Automatic Acquisition of Linguistic Knowledge: From Sinica Corpus to Gigaword Corpus

Chu-Ren Huang
Institute of Linguistics
Academia Sinica
churen@gate.sinica.edu.tw

Abstract

The *raison d'etre* for a corpus, as it was first conceived by Francis and Kucera in 1963, was to provide a body of linguistic facts from which linguistic knowledge could be generalized, [1]. The methods of acquisition have evolved as corpus size and technology have advanced in the past 40 years. Originally corpus-based concordances assisted linguists to form generalizations. This was what Fillmore [2] characterized as a 'computer-aided armchair linguist'. Today, direct, automatic acquisition of linguistic knowledge from a corpus is becoming a reality.

Two trends that are critical to the automatic acquisition of linguistic knowledge from a corpus are the increase in corpus size and the development of technology to extract linguistic relations. The release of the Chinese Gigaword corpus [3] by LDC in 2004 set the stage for a billion (1,000,000,000) word corpora; while the development of Sketch Engine by Adam Kilgarriff and colleagues [4] in the same year provided tools for the automatic acquisition of linguistic knowledge.

Unlike the balanced corpus tradition established by the Brown Corpus and adopted by the Sinica Corpus (1995, the first annotated Chinese corpus) [5], the Gigaword Corpus has a uniform data source. It consists of two sub-corpora: one from the Central News Agency in Taiwan and the other from the Xinhua News Agency in PRC. In other words, the Gigaword corpus is a gargantuan news corpus representing the two major variants of Mandarin Chinese.

The sheer size of the Gigaword Corpus poses both a challenge and an opportunity. First, the challenge lies in how to achieve a high quality corpus annotation with a minimal of human intervention. The current standard procedure of corpus annotation, especially POS tagging, is automatic tagging with human post-editing. It is impractical to adopt the standard post-editing procedure for the Gigaword corpus because of the scale of the undertaking. Instead, we apply the Academia Sinica tagging system developed for the construction of the Sinica Corpus, with the statistical model trained on its complete 5 million word corpus. In addition, lexicon adaptation, unknown word detection and feedback modules are implemented. The result is a truly automatic, highly efficient annotation program that creates a fully tagged Gigaword Corpus.

The Gigaword corpus also provides the possibility of automatic extraction of grammatical relations. Automatic assignment of syntactic structure is a difficult NLP task. A precise structural assignment for a specific sentence or construction at the level that a linguist would desire is still impossible. However, such parochial errors become negligible when grammatical relations are extracted based on significant patterns of a large number of examples. This is the design criteria of the Sketch Engine (Kilgarriff et al. [4]) and the same criteria has been applied to the annotated Gigaword Corpus in order to construct the Chinese Sketch Engine. Our early experiments showed that the grammatical information extracted is generally reliable, although the interpretation of the acquired information must still be carried out with the aid of linguistic expertise.

In conclusion, recent developments in corpus linguistics clearly point toward billion-word size, fully automatic annotation, and automatic acquisition of linguistic knowledge. These developments will shape the construction of future corpora.

1. Introduction

1.1. General Background: Corpus as a tool to acquire linguistic knowledge

When the Brown Corpus was completed in 1964 [6], it was the first modern day corpus for computers. It is interesting to note that Chomsky first proposed his arguments against using empirical data alone or Markov models to acquire linguistic knowledge in 1957 [7] and that such arguments have become widely accepted by linguists in mid 60's. Chomsky's arguments were that observable empirical data are too restricted to exhibit the range of all relevant linguistic facts due to the infinite variations of nature languages; and that the simplistic Markov models are inadequate to express the complex generalizations needed to predict the infinite variations. Given this theoretical background, it is not surprising, yet still unfortunate, that in the first years of corpus linguistics, the field often has to justify itself by arguing that a corpus does provide sufficient evidence for linguists to derive generalizations from. Corpus as a tool to help acquire linguistic knowledge was not widely accepted by linguists until the Nobel lecture series on

corpus and Fillmore's [2] statement that corpus linguists are computer-aided arm-chair linguists.

