A Grammar-informed Corpus-based Sentence Database for Linguistic and Computational Studies

Hongzhi Xu[1], Helen Kaiyun Chen[1], Chu-Ren Huang[1], Qin Lu[2], Tin-Shing Chiu[2], Dingxu Shi[1]

[1]Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University
[2]Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
{hongz.xu, helenkychen}@gmail.com, {churen.huang, csluqin, cstschiu, ctdshi}@inet.polyu.edu.hk

Abstract

We adopt the corpus-informed approach to example sentence selections for the construction of a reference grammar. In the process, a database containing sentences that are carefully selected by linguistic experts including the full range of linguistic facts covered in an authoritative Chinese Reference Grammar is constructed and structured according to the reference grammar. A search engine system is developed to facilitate the process of finding the most typical examples the users need to study a linguistic problem or prove their hypotheses. The database can also be used as a training corpus by computational linguists to train models for Chinese word segmentation, POS tagging and sentence parsing.

Keywords: Chinese Reference Grammar, Sentence Database, Linguistic Study

1. Introduction

From the tradition of linguistics, some linguists rely on their native intuition to make up examples either for discussing certain linguistic phenomena or to use them as illustration in their books on the grammar of certain languages. The problem of made up examples is that they may not reflect the actual use of the language. Corpus-based methods has gained attentions to let linguists be able to observe real data and make judgement based on them. Corpus-informed approaches to lexical computing are well-established in the construction of dictionaries and other lexical resources for over 25 years since the COBUILD project (Sinclair(Ed.), 1987). An approach to corpus-building with specific language resources in mind, however, has a much shorter history (Kilgarriff et al., 2006). The possible synergetic relation between corpora and a comprehensive grammar, however, has yet to be fully explored. In this paper, we introduce a grammar-informed, corpus-driven example sentence database of Mandarin Chinese, based on the pre-screened data extracted from the Chinese Word Sketch Engine system (Kilgarriff et al., 2005). The corpora used in this study are Sinica corpus (Chen et al., 1996) and the Chinese Gigaword Corpus (Hong and Huang, 2006), which contains various sources of newspapers and magazines in traditional Chinese including both written and spoken text. Originally, the example database was designed as a companion to a book describing the Chinese Reference Grammar (CRG), a descriptive grammar of contemporary Mandarin Chinese authored jointly by 20 leading Chinese linguists. A publishing contract of CRG has been assigned with the Cambridge University Press.

This study consists of two critical stages. The first and more conventional stage involves using corpora to inform grammar description and generalizations. In this stage, corpus search and example sentences extraction are directed by linguistic criteria formed by linguistic experts. The Chinese Word Sketch and specially built tools allows selected example sentences to be directly exported to the sentence database together with the grammatical conditions used to select them. At this stage, the database is an authoring tool consisting of preliminarily screened data. After the chapters of the Chinese Reference Grammar are completed and the illustrative examples selected from the database, we enter the second stage. In this stage, the database is given the structure of the reference grammar and annotated with its grammatical points. In other words, the preliminarily screened and roughly structured examples will be mapped to the more thoroughly researched structure of the grammar; and the sentence database become a grammar-informed and knowledge-rich language resource for both linguistic and computational studies. In other words, the CRG will provide the grammatical framework to access the database to verify and elaborate generalizations linguistically, as well as extract training data for computational processing of a specific set of linguistic phenomenon.

The size of Sinica corpus and the Gigaword corpus are providing extensive coverage of grammatical facts with large number of examples. However, the great number of examples presents a challenge when a single illustrative sentence is required for the grammar book. For example, if one wants to find some typical examples of the Chinese preposition'zai (在)' in order to study its syntactic behaviour, a preliminary search using the Sinica corpus yields over 110,000 sentences of 'zai (在)'as preposition. It turns out to be another challenging task to sort through so many examples to make generalizations on the syntactic patterns for the preposition. Although this may be necessary for linguistic studies, it may be unnecessary for all the linguists to do the same thing. Therefore, a carefully selected example database will benefit other linguistic researchers by avoiding lot of repetitive screening work manually.

In addition, we design an indexing system for the example database to make it easier for users to find the examples they look for. More specifically, every sentence is

marked with various linguistic features. Thus users could assign a certain linguistic feature as a query assistant to find the sentences they want. For example, a sentence may be tagged with a (BA) feature indicating that there is a BA-construction in it. Then if a user searches the database with the (BA) feature, this sentence will be retrieved if there is no other query condition.

In what follows, section 2. introduces the example database management system which is designed to facilitate language experts to build the example database. The main principles and methods to build the database are discussed in section 3.. Section 4. briefly discusses potential users and their benefit from the example database. Section 5. is the summary.

