

Computational Tools and Resources for Linguistic Studies

Yu-Ling Una Hsu*, Jing-Shin Chang⁺, Keh-Yih Su⁺

ABSTRACT

This paper presents several useful computational tools and available resources to facilitate linguistic studies. For each computational tool, we demonstrate why it is useful and how can it be used for research. In addition, linguistic examples are given for illustration. First, a very useful searching engine, Key Word in Context (KWIC), is introduced. This tool can automatically extract linguistically significant patterns from large corpora and help linguists discover syntagmatic generalizations. Second, Dynamic Clustering and Hierarchical Clustering are introduced for identifying natural clusters of words or phrases in distribution. Third, statistical measures which could be used to measure the degree of cohesion and correlation among linguistic units are presented. These tools can help linguists identify the boundaries of lexical units. Fourth, alignment tools for aligning parallel texts at the word, sentence and structure levels are presented for linguists who do comparative studies of different languages. Fifth, we introduce Sequential Forward Selection (SFS) and Classification and Regression Tree (CART) for automatic rule ordering. Finally, some available electronic Chinese resources are described to provide reference purposes for those who are interested.

keywords: extraction, clustering, cohesion, alignment, Chinese corpora, electronic dictionary

1. Introduction

Owing to advances in computer technology in providing cheap and fast computation power, and to the increasing availability of machine-readable corpora, corpus-based statistics-oriented (CBSO) approaches [Su 1996] have been gaining prevalence in the community of computational linguistics recently. Many computational and statistical tools have been proposed for building and testing linguistic analyses, and experimental

* Behavior Design Corporation, No.5, 2F, Industrial East Road IV, Science-Based Industrial Park, Hsinchu, Taiwan. E-mail: una@bdc.com.tw

⁺ Department of Electrical Engineering, National Tsing-Hua University, Hsinchu, Taiwan.
E-mail: shin@hermes.ee.nthu.edu.tw; kysu@bdc.com.tw

results have been encouraging. Such computational tools allow people to find more discriminative features and more effective rules for useful natural language applications; they also provide ways to verify conventional linguistic theories with large scale unrestricted texts. This paper introduces several useful computational tools and available resources to facilitate linguistic studies. The functions of these tools include: 1. automatic extraction of linguistic patterns, 2. automatic identification of linguistic classes, 3. automatic evaluation of linguistic cohesion and correlation, 4. automatic comparison of parallel texts, and 5. automatic grammar rule selection. With these tools, linguists can verify their hypotheses with a much broader range of authentic linguistic material in a quicker and more objective way. As G. K. Pullum said: "Some consequences of an entire grammar cannot be seen by the unaided human brain, just as some visual details cannot be seen by the unaided human eyes ([Shieber 1985], p 191)," we believe linguists could give a more systematic, satisfactory, and insightful account of linguistic phenomena if they had access to better tools and larger corpora.

2. Computational Tools

Computational tools can be used to automatically detect generalizations over a large set of data and to help form hypotheses. There are at least four kinds of generalizations that computational tools can help to formulate: 1) syntagmatic: syntactic patterns and correlational generalizations involving co-occurring elements; 2) distributional: the natural clustering in distribution that serves as the basis for categorization; 3) cohesive: the typical internal co-relation of constituents that defines a linguistic unit; and 4) comparative: the correspondence between two languages or texts. We will introduce the computational tools in turn with linguistic examples.

First, we introduce a very useful searching engine: Key Word in Context (KWIC). This tool can help linguists automatically extract their research focuses from large corpora. Linguists can specify their key words with Exact Match, Partial Match or Fuzzy Match, and organize the contexts of the key words to form a useful automatic concordance. Second, we present techniques, such as Dynamic Clustering and Hierarchical Clustering, for finding natural clusters of words or phrases in distribution. Such techniques allow linguists to cluster linguistic units into classes so that linguistic generalizations can be captured effectively in terms of a finite set of classes. Third, statistical measures which can be used to measure the degree of cohesion among linguistic units are presented. In particular, Mutual Information, Dice Metric and Entropy are introduced. Such measures are useful for identifying the correlations among constituents of a known linguistic unit as well as to discover the boundaries of an unknown linguistic unit. Fourth, alignment tools for aligning parallel texts at the word, sentence and structure

levels are presented. These alignment tools are very useful for linguists who do comparative studies on different languages. Finally, we introduce Sequential Forward Selection (SFS) and Classification and Regression Tree (CART). They can provide an objective measure for rule selection.

2.1 Automatic Extraction of Linguistic Patterns : Key Word in Context (KWIC)

The most frequently used computational tools in linguistic studies are searching tools, such as Key Word in Context (KWIC). The KWIC tool is, in essence, an automatic concordancing tool. It can extract desired key words or phrases with their contexts from large corpora. The lines extracted are centered on the key words and show a few preceding and following words. [Huang 1994] presented an interesting KWIC-based corpus linguistic study on the co-occurrences of the negator '不' and auxiliary/aspect '有'。 Most Chinese linguists may be familiar with the transformation rule proposed in [Wang 1965], as shown in Figure 1.

Condition: 不 - 有 X
 1 2 3
Change: 沒 3

Figure 1 Transformation rule which changes '不有' to '沒'
[Wang 1965]

According to the above transformation rule, we should predict that there are no cooccurring pairs of '不有' in Mandarin Chinese. However, with the help of the KWIC tool, [Huang 1994] found 48 instances of '不' occurring before '有' from a 20 million character corpus. Figure 2 shows part of the KWIC result for the key word 「不有」

部，是怕財政負擔太重，而又不能	〈不有〉	[福利]，於是便產生了一個不倫
天下事無奇	〈不有〉	，口湖鄉民王坤芬家裡飼養一隻極
毒梟藏放安非他命的花招無奇	〈不有〉	，台北市刑警大隊偵七隊昨天即查
音機到飛機及軍事武器雷達等，無	〈不有〉	其蹤跡，影響力可謂無遠弗屆。日
的另一大警訊，更迫使 Fed 不得	〈不有〉	所回應。而由 Fed 一反常例，未
銀行的潛在競爭敵人，老銀行不能	〈不有〉	所防範。
起同仇敵愾之感，他們說，警方若	〈不有〉	個交代，到時走著瞧。（許司任）
整治的青草湖，如果上游濫墾問題	〈不有〉	效防範，這筆巨額工程經費，勢必
經官員頻發佈景氣復甦的消息，豈	〈不有〉	欺騙大眾之嫌。

Figure 2 Part of the KWIC result for the key word 「不有」
[Huang 1994]

From the KWIC result, we found that, even though the syntactic suppletion rule of '沒' makes correct predictions regarding an overwhelming majority of the data, it does not account for all the facts. After a careful study of the exceptions, [Huang 1994] suggested that grammaticality cannot be predicted with no-exception syntactic rules but must be captured in terms of a set of lexically governed rules.

As exemplified above, KWIC is a very useful and time-saving pre-filter which helps to automatically extract linguistically significant patterns from large corpora. It is especially helpful in discovering syntagmatic generalizations that range from morphemic to sentential structural patterns. Further, these generalizations can then be abstracted to form theoretical hypotheses, to verify proposed formal accounts, or they can be applied to computational systems.

To be more specific, there are two important dimensions of KWIC: the keyword and the context. It should be noted that the computational tools allow us to manipulate both the keyword and the context. Besides the exact match function exemplified above, Section 2.1.1 goes on to discuss the fact that a keyword can either be partially specified (thus useful for matching affixes or compound components etc.) or be discontinuous (thus useful for looking for patterns that may extend beyond a word or a clause). Section 2.1.2 then deals with the error-tolerance aspect as well as the semantically/pragmatically close not-exact matches of a keyword. Finally, Section 2.1.3 introduces the tools and ideas in automatic organization of contexts (i.e., automatic concordance), which not only put similar examples together, but also make it easy to see which are the most typical uses.

2.1.1 Partial Match

Sometimes the patterns we are searching for are slightly different from one another and we do not want to exhaustively list them as key words, or sometimes the words we are interested in are disconnected, and we do not care what kinds of words or phrases will occur in between. In these cases, what we actually need is a tool capable of matching partially specified key words. A KWIC tool with a partial match function allows users to use wild card characters to extract patterns that are either partially specified or discontinuous; thus, it is a useful tool for linguists to match affixes or compound components etc. or to look for patterns that may extend beyond a word or a clause. For example, if one wants to find some phrases similar to '從今以後', he can simply input '從今 ? 後' as a key word and get a list of candidate phrases as shown on the left side of Figure 3, or if one wants to study the relationship between '從' and '後', or to observe what kinds of phrases can occur between them, he can input '從 * 後' as a key word and automatically get a list of patterns including '從' and '後' as shown on the right side of

Figure 3 (where "?" represents the wild card for matching a single character, and "*" represents the wild card for matching any number of characters) :

「從今 ? 後」 :	<u>從今以後</u> <u>從今之後</u> <u>從今而後</u> <u>從今天後</u> :	「從 * 後」 :	<u>從此後</u> <u>從門後</u> <u>從今以後</u> <u>從此之後</u> <u>從那人身後</u> <u>從她嫁給張三之後</u> :
------------	---	-----------	--

Figure 3 Outputs of a partial match

Such tools are usually implemented by using a finite state machine [Aho 1986] which is capable of describing string patterns in regular expressions. For example, the UNIX 'grep' ("get regular expression") tool and its variants are supported in many computer platforms due to their usefulness in such applications.

2.1.2 Fuzzy Match

When we are interested in finding patterns that are semantically or pragmatically close to a certain key word, either the exact match function or the partial match function may not provide a satisfactory output since they are based on strict substring matching. The fuzzy match function outputs patterns according to a pre-defined distance measure between the searching pattern and the patterns in the corpus. Unlike the exact match function and the partial match function, which require that the extracted pattern contain the input pattern as a substring, the fuzzy match function finds substrings that are "similar" to the input pattern; thus it has the error-tolerance capability. For instance, the string "ADC", "ABD", and "DBC" are all considered sufficiently similar to an input string "ABC" in the sense that only one character in either string is different. The KWIC tools with the fuzzy match function therefore allow users to locate strings that are subject to typographic errors (such as '國民大會' and '國民太會'), writing variants (such as '台積', '台積電' and '台灣積體電路'), and closely related patterns (such as '語言學', '計算語言學', '計算語言學會' and '語言學會會員'). For example, if the distance measure is defined in terms of the number of characters subject to insertion, deletion, and replacement operations with respect to the input pattern, and if the weights for such operations are all assigned to 1, then the fuzzy match outputs of '從今以後' with distance 1 are those listed in Figure 4.

