991) and Chang and Chen (1993). More complex approaches such as the **relaxation** technique have been applied to this problem Fan and Tsai (1988). Note that Chang, Chen, and Chen (1991), in addition to word-frequency information, include a constraint-satisfication model, so their method is really a hybrid approach. Several papers report the use of part-of-speech information to rank segmentations (Lin, Chiang, and Su 1993; Peng and Chang 1993; Chang and Chen 1993); typically, the probability of a segmentation is multiplied by the probability of the tagging(s) for that segmentation to yield an estimate of the total probability for the analysis.

Statistical methods seem particularly applicable to the problem of unknown-word identification, especially for constructions like names, where the linguistic constraints are minimal, and where one therefore wants to know not only that a particular sequence of hanzi might be a name, but that it is likely to be a name with some probability. Several systems propose statistical methods for handling unknown words (Chang et al. 1992; Lin, Chiang, and Su 1993; Peng and Chang 1993). Some of these approaches (e.g., Lin, Chiang, and Su [1993]) attempt to identify unknown words, but do not actually tag the words as belonging to one or another class of expression. This is not ideal for some applications, however. For instance, for TTS it is necessary to know

that a particular sequence of hanzi is of a particular category because that knowledge could affect the pronunciation; consider, for example the issues surrounding the pronunciation of 乾 *gan1/qian2* discussed in Section 1.

Following Sproat and Shih (1990), performance for Chinese segmentation systems is generally reported in terms of the dual measures of precision and recall.[5] It is fairly standard to report precision and recall scores in the mid to high 90% range. However, it is almost universally the case that no clear definition of what constitutes a "correct" segmentation is given, so these performance measures are hard to evaluate. Indeed, as we shall show in Section 5, even human judges differ when presented with the task of segmenting a text into words, so a definition of the criteria used to determine that a given segmentation is correct is crucial before one can interpret such measures. In a few cases, the criteria for correctness are made more explicit. For example Chen and Liu (1992) report precision and recall rates of over 99%, but this counts only the words that occur in the test corpus that also occur in their dictionary. Besides the lack of a clear definition of what constitutes a correct segmentation for a given Chinese sentence, there is the more general issue that the test corpora used in these evaluations differ from system to system, so meaningful comparison between systems is rendered even more difficult.

The major problem for all segmentation systems remains the coverage afforded by the dictionary and the lexical rules used to augment the dictionary to deal with unseen words. The dictionary sizes reported in the literature range from 17,000 to 125,000 entries, and it seems reasonable to assume that the coverage of the base dictionary constitutes a major factor in the performance of the various approaches, possibly more important than the particular set of methods used in the segmentation. Furthermore, even the size of the dictionary per se is less important than the appropriateness of the lexicon to a particular test corpus: as Fung and Wu (1994) have shown, one can obtain substantially better segmentation by tailoring the lexicon to the corpus to be segmented.

4. The Proposal

Chinese word segmentation can be viewed as a stochastic transduction problem. More formally, we start by representing the dictionary D as a Weighted Finite State Transducer (WFST) (Pereira, Riley, and Sproat 1994). Let H be the set of hanzi, p be the set of pinyin syllables with tone marks, and P be the set of grammatical part-of-speech labels. Then each arc of D maps either from an element of H to an element of p, or from ϵ—i.e., the empty string—to an element of P. More specifically, each word is represented in the dictionary as a sequence of arcs, starting from the initial state of D and labeled with an element S of H$\times$p, which is terminated with a **weighted** arc labeled with an element of $\epsilon\times$P. The weight represents the estimated cost (negative log probability) of the word. Next, we represent the input sentence as an unweighted finite-state acceptor (FSA) I over H. Let us assume the existence of a function *Id*, which takes as input an FSA A, and produces as output a transducer that maps all and only the strings of symbols accepted by A to themselves (Kaplan and Kay 1994). We can

5 Recall that precision is defined to be the number of correct hits divided by the total number of items selected; and that recall is defined to be the number of correct hits divided by the number of items that *should have been* selected.

then define the best segmentation to be the cheapest or **best** path in $Id(I) \circ D^*$ (i.e., $Id(I)$ composed with the transitive closure of D).[6]

Consider the abstract example illustrated in Figure 2. In this example there are four "input characters," A, B, C and D, and these map respectively to four "pronunciations" a, b, c and d. Furthermore, there are four "words" represented in the dictionary. These are shown, with their associated costs, as follows:

AB/nc	4.0
ABC/jj	6.0
CD/vb	5.0
D/nc	5.0

The minimal dictionary encoding this information is represented by the WFST in Figure 2(a). An input $ABCD$ can be represented as an FSA as shown in Figure 2(b). This FSA I can be segmented into words by composing $Id(I)$ with D^*, to form the WFST shown in Figure 2(c), then selecting the best path through this WFST to produce the WFST in Figure 2(d). This WFST represents the segmentation of the text into the words AB and CD, word boundaries being marked by arcs mapping between ϵ and part-of-speech labels.

Since the segmentation corresponds to the sequence of words that has the lowest summed unigram cost, the segmenter under discussion here is a zeroth-order model. It is important to bear in mind, though, that this is not an inherent limitation of the model. For example, it is well-known that one can build a finite-state bigram (word) model by simply assigning a state s_i to each word w_i in the vocabulary, and having (word) arcs leaving that state weighted such that for each w_j and corresponding arc a_j leaving s_i, the cost on a_j is the bigram cost of $w_i w_j$. (Costs for unseen bigrams in such a scheme would typically be modeled with a special **backoff** state.) In Section 6 we discuss other issues relating to how higher-order language models could be incorporated into the model.

4.1 Dictionary Representation

As we have seen, the lexicon of basic words and stems is represented as a WFST; most arcs in this WFST represent mappings between hanzi and pronunciations, and are costless. Each word is terminated by an arc that represents the transduction between ϵ and the part of speech of that word, weighted with an estimated cost for that word. The cost is computed as follows, where N is the corpus size and f is the frequency:

$$C = -\log\left(\frac{f}{N}\right) \tag{1}$$

Besides actual words from the base dictionary, the lexicon contains all hanzi in the Big 5 Chinese code,[7] with their pronunciation(s), plus entries for other characters that can be found in Chinese text, such as Roman letters, numerals, and special symbols. Note that hanzi that are not grouped into dictionary words (and are not identified as single-hanzi words), or into one of the other categories of words discussed in this paper, are left unattached and tagged as unknown words. Other strategies could readily

6 As a reviewer has pointed out, it should be made clear that the function for computing the best path is an instance of the Viterbi algorithm.

7 Big 5 is the most popular Chinese character coding standard in use in Taiwan and Hong Kong. It is based on the traditional character set rather than the simplified character set used in Singapore and Mainland China.

