The Chinese-English Contrastive Language Knowledge Base and Its Applications

Xiaojing Bai[1], Christoph Zähner[2], Hongying Zan[3], and Shiwen Yu[4]

[1] Department of Foreign Languages and Literatures, Tsinghua University, Beijing, China
bxj@tsinghua.edu.cn
[2] Language Centre, University of Cambridge, Cambridge, UK
cz201@cam.ac.uk
[3] College of Information Engineering, Zhengzhou University, Zhengzhou, China
iehyzan@zzu.edu.cn
[4] Institute of Computational Linguistics, Peking University, Beijing, China
yusw@pku.edu.cn

Abstract. In this paper, we introduce our ongoing research on a Chinese-English Contrastive Language Knowledge Base, including its architecture, the selection of its entries and the XML-based annotation schemes used. We also report on the progress of annotation. The knowledge base is linguistically motivated, focusing on a wide range of sub-sentential contrasts between Chinese and English. It will offer a new form of bilingual resources for NLP tasks, for use in contrastive linguistic research and translation studies, amongst others. Currently, joint efforts are being made to develop tools for Computer-Assisted Translation and Second Language Acquisition using this knowledge base.

Keywords: Language Knowledge Base. Contrastive Linguistics. Parallel Corpus. Sub-sentential Alignment. Computer-Assisted Translation. Second Language Acquisition

1 Introduction

Language knowledge bases are collections of linguistic knowledge that facilitate the automatic analysis and generation of natural languages in NLP systems. They are indispensable components of NLP systems, and their quality and scale influence the performance of these systems significantly [1].

Starting with a sentence-aligned parallel corpus [2], the Chinese-English Contrastive Language Knowledge Base (CECLKB) aims to provide formal descriptions of the sub-sentential contrast between Chinese and English: How do the two languages express the same notion? What is the nature of the correpondence between the two expressions representing the same notion? What syntactic and semantic constraints are involved? The sub-sentential contrastive knowledge, originally implicit in the parallel corpus, is made explicit and marked up with XML tags in order to support the processing of natural languages in bilingual or multilingual scenarios.

In this paper, we briefly review related research in Section 2, followed by an overview of CECLKB in Section 3. Section 4 describes our plans for employing CECLKB for Computer-Assisted Translation (CAT) and Second Language Acquisition (SLA). The last section summarises the present stage of our research and looks at plans for more future efforts.

2 Related Research

The construction of a contrastive language knowledge base was first motivated by the employment of new metrics for subsegment-level analysis and the prospect they offer for enhancing Translation Memory (TM) [3]. Compared with sentence-level alignment, sub-sentential alignment of parallel texts provides us with a more fine-grained look at the correspondence between matching expressions in different languages.

Despite the fact that sentences function as the operational unit of most TM systems currently in use, there has long been an assumption that where the complete sentence has not been translated before, the identification of corresponding sub-sentence segments by the TM would be of use to the translator [4]. Bowker and Barlow indicate that linguistic repetition occurs most often at the level of expressions or phrases [5], and Macken expects the second-generation TM systems to provide additional translation suggestions for sub-sentential chunks [6].

In parallel text processing, the alignment of parallel texts occurs either at the sentence level or at other levels including words and expressions, clauses and sentence structures, or even document structures [7]. There have been substantial efforts made on word alignment in projects such as Blinker, Arcade, Plug and GALE [8, 9, 10, 11], with the GALE project working on Chinese-English parallel texts in particular. Other research on Chinese-English alignment focuses on the characteristics of word alignment, distinguishing genuine links (strong or weak) from pseudo links [12], or looks at the alignment of senses between the two languages with the help of WordNet [13].

3 An Overview of the Knowledge Base

CECLKB is a formalized and structured collection of contrastive linguistic knowledge. The design of its architecture, the entries included and the annotation scheme for all entries are based on the anticipated applications of the knowledge base and the findings of relevant linguistic research and translation studies concerned with typical contrasting features of Chinese and English.

As the correspondence between Chinese and English can be found at various levels, CECLKB has entries of contrastive knowledge at the word, phrase, chunk and sentence pattern levels. This entails the alignment of parallel texts at all these levels and contrasts with the alignment in previous research, which mainly worked on the word level.