「從今以後」— distance 1 :

從今天以後	1 insertion
從今後	1 deletion
從此以後	1 replacement
從今而後	1 replacement
:	

Figure 4 Outputs of a fuzzy match

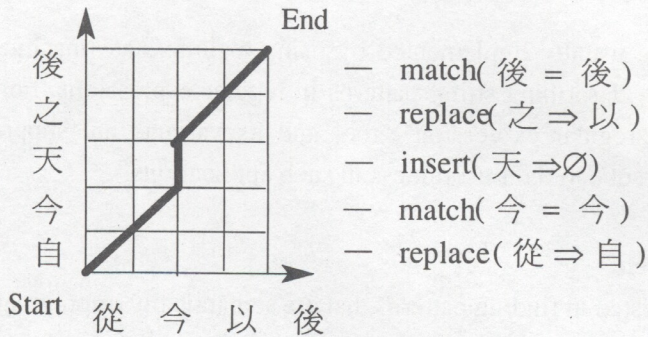


Figure 5 Path to find the minimum distance for a string pair

The evaluation of the distance can be formulated as a "shortest path" searching problem which can be solved using the well-known dynamic programming technique [Denardo 1982]. Figure 5 shows a sample diagram for the dynamic programming problem. Any path along the edges or diagonals from the Start to the End represents a sequence of operations that changes the input searching pattern to the pattern chosen from the corpus. There are three directions in which to go at any position. If we go right, a deletion is performed. If we go upward, an insertion is performed. If we go up-right, either one of the following two cases will happen: when the characters on the two edges of the diagonal are the same, no operation is performed; if, however, they are different, a replacement is performed.

Each path in the figure is associated with a distance according to the distance assigned to the editing operations. The distance between the two input strings is evaluated as the distance which is the minimum among all the paths. For instance, the path in Figure

4 corresponds to a distance of 3; it requires two replacement operations (從 ⇒ 自, and 之 ⇒ 以) and one insertion operation (insert the '天' character after the '今' character), where each such editing operation is assigned a unity distance, to change the input string to a partial string in the corpus. If this path turns out to be the one with minimum distance, we will say that '自今天之後' has a distance of 3 with respect to '從今以後'.

2.1.3 Organizing the Context and Automatic Concordance

In addition to manipulation of the key word, the KWIC tool allows linguists to organize the contexts and, thus, provides them with an automatic concordance for the key word. For example, the KWIC tool used at Academia Sinica allows linguists to specify the size of both the right-hand and left-hand side contexts shown as the search result [Huang 1994, Chen 1996]. It also allows linguists to choose from different sub-corpora to control both the size and the domain for the search. Most important to a linguist, the search result is sorted according to a user-specified context. Linguists can sort the search for the left or right context of a certain keyword; hence, similar examples will be put together, and linguists can easily see which are the most typical uses. Generalizations over sentences involving that specific context can thus be automatically detected.

For example, in the above-mentioned studies on the co-occurrence of '不' and '有' in [Huang 1994], 48 instances of '不有' were found with the help of the KWIC tool from a 20 million character corpus. By sorting the contexts, it could be easily observed that 16 instances of '不有' occurred as the latter part of the idiom chunk '無奇不有', 11 instances involved the double negation constructs '不能不', '不得不', or '安得不', and 13 instances occurred with a preceding negative polarity item '豈', and so on. Linguists can carefully examine the similarities and subtle distinctions among these examples in the KWIC result, and a satisfactory account can be given for each type of co-occurrences of '不' and '有'.

2.2 Automatic Identification of Linguistic Classes

Linguistic studies usually involve formulating rules to correctly account for the behavior of linguistic elements. However, before writing rules, linguists have to group elements with similar characteristics together to form a class. A familiar instance is using grammatical categories, which represents classes of words having similar syntactic behavior, to describe the syntax of a language. Different classifications will lead to different sets of rules and, thus, different kinds of analyses; therefore, the way of classification is very important. In the following, two automatic clustering techniques, namely dynamic clustering and hierarchical clustering [Devijver 1982], are introduced. Clustering corresponds to categorization in linguistics. Clustering techniques automatically group closely related elements together, so they can help linguists identify

linguistically significant categories, but interpretation of the classes has to be done by linguists/theorists.

2.2.1 Dynamic Clustering

The dynamic clustering approach is an iterative algorithm employed to optimize a clustering criterion function, such as the minimum distance among class members in each cluster. In that algorithm, we are given a set of data tokens and a pre-specified number of clusters K to be clustered. In each iteration of the dynamic clustering algorithm, data are assigned to one of the clusters according to a distance (or similarity) function. The representative of each cluster, which is usually defined as the centroid of data in the cluster, is then updated. The new cluster model is then used in the next iteration to reclassify the data. Such an iterative procedure continues until the class members in each cluster do not change anymore or a stopping criterion is satisfied.

For instance, in [張 1994], the dynamic clustering approach with the vector quantization technique [Duda 1973], which will be described later, was adopted for clustering of 5000 frequently used Chinese words into classes, and such classes were used to improve the performance of a speech recognition system. A few examples of the results are given in Figure 6. Every row in Figure 6 represents a specific class.

密集 低迷 老舊 激烈 辛苦 重視 熱烈
 度 屆 季
 遍 步 票 片 頓 趨 劑 陣 系列
 噸 戶 公噸根 顆 隻份 棟 套 句 座
 式 部會 階層 共和國
 局 事務所
 凌晨 清晨 深夜 上午

Figure 6 Part of the output clusters for frequently-used words
 [張 1994]

In the vector quantization process mentioned above, the 'vector' for a word in [張 1994] consists of a number of frequencies, each frequency being the co-occurrence frequency between this word and a particular left hand side neighbor of this word in all word bigram pairs; two words which have similar distribution in their co-occurrence

frequencies with their left neighbors are considered similar to each other. Such similarity is then used for clustering.

A typical procedure for dynamic clustering is K-means clustering [Devijver 1982, Schalkoff 1992], which is described as follows:

Initialization: Arbitrarily partition the data set $Y = \{y_1, y_2, \dots, y_n\}$ into K clusters, C_1, C_2, \dots, C_K , where C_j is represented by its mean vector μ_j over n_j data, $j=1, 2, \dots, K$, and

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ji}, \quad y_{ji} \in C_j,$$

where y_{ji} is the i -th token of class j , and μ_j is the mean of class j . (One way to do this is to randomly pick out K tokens as the initial centroids of the K clusters and then classify the data to the class corresponding to the nearest centroid.)

Step1: Assign each data $y_i, i=1, 2, \dots, n$, to the cluster C_j if

$$j = \operatorname{argmin}_k D(y_i, \mu_k),$$

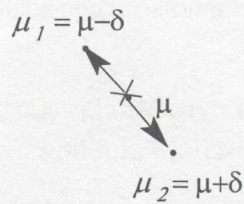
where $D(y_i, \mu_k)$ is the distance between data y_i and the mean of cluster k , and the *argmin* operator returns the argument (k) of the D function which corresponds to the minimum distance.

(Note: the minimum distance criterion could be replaced by other criteria.)

Step2: Recalculate the mean vectors μ_j , as in the Initialization step, for $j=1, 2, \dots, K$.

Step3: Terminate the clustering procedure if the mean vectors remain unchanged or if the convergence criterion is satisfied. Otherwise, go to step 1.

The K-means algorithm can be modified in several ways [Schalkoff 1992]. For instance, we can start with one cluster and generate an additional new cluster in each iteration by splitting one existing cluster until K clusters are obtained. In each iteration, one cluster is selected based on a selection criterion. Two vectors μ_1 and μ_2 are then derived based on the mean vector μ of the selected cluster by slightly shifting the original mean by a small vector δ in two opposite directions, i.e.,

$$\begin{aligned} \mu_1 &= \mu + \delta \\ \mu_2 &= \mu - \delta \end{aligned}$$


Then, all the tokens in the selected cluster are re-classified with respect to these two new means.

2.2.2 Hierarchical Clustering

Another clustering technique, known as the hierarchical clustering technique, acquires clusters by merging two clusters that are most 'similar' in each iteration, i.e., by combining class members that are closely related. [Brown 1992], for instance, used this algorithm to automatically divide 260,741 different English words into 1,000 classes. Figure 7 contains examples of classes that are particularly interesting. Each row in Figure 7 denotes a specific class.

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays ...
 June March July April January December October November September August ...
 down backwards ashore sideways southward northward overboard aloft downwards ...
 water gas coal liquid acid sand carbon steam shale iron ...
 great big vast sudden mere sheer gigantic lifelong scant colossal ...
 American Indian European Japanese German African Catholic Israeli Italian Arab ...
 pressure temperature permeability density porosity stress velocity viscosity gravity ...
 machine device controller processor CPU printer spindle subsystem compiler plotter ...
 John George James Bob Robert Paul William Jim David Mike ...
 anyone someone anybody somebody ...
 had hadn't has would've could've should've must've might've ...

Figure 7 Classes from a 260,741-word vocabulary [Brown 1992]

[Redington 1995] utilized the distributional information and hierarchical clustering technique to identify syntactic categories in Mandarin Chinese. It was revealed that

clusters which resulted from the hierarchical clustering procedure could roughly match the canonical classification (proposed by Academia Sinica [陳 1991]). Several word classification studies with an English corpus have reached similar conclusions [Redington 1995]. It was suggested in [Redington 1995] that the distributional information might be an important source of information for human language learning .

The hierarchical clustering algorithm is performed in a bottom-up fashion, where two of the most similar clusters are merged to form a new cluster at each stage. Since each merging action will reduce the number of clusters by one, this algorithm terminates after $n-1$ steps, where n is the number of data. In addition, the number of clusters in the hierarchical clustering algorithm need not be known *a priori*. The algorithm of the hierarchical clustering algorithm is shown as follows:

Initialization: Each data point in $Y = \{y_1, y_2, \dots, y_n\}$ represents an individual cluster, i.e., $C_j = \{y_j\}$, $j=1, 2, \dots, n$.