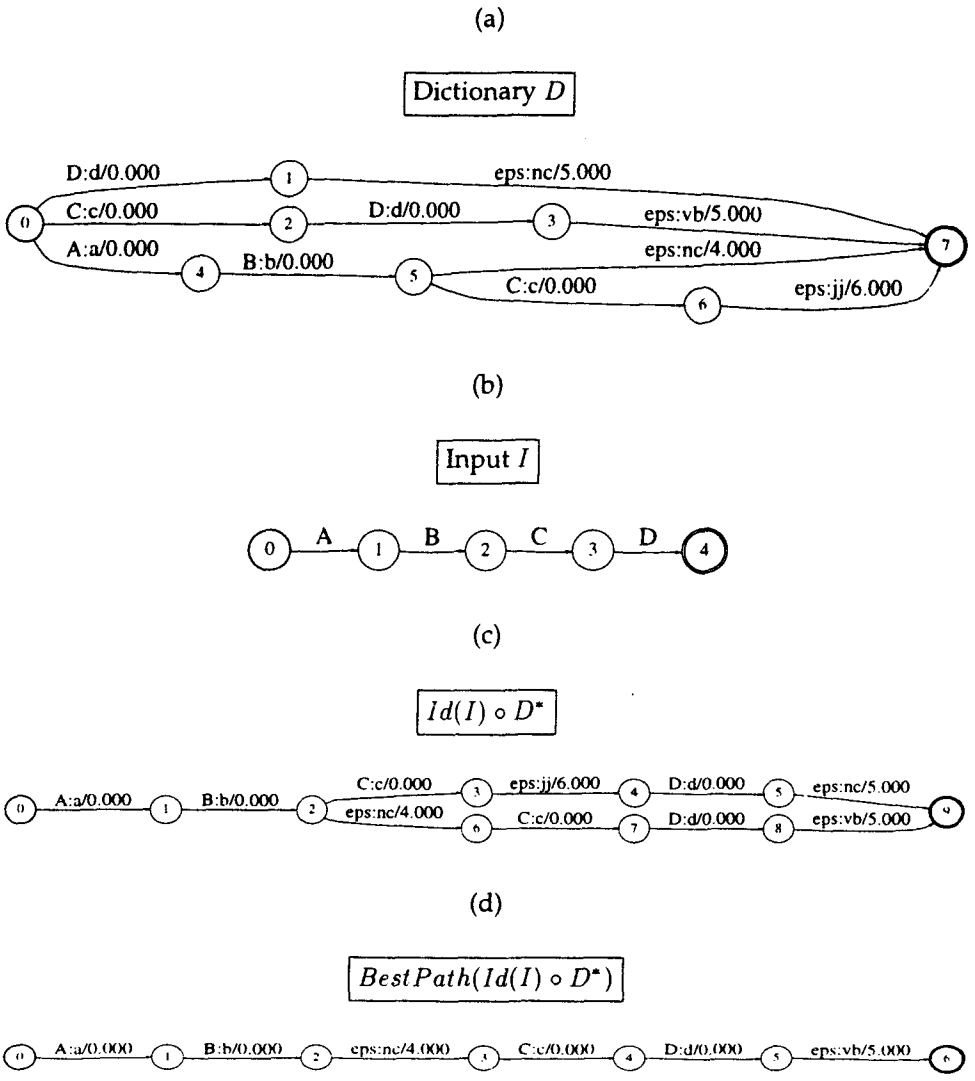

Figure 2
An abstract example illustrating the segmentation algorithm. The transitive closure of the dictionary in (a) is composed with *Id(input)* (b) to form the WFST (c). The segmentation chosen is the best path through the WFST, shown in (d). (In this figure *eps* is ϵ.)

be implemented, though, such as a maximal-grouping strategy (as suggested by one reviewer of this paper); or a pairwise-grouping strategy, whereby long sequences of unattached hanzi are grouped into two-hanzi words (which may have some prosodic motivation). We have not to date explored these various options.

Word frequencies are estimated by a re-estimation procedure that involves applying the segmentation algorithm presented here to a corpus of 20 million words,[8] using

8 Our training corpus was drawn from a larger corpus of mixed-genre text consisting mostly of newspaper material, but also including kungfu fiction, Buddhist tracts, and scientific material. This larger corpus was kindly provided to us by United Informatics Inc., R.O.C.

a set of initial estimates of the word frequencies.[9] In this re-estimation procedure only the entries in the base dictionary were used: in other words, derived words not in the base dictionary and personal and foreign names were *not* used. The best analysis of the corpus is taken to be the true analysis, the frequencies are re-estimated, and the algorithm is repeated until it converges. Clearly this is not the only way to estimate word-frequencies, however, and one could consider applying other methods: in particular since the problem is similar to the problem of assigning part-of-speech tags to an untagged corpus given a lexicon and some initial estimate of the a priori probabilities for the tags, one might consider a more sophisticated approach such as that described in Kupiec (1992); one could also use methods that depend on a small hand-tagged seed corpus, as suggested by one reviewer. In any event, to date, we have not compared different methods for deriving the set of initial frequency estimates. Note also that the costs currently used in the system are actually *string* costs, rather than *word* costs. This is because our corpus is not annotated, and hence does not distinguish between the various words represented by homographs, such as 將, which could be 將/adv *jiang1* 'be about to' or 將/nc *jiang4* '(military) general'—as in 小將 *xiao3-jiang4* 'little general.' In such cases we assign all of the estimated probability mass to the form with the most likely pronunciation (determined by inspection), and assign a very small probability (a very high cost, arbitrarily chosen to be 40) to all other variants. In the case of 將, the most common usage is as an adverb with the pronunciation *jiang1*, so that variant is assigned the estimated cost of 5.98, and a high cost is assigned to nominal usage with the pronunciation *jiang4*. The less favored reading may be selected in certain contexts, however; in the case of 將, for example, the nominal reading *jiang4* will be selected if there is morphological information, such as a following plural affix 們 *men0* that renders the nominal reading likely, as we shall see in Section 4.3.

Figure 3 shows a small fragment of the WFST encoding the dictionary, containing both entries for 將, just discussed, 中華民國 *zhong1-hua2 min2-guo2* (China Republic) 'Republic of China,' and 南瓜 *nan2-gua1* 'pumpkin.'

4.2 A Sample Segmentation Using Only Dictionary Words

Figure 4 shows two possible paths from the lattice of possible analyses of the input sentence 日文章魚怎麼說 'How do you say octopus in Japanese?' previously shown in Figure 1. As noted, this sentence consists of four words, namely 日文 *ri4-wen2* 'Japanese,' 章魚 *zhang1-yu2* 'octopus,' 怎麼 *zen3-me0* 'how,' and 說 *shuo1* 'say.' As indicated in Figure 1(c), apart from this correct analysis, there is also the analysis taking 日 *ri4* as a word (e.g., a common abbreviation for Japan), along with 文章 *wen2-zhang1* 'essay,' and 魚 *yu2* 'fish.' Both of these analyses are shown in Figure 4; fortunately, the correct analysis is also the one with the lowest cost, so it is this analysis that is chosen.

4.3 Morphological Analysis

The method just described segments dictionary words, but as noted in Section 1, there are several classes of words that should be handled that are not found in a standard dictionary. One class comprises words derived by productive morphological processes, such as plural noun formation using the suffix 們 *men0*. (Other classes handled by the current system are discussed in Section 5.) The morphological analysis itself can be handled using well-known techniques from finite-state morphol-

9 The initial estimates are derived from the frequencies in the corpus of the strings of hanzi making up each word in the lexicon *whether or not* each string is actually an instance of the word in question.

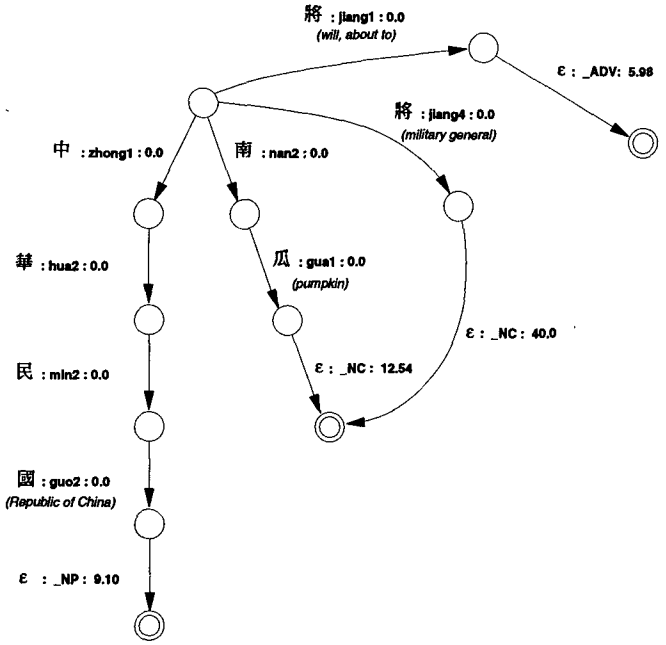

Figure 3
Partial Chinese Lexicon (NC = noun; NP = proper noun).