Each entry in CECLKB contains a Chinese-English sentence pair[1], with one and only one linguistic focus. It focuses on and marks up a specific instance of correspondence between a selected Chinese word, phrase, chunk or sentence pattern and its corresponding expression in English. The markup highlights the syntactic and semantic constraints on the particular instance of correspondence, which adds more dimensions to the cross-language observation than the previous research did. The following is an example of an entry in CECLKB[2].

Example 1:

```
weicheng01.xml
  <a id="160" no="1"><s id="50"><NR SR="EX" COMP_SO="T">周经理</NR><VP
FCS="T" GF="B"><VV>听</VV><DER>得</DER><VA COMP="DG">开心</VA></VP>
，叫主任回信说：</s></a>
  <a id="160" no="1"><s id="80"><S GF="SP" CO="HL"><VP><VBN>Delighted</VBN>
<PP><IN>with</IN><NP><PDT>all</PDT><DT>this</DT></NP></PP></VP></S>, Chou
instructed Wang to reply in the following manner: </s></a>
```

3.1 Architecture of CECLKB

The architecture of CECLKB (Fig.1) shows that there are four types of entries: word entries, phrase entries, chunk entries and sentence pattern entries. We currently work with six sub-categories of word entries and one sub-category for each of the other three types of entries. Subsequently, more sub-categories and entries will be included to enrich the knowledge base and widen its coverage.

The parallel texts in CECLKB include i) Chinese source texts and their English translation, and ii) English source texts and their Chinese translation[3]. This ensures a balanced representation of linguistic phenomena and helps minimize the problems caused by the direction of translation [14]. We start with Chinese (either as the source language or the target language) and establish how Chinese words, phrases, chunks or sentence patterns are structured syntactically, what semantic features they have, and how they are expressed in English. The parallel texts selected are mainly novels or essays relating to culture, ensuring that the knowledge base can support applications for a range of general purposes. The size of the parallel corpus is shown in Table 1.

[1] In a sentence-aligned parallel corpus, a sentence pair is a 2-tuple AS=<Si, Ti>, where both Si (in the source language) and Ti (in the target language) consist of a set of one or more sentences, with Si and Ti being corresponding expressions of each other.

[2] The parallel corpus that our work is based on consists of parallel texts. There are XML tags in these texts already, marking up the alignment at the text, paragraph and sentence levels. In Example 1, the first line specifies the title of the text, from which the sentence pair is taken. The tags <a> and </a> mark up the aligned sentence pair, the value of the attribute *id* is the same for the corresponding Chinese and English sentences, and the value of the attribute *no* specifies the number of sentences involved. Sentences are marked up with <s> and </s>, and the value of the attribute *id* specifies the sentence's location in the text. The other XML tags are the results of the present research, which will be illustrated in Section 3.3.

[3] There are XML tags in the parallel texts, which indicate the direction of translation and are retrievable when needed.

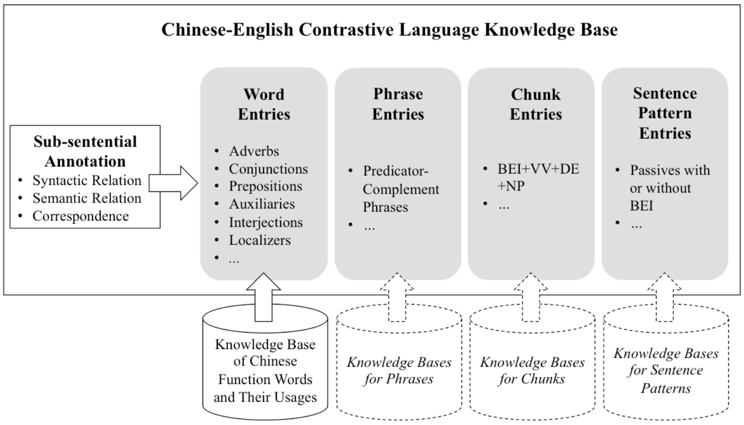
Chinese-English Contrastive Language Knowledge Base
Sub-sentential Annotation
• Syntactic Relation
• Semantic Relation
• Correspondence
Word Entries
• Adverbs
• Conjunctions
• Prepositions
• Auxiliaries
• Interjections
• Localizers
• ...
Phrase Entries
• Predicator-Complement Phrases
• ...
Chunk Entries
• BEI+VV+DE +NP
• ...
Sentence Pattern Entries
• Passives with or without BEI
• ...
Knowledge Base of Chinese Function Words and Their Usages
Knowledge Bases for Phrases
Knowledge Bases for Chunks
Knowledge Bases for Sentence Patterns