Step1: Find C_p and C_r such that $(p, r) = \underset{\forall (j, k), j \neq k}{\operatorname{argmin}} D(C_j, C_k)$, where $D(C_p, C_r)$ is the distance measure between clusters p and r .

Step2: Merge C_p with C_r and delete C_p .

Step3: Go to step1 until the number of clusters is equal to 1.

2.3 Automatic Evaluation of Linguistic Cohesion and Correlation

For linguists or lexicographers who are interested in finding compounds or collocations in a certain language or domain, it is useful to have a tool which can automatically measure the cohesion among lexical elements. Here, we introduce three kinds of statistic features which can be used to measure the correlations among the constituents of a known linguistic unit, as well as to discover the boundaries of an unknown linguistic unit.

2.3.1 Mutual Information

In many applications, it is of interest to find words that are highly associated and that frequently co-occur. For example, 'doctor' often co-occurs with 'nurses', 'patients' or 'hospitals'. A quantitative indication for "word association" is, therefore, useful for finding such terms. Mutual information is useful for measuring the degree of "word association." It compares the joint probability that a group of words will occur together to the probability of each word occurring independently. The bigram mutual information is known as [Church 1989]:

$$I(x;y) \equiv \log_2 \left\{ \frac{P(x,y)}{P(x) \times P(y)} \right\},$$

where x and y are two words in the corpus, $P(x,y)$ is the probability that the two words will co-occur in adjacency or within a range of several words (which is normally 3 to 10 words, e.g., the range was set to 5 words in [Church 1989]), $P(x) \times P(y)$ stands for the probability that these two words will occur independently, and $I(x;y)$ is the mutual information of these two words, indicating the ratio of the probability of co-occurrence to the probability of independent occurrence. If x and y are highly associated, then $P(x,y) \gg P(x) \times P(y)$; thus, $I(x;y) \gg 0$; if x and y are independent, then $P(x,y) = P(x) \times P(y)$, and $I(x;y) \approx 0$, and if x and y are complementarily distributed, then $P(x,y) \ll P(x) \times P(y)$, and $I(x;y) \ll 0$. For instance, the mutual information between "strong" and "tea" is higher than the mutual information between "powerful" and "tea" [Church 1989]. Therefore, "strong tea" is more likely to be a lexical entry than "powerful tea."

The mutual information measure can also be used with other features for extracting interesting terminologies. In [Su 1994], for example, mutual information was used for automatic compound detection. The compound detection problem was formulated as a two class classification problem in which a word n-gram (i.e., a string of n consecutive words) was classified as either a compound or as a non-compound based on the normalized frequency of occurrence of the n-gram, the mutual information among the constituents of the n-gram, and the parts of speech associated with the constituents. The simulation result showed that the approach proposed in [Su 1994] worked well. The testing set performance for the bigram compounds was a 96.2% recall rate and a 48.2% precision rate. For trigrams, the recall and precision rates were 96.6% and 39.6%, respectively. Figure 8 shows the first five most likely bigrams and trigrams for the testing set. Among them, all five bigrams and four out of five trigrams are plausible compounds.

bigram	trigram
dialog box	Word User's guide
mail label	Microsoft Word User's
main document	Template option button
data file	new document base
File menu	File Name box

Figure 8 The first five most likely bigrams and trigrams for the testing set

[Sproat 1990] used mutual information as a measure of character association to group Chinese characters into two-character words. Groupings of characters which are more highly associated (i.e., with higher mutual information scores) are preferred over groupings of characters which are less highly associated (i.e. with lower mutual information scores).

To illustrate, consider the example sentence given below:

我弟弟現在要坐火車回家。

The mutual information scores between successive pairs of characters are given in Figure 9 below.

我 弟	0.00	要 坐	0.00
弟 弟	10.44	坐 火	0.00
弟 現	0.00	火 車	7.31
現 在	4.23	車 回	2.06
在 要	-2.79	回 家	4.69

Figure 9 The mutual information scores between successive pairs of characters [Sproat 1990]

The highest association strength is between the two instances of '弟', so they are grouped together first as '弟弟'. This leaves the singleton '我' on the left, which has nothing to group with, and the remainder of the phrase on the right: '現在要坐火車回家'. Within the right chunk, '火車' has the highest association. We are then left with two further chunks, namely, '現在要坐' and '回家'. Since the lower limit is set as 2.5, below which association strength characters will not be grouped, it follows that of what is left, only '現在' and '回家' will be grouped. This will yield the following bracketing, which is correct:

我 [弟弟] [現在] 要 坐 [火車] [回家].

2.3.2 Dice

The dice metric is commonly used in information retrieval tasks [Salton 1993] for identifying closely related binary relations. It can, therefore, be used as a measure of the word association for two words, x - y , or the association between x and y in two languages. The dice metric for a pair of words x , y is defined as follows [Smadja 1996]:

$$D_2(x, y) = \frac{P(x=1, y=1)}{\frac{1}{2} [P(x=1) + P(y=1)]},$$

where $x=1$ and $y=1$ correspond to events where x appears in first place and y appears in second place, respectively, in a word pair or in an aligned sentence pair. It is, therefore, another indication of word co-occurrence which is similar to the mutual information metric. For instance, [Smadja 1996] described a program called *Champollion*, which, given a pair of parallel corpora in two different languages and a list of collocations in one of them, can automatically produce their translations. In that program, the dice metric is used as the measure of the correlation between the source collocation and its translations. If x represents a term in the source language and y is a possible translation of x in the target language, then it is possible to evaluate the dice coefficient between such a translation pair and to tell whether y is the preferred translation of x . Figure 10 shows some results of a test which gives a set of English collocations taken from the English part of the Hansards corpus (which is a bi-lingual corpus for Canadian Parliament records) and automatically produces their French translations.

English Collocation	French Translation Found by Champollion
additional costs	couts supplementaires
apartheid ... South Africa	apartheid ... afrique sud
affirmative action	action positive
free trade	libre-echange
freer trade	liberalisation ... echanges
employment equity	equite ... matiere ... emploi
make a decision	prendre ... decisions
to take steps	prendre ... mesures
to demonstrate support	prouver ... adhesion

Figure 10 Some Translations produced by Champollion by using dice metric [Smadja 1996]

It was suggested in [Smadja 1996] that the dice measure was a better indication of word co-occurrence than mutual information was in some cases since it discarded fewer informative events corresponding to the $x=0$ and $y=0$ (0-0 match) cases in estimating word co-occurrence.

2.3.3 Entropy

In some NLP applications, we may want to know whether one linguistic unit can freely bind with other arbitrary units in a sentence. If this is the case, then the linguistic unit is unlikely to form a larger unit with the other units. For example, in identifying possible lexical items in a corpus, a substring whose neighbors are randomly distributed is unlikely to form a larger lexical item with the other neighbors. This means that the substring is likely to be a well-defined lexical item by itself. For instance, [Tung 1994] used the entropy measure to automatically identify new Chinese words in a large Chinese corpus. The entropy measure can be used to indicate the degree of randomness or uncertainty for such purposes. It is defined in terms of the probabilities of an event. For example, if $P(w_i)$ represents the probability that the word w_i will occur to the left of a word "W", then the entropy (or the average number of words that could appear to the left of "W" in this particular application) is

$$H(W) = - \sum_{i=1}^V \{ P(w_i) \cdot \log_2 P(w_i) \}, \quad \left(\sum_{i=1}^V P(w_i) = 1 \right),$$

where $P(w_i)$ is the probability of a left neighboring word w_i , and V denotes the cardinality (or the number) of the vocabulary of all the possible words to the left of the word "W". It can be shown that the entropy measure reaches a maximum if all $P(w_i)$ are equally likely, which implies that all words can appear to the left of "W" with equal probability. If only one word, say w_0 , can appear to the left of "W", then $H(W)$ will be zero. In general, if most probability density are contributed by a few words, the entropy will be low, which means that only a few words can appear to the left of "W".

If the experiment is performed N times and the word w_i occurs N_{w_i} times, then the probability of w_i can be estimated as: $P(w_i) = \frac{N_{w_i}}{N}$. In [Tung 1994], for example, if

a string satisfies the following conditions, it has a high probability of being a new word:

1. The number of occurrences of the string str , denoted by $CNT(str)$, is high enough.
2. The entropy of the set of left neighboring characters, denoted by $H_LNC(str)$, is high enough.
3. The entropy of the set of right neighboring characters, denoted by $H_RNC(str)$, is high enough.

Assume that there are four candidate strings '消防隊', '消防', '防隊', and '的消' which occur more than once in a given text. Figure 11 shows different entropies of the sets of left and right neighboring characters of the four strings. For instance, $H_RNC(\text{消防})$, the entropy of the right neighboring characters, is estimated as $[3/6 \log(3/6) + 1/6 \log(1/6) + 1/6 \log(1/6) + 1/6 \log(1/6)] = 0.69$, where the base of the $\log(\cdot)$ function is set to be the number of occurrences of '消防' (i.e., $CNT(\text{消防}) = 6$) so that the entropy is normalized to the range of $[0,1]$. According to the entropies, the string '消防隊' is the most likely new word, and the string '消防' is next most likely. On the other hand, '防隊' and '的消' are not words at all. Experimental results in [Tung 1994] showed that the proposed method could identify a large number of new words from a large corpus in a very short time.