We now know that Chomsky was right to claim that Markov models were not adequate to express explanatory accounts of grammars of natural languages. However, over 40 years of corpus work has shown that a corpus is indeed adequate representative sample of the infinite range of linguistic facts. It was impossible for Chomsky to know in the 50's that we will be able to collect corpus of billions of words and that computers will be able to automatically process and extract meaningful patterns from these data. Hence, it is obvious that the sheer size of corpus data as well as computing power play a central role in the development of corpus studies. There is no longer any doubt that corpora offer a body of linguistic evidence to make generalizations with. The research issue we face now is instead whether such linguistic knowledge can be automatically acquired and expressed from corpus.

It is important to note that, regardless of the theoretical issues, the main considerations guiding the compilation of corpora remain the same: What kind of data? How much? How should it be stored? How can it be accessed? And what knowledge can be extracted? The first consideration is specific to each corpus. That is, a corpus compiler must determiner beforehand the purpose and target of the corpus. The second consideration is determined jointly by the external factors of the availability of data and processing technology. It is fair to say that the more data the better, as long as the data can be efficiently processed. Hence, we will discuss the three remaining considerations of corpus compilation directly related to acquisition of linguistic knowledge. The discussion is based on our experience with corpus compilation accumulated over the past 16 years.

1.2. Issues addressed

1.2.1. How should it be stored: data structure of corpus

The question of 'how should it be stored?' refers to the data structure of the corpus instead of the electronic storage of data. Like genetic data and unlike numerical data, language data are ordered strings with relational and functional interactions among the element of the string. Hence, a corpus must adopt a data structure that is able to represent both the sequential relation and the dependencies among the linguistic elements. Since sequential databases and their processing are still under-developed, a typical corpus is a textual database, with layers of annotation and/or accompanying relational database to help encode sequential and functional information.

Another fundamental decision involving the data structure of a corpus is the basic unit of corpus. It is well-known that linguistic data can be analyzed at different levels: signals at a phonetic level; phonemes at a phonological level; morphemes at a morphological level; words and phrases at a syntactic level; sentences and turns at a discourse level; as well as texts and sub-texts at a textual level. In addition to determining the granulites of representation, a corpus compiler must also decide on which super-imposing and sub-dividing structural information can be encoded and how. In other words, 'how should it be stored' involves how the data can be chunked, and how the original sequential relations of the chunked data can be preserved. It is a common practice in corpus compilation to express non-basic-unit information by annotation. Hence the annotation schemes are central to how data is stored.

1.2.2. How to access it: Inter-operability and reusability of corpus

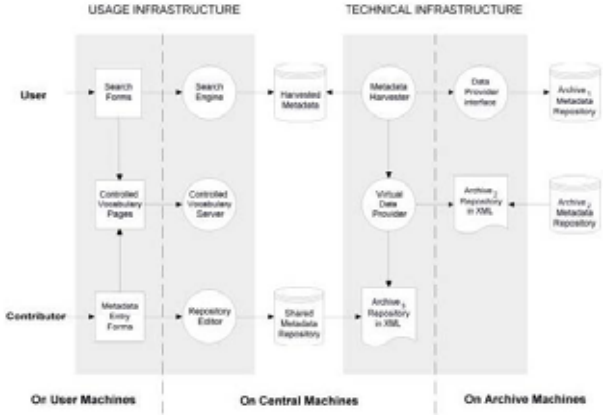
Users who want to access a corpus they need face the same challenges that users of other digital resources face. A series of questions need to be successfully answered before a user can successfully access a corpus: Does such a corpus exist? Where can I find it on the web? Do I have the right to use it? Do I have the necessary tools to use it? And where can I find the find the tool needed if I do not have it already? In order to guarantee successful access to potential users, corpus compilers need to be able to anticipate these questions and provide possible answers. The question is, is it even possible to anticipate and resolve these problems before a corpus is completed?

Bird and Simons [8] surveyed the requirements for language resources to be portable according to seven different aspects of uses. Based on these requirements, they proposed and founded an infrastructure to ensure inter-operability and reusability of language resources: the Open Language Archives Community (OLAC, [9], [10]). The stated mission of OLAC is to create 'a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.' We will follow the OLAC model in our discussion of how to access a corpus.

There are several scenarios where an existing corpus is not accessible to users, and a corpus compiler needs to ensure that these do not happen. One scenario is that a corpus is not even on the internet. Another scenario is that a corpus is on the internet, but the user has no idea it exists so it is not accessible in practice. There are many other possible sub-scenarios. For instance, the user may not be able to find a corpus because different sites have described it in different ways. The user may be overwhelmed with irrelevant

corpora because search terms are significant in other domains too. Tools and advice may play crucial roles too. The user may not be able to use an accessible data file because he is unable to match it with the right tools.