Fig. 1. The Architecture of CECLKB

Table 1. Size of the Parallel Corpus

Translation Direction	Chinese to English	English to Chinese
Total Number of Texts Pairs	250	1221
Total Number of Sentence Pairs	39134	157996
Number of Sentence Pairs from Novels	24929	102774
Number of Sentence Pairs from Essays on Culture	2800	16890

CECLKB is designed to integrate with other existing language knowledge bases, promoting the sharing of formalized linguistic knowledge among different projects. The Chinese Function Word Usage Knowledge Base [15] is the first that has been integrated; it allows the retrieval of linguistic features of Chinese function words directly through CECLKB.

3.2 Entries in CECLKB

Each entry in CECLKB contains a sub-sentential alignment, which exploits parallel texts to obtain fine-grained contrastive knowledge. At present, CECLKB entries focus on: i) adverbs, e.g. 随手 *casually*, ii) predicator-complement phrases, e.g. 吃得很多 *ate quite a bit*, iii) BEI+VV+DE+NP[4], e.g. 被破获的扒手 *captured pickpocket*, and iv) passives with or without BEI, e.g. 啤酒(被)酿造出来 *the beer was brewed*. Below we set out the rationales behind our selection of entries.

There are far fewer function words than content words in Chinese. A function word, however, usually carries much more weight than a content word does [16]. To

[4] BEI stands for the preposition 被, the most important passive marker in Chinese. DE stands for the auxiliary 的, a structural particle usually placed after an attributive modifier. The chunk BEI+VV+DE+NP, in most circumstances, forms a noun phrase with passive attribute.

begin with, we select adverbs as the focus of word entries, which are usually categorized as function words but do have their own lexical meanings.

Predicator-complement phrases come next. They are one of the most frequently used phrase types in Chinese. They have a wide variety of internal and external formal features, and the semantic links[5] of these complements are complex. Our pilot study showed a rich diversity of corresponding English expressions in parallel texts, a further reason for selecting them as the focus for phrase entries.

A BEI+VV+DE+NN structure represents a typical chunk in CECLKB. A chunk is defined as an ordered sequence of words and word categories, which exhibits special lexical, syntactic or semantic features. These special features are usually well captured by the corresponding English expressions in parallel texts. A chunk may also constitute a phrase of a certain type, such as BEI+VV+DE+NN constituting a complex noun phrase. Classified as a chunk, however, the sequence is seen more as a typical combination of particular words (被 and 的) and word categories (VV and NN) than as an ordinary phrase.

Entries of the fourth type deal with cases where corresponding Chinese and English sentence patterns fail to pair up in parallel texts. With passive constructions we first select the Chinese passives with or without BEI, which supposedly correspond to English passives. The focus of our observation is set on: i) Chinese passives with BEI, which are not expressed by English passives; and ii) Chinese passives without BEI or Chinese non-passives, which are expressed by English passives.

3.3 Annotation Schemes

The way in which linguistic knowledge is annotated reflects the nature of the relevant linguistic phenomena and how the formal descriptions are to be used. The annotations in CECLKB mark up the syntactic, semantic and corresponding relations in each entry, adding additional XML tags to the sentence pairs extracted from the corpus.

When designing the annotation schemes, we mainly consider: i) what tags and attributes are needed to mark up the three relations set out below; ii) whether different tags and attributes are needed for different types of entries; iii) whether the annotation schemes are extensible; and iv) whether the annotations are adaptable for use with a range of applications?

For the convenience of explanation, we begin by defining the three kinds of correspondences (abbreviated as HL, CT and NO respectively) in CECLKB, illustrated by five examples of the adverb 随手 *casually* and its corresponding English expressions.

[5] In a sentence, if constituent A is semantically linked to constituent B, then A is immediately related to B in meaning. In Chinese, there is considerable ambiguity about the semantic link of three kinds of sentence constituents: complements, modifiers (particularly adverbials) and predicates [17]. In Example 1, the complement 开心 *delighted* specifies the feeling of the experiencer 周经理 *Chou* – the subject of the clause, and is therefore semantically linked to 周经理 instead of the predicator 听 *listen*.