$LNC(str)$	str	$RNC(str)$		
(心, 1)	消防	(隊, 3)	$CNT(\text{消防}) = 6$ $H_LNC(\text{消防}) = 0.83$ $H_RNC(\text{消防}) = 0.69$	$CNT(str)$: the number of times that a string str occurs in the corpus $LNC(str)$: the left neighboring characters of the string str $RNC(str)$: the right neighboring characters of the string str
(的, 2)		(演, 1)		
(去, 1)		(檢, 1)		
(過, 1)		(工, 1)		
(實, 1)				
(消, 3)	防隊	(街, 1)	$CNT(\text{防隊}) = 3$ $H_LNC(\text{防隊}) = 0$ $H_RNC(\text{防隊}) = 1$	
		(位, 1)		
		(上, 1)		
(心, 1)	消防隊	(接, 1)	$CNT(\text{消防隊}) = 3$ $H_LNC(\text{消防隊}) = 1$ $H_RNC(\text{消防隊}) = 1$	
(的, 1)		(位, 1)		
(去, 1)		(上, 1)		
(次, 1)	的消	(防, 2)	$CNT(\text{的消}) = 2$ $H_LNC(\text{的消}) = 1$ $H_RNC(\text{的消}) = 0$	
(市, 1)				

Figure 11 The set of left and right neighboring characters of four strings and their corresponding entropies [Tung 1994]

2.4 Automatic Comparison of Parallel Texts

Parallel corpora, such as the Hansards corpus [Brown 1991a, Gale 1991a], are very useful knowledge sources for automatic acquisition of bi-lingual (and monolingual) knowledge. In the field of computational linguistics, a variety of researches have investigated the use of bilingual corpora, including sentence alignment [Wu 1994], word correspondence [Dagan 1993], collocation correspondence [Smadja 1996], word sense disambiguation [Brown 1991b, Dagan 1991, Gale 1992] and machine translation [Brown 1990, Su 1995, Wu 1995]. Bi-lingual material is also valuable for linguists who are interested in comparative studies of different languages, or in doing bi-lingual lexicography for translation. Here, we introduce three kinds of techniques which can automatically align parallel material and enable linguists to identify the most interesting data more easily. These techniques in fact can be applied to any parallel texts, such as in text criticism of two different editions of the same text. But the most common application now is for bi-lingual or multi-lingual texts.

2.4.1 Sentence Alignment

The purpose of sentence alignment is to identify correspondences between sentences in one language and sentences in another language. Figure 12 presents part of the output of a sentence alignment tool [Wu 1994]. Note that most of the time, a sentence is matched with one or two sentences in the other language. In addition, short source sentences tend to be translated into short sentences, and long sentences tend to be translated into long sentences or a combination of several short sentences. Therefore, the sentence alignment problem can be formulated as a shortest path searching problem as described in the Fuzzy Match section (2.1.2). In this case, the distance depends on how likely such a match is. If the probability of such a match is high (given the known sentence lengths of the members of the sentence pair), then the corresponding 'distance' is small; otherwise, the distance for such a match will be large. The minimum distance path thus acquired will then identify the most likely alignment between two text corpora.

1. MR FRED LI (in Cantonese): ↓	李華明議員問： ↓
2. I would like to talk about public assistance. ↓	我想談及公共援助問題。 ↓
3. I notice from your address that under the Public Assistance Scheme, the basic rate of \$825 a month for a single adult will be increased by 15% to \$950 a month. ↓	施政報告提到提高單身人士的公共援助基本金額，由每月 825 元提高至 950 元，即加幅是 15%。 ↓
4. However, do you know that the revised rate plus all other grants will give each recipient no more than \$2000 a month? On average, each recipient will receive \$1600 to \$1700 a month. ↓	但你知否經過調整後，即使加上所有其他津貼，每名受助者每月所得到的公共援助都不會超過 2000 元，平均來說，他們每月所得的是 1600 至 1700 左右。 ↓
5. In view of Hong Kong's prosperity and high living cost, this figure is very ironical. ↓	以香港的繁榮和生活水平之高，這數字根本是一個很大的諷刺。 ↓
6. May I have your views and that of the Government? ↓	請問政府或總督先生，你有何看法，是否覺得應全面探討公共援助的計算方式？ ↓
7. Do you think that a comprehensive review should be the method of calculating public assistance? ↓	因為基數這麼低，就算加多 20% 至 30%，仍是遠遠落後於現時的生活水平。 ↓
8. Since the basic rate is so low, it will still be far below the current level of living even if it is further increased by 20% to 30%. If no comprehensive review is carried out in this aspect, this "safety net" cannot provide any assistance at all for those who are really in need. ↓	若不全面檢討公共援助的計算方法，這安全網根本不能為那些真正有需要的人士提供協助。 ↓
9. I hope Mr Governor will give this question a serious response. ↓	希望總督先生認真回應這問題。 ↓
10. THE GOVERNOR: ↓	總督答（譯文）： ↓
11. It is not in any way to belittle the importance of the point that the Honourable Member has made to say that, when at the outset of our discussions I said that I did not think that the Government would be regarded for long as having been extravagant yesterday, I did not realize that the criticisms would begin quite as rapidly as they have. ↓	我在昨天的討論開始時說，我相信政府不會長期被指為揮霍無度。當時我沒有料到批評會來得這麼快。 ↓
12. The proposals that we make on public assistance, both the increase in scale rates, and the relaxation of the absence rule, are substantial steps forward in Hong Kong which will, I think, be very widely welcomed. ↓	我說這句話，絕對無意貶低這位議員剛才所提意見的重要性。我們對公共援助提出的建議，不論是增加援助金額或是放寬離港期限的規定，對本港來說，可說是向前跨進一大步，我相信普遍會受到歡迎。 ↓
13. But I know that there will always be those who, I am sure for very good reason, will say you should have gone further, you should have done more. ↓	不過，我知道有些人一定會說，你應更向前邁進一步，你應該做多一些，我肯定他們這樣說是有非常充分的理由。 ↓
14. Societies customarily make advances in social welfare because there are members of the community who develop that sort of case very often with eloquence and verve. ↓	很多社會慣於改善其社會福利，原因是有些人經常利用動聽的說話及凌厲的詞鋒，提出這方面的意見。 ↓

Figure 12 A sample of length-based alignment output [Wu 1994]

In most current research works, two factors were used to estimate the likelihood of a particular alignment, namely, the matching type of aligned passages (such as 1-2 matching, i.e., one source sentence matching two target sentences); and the lengths of the aligned source-target passages (say, 20 words in the source passage versus 24 words in the target passage). The alignment which has the maximum likelihood of being aligned according to these two factors is then considered to be the best alignment.

[Wu 1994] used the above mentioned length-based algorithm in automatic alignment of sentences in parallel English-Chinese texts, and 86.4% accuracy was reported (95.2% could be achieved if the introductory session headers were discarded.)

Besides the match type and sentence length information, other useful information, such as bilingual lexicon information (i.e., corresponding translations of the words within the passages), syntactic information (such as the parts of speech and parse trees of the sentences) and semantic information (such as case information and word sense information) would be also helpful for aligning sentences [Su 1996].

2.4.2 Word Correspondence

The purpose of word correspondence is to find the correspondence between a word in one language and its counterpart in another language. The output of word correspondence can be used to build a bi-lingual lexicon for translation. The word correspondence problem can also be formulated as a shortest path searching problem, where a path corresponds to a possible set of correspondences between words in a sentence pair. However, we can not simply use the length information of the words for such purposes. Instead, we should take advantage of the fact that a word in the source text is likely to correspond to another word in the target text that is highly associated with the source word. In addition, we must consider the relative position of the corresponding word in the other language, which also provides useful information due to the locality phenomena.

For instance, Gale and Church (1991b) used the ϕ^2 statistic, a χ^2 -like statistic [Hoel 1971], as a measure of the association of pairs of words to find the possible correspondence among words which had high word association. (Interested readers are referred to [Gale 1991b, Su 1996] for definition of the ϕ^2 statistic.) This possible correspondence is equivalent to the possible paths in a shortest path problem. The path with the maximum probability of alignment score then identifies the best correspondence.

Once a set of word pairs which have high ϕ^2 is selected, a matching procedure is used to match English and French words within the aligned regions using the selected pairs. When there are several possibilities for matching one source word with a target

word at different target word positions, the matching procedure will select the best correspondence based on a correspondence score. Intuitively, the correspondence score prefers a correspondence which has less change in the word order of the source words ; such a change in word order is defined in terms of the difference between the word index of the current target word and the word index of the preceding target word, referred to as the *slope* of the current target word. The correspondence score for J source-target pairs is, thus, defined as

$$\sum_{j=1}^J \log P(\text{slope}_j / \text{type-of-match}) P(\text{type-of-match}),$$

where J is the number of source words which have non-null correspondence on the target side, $P(\text{type-of-match})$ is the prior probability of the number of source words which could be matched with a target word (the number of matching source words is classified into three types: *type-of-match*= 1, 2 or 'many', and where $P(\text{slope}_j / \text{type-of-match})$ indicates how likely it is that the j -th target word will have a slope of slope_j if the number of source words which could be mapped to the j -th target word is known. For simplicity, the best correspondence is obtained by using the dynamic programming technique [Denardo 1982] for all possible correspondences. The performance was evaluated on a sample of 800 sentences, where 61% of the English words were matched with some French words, and about 95% of these pairs were judged as being correctly matched. Readers who are interested in the details are referred to [Gale 1991b].

[Wu 1995] also used similar automatic word correspondence techniques to build a probabilistic English-Chinese lexicon. Figure 13 shows some example alignment outputs. (Interested readers are referred to [Wu 1995, Su 1996] for more details.)