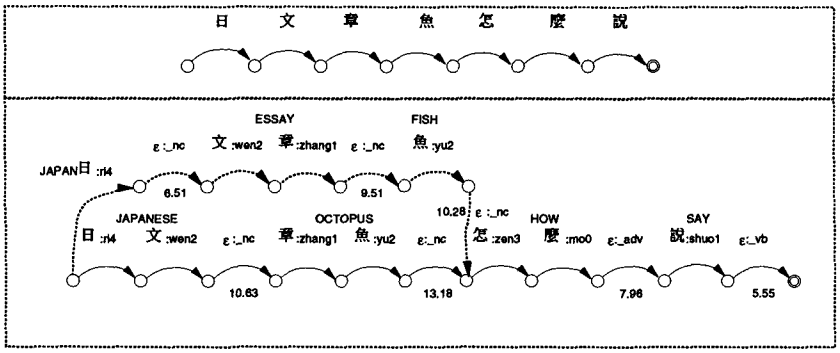

Figure 4
Input lattice (top) and two segmentations (bottom) of the sentence 'How do you say octopus in Japanese?'. A non-optimal analysis is shown with dotted lines in the bottom frame.

ogy (Koskenniemi 1983; Antworth 1990; Tzoukermann and Liberman 1990; Karttunen, Kaplan, and Zaenen 1992; Sproat 1992); we represent the fact that 們 attaches to nouns by allowing ϵ-transitions from the final states of all noun entries, to the initial state of the sub-WFST representing 們. However, for our purposes it is not sufficient to represent the morphological decomposition of, say, plural nouns: we also need an estimate of the cost of the resulting word. For derived words that occur in our corpus we can estimate these costs as we would the costs for an underived dictionary entry. So, 學生們 *xue2-sheng1+men0* (student+PL) 'students' occurs and we estimate its cost at 11.43; similarly we estimate the cost of 將們 *jiang4+men0* (general+PL) 'generals' (as in 小將們 *xiao3-jiang4+men0* 'little generals'), at 15.02.

But we also need an estimate of the probability for a non-occurring though possible plural form like 南瓜們 *nan2-gua1-men0* 'pumpkins.'[10] Here we use the Good-Turing estimate (Baayen 1989; Church and Gale 1991), whereby the aggregate probability of previously unseen instances of a construction is estimated as n_1/N, where N is the total number of observed tokens and n_1 is the number of types observed only once. Let us notate the set of previously unseen, or novel, members of a category X as *unseen*(X); thus, novel members of the set of words derived in 們 *men0* will be denoted *unseen*(們). For 們, the Good-Turing estimate just discussed gives us an estimate of $p(unseen(們) \mid 們)$—the probability of observing a previously unseen instance of a construction in 們, *given that we know that we have a construction in* 們. This Good-Turing estimate of $p(unseen(們) \mid 們)$ can then be used in the normal way to define the probability of finding a novel instance of a construction in 們 in a text: $p(unseen(們)) = p(unseen(們) \mid 們)\ p(們)$. Here $p(們)$ is just the probability of any construction in 們, as estimated from the frequency of such constructions in the corpus. Finally, assuming a simple bigram backoff model, we can derive the probability estimate for the particular unseen word 南瓜們, as the product of the probability estimate for 南瓜, and the probability estimate just derived for unseen plurals in 們: $p(南瓜們) \approx p(南瓜)p(unseen(們))$. The cost estimate, $cost(南瓜們)$, is computed in the obvious way by summing the negative log probabilities of 南瓜 and 們.

Figure 5 shows how this model is implemented as part of the dictionary WFST. There is a (costless) transition between the NC node and 們. The transition from 們 to a final state transduces ϵ to the grammatical tag PL with cost $cost(unseen(們))$: $cost(南瓜們) = cost(南瓜) + cost(unseen(們))$, as desired. For the seen word 將們 'generals,' there is an ϵ:NC transduction from 將 to the node preceding 們; this arc has cost $cost(將們) - cost(unseen(們))$, so that the cost of the whole path is the desired $cost(將們)$. This representation gives 將們 an appropriate morphological decomposition, preserving information that would be lost by simply listing 將們 as an unanalyzed form. Note that the backoff model assumes that there is a positive correlation between the frequency of a singular noun and its plural. An analysis of nouns that occur in both the singular and the plural in our database reveals that there is indeed a slight but significant positive correlation—$R^2 = 0.20$, $p < 0.005$; see Figure 6. This suggests that the backoff model is as reasonable a model as we can use in the absence of further information about the expected cost of a plural form.

10 Chinese speakers may object to this form, since the suffix 們 *men0* (PL) is usually restricted to attaching to terms denoting human beings. However, it is possible to personify any noun, so in children's stories or fables, 南瓜們 *nan2-gua1+men0* 'pumpkins' is by no means impossible.

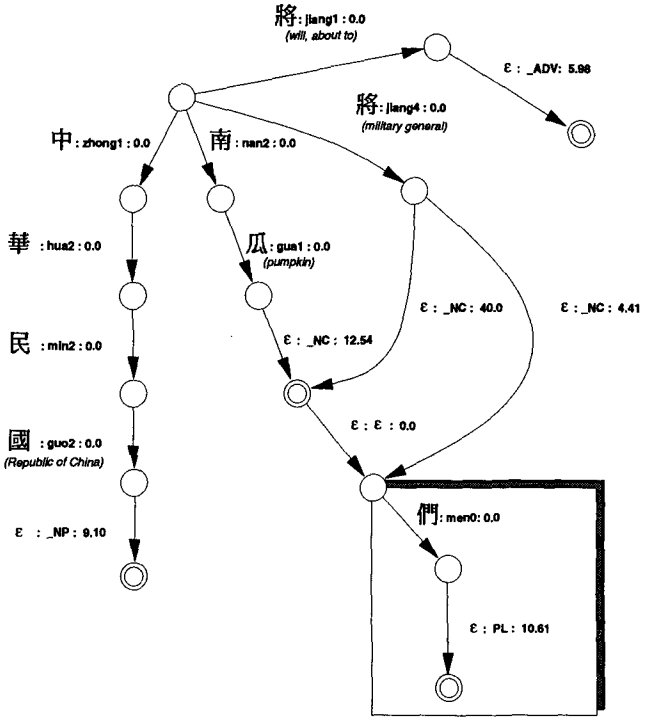

Figure 5
An example of affixation: the plural affix.

4.4 Chinese Personal Names

Full Chinese personal names are in one respect simple: they are always of the form *family+given*. The family name set is restricted: there are a few hundred single-hanzi family names, and about ten double-hanzi ones. Given names are most commonly two hanzi long, occasionally one hanzi long: there are thus four possible name types, which can be described by a simple set of context-free rewrite rules such as the following:

1. word ⇒ name
2. name ⇒ 1-hanzi-family 2-hanzi-given
3. name ⇒ 1-hanzi-family 1-hanzi-given
4. name ⇒ 2-hanzi-family 2-hanzi-given
5. name ⇒ 2-hanzi-family 1-hanzi-given
6. 1-hanzi-family ⇒ hanzi_i
7. 2-hanzi-family ⇒ hanzi_i hanzi_j
8. 1-hanzi-given ⇒ hanzi_i
9. 2-hanzi-given ⇒ hanzi_i hanzi_j

The difficulty is that given names can consist, in principle, of any hanzi or pair of hanzi, so the possible given names are limited only by the total number of hanzi, though some hanzi are certainly far more likely than others. For a sequence of hanzi that is a *possible* name, we wish to assign a probability to that sequence *qua* name. We can model this probability straightforwardly enough with a probabilistic version of the grammar just given, which would assign probabilities to the individual rules. For example, given a sequence $F_1G_1G_2$, where F_1 is a legal single-hanzi family name, and

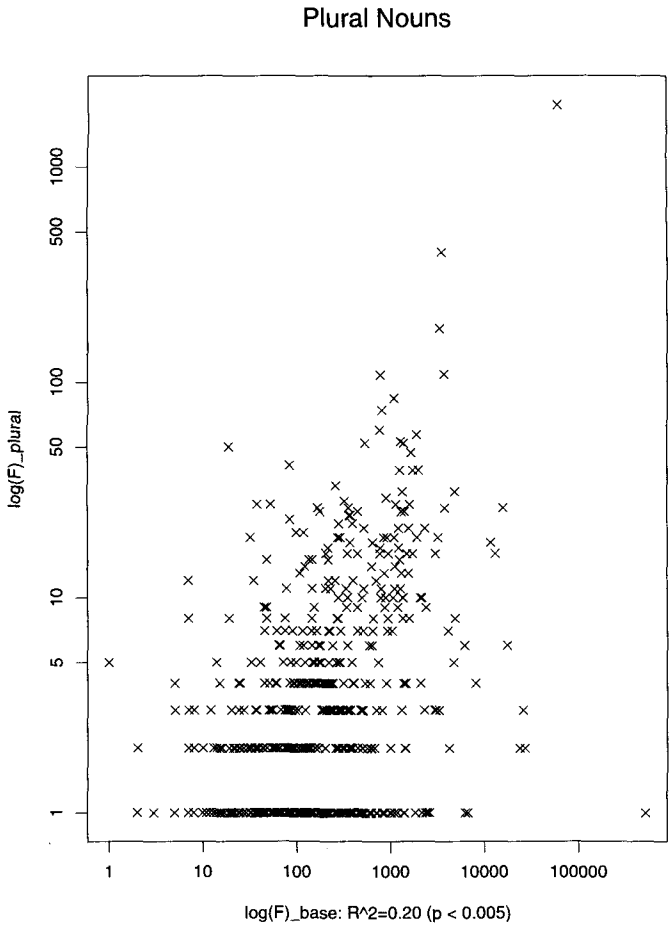

Figure 6
Plot of log frequency of base noun, against log frequency of plural nouns.