- HL: Highlighted Correspondence (Examples 2 and 3) is a 2-tuple in a sentence pair, written as HL=<CSeg, ESeg>, where CSeg is the focused Chinese word, phrase, chunk or sentence pattern, and ESeg is an English expression, with CSeg and ESeg being the corresponding expressions of each other.
- CT: Contextual Correspondence (Examples 4 and 5) is also a 2-tuple in a sentence pair, written as CT=<CSeg, ESeg>, where CSeg is the combination of the focused Chinese word, phrase, chunk or sentence pattern and its context, and ESeg is an English expression, with CSeg and ESeg being the corresponding expressions of each other. Further, there is no sub-segment in ESeg, which corresponds, on its own, to the focused Chinese word, phrase, chunk or sentence pattern.
- NO: No Correspondence (Example 6) is assumed when there does not exist an English expression in a sentence pair, which corresponds, in either of the two ways mentioned above, to the focused Chinese word, phrase, chunk or sentence pattern.

Example 2: **随手**翻开第二本的扉页，大叫道：“辛楣，你看见这个没有？”
He **casually** opened to the flyleaf of the other book and exclaimed, "Hey, Hsin-mei, did you see this one?"

Example 3: 范博文接过香来，**随手**又丢在地下，看见人堆里有一条缝，他就挤进去了。
Fan Po-wen took one and **immediately** let it drop to the ground, then, seeing a gap in the crowd, he pushed his way in.

Example 4: 马丁一言不发，也没有打什么招呼，就走了出去，悄悄地**随手**关上了门。
Without speaking or giving any kind of salutation, Martin went out, **closing the door** silently **behind him**.

Example 5: 福尔摩斯在他的一张名片背后**随手**写了几个字，扔给雷斯垂德。
Holmes **scribbled** a few words upon the back of one of his visiting cards and threw it over to Lestrade.

Example 6: 我们不打算趁四周无人时**随手**借它一只，就象我爸爸当年干的那个样子，因为那么一来，就会有人在后面追我们。
We warn't going to borrow it when there warn't anybody around, the way pap would do, for that might set people after us.

XML tags and attributes are designed to mark up i) the Chinese segments in focus, ii) the context of the focused segments, and iii) the corresponding expressions of the focused segments in English. For entries of different types or sub-categories, bilingual correpondence may involve a diversity of syntactic and semantic constrains. It is therefore necessary to have XML tags and attributes that apply to all entries and the special ones that apply to some entries only.

We use the Stanford parser for the pre-annotation of syntactic relations and have therefore adopted its inventory of POS and phrasal-category tags [18, 19], supplemented by an inventory of attributes and values (see Appendix for a selected list)[6].

Example 7 in Table 2 illustrates how predicator-complement phrase entries are annotated. The left column of the table shows how the phrase entry is annotated, and the right column zooms in on the three main targets of annotation.

[6] There are detailed annotation schemes for different categories and sub-categories of entries. For the sake of brevity, we only introduce attributes and values related to the examples here.

Table 2. A Predicator-Complement Phrase Entry (Example 7)

Stored Annotation	Targets of Annotation
1984-1.xml <a id="22" no="1"> <s id="1"> 玻璃**<NN SR="PT">窗</NN><VP FCS="T" GF="B"><VV COMP_SO="T">关</VV><DER>得</DER><VP COMP="DG"><ADVP><AD>很</AD></ADVP><VP><VA>严实</VA></VP></VP></VP>**，可是朝窗外望一眼，依然觉出外面冷得紧。</s></a> <a id="22" no="1"> <s id="1"> Outside, even through the **<ADJP GF="ATM" CO="HL"><VBN>shut</VBN></ADJP>** window-pane, the world looked cold.</s></a>	• **Chinese segment in focus** <VP FCS="T" GF="B"> <VV COMP_SO="T">关</VV> <DER>得</DER> <VP COMP="DG"> <ADVP><AD>很</AD></ADVP> <VP><VA>严实</VA></VP> </VP> </VP> • **Context of the focused segment** <NN SR="PT">窗</NN> • **English expression corresponding to the focused segment** <ADJP GF="ATM" CO="HL"> <VBN>shut</VBN> </ADJP>

In syntactic annotation we mark up:

- the focused Chinese phrase (关得严实), its syntactic structure and its grammatical function in the clause by i) tagging the POS of each word constituent and the syntactic tree of the phrase, and ii) adding the attribute *FCS* (its value being “T” to signal a focused phrase) and the attribute *GF* (its value being the code of the grammatical function) to the phrasal-category tag of the phrase;
- the context (窗 *window*, the subject) of the focused phrase, its syntactic structure and its semantic role in the clause by i) tagging its POS or phrasal category, and ii) adding the attribute *SR* (its value being the code of the semantic role) to the POS or phrasal-category tag of the context; and
- the corresponding English expression (*shut*) of the focused phrase, its syntactic structure and its grammatical function in the clause by i) tagging the POS of each word constituent and the syntactic tree of the expression, and ii) adding the attribute *GF* to the phrasal-category tag of the expression.

In semantic annotation we mark up:

- the type of the complement by adding the attribute *COMP* (its value being the code of the complement type) to the phrasal-category tag of the complement; and
- the semantic link of the complement, by adding the attribute *COMP_SO* (its value being “T”) to the constituent in the clause that the complement specifies.

In correspondence annotation we mark up:

- the highlighted correspondence, by adding the attribute *CO* (its value being “HL”) to the tag of the corresponding English expression;

- the contextual correspondence, by adding the attribute *CO* (its value being "CT") to the tag of the corresponding English expression; and
- the lack of parallelism, by adding the attribute *CO* (its value being "NO") to the tag of the focused Chinese phrase.

3.4 Progress of Annotation

Human annotators following a strict set of guidelines annotate the corresponding texts, with the support of annotation tools. The annotators are researchers, graduate students and trained undergraduates. They have a background in linguistics and are native speakers of Chinese, who speak English. The guidelines consist of the general principles of annotation and the detailed rules, suggestions and samples for different categories and sub-categories of entries. Annotations are checked for adherence to guidelines and consistency by two chief annotators. New rules and samples are added to the guidelines when annotators encounter examples not yet covered.

At the present stage, the tools we use are the Stanford parser[7] (for syntactic pre-processing) and the UAM CorpusTool[8] (for markup). An annotation tool is being developed, which will integrate the automatic pre-processing with the human markup process. It will also assist XML validation, tree display and quantitative analysis. Once implemented, we expect to make the tool available to other projects.

For the pilot annotation covering all four major entry types, 4000 sentence pairs have been extracted from the parallel corpus. We have completed the second round of annotation for 100 entries in the adverb sub-category and 100 entries in the predicator-complement sub-category. In the first round, we also completed the annotation for 230 entries in the BEI+VV+DE+NP sub-category.

4 Applications in CAT and SLA

The complexity of language makes it extremely demanding to process natural languages automatically and precisely. This being the case, we have designed the contrastive language knowledge base as a collection of well-understood and formally described facts about language, which extend the human intelligence using CAT and SLA tools rather than attempting to substitute it.

4.1 CECLKB and TM Tools in CAT

The success of TM is a question of its usefulness, that is to what extent can translations extracted from a TM tool be of use to a human translator [4]. The needs of translators vary when they search for contrastive linguistic knowledge. The TM tool that we are designing incorporates CECLKB and is therefore able to allow translators to

[7] http://nlp.stanford.edu/software/lex-parser.shtml

[8] http://www.wagsoft.com/CorpusTool/

highlight certain types of entries, which will then be given more weight when the TM tool extracts previous translations.

Take the Chinese adverb 随手 as an example. The Contemporary Chinese Dictionary [20] gives the following information about this word, which includes a bilingual definition and a contextual translation:

- 顺手 sth. done at sb.'s convenience; without extra effort; sth. that can be done handily along with sth. else: 出门时请～关门。*Please shut the door as you leave.*

With the support of CECLKB, a TM tool can amongst other things provide the following:

- ways of expressing this adverb in English when it is found in different contexts (see Examples 2 to 6), and particularly, how this adverb is expressed in contextual correspondences (see Examples 4 and 5), something which is usually not available in bilingual dictionaries; and
- probabilistic rules based on the formal descriptions, to assist the translator with his choices and decisions in the process of translation.