[These/ 這些 arrangements/ 安排 will/ε ε/ 可 enhance/ 加強 our/ 我們 ([ε/ 的 ability/ 能力]
[to/ε ε/ 日後 maintain/ 維持 monetary/ 金融 stability/ 穩定 in the year to come/ε)] ⊥ °]
[The/ε Authority/ 管理局 will/ 將會 ([be/ε accountable/ 負責] [to the/ε ε/ 向 Financial/ 財政
Secretary/ 司]) ⊥ °]
[They/ 他們 (are/ε right/ 正確 ε/ 十分 to/ε do/ 做 ε/ 這樣 so/ε)] ⊥ °]
[([Even/ε more/ 更 important/ 重要] [Lε however/ 但]) [Lε ε/ 的 , is/ 是 to make the very
best of our/ε ε/ 善用香港 own/ 本身 ε/ 的 talent/ 人才] ⊥ °]
I/ 我 hope/ε ε/ ◇望 employers/ 僱主 will/ 會 make full/ε ε/ 充分善 use/ 用 [of/ε those/ 那些]
(([ε/ 的工 who/ 人] [have acquired/ε ε/ 學到 new/ 新 skills/ 技能]) [through/ 透過 this/ 這個
programme/ 計劃]) ⊥ °]
[I/ 我 have/ 已 ◇ at/ε length/ 詳細 (on/ε how/ 怎樣 we/ 我們 ε/ 講述) [can/ 可以 boost/ε ε/ 促進
our/ 本港 ε/ 的 prosperity/ 繁榮] ⊥ °]

Figure 13 Word alignment output examples [Wu 1995]
(◇ =unrecognized input token)

2.4.3 Structure Alignment

Structure alignment is used to identify phrasal structure correspondences. For instance, a structure alignment algorithm was presented in [Wu 1995] which performed structure alignment for identifying sub-sentential phrasal translation examples in English-Chinese parallel bilingual corpora. Since the number of structure alignment patterns was large, he defined an Inversion Transduction Grammar (ITG) formalism to limit the number of alignment patterns. The ITG is a bi-lingual version of the context-free grammar that generates two children nodes for each parent. Each such phrase structure rule is associated with a matching score that associates the corresponding tokens and constituents of each string. The best alignment is identified by using this score and the dynamic programming technique as described in a previous section. The output of ITG is illustrated in Figure 14.

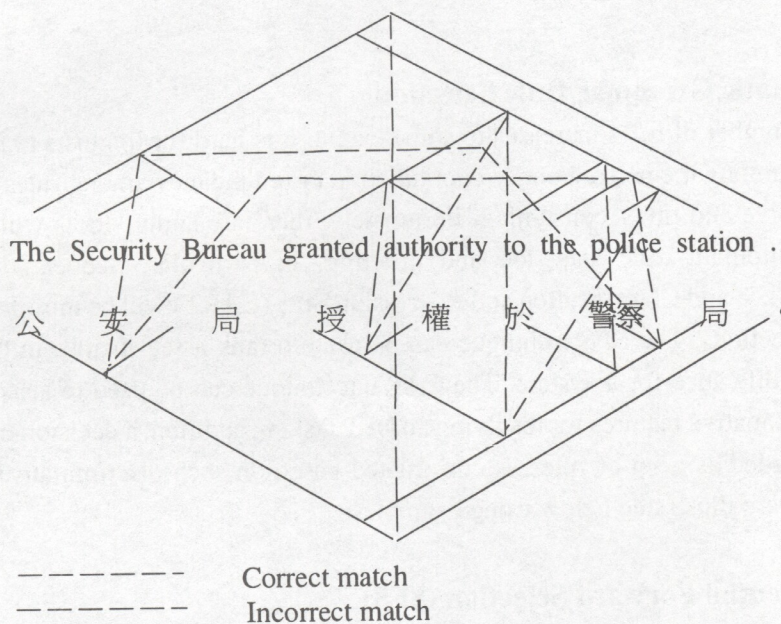


Figure 14 Constituent matching in ITG [Wu 1995]

All the possible constituent matchings are then filtered by using some constraints and a probabilistic English-Chinese translation lexicon (built using the word correspondence techniques). Some the extracted phrasal translations are shown in Figure 15.

have the right to decide our	有權決定我
in what way the Government would increase their job opportunities; and	政府如何增加他們的就業機會 ; 及
last month	上個月
never to say "never"	不要說 " 永不 "
reserves and surpluses	儲備和盈餘
starting point for this new policy	為這項新政策的起點
there will be many practical difficulties in terms of implementation	實行時會有很多實際困難
year ended 31 March 1991	截至一九九一年三月三十一日

Figure 15 Examples of extracted phrasal translations [Wu 1995]

2.5 Automatic Grammar Rule Selection

When the number of rules increases to some extent, it is hard for linguists to handle the interaction among them. To identify contradiction or redundancy among rules is usually labor-intensive and time-consuming. Fortunately, there are simple tools which can be used for automatic rule selection and ordering. In particular, Sequential Forward Selection (SFS) and Classification and Regression Tree (CART) will be introduced in the following sections. The SFS technique can be used to rank a set of rules in decreasing order of significance for a system. The CART technique can be used to select a set of most discriminative features for resolving an NLP task; in addition, a decision tree, which can be regarded as a set of rules, is constructed based on such discriminative features. These tools are illustrated below using examples.

2.5.1 Sequential Forward Selection (SFS)

SFS is a simple bottom-up searching procedure which finds the best rule sequence sequentially [Devijver 1982]. The same technique can also be used to find the best rule order for a set of rules. Initially, there are no rules in the rule set. At each iteration, a new rule is selected from the remaining rules not in the rule set so that the newly formed rule set yields the best system performance. Rules selected earlier are, thus, more significant than rules that are selected later, and redundant or contradictory rules tend to be ranked at the tail end of the rule sequence.

For instance, in a grammar checker application [Liu 1993], 127 pattern rules are used to detect ungrammatical errors. At the first iteration, each of the 127 rules is applied independently to detect errors. The rule which maximizes a pre-defined score (corresponding to the number of detected errors minus the number of false alarms) is

added to rule set A_1 , and the other 126 rules are left in the remaining rule set R_1 . At the second iteration, all the rules in the set R_1 are combined one by one with all the rules in A_1 (which contains only one rule in this case); the score of each combination is examined. The rule with the highest score in combination with rule set A_1 is added to rule set A_1 to form A_2 (A_2 now contains two rules). This procedure repeats until a pre-defined number of rules in Φ is selected or when the score begin to decrease as new rules are incorporated.

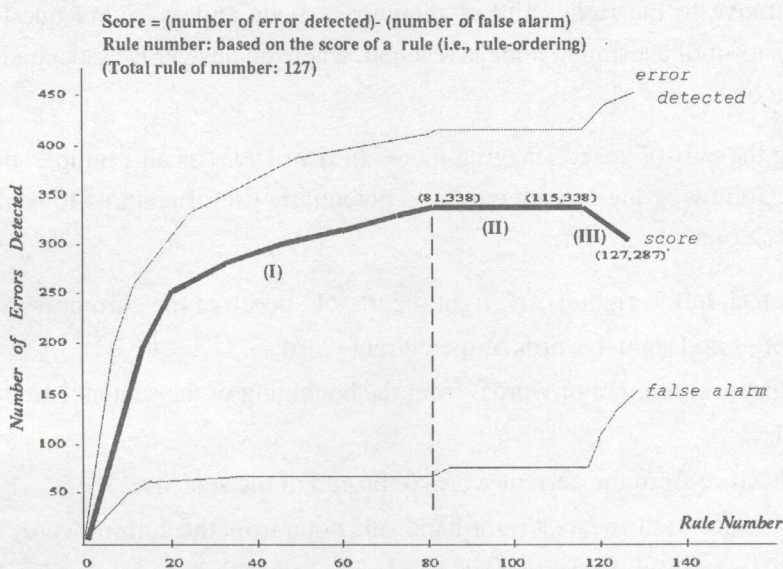


Figure 16 Number of Rules vs Overall Score in SFS [Liu 1993]

As suggested in Figure 16, the score increases monotonically with the number of rules until about 80 rules are applied (region I). When more than 80 rules are applied, the performance no longer improves, which means that there is redundancy among rules (region II). When more than about 120 rules are applied, the performance begins to degrade, which implies that extra rules might be contradictory to previous rules, or that there are rules which introduce more false alarms (region III). We can, therefore, use the first 80 rules to achieve the best performance and discard other redundant or contradictory rules. The SFS technique thus provides a simple and objective way for linguists to arrange their rules in decreasing order of significance.

2.5.2 Classification and Regression Trees (CART)

The simplest form of linguistic rules is a sequence of yes/no questions for making decisions. For example, we may have a rule for determining the part of speech of the word 'is' as shown below :

if (the next word of 'is' is not a verb)
 then ('is' is a verb)
 else ('is' is an auxiliary verb).

In general, such questions can be organized in the form of a decision tree or a classification tree [Breiman '84]. Each node in the decision tree is associated with a question. If the answer to the question is yes, then we move to the left child of the current node and ask further question(s) associated with the left child (and its children); if otherwise, we move to the right child of the current node and ask other questions. This process repeats until a terminal node is reached, where an answer is associated with such a terminal node.

Taking the part-of-speech tagging model in [Lin 1995] as an example, the features listed in the following are considered to be potentially useful features for choosing the part-of-speech of a word:

- the left-2, left-1, right-1, and right-2 parts-of-speech of the current word;
- the left-1 and right-1 words of the current word;
- the distance (number of words) from the beginning of the sentence to the current word;
- the distance from the current word to the end of the sentence;
- the distance to the nearest right-hand side noun from the current word;
- the distance from the nearest left-hand side verb to the current word;
- the distance to the nearest right-hand side verb from the current word.

The following decision tree (after non-discriminative features and questions are pruned) is constructed using the CART technique to determine whether the part of speech of 'out' is IN ('general preposition') or RP ('prepositional adverb which is also particle').

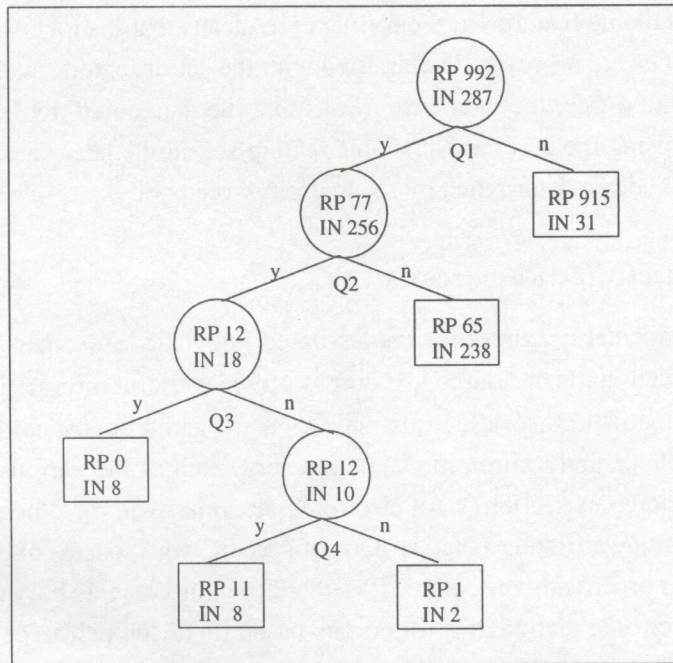


Figure 17 Example of the pruned classification tree for the word 'out' [Lin 1995]

Here, the questions Q1 to Q4 associated with the nodes are listed as follows:

(Q1) Is the next word 'of'?