Fig. 1 Overview of the OLAC Infrastructures (Simons and Bird [9])



The inclusion of metadata information following an established standard is essential to the accessibility of any language resources. Metadata by definition is the formatted description of a digital resource. Standardized metadata allows efficient and accurate resources-discovery on the web. The OLAC Metadata is based on the Dublin Core Metadata. The Dublin Core Metadata Initiative began in 1995 [11] to develop conventions for resource discovery on the web. The Dublin Core metadata elements represent a broad, interdisciplinary consensus about the core set of elements that are likely to be widely useful to support resource discovery. There are two ways in which a standardized set of metadata makes a corpus accessible.

First, in theory, a web-based resources search should search the metadata of resources first. The Dublin Core consists of 15 metadata elements, where each element is optional and repeatable: *Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage* and *Rights*. Hence, a corpus can be searched and identified by one or more of the elements, not simply by its name. Second, a standardized metadata set means that descriptions of different corpora from different owners can be collected together at one site, called a repository. Repositories from different sites can be unified, not unlike a union catalogue in a library. This will allow users to search for all relevant resources simultaneously, regardless of their physical locations or their different status of rights.

It is important to note that target units of metadata description can be as small as a text. Take a balanced corpus, for example, in addition to the metadata markup of the corpus as a whole, it is also possible to mark up each and every text with its own metadata. Detailed and hierarchical metadata markup as such will allow more versatile access to the corpus at various hierarchical levels. For instance, a query for a corpus whose *Subject* is natural science will not return a balanced corpus itself but could return relevant sub-corpora that are included in the balance corpus.

The Dublin Core is the standard for web resources description and adopted by many ISO standards. The OLAC Metadata is widely adopted for searching major linguistic archives, such as the Linguist List [12] and LDC [13]. OLAC Metadata and other related components have also been adopted for Asian language resources [14], [15].

1.2.3. What knowledge can be extracted: Ontology and tools of linguistic knowledge description

The last and central issue is the automatic acquisition of linguistic information from corpora. There are two sides to the coin: the content and representation of linguistic knowledge, as well as the tools to extract linguistic knowledge based on these annotations.

A theory describing the content and representation of linguistic knowledge is referred to as a linguistic ontology. In short, a linguistic ontology defines basic concepts in linguistics, as well as how these concepts are organized, and how they are logically related to one another. GOLD (General Ontology of Linguistics Description [16]) is the only available linguistic ontology now. GOLD was originally designed to resolve the differences among different annotation schemes under the E-MELD project [17], but evolved to a general framework of linguistic knowledge. From the perspective of corpus compilation, it is easy to see that each corpus may choose a different annotation scheme. Hence, we face the same dilemmas again as in other annotated digital resources: How can we find out which two tags in different annotation schemes correspond to each other? How can we find out if the same tags in two different schemes have the same meaning or not? GOLD defines all the basic concepts in linguistics. Hence, each annotation scheme is either adopts GOLD, or can be mapped to GOLD. GOLD makes it possible to merge and exchange of linguistic information across different corpora.

The tools which help to extract linguistic knowledge can be divided into two categories. The basic and more established type of tools are visualization tools, such as KWIC (KeyWord-In-Context). A KWIC program does not manipulate the content of the corpus. What is does is to present the corpus in such a way that it will be easy for linguists to observe salient patterns of distribution. Since KWIC is a well-accepted basic tool in corpus applications, we will not go into further details in this paper. The second type of tool uses distributional information to automatically extract linguistic information. The typical function of these sorts of tools

is to extract collocational patterns. Mutual Information (MI) and frequency are the most often used statistics. We will discuss later in this paper the Sketch Engine work, which adopts an updated version of MI-based statistic to extract and evaluate saliency of collocational patterns from very large corpora.

1.3. Towards International Standards

It is important to note that there is a world-wide initiative to establish standards for language processing and resources under the International Standard Organization. ISO TC37 SC4 [18] is entitled 'Language Resources Management' and is drafting ISO standards that will apply to all different languages in the world. The establishment of these ISO standards will enable all corpora in the world to map to the same representational schemes, hence allows the user to correctly and efficiently interpret the data from different web pages. Standards that are widely used by the community, such as the OLAC Metadata and the GOLD Ontology, will be adopted as part of the ISO.