G_1 and G_2 are hanzi, we can estimate the probability of the sequence being a name as the product of:

- the probability that a word chosen randomly from a text will be a name—$p(rule\ 1)$, and
- the probability that the name is of the form 1-hanzi-family 2-hanzi-given—$p(rule\ 2)$, and
- the probability that the family name is the particular hanzi F_1—$p(rule\ 6)$, and
- the probability that the given name consists of the particular hanzi G_1 and G_2—$p(rule\ 9)$

This model is essentially the one proposed in Chang et al. (1992). The first probability is estimated from a name count in a text database, and the rest of the probabilities

are estimated from a large list of personal names.[11] Note that in Chang et al.'s model the $p(rule\ 9)$ is estimated as the product of the probability of finding G_1 in the first position of a two-hanzi given name and the probability of finding G_2 in the second position of a two-hanzi given name, and we use essentially the same estimate here, with some modifications as described later on.

This model is easily incorporated into the segmenter by building a WFST restricting the names to the four licit types, with costs on the arcs for any particular name **summing** to an estimate of the cost of that name. This WFST is then summed with the WFST implementing the dictionary and morphological rules, and the transitive closure of the resulting transducer is computed; see Pereira, Riley, and Sproat (1994) for an explanation of the notion of summing WFSTs.[12]

Conceptual Improvements over Chang et al.'s Model. There are two weaknesses in Chang et al.'s model, which we improve upon. First, the model assumes independence between the first and second hanzi of a double given name. Yet, some hanzi are far more probable in women's names than they are in men's names, and there is a similar list of male-oriented hanzi: mixing hanzi from these two lists is generally less likely than would be predicted by the independence model. As a partial solution, for pairs of hanzi that co-occur sufficiently often in our namelists, we use the estimated bigram cost, rather than the independence-based cost.

The second weakness is purely conceptual, and probably does not affect the performance of the model. For previously unseen hanzi in given names, Chang et al. assign a uniform small cost; but we *know* that some unseen hanzi are merely accidentally missing, whereas others are missing for a reason—for example, because they have a bad connotation. As we have noted in Section 2, the general semantic class to which a hanzi belongs is often predictable from its semantic radical. Not surprisingly some semantic classes are better for names than others: in our corpora, many names are picked from the GRASS class but very few from the SICKNESS class. Other good classes include JADE and GOLD; other bad classes are DEATH and RAT.

We can better predict the probability of an unseen hanzi occurring in a name by computing a **within-class** Good-Turing estimate for each radical class. Assuming unseen objects *within each class* are equiprobable, their probabilities are given by the Good-Turing theorem as:

$$p_0^{cls} \propto \frac{E(n_1^{cls})}{N * E(N_0^{cls})} \tag{2}$$

where p_0^{cls} is the probability of one unseen hanzi in class *cls*, $E(n_1^{cls})$ is the expected number of hanzi in *cls* seen once, N is the total number of hanzi, and $E(N_0^{cls})$ is the expected number of unseen hanzi in class *cls*. The use of the Good-Turing equation presumes suitable estimates of the unknown expectations it requires. In the denomi-

11 We have two such lists, one containing about 17,000 full names, and another containing frequencies of hanzi in the various name positions, derived from a million names.

12 One class of full personal names that this characterization does not cover are married women's names where the husband's family name is optionally prepended to the woman's full name; thus 許林鹽海 *xu3-lin2-yan2-hai3* would represent the name that Ms. Lin Yanhai would take if she married someone named Xu. This style of naming is never required and seems to be losing currency. It is formally straightforward to extend the grammar to include these names, though it does increase the likelihood of overgeneration and we are unaware of any working systems that incorporate this type of name.

We of course also fail to identify, by the methods just described, given names used without their associated family name. This is in general very difficult, given the extremely free manner in which Chinese given names are formed, and given that in these cases we lack even a family name to give the model confidence that it is identifying a name.

Table 1
The cost as a novel given name (second position) for hanzi from various radical classes.

JADE	GOLD	GRASS	SICKNESS	DEATH	RAT
14.98	15.52	15.76	16.25	16.30	16.42

nator, the N_0^{cls} can be measured well by counting, and we replace the expectation by the observation. In the numerator, however, the counts of n_1^{cls} are quite irregular, including several zeros (e.g., RAT, none of whose members were seen). However, there is a strong relationship between n_1^{cls} and the number of hanzi in the class. For $E(n_1^{cls})$, then, we substitute a smooth S against the number of class elements. This smooth guarantees that there are no zeroes estimated. The final estimating equation is then:

$$p_0^{cls} \propto \frac{S(n_1^{cls})}{N * N_0^{cls}} \tag{3}$$

Since the total of all these class estimates was about 10% off from the Turing estimate n_1/N for the probability of *all* unseen hanzi, we renormalized the estimates so that they would sum to n_1/N.

This class-based model gives reasonable results: for six radical classes, Table 1 gives the estimated cost for an unseen hanzi in the class occurring as the second hanzi in a double GIVEN name. Note that the good classes JADE, GOLD and GRASS have lower costs than the bad classes SICKNESS, DEATH and RAT, as desired, so the trend observed for the results of this method is in the right direction.

4.5 Transliterations of Foreign Words

Foreign names are usually transliterated using hanzi whose sequential pronunciation mimics the source language pronunciation of the name. Since foreign names can be of any length, and since their original pronunciation is effectively unlimited, the identification of such names is tricky. Fortunately, there are only a few hundred hanzi that are particularly common in transliterations; indeed, the commonest ones, such as 巴 *ba1*, 爾 *er3*, and 阿 *a1* are often clear indicators that a sequence of hanzi containing them is foreign: even a name like 夏米爾 *xia4-mi3-er3* 'Shamir,' which is a legal Chinese personal name, retains a foreign flavor because of 爾. As a first step towards modeling transliterated names, we have collected all hanzi occurring more than once in the roughly 750 foreign names in our dictionary, and we estimate the probability of occurrence of each hanzi in a transliteration ($p_{TN}(hanzi_i)$) using the maximum likelihood estimate. As with personal names, we also derive an estimate from text of the probability of finding a transliterated name of any kind (p_{TN}). Finally, we model the probability of a new transliterated name as the product of p_{TN} and $p_{TN}(hanzi_i)$ for each $hanzi_i$ in the putative name.[13] The foreign name model is implemented as an WFST, which is then summed with the WFST implementing the dictionary, morpho-

13 The current model is too simplistic in several respects. For instance, the common "suffixes," *-nia* (e.g., *Virginia*) and *-sia* are normally transliterated as 尼亞 *ni2-ya3* and 西亞 *xi1-ya3*, respectively. The interdependence between 尼 or 西, and 亞 is not captured by our model, but this could easily be remedied.

logical rules, and personal names; the transitive closure of the resulting machine is then computed.

5. Evaluation

In this section we present a partial evaluation of the current system, in three parts. The first is an evaluation of the system's ability to mimic humans at the task of segmenting text into word-sized units; the second evaluates the proper-name identification; the third measures the performance on morphological analysis. To date we have not done a separate evaluation of foreign-name recognition.