(Q2) Is the part of speech of the previous word 'VBN'?

(Q3) Is the distance to the nearest verb on the right-hand side less than or equal to 8?

(Q4) Is the distance from the nearest verb on the left-hand side less than or equal to 8?

In Figure 17, the numbers to the right of the part of speech RP and IN stand for the number of tokens in the training tokens, which satisfy (or do not satisfy) the questions associated with its ancestor nodes. The square boxes are terminal nodes whose answers associated with such nodes are given by means of 'majority vote' to minimize errors. For instance, if all the answers to Q1, Q2 and Q3 are YES, then we will decide that such an 'out' is an IN since in this case none of the usage of 'out' is RP while there are 8 instances of the IN usage in the training tokens. If the answers to Q1 is YES but to Q2 is NO, then it is also tagged as IN according to the majority vote principle (238 IN vs. 65 RP).

In more general cases, CART is constructed by repeatedly splitting the tree nodes according to the most significant feature which minimizes a criterion function, usually referred to as an impurity measure. The tree grows until all the terminal nodes are either

pure or contain tokens which cannot be further differentiated into different classes with the currently available feature set; the former case means that the tokens associated with the terminal nodes are all correctly classified with the set of features along the branches of the classification tree; the later case means that the data could not be classified into correct classes using the currently available feature set. In the later case, the class associated with the node is determined by the majority-vote policy.

3. Chinese Electronic Resources

The most fundamental resources for corpus-based linguistic researches are text corpora and electronic dictionaries. A large text corpus provides useful information for inducing and verifying linguistic theories. It also allows us to gather statistical information for interesting problems and to train model parameters (such as the various kinds of probabilities in the previous sections). An electronic dictionary, on the other hand, provides us with the primitive attributes (such as parts of speech, word senses) of the lexical items of a language for processing the corpus. It is, therefore, important to have appropriate text corpora and electronic dictionaries for corpus-based linguistic researches. Furthermore, owing to the fast growth of the Internet, the single greatest development for Chinese computational resources is the increasing availability of various resources on the World Wide Web (WWW). We will also include some information on WWW Chinese Resources to provide linguists with the most up-to-date on-line resources.

3.1 Text Corpora

The text corpora introduced in the following sections include the modern and classical Chinese corpora available from the Academia Sinica, and the modern Chinese corpora available from the ROCLING society.

3.1.1 Text Corpora Available from Academia Sinica

The text corpora available from the Academia Sinica include both modern and classical Chinese corpora. The modern Chinese corpus, called the Academia Sinica Balanced Corpus (Sinica Corpus), is the first balanced Chinese corpus with part-of-speech tagging [Chen 1996]. The current size of the corpus is 3.5 million words, and the immediate expansion target is five million words. Each text in the corpus is classified and marked according to five criteria: genre, style, mode, topic, and source. The feature values of these classifications are assigned in a hierarchy. Subcorpora can be defined with a specific set of attributes to serve different research purposes. Texts in the corpus are segmented according to the word segmentation standard proposed by ROCLING [Huang 1996a]. Each segmented word is tagged with its part-of-speech. Linguistic patterns and language structures can be extracted from the tagged corpus via a corpus inspection

program which has the functions of KWIC searching, filtering, statistics, printing and collocation. Readers who are interested can access Sinica corpus and its inspection tools through WWW on <http://www.sinica.edu.tw/ftms-bin/kiwi.sh>. As for the classical Chinese corpora, please refer to Wei et. al. and Hsieh et. al.'s papers in this volume.

3.1.2 ROCLING Text Corpora

The ROCLING Text Corpora is a collection of individual text materials collected or compiled by various research institutes among the ROCLING members. It is the result of an effort to share research resources among local research institutes.

The text corpora were tagged with an SGML (Standard Generalized Markup Language, [Bryan 1988]) compliant markup standard, referred to as the ROCLING Text Corpus Exchange Formats ([Chang 1993]). The markup standard was later modified as a proposal for CJK (Chinese/Japanese/Korean) corpus encoding and exchange standard. The markup tags are highly compliant with the TEI (Text Encoding Initiative) P1 standard ([Sperberg 1990]) published by LDC (Linguistics Data Consortium, U.S.A.). Unlike TEI.1, however, the tag set is highly simplified so that text materials most commonly of interest can be made immediately available without requiring that much time be spent on tagging rarely used attributes. Currently, there are four types of text corpora in the collection. The sizes of the sub-corpora were listed in Figure 18.

	Computer Manual		Magazine Articles	Chinese Names	News
	Mono-	Bi-lingual			
No. of Files	347		20	10	1
Languages	English or Chinese	English and Chinese	English and Chinese	Chinese	Chinese
Tags and Attributes	<p> (paragraph), <s> (sentence)		<p> (paragraph)	<propname, sex=FIMI>	<p> (paragraph)
Word Segmentation	N/A				
File Size	14.4 M bytes		224K	27M	9.12M
No. of Tokens*	951,982 ew, 1,389,554 cc		716 cs 30,551 cw 1,063 es 22,624 ew	1,000,000 entries	48,122 p 3,927,905 cc
*. es/ew: English sentences/words *. cs/cw/cc: Chinese sentences/words/characters *. p: paragraphs N/A: not available					

Figure 18 Statistics of the ROCLING Text Corpora

Since the above collection includes both monolingual and bilingual text materials of different domains, it is useful for extracting lexical information, observing syntactic structures, acquiring translation knowledge, and so on. The proper name corpus is particularly useful for locating proper names in Chinese texts. More information is available at roxf@bdc.com.tw.

Besides the text corpora mentioned above, a speech corpus and a newly released 1,000 sentence standard segmentation corpus are also available from the ROCLING society. Interested readers are referred to Cheng et. al in this volume for the speech corpus and to [Huang 1996a] for the segmentation corpus.

3.2 Electronic Dictionaries

Electronic dictionaries provide to a varying extent information on the attributes of the lexical items of a language. A Chinese electronic dictionary, in particular, is useful in identifying the word boundaries of Chinese texts, which is essential before processing at the syntactic and semantic levels can be applied. Two known electronic dictionaries for Chinese language processing are introduced in the following.

3.2.1 CKIP Chinese Electronic Dictionary

The CKIP Chinese Electronic Dictionary was developed by the Chinese Knowledge Information Processing Group at the Institute of Information Science, Academia Sinica, Taipei, the R.O.C. The dictionary currently contains about 100,000 Chinese entries. Each entry contains the following information:

- Chinese entry,
- frequency,
- phonetic transcription,
- syntactic or semantic category,
- semantic features.

Some sample entries and the above-mentioned information are shown in Figure 19.

[Field 1]	[Field 2][Field 3][Field 4]	[Field 5] ⁺
熱乎乎	0	ㄇㄨˋ ㄉㄨˋ ㄉㄨˋ	VH11	*
人行道	97	ㄇㄨˋ ㄉㄨˋ ㄉㄨˋ ㄉㄨˋ	Ncb	+terrains
認可	114	ㄇㄨˋ ㄉㄨˋ	VC2	*
讓步	239	ㄇㄨˋ ㄉㄨˋ	VA4	*
容器	86	ㄇㄨˋ ㄉㄨˋ ㄉㄨˋ	Nab	+equipments
入學	151	ㄇㄨˋ ㄉㄨˋ	VA13	*
入伍	46	ㄇㄨˋ ㄉㄨˋ	VA13	*
滋長	6	ㄇㄨˋ ㄉㄨˋ	VC31	*

雜糧	63	ㄖㄩˇ ㄍㄨㄛˇ	Naa	+meals
總而言之	11	ㄖㄨㄛˊ ㄇㄨˊ ㄌㄨˊ ㄦˇ ㄧㄣˇ ㄉㄨ	Dk	*
阻滯	13	ㄖㄨˊ ㄇㄨˊ ㄉㄨˋ	VC2	*
測偵所	0	ㄘㄨㄛˋ ㄉㄨㄥ ㄌㄨㄟ ㄘㄨˊ	Ncb	+organizations
財物	509	ㄘㄨㄛˋ ㄉㄨˋ	Naeb	+physical
操場	92	ㄘㄨㄛˋ ㄟㄨˊ	Ncb	+regions
草稿	9	ㄘㄨㄛˋ ㄍㄨㄛˋ	Nab	+creation ⁺

⁺ [Field 1] is the Chinese word entry, [Field 2] is the number of occurrences of this entry in CKIP's corpus, [Field 3] is the phonetic transcription of this entry, and [Field 4] and [Field 5] are the category label and semantic feature of this Chinese entry, respectively.

Figure 19 Sample entries of the CKIP Chinese Dictionary

Since this dictionary is specially designed for Chinese natural language processing, verbs have more detailed information, including the possible forms and linear orders for their arguments and adjuncts. This information forms the basis of the information-based case grammar proposed by CKIP [陳 1991]. More information can be acquired from rocling@rocling.iis.sinica.edu.tw.

3.2.2 Behavior Chinese-English Electronic Dictionary

The Behavior Chinese-English Electronic Dictionary was developed by the Behavior Design Corporation. Currently, it contains over 110,000 Chinese entries (basic form). Variants are stored in a separate variant table (e.g. 證 vs. 証, 連絡 vs. 聯絡) and can be expanded and added to the base dictionary when necessary. Each Chinese entry contains the following information:

- Chinese entry,
- phonetic transcription (in BoPoMoFo symbols or PinYin symbols with tone information),
- Chinese part of speech,
- domain,
- frequency,
- English translation(s),
- Chinese synonym(s).

The dictionary has two versions, namely a Big5 version and a GB version. The entries in the dictionary are sorted in the order of their phonetic symbols in the Big5 version and in alphabetic order using PinYin in the GB version. Some sample entries in the Big5 version are shown in Figure 18.