We are suggesting that a translator "borrows", but not necessarily "follows", these probabilistic rules. Information obtained from the analysis of tags, attributes and their values may offer additional choices to the translator. For instance, in CECLKB the adverb entries of 随手 exhibit the bilingual correspondence as follows[9]:

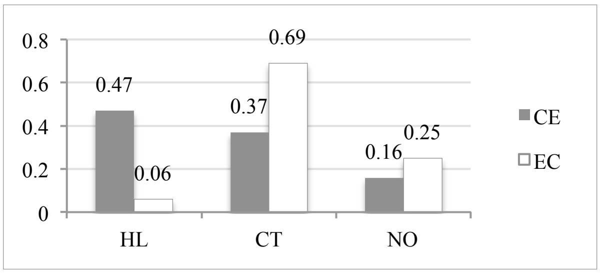

Fig. 2. Bilingual Correspondence in the Adverb Entries of 随手

As the annotation is still going on, the sample currently available for analysis is quite small[10]. It can be seen, however, that in the CE sentence pairs, the meanings of 随手 are more often expressed by corresponding English adverbs, such as *immediately*, *casually*, *idly*, etc. In contrast, the EC sentence pairs give us more English verbs, such as *scribble*, *jam*, *shove*, *snatch*, *toss*, etc., which describe not only the action but also the manner in which the action is carried out. This is just one example, but there

[9] In Fig. 2, CE stands for the sentence pairs with Chinese as the source language and English as the target language, while EC stands for the sentence pairs with English as the source language and Chinese as the target language.

[10] In Fig. 2, only 19 CE sentence pairs and 36 EC sentence pairs are observed.

will be much more contrastive linguistic knowledge that a translator can obtain from CECLKB through TM tools.

4.2 CECLKB and Language Tools in SLA

While acquiring a new language the learner needs to understand the range of subtle semantic, stylistic and rhetorical meanings associated with any new expression she encounters. The learner also needs to understand the range of expressions available to her when voicing her own ideas and feelings.

For instance, with the predicator-complement phrases (see Example 7), an SLA learner of Chinese needs to learn that these phrases describe events and that they express a sense of degree or the potential of the events. Specifically, she must understand that (窗)关得严实 *(window) be shut firmly* is different from (窗)关得上 *(window) can be shut* in that the former indicates how tight the window is shut (the degree of the event), while the latter indicates if the window can be closed (the potential of the event). Further, she needs to know that (窗)关得严实 is also different from (窗)做得结实 *(window) be solidly built*. The complement 严实 *firm(ly)* elaborates on the action 关 *shut*, while the complement 结实 *solid(ly)* comments on the state of 窗 *window*. This kind of information is available in CECLKB, with the types of complements and their semantic links annotated for all predicator-complement phrases.

The learner not only learns to use a predicator-complement phrase to describe an event and its degree or potential, she also needs to know how to use and structure the phrase. Table 3 shows how the annotated 100 predicator-complement phrases function grammatically: 81 of them acting as the predicates in subject-predicate constructions, e.g. 窗关得严实. Further, it shows that 11 phrases with potential complements, e.g. 看得见(整个房间) *could command (the whole room)*, act as the predicators in predicator-object constructions, while phrases with degree complements are not found used in this way. See Appendix for the details of the tags for grammatical functions.

Table 3. Predicator-Complement Phrases: Grammatical Functions vs. Complement Types

Grammatical Functions	Degree Complements	Potential Complements
A	0	1
B	73	8
C	0	11
D	1	0
G	1	0
K	5	0

There are other ways in which the annotations can be exploited in SLA learning scenarios. The SLA language tool we are developing helps non-native Chinese learners improve their reading skills. An understanding of function words, phrases, chunks and sentence patterns plays an important role in developing a level of proficiency in reading Chinese texts. With pre-designed learning scenarios, training data can be retrieved automatically and dynamically from the knowledge base.

5 Conclusion

It is our expectation that a contrastive knowledge base with its collection of rich, in-depth and formalized knowledge about two (or more) languages will be of significant use to NLP technology. CECLKB is linguistically motivated and can provide dynamic and diversified language assistance for CAT and SLA applications. It achieves this by i) including four different types of knowledge entries, ii) selecting diversified sub-categories of knowledge entries with special linguistic focuses, iii) describing the bilingual and sub-sentential contrast from three perspectives in general, and iv) specifying annotations typically applying to some sub-categories only. Further efforts will focus on the annotation of more data, the statistical analysis of the annotated data, the revision of annotation guidelines, the development of the annotation tool and the building of related CAT and SLA tools.