2. Acquisition of linguistic knowledge from a balance corpus

2.1. Sinica Corpus (Academia Sinica Balanced Corpus of Modern Mandarin Chinese)
http://www.sinica.edu.tw/SinicaCorpus/

Sinica Corpus is developed by the CKIP (Chinese Knowledge Information Processing) group of Academia Sinica and is considered to be the first modern day Chinese corpus. One million words were collected in 1992 for language processing as well as for linguistic research [19]. KWIC was introduced to Chinese text corpora then. Segmentation and tagging were added on later and completed in 1995. In November 1996, a version of Sinica Corpus was put on the web for public search access. It is likely the world's first fully web-searchable corpus as well as the corpus that has the longest web-life.

The current version (Sinica Corpus 3.1.), collected between 1990 and 1996, contains 5.2028 million words (7.8927 million characters). Over 141,000 word types are used in the corpus. A 10 million word version (Sinica Corpus 4.0.) has been completed and should be available in 2006.

The Sinica Corpus adopts the segmentation standard CNS 14366, also developed by CKIP under aegis of ROCLING and commissioned by Taiwan's Standardization Bureau. The 46 tag tagset is adapted from Chao's [20] framework and described in CKIP [21]. Each of the 9228 texts is marked with textual description.

The Sinica Corpus is a balanced corpus following Brown Corpus's format. We did take one step further in balancing the content of the corpus. That is, each text is classified and encoded according to four different criteria: Genre, Style, Media, and Topic. Although we follow Brown Corpus and designed the Sinica Corpus to be balanced mainly by topic, the corpus is also loosely balanced by genre, styles, and media. This innovative design allows multiple and versatile use of the same corpus.

The Sinica Corpus is a participating archive of OLAC and its metadata conforms to the OLAC Metadata standard. The corpus data were originally encoded in Big-5. An Unicode version has already successfully been converted to broaden the accessibility. Sinica Corpus is available both for web-based search as well as for licensing. It is also available through participation in international tasks such as the ACL Sighan Chinese segmentation bakeoff. A 2 million word subset used in developing CNS14366 segmentation standard is also available for free license.

2.2. Innovations in Sinica Corpus

Even though the Sinica Corpus follows the well-established tradition of a balanced corpus, it also contains several innovations to enable more versatile access of corpus content as well as more robust extraction of linguistic information.

2.2.1. Un-balancing a balance corpus

A balanced corpus is designed to be a representative sampling of general language use. However, we observed that a balanced corpus requires each text to be identified by its topic. Hence, a balanced corpus design has the inherent versatility of sub-corpora being created according to its topics. Since each text in Sinica Corpus is marked with four fields of information, we have the further versatility of allowing a user to create sub-corpora according to four different criteria: Genre, Style, Media, and Topic. Hence a user also has the option to study the interaction of genre, style, media, topic, and language use.

2.2.2. KWIC for tagged text

Another relatively straightforward, yet crucial innovation in the web-interface for Sinica Corpus was the fact that we allowed the KWIC result to show tagged keywords. In other words, not only can the user specify the POS of the keyword that s/he wants to study; it is also possible for a user to specify sorting and screening conditions on the left and right contexts of the keyword based on both collecting words and their POS. For instance, it is straightforward to stipulate that we only want to see the examples of a keyword when it occurs before a transitive verb. A sample showing the use of *ci4yao4* 'secondary' when it is used as an adjective is given in Figure 2 for illustration.

Figure II: Tagged KWIC of Sinica Corpus

2.1.3. Innovations in Linguistic Knowledge Acquisition

Sinica Corpus was designed to give linguists more opportunities and tools to extract grammatical information. In Sinica Corpus, given the constraints of computer more than 10 years ago, our design criteria were to utilize as much existing information as possible, at the same time facilitating the combination and extraction of data. The tools to extract information knowledge are collected on the advanced processing page of Sinica Corpus, as seen in Figure III. Note that the target of advanced processing is the data extracted by a standard KWIC search.