2. The Example Database Management System

To facilitate language experts to construct the example database, a management system is built to help extract example sentences from the Sinica and Gigaword corpora and tag the sentences with linguistic features. These sentences can be imported one at a time or a list in a batch into the database manually. The system uses web-based client-server model for the design of implementation. The system consists of three modules: the Example Import Module, the Management Interface Module, and the Kernel. Figure 1 shows the system architecture of the CRG Management System.

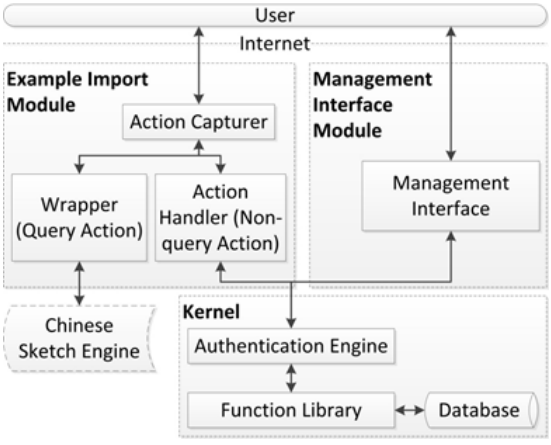

Figure 1: CRG Management System Architecture.

The system works as follows: The Example Import Module is responsible for importing example sentences. The Action Capturer identifies the action of the user and classifies the action as either a query action or a non-query action. For a query action, i.e. when a user wants to search example sentences from the Chinese Sketch Engine, a wrapper service redirects the user query to the Chinese Sketch Engine and gets the results. The wrapper will then inject some execution code to the page returned from the Chinese Sketch Engine. The injected page will then be sent back to users with all available controls of the Action Handler. Then, users can thus select individual items or a list of items to import to the database by making a non-query action, e.g. a click. Authentication is required for inserting examples to the database. The Action Handler will pass these requests to the Kernel to handle them properly. The Management Interface Module provides a web interface for users to manipulate the previous stored examples in the database. The kernel also provides the underlying functions to support the user operations for editing the example sentences. For example, users can add their own example sentences, delete an existing sentence or add linguistic facts/features to identify the sentences as we will discuss in the next section.

3. Build the Example Database

In this section, we will give an introduction how we build our database in terms of the methodology, linguistic principle and how to index the final database.

3.1. Methodology

As introduced above, the example sentences collected in the database are based on the linguistic analysis of Mandarin Chinese by linguists while writing the book of CRG. They seek substantiation for each basic linguistic structure and pattern from the Chinese language corpus resources through the Chinese Sketch Engine in order to confirm that the grammar points actually reflect the usage of Chinese in reality. Thus the example sentences used through the corpora search and selection to illustrate each grammar point described in the volume are not based on the authors' imaginary intuition, Instead, they are derived from existing corpora of authentic Chinese text. Finally, searches of additional examples that fit in with all the grammar points are conducted, and all example sentences are saved into the database through the example management system.

In case there is a lack of examples for some particular linguistic issues, such as an interesting use of a new word, the system should allow experts to add their own examples outside the corpora accessed through the Chinese Sketch Engine. This is very important since nowadays with the development of the internet, people tend to invent new words or use new ways to express their emotions and ideas, and it is true that some of these new expressions are widely accepted and used by other people. Consequently, linguists should pay special attentions to them, especially in terms of extended meanings and pragmatic uses.

In the example database, one sentence is tagged with different linguistic features when it is selected. Following the subfields of linguistics, i.e. syntax, semantics and lexicology, each deal with different facets of Chinese. In general, three different types of features are added to tag the example sentences including word class features, syntactic structure features and semantic/pragmatic features. We should notice that, there are many different opinions regarding the open questions of Chinese grammar, e.g. whether there is tense in Chinese or not, how many different kinds of adjectives are there. For such problems, we just follow the Chinese Reference Grammar described in the corresponding book. In other words, the linguistic features used for annotating the example sentences are actually identical to the organization of the book with each chapter dealing with various aspects of the linguistic issues. Word class features are mainly used to discriminate words whether homographic or

not with different POS tags. Syntactic structure features are used to distinguish sentences with different syntactic structures, such as clause, negation, comparative constructions or Ba (把) construction in Chinese. Pragmatic features are used to tag the sentences with different pragmatic features, such as anaphora, entailment and implicature etc.
To avoid the situation when some sentences that are selected by pragmatic experts and thus only tagged with pragmatic features which may also have syntactic features, language experts can search examples first within the example database to find examples they want. If there are no or not enough examples, they can then search through the Sketch Engine. There may still be some sentences that are not tagged with all their linguistic features. However, this should not be a major issue since the linguistic features are now only used for indexing the examples in the system to help users to find the examples they want after the database is built. Recall is not such an important issue here. If one sentence is imported more than one time by different linguists and annotated with different linguistic features, all the linguistic features will be simply combined.