[Field 1]						
[Field 2]	[[Field 3]]	[[Field 4]]	[[Field 5] [Field 6][Field 7] ⁺
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ]						
拜謝	v	GL	2	(v 000 thank humbly)		
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ]						
敗興	v	GL	3	(v 000 feel disappointed)	掃興，殺風景，興致索然	
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ ㄨ ㄛ ㄨ ㄛ]						
敗興而歸	v	GL	1	(v 000 return in low spirits)		
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ]						
敗絮	nc	GL	4	(n 000 old cotton wool)		
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ ㄨ ㄛ]						
拜占庭	np	GL	3	(a 000 Byzantine; n 000 Byzantium)		
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ]						
敗陣	v	ML	3	(v 000 defeat)	戰敗，敗北，失利，潰敗	
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ]						
拜壽	v	GL	1	(v 000 congratulate on birthday)		
[ㄅ ㄛ ㄨ ㄛ ㄉ ㄟ ㄛ ㄨ ㄛ]						
拜謁	v	GL	2	(v 000 call to pay respects)	拜訪，拜會，造訪	

⁺ Each Chinese entry occupies two lines. The first line is the Phonetic symbol line, denoted as [Field 1]. Each phonetic unit for a Chinese character contains four parts, including initial (consonant), glide, final (vowel), and tone. The second line is the word entry line. [Field 2] is the Chinese word entry. [Field 3] is the Chinese Part of Speech of this entry. [Field 4] is the domain of this entry; e.g., 'GL' stands for general domain, and 'ML' stands for military domain. [Field 5] is the frequency of occurrence of this entry, which has been intuitively ranked on a scale of 1-4 by lexicographers. '0' stands for the highest frequency, and '4' stands for the lowest frequency. [Field 6] is the English translation part, including the English Part of Speech of the English translation, a reserved field '000', and the English translation of the Chinese entry. [Field 6] is immediately followed by [Field 7], which contains Chinese synonyms of the Chinese entry.

Figure 20 Sample Entries of Behavior Chinese-English Electronic Dictionary

The Behavior Chinese-English Electronic Dictionary was the first large scale Chinese-English electronic dictionary developed in Taiwan and is used by several well-known research laboratories around the world. It is a very good reference for Chinese-English comparative studies and provides useful information for bi-lingual word alignment. Furthermore, the Chinese phonetic transcriptions can be used in a Chinese text-to-speech system and the Chinese synonyms can be used to build Chinese semantic

networks. The dictionary is a commercial product. Currently, it is available at very low cost to research institutes. More information is available from edic@bdc.com.tw.

3.3 World Wide Web (WWW) Resources

During the last few years, the rapid growth of the Internet community has made it possible to access a large volume of text and speech resources through the Internet at very little cost. The research institutes in the natural language processing community thus have more opportunities to practice real-world tasks using such network resources than they did earlier.

The only problem is that the legal status of such publicly available resources is still subject to legal debate. For instance, some electronic lexicons which were once available on the Internet no longer are due to legal problems.

In spite of these possible legal problems, we will briefly include a few known Chinese resources that are publicly accessible. Researchers are encouraged to contact the resource providers and to use such resources as far as the laws of the respective countries permit.

Since the list of accessible online resources is increasing day-by-day, the following information will soon be out-of-date. The authors will try to maintain an up-to-date list of such resources at the ROCLING home page (currently) at <http://www.bdc.com.tw/~rocling/> (including both Chinese and non-Chinese resources). The following paragraphs will, therefore, list only a few kinds of typical resources so that new NLP researchers can have a good starting point for finding the resources relevant to his/her tasks. (Most of the resources outside the Taiwan area are not listed here due to the authors' limited knowledge of these resources.) Briefly, the following types of resources are accessible from the Internet.

3.3.1 Online Electronic News

Many major Chinese News paper providers, radio/cable TV news departments (in Taiwan, Hong-Kong, China, Singapore, Malaysia, and the United States) have transferred their publications from paper or voice broadcast to electronic form (including text and speech).

Such text resources provide the largest volume of frequently updated and well organized articles on politics, economics, recreation, literature, and science and technology. Therefore, they are always the first choices for most NLP researchers.

Most news providers have monolingual text resources. Therefore, it is appropriate to use such resources for mono-lingual research. However, much non-local news in local

newspapers is simply translated from news provided by international news agencies. Therefore, it will be possible in the future to use such resources and their counterparts in other languages for multi-lingual research.

A few major Chinese news providers are listed as follows for reference. The reader can easily find other links by starting at such sites, by using a searching engine, or by entering the home pages of the major Internet Service Providers (ISP) of the various countries.

[China]

- <http://www.peopledaily.co.cn/> (GB)
http://www.egis.com/gb/people_daily/ (GB)
http://www.egis.com/big5/people_daily/ (Big5)
 - People Daily
<http://www.asia1.com.sg/gzbao/> (GB)
 - Guangzhou Daily
http://info.bta.net.cn/young/you_main.htm (GB)
 - Beijing Youth Daily

[Taiwan]

- <http://www.chinatimes.com.tw/> (Big5)
 - China Times (中國時報) Group
 (including Commercial Times (工商時報), Infotimes (時報資訊))
<http://uen.globalnet.com.tw/> (Big5)
<http://www.sinanet.com/minsheng/> (Big5)
 - United Daily News (聯合報) Group
 (including Ming-Sheng Daily (民生日報), United Evening News (聯合晚報))
<http://www.libertytimes.com.tw/> (Big5)
<http://www.nsysu.edu.tw/> (Big5)
 - The Liberty Times (自由時報)
<http://www.cna.com.tw/> (Big5, GB)
<http://www.sinanet.com/rtn/> (GIF)
 - Central News Agency (中央社) Real Time News
<http://www.tpg.gov.tw/twnews/> (Big5)
 - Taiwan Shin Wen Daily News (台灣新聞報)
<http://www.aide.gov.tw/> (Big5)

- Ming-Sheng Daily, United Daily, Liberty Times

<http://www.era.com.tw/>

- The Era (TVBS) Cable TV News

<http://www.cts.com.tw/>

- The Chinese TV System News

<http://www.bcc.com.tw/>

(Real Audio)

- Broadcasting Corporation of China

[Hong Kong]

<http://www.mingpao.com/newspaper/>

(Big5)

- Ming-Pao

<http://www.singtao.com/>

(Big5)

- Sing Tao Electronic Daily

<http://www.chinanews.com/>

(Big5)

http://www.chinanews.com/project/group_list/

- China News Service, Hong Kong China News Agency
- a large list of Chinese media (traditional or electronic) is being constructed here

[Singapore]

<http://www.asia1.com.sg/zaobao/>

(GB)

<http://www.asia1.com.sg/cgi-bin/cweb/g2b.pl>

(Big5)

- Lian-Hao Zhaobao

[Malaysia]

<http://www.founder.net.my/sinchew/> <http://web3.asia1.com.sg/sinchew/> (Big5, GB)

- Sin Chew Jit Poh

<http://www.asia-online.com/nsp/>

(GB)

- Nanyang Siang Pau

3.3.2 Online Electronic Magazines

Online electronic magazines represent another kind of well-organized but less frequently updated text resource. Most such magazines, in paper form, focus on a particular sub-domain for a particular type of reader. These domains may include personal computers, political commentary, recreation (such as cars, sports, music) and so on. Such resources are, therefore, useful for acquiring domain-specific information.

For instance, a few E-magazines accessible through the Internet are listed as follows:

- <http://www.cw.com.tw/> (Economy, Politics)
 - Common Wealth Magazine (天下雜誌)
- <http://www.infopro.com.tw/> (PC)
 - PC Week, 資訊傳真, PC Magazine, etc.
- <http://udn.com.tw/service/pcnews/infoweekly/> (PC)
 - United Daily Info Weekly (聯合報資訊專刊)
- <http://www.cnd.org/> <http://www.cnd.org:8009/HXWZ/> (Big5, GB, HZ)
 - China News Digest, Hwa Xia Wen Zhai
- <http://www.rpi.edu/~cheny6/java.html> (GB, Big5)
 - Chinese Poetry Magazine, with links to many E-News and Magazines

3.3.3 News Groups, Mailing Lists and Bulletin Board Systems

There are thousands of news groups, mailing lists (discussion lists) and bulletin board systems (BBS), which provide, mostly, dialogue-based articles on the Internet. Each newsgroup, list, or board represents a subdomain that is even more narrow in readership than are those of E-magazines. Many of the subdomains rarely appear in newspapers or magazines. Therefore, such resources are potential candidates for very special sub-languages. Characteristic of such articles is the use of very new vocabulary and slang that may never appear in more formal articles. Since such resources are dialogue-based, they provide good scripts for real-world dialogue and question-answering systems.

A particular application for such text materials is the training of error models of error detection (or correction) systems since such articles contain various types of typographic errors. For instance, it is easy to find typing errors (either intentional or un-intentional) which result from homophonic Chinese characters in a Chinese BBS.

3.3.4 Search Engines

Because there are so many articles on the Internet, it is difficult to find relevant materials for research if one does not have a list of resources such as that given above or if the list is too short to fit the general interests of the NLP community. In that case, a search engine will be very helpful to find relevant articles and information providers. In fact, a search engine by itself can be used by researchers to find the contexts of particular words. A search engine is also associated with a medium or large corpus behind the engine. Therefore, search engines for NLP research represent a way to gather language information without collecting a large corpus.