Figure III: Extracted Linguistic Information from Sinica Corpus

In addition to standard filtering functions allowing users to manipulate the right and left contexts of the keyword, three types of distributional information can be obtained with tools provided in Sinica Corpus advanced processing. First, sub-lexical information can be utilized. Since Chinese writing are character based, it is easy to identify affixes and suffixes of a complex word by simply looking at the first or the last character of that word. Hence we developed a sorter function to allow users to sort information, not only by keywords and their contexts, but also by the affixes of either a keyword or its context. Second, users view position based POS and keyword distribution in relation to the keyword. This allows a user to view directly if there is any structural dependency between a keyword and a certain POS. Last but not the least, Sinica Corpus was probably the first corpus to allow users full access of all collocational information between all word pairs in context, based on both frequency and mutual information (MI). These two pieces of information can be calculated based on word pairs, or based on a keyword and the POS of collocating words.

2.3. Summery

Altough Sinica Corpus follows a typical balanced corpus design, we have showed that with innovative corpus design and suitable tools, it allows very robust extraction of linguistic knowledge that will be of great interests to linguists.

3. Gigaword Corpus and their implications

The Chinese Gigaword Corpus published by LDC has the potential of changing the landscape of corpus-based NLP. First, its sheer size of over 111 million characters allows it to exhibit a wider and much more comprehensive range of grammatical behaviours. Second, the fact that it is divided into two major sub-corpora of the language in Mainland China and Taiwan allows in-depth study of language variation as well as NLP applications to overcome them. The Chinese Gigaword Corpus contains over 735 million characters from Taiwan's Central News Agency, and over 382 million characters from China's Xinhua News Agency. Although the original texts were in Big-5 and GB codes respectively, the released data has been converted to Unicode (UTF-8) to allow simultaneous processing of the data from both sources. The corpus is presented in SGML form and contains details textual information, including dateline and a rough 4-class topic classification.

The size of the Gigaword Corpus makes it impractical to tag with substantial human intervention, as was done with all previous Chinese corpora. In fact, the size of the Chinese Gigaword Corpus makes it easier for NLP researchers to ignore the fact that it is neither segmented nor tagged, although the texts are carefully formatted and annotated. It poses a new challenge and new opportunity of high-quality fully automatic tagging of Chinese texts. Both the Academia Sinica team and the Peking University team are taking on this challenge and the Academia Sinica has already produced preliminary results [22].

The Chinese GigaWord Corpus can be licensed from LDC.

4. Word Sketch Engine and Processing of Gigaword Corpus

4.1. Initial Implementation and Design of the Sketch Engine

The Sketch Engine is a corpus processing system developed in 2002 [23], [4]. The main components of the Sketch Engine are KWIC concordances, word sketches, grammatical relations, and a distributional thesaurus. In its first implementation, it takes as input basic BNC (British National Corpus [24]) data: the annotated corpus, as well as list of lemmas with frequencies. In other words, the Sketch Engine has a relatively low threshold for the complexity of the input corpus.

The Sketch Engine has a versatile query system. Users can restrict their query in any sub-corpus of BNC. A query string may be a word (with or without POS specification), or a phrasal segment. A query can also be performed using Corpus Query Language (CQL). The output display format can be adjusted, and the displayed window of a specific item can be freely expanded left and right. The most relevant feature is that the Sketch Engine produces a word sketch [23] that is an automatically generated grammatical description of a lemma in terms of corpus collocations. All items in each collocation are linked back to the original corpus data.

A Word Sketch is a one-page list of a keyword's functional distribution and collocation in the corpus. The functional distribution includes: subject, object, prepositional object, and modifier. Its collocations are described by a list of linguistically significant patterns in the language. Word Sketch uses regular expressions over POS-tags to formalize rules of collocation patterns, e.g. (1) is used to retrieve the verb-object relation in English:

(1) . 1:"V" "(DET|NUM|ADJ|ADV|N)"* 2:"N"

The expression in (1) states that: extract the data containing a verb followed by a noun regardless of how many determiners, numerals, adjectives, adverbs and nouns preceding the noun. It can extract data containing *cook meals* and *cooking a five-course gala dinner*, and *cooked the/his/two surprisingly good meals* etc.