Evaluation of the Segmentation as a Whole. Previous reports on Chinese segmentation have invariably cited performance either in terms of a single percent-correct score, or else a single precision-recall pair. The problem with these styles of evaluation is that, as we shall demonstrate, even human judges do not agree perfectly on how to segment a given text. Thus, rather than give a single evaluative score, we prefer to compare the performance of our method with the judgments of *several* human subjects. To this end, we picked 100 sentences at random containing 4,372 total hanzi from a test corpus.[14] (There were 487 marks of punctuation in the test sentences, including the sentence-final periods, meaning that the average inter-punctuation distance was about 9 hanzi.) We asked six native speakers—three from Taiwan (T1–T3), and three from the Mainland (M1–M3)—to segment the corpus. Since we could not bias the subjects towards a particular segmentation and did not presume linguistic sophistication on their part, the instructions were simple: subjects were to mark all places they might plausibly pause if they were reading the text aloud. An examination of the subjects' bracketings confirmed that these instructions were satisfactory in yielding plausible word-sized units. (See also Wu and Fung [1994].)

Various segmentation approaches were then compared with human performance:

1. A **greedy** algorithm (or maximum-matching algorithm), **GR**: proceed through the sentence, taking the longest match with a dictionary entry at each point.
2. An **anti-greedy** algorithm, **AG**: instead of the longest match, take the shortest match at each point.
3. The method being described—henceforth **ST**.

Two measures that can be used to compare judgments are:

1. **Precision**. For each pair of judges consider one judge as the standard, computing the precision of the other's judgments relative to this standard.
2. **Recall**. For each pair of judges, consider one judge as the standard, computing the recall of the other's judgments relative to this standard.

Clearly, for judges J_1 and J_2, taking J_1 as standard and computing the precision and recall for J_2 yields the same results as taking J_2 as the standard, and computing for J_1,

14 All evaluation materials, with the exception of those used for evaluating personal names were drawn from the subset of the United Informatics corpus not used in the training of the models.

Table 2
Similarity matrix for segmentation judgments.

Judges	AG	GR	ST	M1	M2	M3	T1	T2	T3
AG		0.70	0.70	0.43	0.42	0.60	0.60	0.62	0.59
GR			0.99	0.62	0.64	0.79	0.82	0.81	0.72
ST				0.64	0.67	0.80	0.84	0.82	0.74
M1					0.77	0.69	0.71	0.69	0.70
M2						0.72	0.73	0.71	0.70
M3							0.89	0.87	0.80
T1								0.88	0.82
T2									0.78

respectively, the recall and precision. We therefore used the arithmetic mean of each interjudge precision-recall pair as a single measure of interjudge similarity. Table 2 shows these similarity measures. The average agreement among the human judges is .76, and the average agreement between **ST** and the humans is .75, or about 99% of the interhuman agreement.[15] One can better visualize the precision-recall similarity matrix by producing from that matrix a distance matrix, computing a classical metric multidimensional scaling (Torgerson 1958; Becker, Chambers, Wilks 1988) on that distance matrix, and plotting the first two most significant dimensions. The result of this is shown in Figure 7. The horizontal axis in this plot represents the most significant dimension, which explains 62% of the variation. In addition to the automatic methods, **AG**, **GR**, and **ST**, just discussed, we also added to the plot the values for the current algorithm *using only dictionary entries* (i.e., no productively derived words or names). This is to allow for fair comparison between the statistical method and **GR**, which is also purely dictionary-based. As can be seen, **GR** and this "pared-down" statistical method perform quite similarly, though the statistical method is still slightly better.[16] **AG** clearly performs much less like humans than these methods, whereas the full statistical algorithm, including morphological derivatives and names, performs most closely to humans among the automatic methods. It can also be seen clearly in this plot that two of the Taiwan speakers cluster very closely together, and the third Taiwan speaker is also close in the most significant dimension (the x axis). Two of the Mainlanders also cluster close together but, interestingly, not particularly close to the Taiwan speakers; the third Mainlander is much more similar to the Taiwan speakers.

The breakdown of the different types of words found by **ST** in the test corpus is given in Table 3. Clearly the percentage of productively formed words is quite small (for this particular corpus), meaning that dictionary entries are covering most of the

15 **GR** is .73 or 96%.

16 As one reviewer points out, one problem with the unigram model chosen here is that there is still a tendency to pick a segmentation containing fewer words. That is, given a choice between segmenting a sequence *abc* into *abc* and *ab*, *c*, the former will always be picked so long as its cost does not exceed the summed costs of *ab* and *c*: while; it is possible for *abc* to be so costly as to preclude the larger grouping, this will certainly not usually be the case. In this way, the method reported on here will necessarily be similar to a greedy method, though of course not identical.

As the reviewer also points out, this is a problem that is shared by, e.g., probabilistic context-free parsers, which tend to pick trees with fewer nodes. The question is how to normalize the probabilities in such a way that smaller groupings have a better shot at winning. This is an issue that we have not addressed at the current stage of our research.

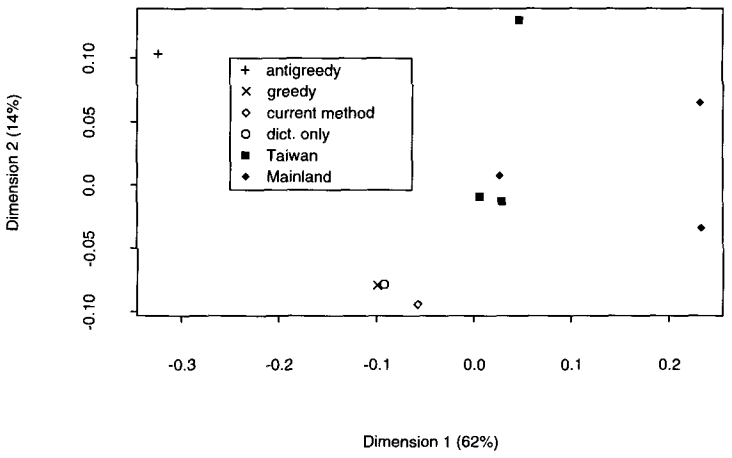

Figure 7
Classical metric multidimensional scaling of distance matrix, showing the two most significant dimensions. The percentage scores on the axis labels represent the amount of variation in the data explained by the dimension in question.

Table 3
Classes of words found by **ST** for the test corpus.

Word type	N	%
Dictionary entries	2,543	97.47
Morphologically derived words	3	0.11
Foreign transliterations	9	0.34
Personal names	54	2.07

cases. Nonetheless, the results of the comparison with human judges demonstrates that there is mileage being gained by incorporating models of these types of words.

It may seem surprising to some readers that the interhuman agreement scores reported here are so low. However, this result is consistent with the results of experiments discussed in Wu and Fung (1994). Wu and Fung introduce an evaluation method they call *nk-blind*. Under this scheme, n human judges are asked independently to segment a text. Their results are then compared with the results of an automatic segmenter. For a given "word" in the automatic segmentation, if at least k of the human judges agree that this is a word, then that word is considered to be correct. For eight judges, ranging k between 1 and 8 corresponded to a precision score range of 90% to 30%, meaning that there were relatively few words (30% of those found by the automatic segmenter) on which all judges agreed, whereas most of the words found by the segmenter were such that one human judge agreed.

Proper-Name Identification. To evaluate proper-name identification, we randomly selected 186 sentences containing 12,000 hanzi from our test corpus and segmented the text automatically, tagging personal names; note that for names, there is always a single unambiguous answer, unlike the more general question of which segmentation is correct. The performance was 80.99% recall and 61.83% precision. Interestingly, Chang et al. report 80.67% recall and 91.87% precision on an 11,000 word corpus: seemingly, our system finds as many names as their system, but with four times as many false hits. However, we have reason to doubt Chang et al.'s performance claims. Without using the same test corpus, direct comparison is obviously difficult; fortunately, Chang et al. include a list of about 60 sentence fragments that exemplify various categories of performance for their system. The performance of our system on those sentences appeared rather better than theirs. On a set of 11 sentence fragments—the A set—where they reported 100% recall and precision for name identification, we had 73% recall and 80% precision. However, they list two sets, one consisting of 28 fragments and the other of 22 fragments, in which they had 0% recall and precision. On the first of these—the B set—our system had 64% recall and 86% precision; on the second—the C set—it had 33% recall and 19% precision. Note that it is in precision that our overall performance would appear to be poorer than the reported performance of Chang et al., yet based on their published examples, our system appears to be doing better precisionwise. Thus we have some confidence that our own performance is at least as good as that of Chang et al. (1992).