3.2. Label Sentences with Linguistic Features

As we have said, we use the CRG grammar system as the linguistic features, we would like give a brief description on it in order to let users of the example database to get an overview about how the database is organized. There are mainly ten Chinese linguistic problems as shown in Table 1. Each problem corresponds with one or two chapters in the book. For the ten general problems. The labels can be easily generated when the linguists add examples to the database, because different linguists deal with different problems, and the user information has been saved corresponding to each sentence.

Sentence Types
Aspect
Clause
Prepositions and Preposition Phrases
Comparative Construction
Classifier
Noun/Noun Phrases
Adjective and adverb
Lexical word formation
Deixis and Anaphora

Table 1: Linguistic Features in CRG

Using such general terms to annotate the sentences is not specific enough to be helpful to the users of the example database, especially for the researchers who want to fetch examples on a particular linguistic phenomena. So, a better way is to give an abstract or description as which linguistic problems or terms are related to a certain sentence. However, it is unnecessary to describe it with a paragraph which will evoke another problem of indexing. In stead, we could only use several special keywords to indicate the related terms. The terms could be sub problems of a general problem. For example, for aspect of Chinese, we could use 'progressive', 'continuous' or 'perfect' to describe a sentence. Meanwhile, the keywords that the linguists used to extract the sentences are also maintained as a potential useful information. In the following, we will give more information about each of the above ten general linguistic problems.

Sentence Types Based on communicative functions of sentences, there are five different sentence types: declarative sentence, exclamatory sentence, interrogative sentence, imperative sentence, vocative or responsive sentence. Based on grammatical structure of sentences, there are three different sentence types: a canonical simplex sentence contains one clause of a subject-predicate construction; the subject or the predicate may be omitted, which leads to a reduced form; the subject and the predicate may have other clauses embed, which leads to an expanded form.

Aspect There are limited number of aspectual markers in Chinese. For progressive aspect, 'zai(在)' is usually used; for continuous aspect, 'zhe(着)' is usually used, while for perfect aspect, 'le(了)' or 'guo(过)' is usually used. Some other linguists may prefer to adopt experiential aspect as well, in which case, 'guo(过)' is usually used and leave 'le(了)' only for the perfect aspect. Experiential aspect is not included in our database. However, users can still use 'le(了)' and 'guo(过)' to discriminate the two different cases.

Clause For clauses, there are relative Clauses, noun-modifying clauses. There are also four different kinds of semantic relation of clauses, conjunctive and disjunctive, contrastive and concessive, conditional and suppositive, causative and purposive clauses.

Prepositions and Preposition Phrases Based on the distribution and function of preposition phrases (PP), there are PPs in preverbal position, PPs in sentence-initial position, PPs in postverbal position, PPs in noun modifier position. Based on semantic classification of prepositions, we have prepositions for space, space extensions, involved parties, prepositions for topic, reference, condition, cause and so on. Ba(把), gei(给), rang(让), jiao(叫) constructions are also included here as linguistic features.

Comparative Construction Comparative constructions usually contain words such as bi(比), xiang(像)...yi yang(一样), yue(越)..yue(越). For comparative constructions, no special linguistic features are included except for the keywords themselves. For the separated features, we can provide the function of the combination of linguistic keywords to allow users to do such queries.

Classifier For classifiers, there are two subtypes: sortal classifier and measure words. Sortal classifiers can be further divided into individual classifiers, e.g. ge(个), zhang(张), event classifiers, e.g. chang(场), ci(次), kind classifiers, e.g. zhong(种), yang(样). measure words can also be further divided into three sub classes: container measure words, e.g. wan(碗), ping(瓶), standard measure words, e.g. mi(米), nian(年) and approximation measure words, e.g. dui(堆), shen(身). The subtype names are used in our database as the fine-grained linguistic features.

Noun/Noun Phrases For noun phrases, different kinds of relations of the components/morphemes constructing them

are used to identify them including subject-verb, modifying, coordination, verb-object and so on.

Adjective and adverb For adjectives, there are mainly attributive adjective modifying other words and predicative adjectives used as predicates. Adverbs include four different subtypes: temporal adverbs, e.g. jiang(将), gang(刚), hai(还), degree adverbs, e.g. fei chang(非常), tai(太), scope adverbs, e.g. quan(全), dou(都), attitudinal adverbs, e.g. qian wan(千万), nan dao(难道), hebi(何必). So, we mainly use the the different types of adjectives or adverbs as the linguistic features.