The Sketch Engine also produces thesaurus lists, for an adjective, a noun or a verb, the other words most similar to it in their use in the language [23]. For instance, the top five synonym candidates for the verb *kill* are *shoot* (0.249), *murder* (0.23), *injure* (0.229), *attack* (0.223), and *die* (0.212). It also provides direct links to the Sketch Differences which lists the similar and different patterns between a keyword and its similar word. For example, both *kill* and *murder* can occur with objects such as *people* and *wife*, but *murder* usually occurs with personal proper names and seldom selects animal nouns as complement whereas *kill* can take *fox*, *whale*, *dolphin*, and *guerrilla*, etc. as its object.

Sketch Engine adopts Mutual Information (MI) to measure the salience of a collocation. Salience data are shown against each collocation in Word Sketches and other Sketch Engine output. MI provides a measure of the degree of association of a given segment with others. Pointwise MI, calculated by Equation (2), is what is used in lexical processing to return the degree of association of two words *x* and *y* (a collocation).

(2). $I(x;y) = \log \frac{P(x \mid y)}{P(x)}$

4.2. Application to Chinese Corpus

In order to show the cross-lingual robustness of the Sketch Engine as well as to propose a powerful tool for collocation extraction based on large scale corpus with minimal pre-processing; we constructed Chinese Sketch Engine (CSE) by loading the Chinese Gigaword to the Sketch Engine [25], [26]. The Chinese Gigaword contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency, and 380 million characters from China's Xinhua News Agency. Before loading Chinese Gigaword into Sketch Engine, all the simplified characters were converted into traditional characters, and the texts were segmented and POS tagged using the Academia Sinica segmentation and tagging system [27]. An array of machine was used for to process the 1.12 million characters, which took over 3 days to perform. All components of the Sketch Engine were implemented, including Concordance, Word Sketch, Thesaurus and Sketch Differences. We show the thesaurus results of two near synonyms 認為 'think, to hold the opinion' and 以為 'think, to have a (mistaken) impression' are given in figures IV and V to illustrate one of the unique function of the WordSketch Engine.

Figure IV. Thesaurus of 認為

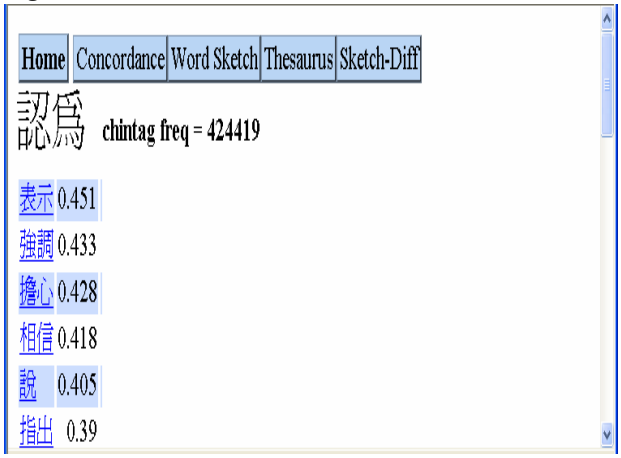
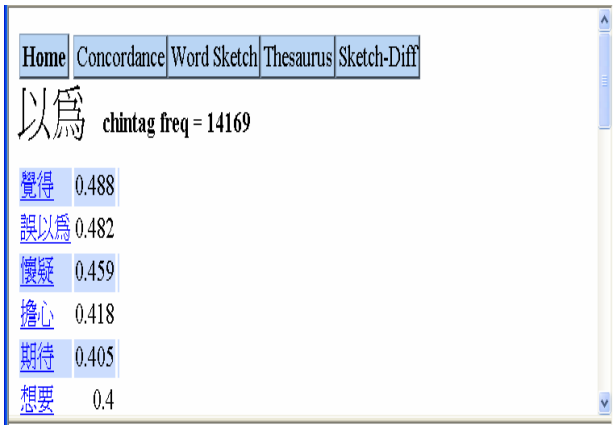


Figure V. Thesaurus of 以為



One characteristics that jumps out when we examine the Word Sketch results is the strong empirical base of evidence. The total number of examples that the generalizations are based from are 424,419 sentences containing 認為, and 14,169 sentences containing 以為. The fact that the distribution of 認為 is almost forty times as frequent as 以為 is a strong indication that is the more basic term. Lastly, the cluster of synonyms are scored, indicating their distance to the keyword. We can easily surmise from their different sets of synonyms the semantic difference of these two hard to differentiate near synonyms.