In a more recent study than Chang et al., Wang, Li, and Chang (1992) propose a surname-driven, non-stochastic, rule-based system for identifying personal names.[17] Wang, Li, and Chang also compare their performance with Chang et al.'s system. Fortunately, we were able to obtain a copy of the full set of sentences from Chang et al. on which Wang, Li, and Chang tested their system, along with the output of their system.[18] In what follows we will discuss all cases from this set where our performance on names differs from that of Wang, Li, and Chang. Examples are given in Table 4. In these examples, the names identified by the two systems (if any) are underlined; the sentence with the correct segmentation is boxed.[19]

The differences in performance between the two systems relate directly to three issues, which can be seen as differences in the tuning of the models, rather than representing differences in the capabilities of the model per se. The first issue relates to the completeness of the base lexicon. The Wang, Li, and Chang system fails on fragment (b) because their system lacks the word 幽幽 *you1-you1* 'soberly' and misinterpreted the thus isolated first 幽 *you1* as being the final hanzi of the preceding name; similarly our system failed in fragment (h) since it is missing the abbreviation 台獨 *tai2-du2* 'Taiwan Independence.' This is a rather important source of errors in name identification, and it is not really possible to objectively evaluate a name recognition system without considering the main lexicon with which it is used.

17 They also provide a set of title-driven rules to identify names when they occur before titles such as 先生 *xian1sheng1* 'Mr.' or 台北市長 *tai2bei3 shi4zhang3* 'Taipei Mayor.' Obviously, the presence of a title after a potential name *N* increases the probability that *N* is in fact a name. Our system does not currently make use of titles, but it would be straightforward to do so within the finite-state framework that we propose.

18 We are grateful to Chao-Huang Chang for providing us with this set. Note that Wang, Li, and Chang's set was based on an earlier version of the Chang et al. paper, and is missing 6 examples from the A set.

19 We note that it is not always clear in Wang, Li, and Chang's examples which segmented words constitute names, since we have only their segmentation, not the actual classification of the segmented words. Therefore in cases where the *segmentation* is identical between the two systems we assume that tagging is also identical.

Table 4
Differences in performance between our system and Wang, Li, and Chang (1992).

	Our System	Wang, Li, and Chang	Transliteration/Translation
a.	陳中 申曲	陳中申曲	*chen2-zhong1-shen1 qu3* 'music by Chen Zhongshen '
b.	黃蓉幽幽的道	黃蓉幽幽的道	*huang2-rong2 you1-you1 de dao4* 'Huang Rong said soberly'
c.	張群	張群	*zhang1 qun2* Zhang Qun
d.	縣長尤清上 任後	縣長尤清上任後	*xian4-zhang3 you2-qing1 shang4-ren2 hou4* 'after the county president You Qing had assumed the position'
e.	林全	林全	*lin2 quan2* 'Lin Quan'
f.	王建	王建	*wang2 jian4* 'Wang Jian'
g.	歐陽克	歐陽克	*ou1-yang2-ke4* 'Ouyang Ke'
h.	因其不可能容許台獨而	因其不可能許容台獨而	*yin1 qi2 bu4 ke2-neng2 rong2-xu3 tai2-du2 er2* 'because it cannot permit Taiwan Independence so'
i.	司法院長林洋港	司法院長林洋港	*si1-fa3-yuan4-zhang3 lin2-yang2-gang3* 'president of the Judicial Yuan, Lin Yanggang'
j.	林章湖將做現場解說	林章湖將做現場解說	*lin2-zhang1-hu2 jiang1 zuo4 xian4-chang3 jie3-shuo1* 'Lin Zhanghu will give an explanation live'
k.	近兩年內撒下的金錢會停止	近兩年內撒下的金錢會停止	*jin4 liang3 nian2 nei4 sa3 xia4 de jin1-qian2 hui4 ting2-zhi3* 'in two years the distributed money will stop'
l.	高湯大匙椰子粉	高湯大匙椰子粉	*gao1-tang1 da4-chi2 ye1-zi0 fen3* 'chicken stock, a tablespoon of coconut flakes'
m.	尤清入主縣府後	尤清入主縣府後	*you2-qing1 ru4-zhu3 xian4-fu3 hou4* 'after You Qing headed the county government'

Table 5
Performance on morphological analysis.

Affix	Pron	Base category	*N* found	*N* missed (recall)	*N* correct (precision)
不下	*bu2-xia4*	verb	20	12 (63%)	20 (100%)
不下去	*bu2-xia4-qu4*	verb	30	1 (97%)	29 (97%)
不了	*bu4-liao3*	verb	72	15 (83%)	72 (100%)
得了	*de2-liao3*	verb	36	11 (77%)	36 (100%)
們	*men0*	noun	141	6 (96%)	139 (99%)

The second issue is that rare family names can be responsible for overgeneration, especially if these names are otherwise common as single-hanzi words. For example, the Wang, Li, and Chang system fails on the sequence 年內撒 *nian2 nei4 sa3* in (k) since 年 *nian2* is a possible, but rare, family name, which also happens to be written the same as the very common word meaning 'year.' Our system fails in (a) because of 申 *shen1*, a rare family name; the system identifies it as a family name, whereas it should be analyzed as part of the given name.

Finally, the statistical method fails to correctly group hanzi in cases where the individual hanzi comprising the name are listed in the dictionary as being relatively high-frequency single-hanzi words. An example is in (i), where the system fails to group 林洋港 *lin2yang2gang3* as a name, because all three hanzi can in principle be separate words (林 *lin2* 'wood'; 洋 *yang2* 'ocean'; 港 *gang3* 'harbor'). In many cases these failures in recall would be fixed by having better estimates of the actual probabilities of single-hanzi words, since our estimates are often inflated. A totally nonstochastic rule-based system such as Wang, Li, and Chang's will generally succeed in such cases, but of course runs the risk of overgeneration wherever the single-hanzi word is really intended.

Evaluation of Morphological Analysis. In Table 5 we present results from small test corpora for the productive affixes handled by the current version of the system; as with names, the segmentation of morphologically derived words is generally either right or wrong. The first four affixes are so-called resultative affixes: they denote some property of the resultant state of a verb, as in 忘不了 *wang4-bu4-liao3* (forget-not-attain) 'cannot forget.' The last affix in the list is the nominal plural 們 *men0*.[20] In the table are the (typical) classes of words to which the affix attaches, the number found in the test corpus by the method, the number correct (with a precision measure), and the number missed (with a recall measure).

6. Discussion

In this paper we have argued that Chinese word segmentation can be modeled effectively using weighted finite-state transducers. This architecture provides a uniform framework in which it is easy to incorporate not only listed dictionary entries but also morphological derivatives, and models for personal names and foreign names in transliteration. Other kinds of productive word classes, such as company names, abbreviations (termed 縮寫 *suo1-xie3* in Mandarin), and place names can easily be

20 Note that 了 in 忘不了 is normally pronounced as *le0*, but as part of a resultative it is *liao3*.

handled given appropriate models. (For some recent corpus-based work on Chinese abbreviations, see Huang, Ahrens, and Chen [1993].)

We have argued that the proposed method performs well. However, some caveats are in order in comparing this method (or *any* method) with other approaches to segmentation reported in the literature. First of all, most previous articles report performance in terms of a single percent-correct score, or else in terms of the paired measures of precision and recall. What both of these approaches presume is that there is a single correct segmentation for a sentence, against which an automatic algorithm can be compared. We have shown that, at least given independent human judgments, this is not the case, and that therefore such simplistic measures should be mistrusted. This is not to say that a set of standards by which a particular segmentation would count as correct and another incorrect could not be devised; indeed, such standards have been proposed and include the published PRCNSC (1994) and ROCLING (1993), as well as the unpublished Linguistic Data Consortium standards (ca. May 1995). However, until such standards are universally adopted in evaluating Chinese segmenters, claims about performance in terms of simple measures like percent correct should be taken with a grain of salt; see, again, Wu and Fung (1994) for further arguments supporting this conclusion.

Second, comparisons of different methods are not meaningful unless one can evaluate them on the same corpus. Unfortunately, there is no standard corpus of Chinese texts, tagged with either single or multiple human judgments, with which one can compare performance of various methods. One hopes that such a corpus will be forthcoming.