Lexical word formation Chinese word formation in terms of bound or free morphemes can be divided into three different types: bound + bound, bound + free, free + free. From the affix point of view, there can be prefix, suffix and infix. So, this kind of features are provided as a deeper level than word or sentence based features. A special kind of words are also considered, that are the separated words, such as you yong(游泳) can be used separately (e.g. you le yi ci yong/游了一次泳). This kind of words are more difficult to be found by keywords-only based systems. In such cases, a linguistic feature 'separated' will be added plus the original word you yong(游泳).

Deixis and Anaphora Deixis and anaphora are mainly based on pragmatic point of view. There words include zhe(这), na(那), ta(它), ni(你), ta(他) etc. Some of them can be used in both deixis and anaphora, some can not, such as ni(你). The related terms will be definite and indefinite regarding whether the word refers to a fixed object or not.

3.3. Indexing the Database

Based on the above description, three kinds of features are used to index the database. The first is the traditional information as adopted by most online corpora, such as words and POS tags, with wildcard functions to allow users to match sequences with generalized patterns. The second is the ten general features shown in Table 1. This kind of feature can be implemented with a drop-down list or check box component to let users to select one or several features. The third is the linguists added keywords as shown above including the descriptive keywords and that the linguists used to extract them from Chinese Sketch Engine. Users are allowed to input keywords in a separate input box from the one used to search words in sentences. The following is a example sentence annotated in the database. If a user inputs linguistic keywords '那', and select classifier as the general linguistic feature, this sentence will be retrieved.

那	个	屋子	里	到处	是	金子	。
na	ge	wuzi	li	daochu	shi	jinzi	.
that		room	in	everywhere	be	gold	.

Linguistic Features:
Dexis, na(那)
Classifier, ge(个), sortal, individual
Suffix, zi(子)

4. The Users of the Example Database

Since the example sentences collected in the database have been pre-screened and pre-selected, and are generated from large scale corpora of Mandarin Chinese, the current database has the advantage of providing rather specific and precise examples to each pattern and/or construction described in the CRG. For linguists, readers and general public who desire to obtain some organized data of specific linguistic constructions or patterns, the corpus in the database will be more useful because users are able to get further refined search results without the need to deal with noisy data from the Sketch Engine. Besides, indexing provided for the tagged sentences with linguistic features also makes it easier than only keyword-based search for users to find desired examples.

Although this example database is originally built to support the book of a Chinese reference grammar, it can serve as an extension of examples for readers of the book, should they wish to get additional examples to understand a particular linguistic point. The database can also be used by users who want to compile their own books and need exemple sentences for their explanations. In example database management system, a small useful function is also provided to translate the selected sentences into Chinese pinyin and English word by word, as shown in the above example sentence.

For computational linguists, the example database can be used as a training corpus to train their models for Chinese word segmentation and POS tagging. The linguistic features added by linguists can also serve as important information which may potentially improve the performance of the task. The syntactic structure information are extracted from the Sinica Tree Bank (Huang et. al., 2000) which contains the same set of sentences, so the database can also be used to train syntactic parsers. However, for the sentences that are added by experts out of the Sinica corpus, syntactic structure information must be added manually or semi-automatically with an existing parser.

5. Conclusion

This paper presented an innovative approach to build a grammar-informed sentence database for computational and linguistic studies. It started with the well-established corpus-driven approach towards selections of relevant example sentences for certain linguistic facts and ended up with a feature-rich sentence database through the Chinese Reference Grammar with comprehensive coverage of the linguistic facts of Chinese. The database can be used by various kinds of users including readers of the reference grammar book, linguistic researchers and computational linguists.

However, the linguistic features in each sentence may not complete since each linguist deal with one particular linguistic issue. As well, the linguistic features provided by linguists are only based on keywords rather than well-structured. In the future, we will focus on tagging each sentence with all linguistic features it has and make the syntactic information available for all the sentences in the database. We will also keep this database updated with new examples for new language problems and issues. More importantly, we will try to develop methods to make the linguistic features more helpful by providing more structured information.

6. References

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asian Conference on Language, Information and Computation (PACLIC'11)*, pages 167–176.

Jia-Fei Hong and Chu-Ren Huang. 2006. Using chinese gigaword corpus and chinese word sketch in linguistic research. In *Proceedings of the 20th Pacific Asian Conference on Language, Information and Computation (PACLIC'20)*, Wuhan, China, November.

Adam Kilgarriff, Chu-Ren Huang, Pavel Rychly, Simon Smith, and David Tugwell. 2005. Chinese word sketches. In *Proceedings of 2nd Asialex*, Singapore.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the new corpus for ireland. *Language Resources and Evaluation Journal*, 40(2):127–152.

John Sinclair(Ed.). 1987. *Looking Up: Account of the COBUILD Project in Lexical Computing*. Collins COBUILD.