Most search engines provide exact string matches, case-insensitive string matches, and AND/OR operators for combining queries; more advanced search engines also

provide natural language query. A few search engines in Taiwan for Chinese text searches are list here for reference:

<http://csmart.iis.sinica.edu.tw/cna.html/>

- Csmart search for CNA News
- provides natural language query

<http://www.sinica.edu.tw/csmart/>

- Csmart search for Chinese lexicons and other databases

<http://www.sinica.edu.tw/ftms-bin/kiwi.sh>

- The Academia Sinica Balanced Corpus searching engine
- search by keywords with other specifications such as part-of-speech and semantic features

<http://gais.cs.ccu.edu.tw/cgais.html>

- Global Area Information System
- search for general internet text resources such as BBS articles (in Taiwan and Asia)

<http://udn.com.tw/>

- United Daily Full Text Indexing for Info Weekly

<http://taiwan.yam.org.tw/b5/yam/> <http://www.hello.com.tw/> <http://www.whatsite.com.tw/>

- a few commonly used commercial search engines

3.3.5 Special Online Resources

Most of the above are text resources. However, natural language may take other forms, such as speech. For instance, a Mandarin Chinese Text-to-Speech system was announced recently at :

<http://www.bell-labs.com/project/tts/mandarin.html> (Big5 page),

and <http://www.bell-labs.com/project/tts/mandarin-gb.html> (GB page)

4. Concluding Remarks

Owing to the increasing availability of machine-readable corpora, computational tools are becoming more and more important in linguistic studies. With these tools, linguists could verify their linguistic hypotheses more quickly using a much broader range of authentic material and, thus, can have more time for high level model construction and argumentation. The aim of this paper has thus been two-fold. On the one hand, we hope that linguists will be able to take advantage of these techniques and resources, and carry out complete, precise, and insightful linguistic analyses on their own interesting research topics. On the other hand, we hope that more and more linguists will become interested

in the interdisciplinary field of computational linguistics and will contribute their knowledge to the field of natural language processing.

In summary, this paper has presented a brief introduction to several well-established computational tools for extracting linguistically significant patterns, identifying linguistic classes, measuring the cohesion and correlation among linguistic units, and organizing linguistic rules in their order of significance. We have also introduced several available electronic resources and a few techniques for preparing aligned corpora for research. Interested readers are referred to [Su 1996] for more technical issues in applying such computational tools or techniques to text corpora.

Acknowledgements

We would like to express our gratitude to the reviewers for their valuable comments, and especially to the guest editor, Dr. Chu-Ren Huang, who provided substantial suggestions which make this article more immediately relevant and explanatory for linguists.

References

- Aho, A.V., R. Sethi and J.D. Ullman, *Compilers: Principles, Techniques, and Tools*, Addison Wesley, 1996.
- Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification And Regression Trees*, Wadsworth Inc., CA, USA, 1984.
- Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin, "A Statistical Approach to Machine Translation," *Computational Linguistics*, Vol. 16, No. 2, June 1990, pp. 79-85.
- Brown, P. *et al.*, "Aligning Sentences in Parallel Corpora," *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics(ACL-29)*, California, USA, June 18-21, 1991, pp. 169-184.
- Brown, P. *et al.*, "Word-Sense Disambiguation Using Statistical Methods," *Proceedings of ACL-29*, California, USA, June 18-21, 1991, pp. 264-270.
- Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai and Robert L. Mercer, "Class-Based N-gram Models of Natural Language," *Computational Linguistics*, Vol. 18, No.4, 1992, pp. 467-479.
- Bryan, Martin, *SGML -- An Author's Guide to the Standard Generalized Markup Language*, Addison-Wesley Publishing Company, 1988.
- Chang, Jing-Shin, "ROCLING Text Corpus Exchange Formats," (Rev. 1.2.1), Technical Report, 1993. (Revised version available at: <ftp://www.bdc.com.tw/pub/rocling/RXF.1.2.2.ps.Z>.)
- Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang and Hui-li Hsu, "Sinica Corpus: Design

- Methodology for Balanced Corpus," *Proceeding of the 11th Pacific Asia Conference on Language, Information, and Computation*, Seoul: Kyung Hee University, 1996, pp. 167-176.
- Church, K. and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Proceedings of ACL-27*, Vancouver, 1989, pp.76-83.
- Dagan, I., A. Itai and U. Schwall, "Two Language Are More Informative Than One," *Proceedings of ACL-29*, California, USA, June 18-21, 1991, pp. 130-137.
- Dagan, Ido, Kenneth W. Church and William A. Gale, "Robust Bilingual Word Alignment for Machine -Aided Translation," *Proceedings of Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, 1993, pp. 1-8.
- Denardo, E.V., *Dynamic Programming: Models and Applications*, Prentice-Hall, N.J., 1982.
- Devijver, P.A., and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall, London, 1982.
- Duda, O.R., P.E. Hart, *Pattern Classification and Scene Analysis*, Hohn Wiley and Sons, Inc., USA, 1973.
- Gale, William A. and Kenneth W. Church, "A Program for Aligning Sentences in Bilingual Corpora," *Proceedings of ACL-29*, California, USA, June 1991, pp. 177-184.
- Gale, William A. and Kenneth W. Church, "Identifying Word Correspondences in Parallel Texts," *Proceedings of DARPA Speech and Natural Language Workshop*, Pacific Grove, California, USA, 1991, pp. 152-157.
- Gale, William A., Kenneth W. Church and D. Yarowsky, "Using Bilingual Materials to Develop Word Sense Disambiguation Methods," *Proceedings of the 4th Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, Montreal, Canada, June 25-27, 1992, pp. 101-112.
- Hoel, P.G., S.C. Port and C.J. Stone, *Introduction to Statistical Theory*, Houghtone Mifflin Company, USA, 1971.
- Huang, Chu-Ren, "Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results," in Matthew Y. Chen and Ovid J-L. Tzeng Eds. In *Honor of William S.-Y. Wang. Interdisciplinary Studies on Language and Language Change*, Taipei: Pyramid, 1994, pp. 165-186.
- Huang, Chu-Ren, "The Morpho-lexical Meaning of Mutual Information: A Corpus-based Approach Towards a Definition of Mandarin Words," Presented at the 1995 Linguistics Society of America Annual Meeting, New Orleans, January 5-8, 1995.
- Huang, Chu-Ren, Keh-jiann Chen, Li-ping Chang and Hiu-li Hsu, "An Introduction to Academia Sinica Balanced Corpus [In Chinese]," *Proceedings of ROCLING VIII*, pp. 81-99.

- Huang, Chu-Ren, Keh-Jiann Chen and Lili Chang, "Segmentation Standard for Chinese Natural Language Processing," *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, Aug. 5-9. 1996, pp. 1045-1048.
- Huang, Chu-Ren and Keh-Jiann Chen, "Issues and Topics in Chinese Natural Language Processing," in C.R. Huang et al. Eds. *Readings in Chinese Natural Language Processing*, JCL Monograph No. 9. Berkeley: Journal of Chinese Linguistics, 1996, pp. 1-22.
- Lin, Yi-Chung, Tung-Hui Chiang and Keh-Yih Su, "The effects of Learning, Parameter Tying, and Model Refinement for Improving Probabilistic Tagging," *Computer Speech and Language*, Vol 9, 1992, pp. 37-61.
- Liu, Yuan-Ling, Shih-Ping Wang and Keh-Yih Su, "Corpus-based Automatic Rule Selection in Designing a Grammar Checker," *Proceedings of ROCLING VI*, Sept. 1993, pp. 161-171.
- Redington, Martin, Nick Chater, Chu-Ren Huang, Li-Ping Chang, Steve Finch and Keh-jiann Chen, "The Universality of Simple Distributional Methods: Identifying Syntactic Categories in Mandarin Chinese," presented at the International Conference on Cognitive Science and Natural Language Processing, July 4- 11, Dublin City University, 1995.
- Gerard Salton and Michael J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, 1983.
- Schalkoof, R., *Pattern Recognition: Statistical, Structural and Neural approaches*, John Wiley & Sons, Inc. Singapore, 1992.
- Shieber, Stuart M. , "Criteria for Designing Computer Facilities for Linguistic Analysis," *Linguistics* 23, 1985, pp. 189-211.
- Smadja, F., K.R. McKeown and V. Hatzivassiloglou, "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, Vol 22, No. 1, 1996.
- Sperberg-McQueen, C.M. and Lou Burnard (eds.), *Guidelines For the Encoding and Interchange of Machine-Readable Texts* (TEI P1), ACH, ACL, ALLC, 1990.
- Sproat, Richard and Chilin Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages*, 4.4: 336-351.
- Su, Keh-Yih, Yu-Ling Hsu and Clair Saillard, "Constructing a Phrase Structure Grammar by Incorporating Linguistic Knowledge and Statistical Log-Likelihood Ratio," *Proceedings of ROCLING IV*, PingTung, ROC, Aug. 18-20, 1991, pp. 257-273.
- Su, Keh-Yih, Ming-Wen Wu and Jing-Shin Chang, "A Corpus-based Approach to Automatic Compound Extraction," *Proceedings of ACL-32*, 32nd Annual, USA, 27 June - 1 July, 1994, pp. 242-247.
- Su, Keh-Yih, Jing-Shin Chang and Yu-Ling Una Hsu, "A Corpus-based Two-Way Design for Parameterized MT Systems: Rationale, Architecture and Training Issues," *Proceedings of*

TMI 95, Vol. 2, Belgium, July 5-7, 1995, pp. 334-353.

Su, Keh-Yih, Tung-Hui Chiang and Jing-Shin Chang, "An Overview of Corpus-Based Statistics-Oriented (CBSO) Techniques for Natural Language Processing," *Intl. Journal of Computational Linguistics and Chinese Language Processing (CLCLP)*, Vol. 1, No. 1, Taipei, August 1996, pp. 101-157.

Tung, Cheng-Huang and Hsi-Jian Lee, "Identification of Unknown Words from Corpus," *Computer Processing of Chinese & Oriental Languages*, Vol. 8, Dec. 1994, pp. 131-145.

Wu, Dekai, "Aligning A Parallel English-Chinese Corpus Statistically With Lexical Criteria," *Proceedings of ACL-32*, New Mexico, USA, June 27-30, 1994, pp. 80-87.

Wu, Dekai, "Grammarless Extraction of Phrasal Translation Examples from Parallel Texts," *Proceedings of TMI-95*, Leuven, Belgium, July 5-7, 1995, pp. 354-371.

張元貞，林頌堅，簡立峰，陳克健，李琳山，「國語語音辨認中詞群語言模型之分群方法與應用」，中華民國八十三年第七屆計算語言學研討會論文集，新竹，清華大學。1994, pp.17-34.

陳克健，中文詞知識庫小組，「中文詞知識庫計劃與中文電子詞典」，第四屆中日雙邊資訊研討會論文集，台灣，台北。1991, pp.19-37.