Finally, we wish to reiterate an important point. The major problem for our segmenter, as for all segmenters, remains the problem of unknown words (see Fung and Wu [1994]). We have provided methods for handling certain classes of unknown words, and models for other classes could be provided, as we have noted. However, there will remain a large number of words that are not readily adduced to any productive pattern and that would simply have to be added to the dictionary. This implies, therefore, that a major factor in the performance of a Chinese segmenter is the quality of the base dictionary, and this is probably a more important factor—from the point of view of performance alone—than the particular computational methods used.

The method reported in this paper makes use solely of unigram probabilities, and is therefore a zeroeth-order model: the cost of a particular segmentation is estimated as the sum of the costs of the individual words in the segmentation. However, as we have noted, nothing inherent in the approach precludes incorporating higher-order constraints, provided they can be effectively modeled within a finite-state framework. For example, as Gan (1994) has noted, one can construct examples where the segmentation is locally ambiguous but can be determined on the basis of sentential or even discourse context. Two sets of examples from Gan are given in (1) and (2) (= Gan's Appendix B, exx. 11a/11b and 14a/14b respectively). In (1) the sequence 馬路 *ma3-lu4* cannot be resolved locally, but depends instead upon broader context; similarly in (2), the sequence 才能 *cai2-neng2* cannot be resolved locally:

1. (a)	這	匹	馬	路	上	病	了
	zhe4	*pi1*	**ma3**	**lu4**	*shang4*	*bing4*	*le0*
	this	CL(assifier)	horse	way	on	sick	ASP(ect)

'This horse got sick on the way'

(b)

這	條	馬路	很	少	車	經過
zhe4	*tiao2*	**ma3-lu4**	*hen3*	*shao3*	*che1*	*jing1-guo4*
this	CL	road	very	few	car	pass by

'Very few cars pass by this road'

2. (a)

甚麼	時候	我	才	能	克服	這	個	困難
shen3-me0	*shi2-hou4*	*wo3*	**cai2**	**neng2**	*ke4-fu2*	*zhe4*	*ge4*	*kun4-*
what	time	I	just	be able	overcome	this	CL	diffic

'When will I be able to overcome this difficulty?'

(b)

他	的	才能	很	高
ta1	*de*	**cai2-neng2**	*hen3*	*gao1*
he	DE	talent	very	high

'He has great talent'

While the current algorithm correctly handles the (b) sentences, it fails to handle the (a) sentences, since it does not have enough information to know not to group the sequences 馬路 *ma3-lu4* and 才能 *cai2-neng2* respectively. Gan's solution depends upon a fairly sophisticated language model that attempts to find valid syntactic, semantic, and lexical relations between objects of various linguistic types (hanzi, words, phrases). An example of a fairly low-level relation is the **affix relation**, which holds between a stem morpheme and an affix morpheme, such as 們 *-men0* (PL). A high-level relation is **agent**, which relates an animate nominal to a predicate. Particular instances of relations are associated with goodness scores. Particular relations are also consistent with particular hypotheses about the segmentation of a given sentence, and the scores for particular relations can be incremented or decremented depending upon whether the segmentations with which they are consistent are "popular" or not.

While Gan's system incorporates fairly sophisticated models of various linguistic information, it has the drawback that it has only been tested with a very small lexicon (a few hundred words) and on a very small test set (thirty sentences); there is therefore serious concern as to whether the methods that he discusses are scalable. Another question that remains unanswered is to what extent the linguistic information he considers can be handled—or at least approximated—by finite-state language models, and therefore could be directly interfaced with the segmentation model that we have presented in this paper. For the examples given in (1) and (2) this certainly seems possible. Consider first the examples in (2). The segmenter will give both analyses 才 能 *cai2 neng2* 'just be able,' and 才能 *cai2-neng2* 'talent,' but the latter analysis is preferred since splitting these two morphemes is generally more costly than grouping them. In (2a), we want to split the two morphemes since the correct analysis is that we have the adverb 才 *cai2* 'just,' the modal verb 能 *neng2* 'be able' and the main verb 克服 *ke4-fu2* 'overcome'; the competing analysis is, of course, that we have the noun 才能 *cai2-neng2* 'talent,' followed by 克服 *ke4-fu2* 'overcome.' Clearly it is possible to write a rule that states that if an analysis Modal + Verb is available, then that is to be preferred over Noun + Verb: such a rule could be stated in terms of (finite-state) **local grammars** in the sense of Mohri (1993). Turning now to (1), we have the similar problem that splitting 馬路 into 馬 *ma3* 'horse' and 路 *lu4* 'way' is more costly than retaining this as one word 馬路 *ma3-lu4* 'road.' However, there is again local grammatical information that should favor the split in the case of (1a): both 馬 *ma3* 'horse' and 馬路 *ma3-lu4* are nouns, but only 馬 *ma3* is consistent with the classifier 匹 *pi1*, the classifier

for horses.[21] By a similar argument, the preference for not splitting 馬路 could be strengthened in (1b) by the observation that the classifier 條 *tiao2* is consistent with long or winding objects like 馬路 *ma3-lu4* 'road' but not with 馬 *ma3* 'horse.' Note that the sets of possible classifiers for a given noun can easily be encoded on that noun by grammatical features, which can be referred to by finite-state grammatical rules. Thus, we feel fairly confident that for the examples we have considered from Gan's study a solution can be incorporated, or at least approximated, within a finite-state framework.

With regard to purely morphological phenomena, certain processes are not handled elegantly within the current framework. Any process involving reduplication, for instance, does not lend itself to modeling by finite-state techniques, since there is no way that finite-state networks can directly implement the copying operations required. Mandarin exhibits several such processes, including A-not-A question formation, illustrated in (3a), and adverbial reduplication, illustrated in (3b):

3. (a) 是 *shi4* 'be' ⇒ 是不是 *shi4-bu2-shi4* (be-not-be) 'is it?'
高興 *gao1-xing4* 'happy' ⇒ 高不高興 *gao1-bu4-gao1-xing4* (hap-not-happy) 'happy?'

(b) 高興 *gao1-xing4* 'happy' ⇒ 高高興興 *gao1-gao1-xing4-xing4* 'happily'

In the particular form of A-not-A reduplication illustrated in (3a), the first syllable of the verb is copied, and the negative marker 不 *bu4* 'not' is inserted between the copy and the full verb. In the case of adverbial reduplication illustrated in (3b) an adjective of the form *AB* is reduplicated as *AABB*. The only way to handle such phenomena within the framework described here is simply to expand out the reduplicated forms beforehand, and incorporate the expanded forms into the lexical transducer.

7. Conclusions

Despite these limitations, a purely finite-state approach to Chinese word segmentation enjoys a number of strong advantages. The model we use provides a simple framework in which to incorporate a wide variety of lexical information in a uniform way. The use of weighted transducers in particular has the attractive property that the model, as it stands, can be straightforwardly interfaced to other modules of a larger speech or natural language system: presumably one does not want to segment Chinese text for its own sake but instead with a larger purpose in mind. As described in Sproat (1995), the Chinese segmenter presented here fits directly into the context of a broader finite-state model of text analysis for speech synthesis. Furthermore, by inverting the transducer so that it maps from phonemic transcriptions to hanzi sequences, one can apply the segmenter to other problems, such as speech recognition (Pereira, Riley, and Sproat 1994). Since the transducers are built from human-readable descriptions using a lexical toolkit (Sproat 1995), the system is easily maintained and extended. While size of the resulting transducers may seem daunting—the segmenter described here, as it is used in the Bell Labs Mandarin TTS system has about 32,000 states and 209,000 arcs—recent work on minimization of weighted machines and transducers (cf.

21 In Chinese, numerals and demonstratives cannot modify nouns directly, and must be accompanied by a classifier. The particular classifier used depends upon the noun.

Mohri [1995]) shows promise for improving this situation. The model described here thus demonstrates great potential for use in widespread applications. This flexibility, along with the simplicity of implementation and expansion, makes this framework an attractive base for continued research.

Acknowledgments

We thank United Informatics for providing us with our corpus of Chinese text, and BDC for the 'Behavior Chinese-English Electronic Dictionary.' We further thank Dr. J.-S. Chang of Tsinghua University, Taiwan, R.O.C., for kindly providing us with the name corpora.