In our initial in-house testing of this prototype of the Chinese Sketch Engine, it does produce the expected results with an easy to use interface. For instance, the Chinese Word Sketch correctly shows that the most common and salient object of *dai.bu* 逮捕 'to arrest' is *xian.fan* 嫌犯 'suspect'; the most common subject *jing.fang* 警方 'police'; and the most common modifier *dang.chang* 當場.

The output data of Thesaurus correctly verify the following set of synonyms from the Chinese VerbNet Project: that *ren.wei* 認為'think' behaves most like *biao.shi* 表示 'to express, to state' (salience 0.451), while *yi.wei* 以為 'to take somebody/something as' is more like *jue.de* 覺得 'to feel, think' (salience 0.488). The synonymous relation can be illustrated by (4) and (5).

4a. 他認為到海外投資有一個觀念很重要，就是要知道當地的遊戲規則，接受這些條件。
ta ren.wei dao hai.wai tou.zi you yi ge guan.nian hen zhong.yao, jiu shi yao zhi.dao dang.di de you.xi gui.ze
'He believes that for those investing overseas, there is a very important principle-one must know the local rules of the game, and accept them.'

b. 執政黨也表示，由於公視爭議太大，恐怕無法全力支持。
zhi.zheng.dang ye biao.shi, you.yu gong.shi zheng.yi tai da, kong.pa wu.fa quan.li zhi.chi
'The KMT also commented that due to the many controversies surrounding PTV, it could not wholeheartedly support it either.'

5a. 何家駒就認為：「電視有基本語言和文法，要講究賣點和市場。」
he.jia.ju jiu ren.wei：「dian.shi you ji.ben yu.yan he wen.fa, yao jiang.jiu mai.dian he shi.chang。」
'Ho Chia-chu says, "Television has its own fundamental language and grammar. You must consider selling points and the market."'

b. 她表示：「我希望佛教徒能瞭解，父權社會與覺悟的社會是不相和的。」
ta biao.shi：「wo xi.wang fuo.jiao.tu neng liao.jie, fu.quan she.hui yu jue.wu de she.hui shi bu xiang.he de。」
'She says "I hope that followers of Buddhism can realize that a patriarchal society is incompatible with an enlightened society."'

The above examples show that *ren.wei* and *biao.shi* can take both direct and indirect quotation. *Yi.wei* and *jue.de*, on the other hand, can only be used in reportage and cannot introduce direct quotation.

Distinction between near synonymous pairs can be obtained from Sketch Difference. This function is verified with results from Tsai et al.'s study on *gao.xing* 高興 'glad' and *kuai.le* 快樂 'happy' [28]. *Gao.xing* 'glad' specific patterns include the negative imperative *bie* 別 'don't'. It also has a dominant collocation with the potentiality complement marker *de* 得 (e.g. *ta gao.xing de you jiao you tiao* 她高興得又叫又跳 'she was so happy that she cried and danced'). In contrast, *kuai.le* has the specific collocation with holiday nouns such as *qiu.jie* 秋節 'Autumn Festival. The Sketch Differences result is consistent with the account that *gao.xing/kuai.le* contrast is that inchoative state vs. homogeneous state.

5. Conclusion

In this paper, we used two Chinese corpora to discuss the evolution of automatic linguistic knowledge acquisition. We first introduced the main considerations in corpus compilation that will directly affect linguistic knowledge acquisition. In particular, we discussed how should a corpus be stored, accessed, and what knowledge can be extracted. In the discussion, it becomes clear that the there need to be general international consensus on the standard for specification, exchange, and format for language resources including corpus. The most urgent standards that are needed are metadata and annotation schemes. The current efforts by OLAC and ISO TC37 SC4 are introduced.

With regard to acquiring linguistic knowledge, we introduced two approaches. The first approach is based on Sinica Corpus, where versatility of use is added on to a typical balance corpus structure. The

innovations include allowing the corpus to be balanced with multiple criteria, which facilitates the creation of multiple sub-corpora for comparative studies; as well as the flexible use of collocational tools on both keywords and POS. The second approach is based on the Gigaword Corpus and Sketch Engine. This approach takes a large modern corpus where the scale dictates that human post-editing must be kept to a minimum. However this approach also underlines that central tenet of corpus linguistics: that linguistic generalizations can be acquired directly from distributional information of a corpus. Our work shows that, given the right tools, the richness of linguistic knowledge that can be acquired is largely dependent on corpus size.

Acknowledgements

I would like to thank all colleagues who have worked with me on the work reported here. In particular, Steven Bird, Keh-jiann Chen, and Adam Kilgarriff. They should be given the main credits for many of the reported work. I would like to express thanks to all the CKIPers, who have constructed all the resources reported here. In particular, I would like to thank Ru-Yng Chang, Wei-yun Ma, and Yiching Wu, for their direct contributions to various projects reported in the paper. Last, but not the least, I would like to thank Kathleen Ahrens for her continuing intellectual stimulus and support in reading and commenting on all my papers.

6. References

[1] http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
[2] Fillmore, C. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In Jan Svartvik (ed.) *Directions in Corpus Linguistics*. (Proceedings of Nobel Symposium 82), Berlin: Mouton de Gruyter
[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09
[4] Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. The Sketch Engine. Proceedings of EURALEX, Lorient, France (2004) (http://www.sketchengine.co.uk/)
[5] Huang, C., Chen, K., Chang, L. and Hsu, H. (1995). An Introduction to Academia Sinica Balanced Corpus. [In Chinese]. *Proceedings of ROCLING VIII*. 81-99.
[6] Francis, W.N., and Kucera, H. (1964/1971/1979). Brown Corpus Manual. http://helmer.aksis.uib.no/icame/brown/bcm.html
[7] Chomsky, N. (1957). Syntactic Structures. The Hague: Mouton.
[8] Bird, S. and Simons, G. (2003). Seven Dimensions of Portability for Language Documentation and Description. Language 79/3: 557-582.
[9] Simons, G. and Bird, S. (2003). The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing*, 8(4), 259-65.
[10] http://www.language-archives.org
[11] http://dublincore.org
[12] http://linguistlist.org/olac/
[13] http://www.language-archives.org/tools/search/
[14] Bird, S. Simons, G. and Huang C. (2001). The Open Language Archives Community and Asian Language Resources. *Proceedings of the Workshop on Language Resources in Asia*
[15] Chang, R. Huang, C. and Cheng C. (2002) OLACMS: Comparisons and Applications in Chinese and Formosan Languages. *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization.* A COLING2002 post-conference workshop.
[16] http://www.linguistics-ontology.org
[17] http://emeld.org/index.cfm
[18] http://www.tc37sc4.org/
[19] Huang, C. and Chen, K. (1992). A Chinese Corpus for Linguistic Research. *Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92).* 1214-1217 Nantes, France.
[20] Chao, Y. (1968). A Grammar of Spoken Chinese. Berkeley: University of California Press.
[21] Chinese Knowledge Information Processing. (1993). The Categorical Classification of Chinese. 3rd Edition. [In Chinese] CKIP Technical Report 93-05. Nankang, Academia Sinica.
[22] Ma, Wei-Yun and Huang Chu-Ren. (2006). Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. To be presented at the Fifth International Conference on Language Resources and Evaluation. Genoa, Italy.
[23] Kilgarriff, Adam and Tugwell, David. Sketching Words. In Marie-Hélène Corréard (ed.): Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Euralex (2002)
[24] Leech, Geoffrey. 100 million words of English: the British National Corpus (BNC). Language Research 28.1. (1992) 1-13
[25] Kilgarriff, A., Huang, C., Rychly, P., Smith, S., and Tugwell, D. (2005). Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. June 1-3. Singapore.
[26] Huang, **C.,** Kilgarriff, A., Wu, Y., Chiu, C., Smith, S., Rychly, P., Bai, M., and Chen, K. (2005). Chinese Sketch Engine and the Extraction of Collocations. Proceedings of the Fourth SigHan Workshop on Chinese Language Processing. October 14-15. Jeju, Korea.)
[27] Huang, C., Chen, K., and Chang, L. (1997). Segmentation Standard for Chinese Natural Language Processing. *Computational Linguistics and Chinese Language Processing.* 2.2.47-62.
[28] Tsai, M., Huang, C., Chen, K., and Ahrens, K. 1998. Towards a Representation of Verbal Semantics--An Approach Based on Near Synonyms. *Computational Linguistics and Chinese Language Processing.* 3.1.61-74..