We also thank Chao-Huang Chang, reviewers for the 1994 ACL conference, and four anonymous reviewers for *Computational Linguistics* for useful comments.

References

Antworth, Evan. 1990. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*. Occasional Publications in Academic Computing, 16. Summer Institute of Linguistics, Dallas, TX.

Baayen, Harald. 1989. *A Corpus-Based Approach to Morphological Productivity: Statistical Analysis and Psycholinguistic Interpretation*. Ph.D. thesis, Free University, Amsterdam.

Becker, Richard, John Chambers, and Allan Wilks. 1988. *The New S Language*. Wadsworth and Brooks, Pacific Grove.

Chang, Chao-Huang and Cheng-Der Chen. 1993. A study on integrating Chinese word segmentation and part-of-speech tagging. *Communications of the Chinese and Oriental Languages Information Processing Society*, 3(2):69–77.

Chang, Jyun-Shen, C.-D. Chen, and Shun-De Chen. 1991. Xianzhishi manzu ji jilu zuijiahua de zhongwen duanci fangfa [Chinese word segmentation through constraint satisfaction and statistical optimization]. In *Proceedings of ROCLING IV*, pages 147–165, Taipei. ROCLING.

Chang, Jyun-Shen, Shun-De Chen, Ying Zheng, Xian-Zhong Liu, and Shu-Jin Ke. 1992. Large-corpus-based methods for Chinese personal name recognition. *Journal of Chinese Information Processing*, 6(3):7–15.

Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley, CA.

Chen, Keh-Jiann and Shing-Huan Liu. 1992. Word identification for Mandarin Chinese sentences. In *Proceedings of COLING-92*, pages 101–107. COLING.

Church, Kenneth and William Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5(1):19–54.

Church, Kenneth and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Morristown, NJ. Association for Computational Linguistics.

DeFrancis, John. 1984. *The Chinese Language*. University of Hawaii Press, Honolulu.

Fan, C.-K. and W.-H. Tsai. 1988. Automatic word identification in Chinese sentences by the relaxation technique. *Computer Processing of Chinese and Oriental Languages*, 4:33–56.

Fung, Pascale and Dekai Wu. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *WVLC-94, Second Annual Workshop on Very Large Corpora*.

Gan, Kok-Wee. 1994. *Integrating Word Boundary Disambiguation with Sentence Understanding*. Ph.D. thesis, National University of Singapore.

Gu, Ping and Yuhang Mao. 1994. Hanyu zidong fenci de jinlin pipei suanfa ji qi zai QHFY hanying jiqi fanyi xitong zhong de shixian [The adjacent matching algorithm of Chinese automatic word segmentation and its implementation in the QHFY Chinese-English system]. In *International Conference on Chinese Computing*, Singapore.

Hirschberg, Julia. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63:305–340.

Huang, Chu-Ren, Kathleen Ahrens, and Keh-jiann Chen. 1993. A data-driven approach to psychological reality of the mental lexicon: Two studies on Chinese corpus linguistics. Presented at the conference on Language and its Psychobiological Bases, December.

Kaplan, Ronald and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20:331–378.

Karttunen, Lauri, Ronald Kaplan, and Annie Zaenen. 1992. Two-level

morphology with composition. In *COLING-92*, pages 141–148. COLING.
Koskenniemi, Kimmo. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki, Helsinki.
Kupiec, Julian. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*. Submitted.
Li, B.Y., S. Lin, C.F. Sun, and M.S. Sun. 1991. Yi zhong zhuyao shiyong yuliaoku biaoji jinxing qiyi jiaozheng de zuida pipei hanyu zidong fenci suanfa sheji [A maximum-matching word segmentation algorithm using corpus tags for disambiguation]. In *ROCLING IV*, pages 135–146, Taipei. ROCLING.
Li, Charles and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley, CA.
Liang, Nanyuan. 1986. Shumian hanyu zidong fenci xitong-CDWS [A written Chinese automatic segmentation system–CDWS]. *Journal of Chinese Information Processing*, 1(1):44–52.
Lin, Ming-Yu, Tung-Hui Chiang, and Keh-Yi Su. 1993. A preliminary study on unknown word problem in Chinese word segmentation. In *ROCLING 6*, pages 119–141. ROCLING.
Mohri, Mehryar. 1993. *Analyse et représentation par automates de structures syntaxiques composées*. Ph.D. thesis, University of Paris 7, Paris.
Mohri, Mehryar. 1995. Minimization algorithms for sequential transducers. *Theoretical Computer Science*. Submitted.
Nagata, Masaaki. 1994. A stochastic Japanese morphological analyzer using a forward-DP backward A* N-best search algorithm. In *Proceedings of COLING-94*, pages 201–207. COLING.
Nie, Jian-Yun, Wanying Jin, and Marie-Louise Hannan. 1994. A hybrid approach to unknown word detection and segmentation of Chinese. In *International Conference on Chinese Computing*, Singapore.
Peng, Z.-Y. and J-S. Chang. 1993. Zhongwen cihui qiyi zhi yanjiu—duanci yu cixing biaoshi [Research on Chinese lexical ambiguity—segmentation and part-of-speech tagging]. In *ROCLING 6*, pages 173–193. ROCLING.
Pereira, Fernando, Michael Riley, and Richard Sproat. 1994. Weighted rational transductions and their application to human language processing. In *ARPA Workshop on Human Language Technology*, pages 249–254. Advanced Research Projects Agency, March 8–11.
PRCNSC, 1994. *Contemporary Chinese Language Word Segmentation Specification for Information Processing*. People's Republic of China National Standards Committee. In Chinese.
ROCLING. 1993. Jisuan yuyanxue tongxun [Computational linguistics communications]. Newsletter of the Republic of China Computational Linguistics Society (ROCLING), April. In Chinese.
Shih, Chilin. 1986. *The Prosodic Domain of Tone Sandhi in Chinese*. Ph.D. thesis, UCSD, La Jolla, CA.
Sproat, Richard. 1992. *Morphology and Computation*. MIT Press, Cambridge, MA.
Sproat, Richard. 1994. English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language*, 8:79–94.
Sproat, Richard. 1995. A finite-state architecture for tokenization and grapheme-to-phoneme conversion for multilingual text analysis. In Susan Armstrong and Evelyne Tzoukermann, editors, *Proceedings of the EACL SIGDAT Workshop*, pages 65–72, Dublin, Ireland. Association for Computational Linguistics.
Sproat, Richard and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4:336–351.
Sproat, Richard and Chilin Shih. 1995. A corpus-based analysis of Mandarin nominal root compounds. *Journal of East Asian Linguistics*, 4(1):1–23.
Tang, Ting-Chih. 1988. *Hanyu Cifa Jufa Lunji [Studies on Chinese Morphology and Syntax]*, volume 2. Student Book Company, Taipei. In Chinese.
Tang, Ting-Chih. 1989. *Hanyu Cifa Jufa Xuji [Studies on Chinese Morphology and Syntax: 2]*, volume 2. Student Book Company, Taipei. In Chinese.
Torgerson, Warren. 1958. *Theory and Methods of Scaling*. Wiley, New York.
Tzoukermann, Evelyne and Mark Liberman. 1990. A finite-state morphological processor for Spanish. In *COLING-90, Volume 3*, pages 3: 277–286. COLING.
Wang, Liang-Jyh, Wei-Chuan Li, and Chao-Huang Chang. 1992. Recognizing unregistered names for Mandarin word identification. In *Proceedings of COLING-92*, pages 1239–1243. COLING.
Wang, Michelle and Julia Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and*

Language, 6:175–196.
Wang, Yongheng, Haiju Su, and Yan Mo. 1990. Automatic processing of Chinese words. *Journal of Chinese Information Processing*, 4(4):1–11.
Wieger, L. 1965. *Chinese Characters*. Dover, New York. Republication of second edition, published 1927 by Catholic Mission Press.
Wu, Dekai and Pascale Fung. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 180–181, Stuttgart, October.
Wu, Zimin and Gwyneth Tseng. 1993. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, 44(9):532–542.
Yeh, Ching-long and Hsi-Jian Lee. 1991. Rule-based word identification for Mandarin Chinese sentences—a unification approach. *Computer Processing of Chinese and Oriental Languages*, 5(2):97–118.