We are considering making part of the knowledge base available to NLP, SLA and other relevant research once a significant part of it is completed and validated. Our current development of an SLA demonstrator will make a small portion of the knowledge base accessible online.

Acknowledgements. This paper is supported by the National Key Basic Research Program of China 2014CB340504, the National Social Science Fund of China 14BYY096, and the Philosophy and Social Science Fund of the Ministry of Education of China 12JHQ046.

References

1. Yu, S., Duan, H., Zhu, X., Zhang, H.: The Construction and Utilization of a Comprehensive Language Knowledge-base. Journal of Chinese Information Processing, 18(5), pp. 1-10 (2004)
2. Bai, X., Chang, B., Zhan, W.: The Construction of a Large-scale Chinese-English Parallel Corpus. In: Huang, H. (ed.) Recent Development in Machine Translation Studies - Proceedings of the National Conference on Machine Translation 2002, pp. 124-131. Publishing House of Electronics Industry, Beijing (2002)
3. Benito, D.: Future Trends in Translation Memory. Tradumàtica. 7 - ISSN 1578-7559, http://www.fti.uab.cat/tradumatica/revista/hemeroteca.htm (2009)
4. Simard, M., Langlais, P.: Sub-sentential Exploitation of Translation Memories. In: Proceedings of Machine Translation Summit VIII: Machine Translation in the Information Age, pp. 335-339 (2001)
5. Bowker, L., Barlow, M.: Bilingual Concordancers and Translation Memories: A Comparative Evaluation. In: Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training, pp. 70-79, (2004)
6. Macken, L.: Sub-Sentential Alignment of Translational Correspondences. University of Antwerp (2010)
7. Véronis J. (ed.): Parallel Text Processing – Alignment and Use of Translation Corpora. Kluwer Academic Publishers, The Netherlands (2003)
8. Melamed, D.: Annotation Style Guide for the Blinker Project – Version 1.0.4. http://arxiv.org/pdf/cmp-lg/9805004v1.pdf (2008)

9. Véronis, J.: ARCADE: Tagging Guidelines for Word Alignment – Version 1.0. http://aune.lpl.univ-aix.fr/projects/arcade/2nd/word/guide/ (1998)
10. Merkel, M.: Annotation Style Guide for the PLUG Link Annotator. Linköping University (1999)
11. Linguistic Data Consortium. Guidelines for Chinese-English Word Alignment – Version 4.0. http://catalog.ldc.upenn.edu/docs/LDC2012T24/GALE_Chinese_alignment_guidelines_v4.0.pdf (2009)
12. Zhao, H., Liu, Q., Zhang, R., Lv, Y., Sumita, E.: Guidelines for Chinese-English Word Alignment. Journal of Chinese Information Processing, 23(3), pp. 65-87 (2009)
13. Bond, F., Wang, S.: Issues in building English-Chinese parallel corpora with WordNets. In: Orav H., Fellbaum, C., Vossen, P. (eds.) Proceedings of The Seventh Global WordNet Conference (GWC-7), pp 391–399. Tartu, Estonia (2014)
14. Johansson, S.: Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies. John Benjamins Publishing, Amsterdam (2007)
15. Zan, H., Zhang, K., Zhu, X., Yu S.: Research on the Chinese Function Word Usage Knowledge Base. International Journal on Asian Language Processing 21 (4), pp. 185-198 (2011)
16. Lu, J., Ma, Z.: Some Comments on the Function Words in Contemporary Chinese. Beijing: Language & Culture Press (1999)
17. Lu, J.: On Semantic Link Analysis. Essays on Linguistics, 1. Beijing: Beijing Language and Culture University Press (1997)
18. Santorini, B.: Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision). http://repository.upenn.edu/cis_reports/570/ (1990)
19. Xia, F.: The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0). http://repository.upenn.edu/ircs_reports/38/ (2000)
20. Institute of Linguistics, Chinese Academy of Social Science. The Contemporary Chinese Dictionary [Chinese-English Edition]. Beijing: Foreign Language Teaching and Research Press (2002)

Appendix. XML Attributes and Their Values - A Selected List

Major Grammatical Functions of Chinese Constituents

A: Subject in Subject-Predicate Construction
B: Predicate in Subject-Predicate Construction
C: Predicator in Predicator-Object Construction
D: Object in Predicator-Object Construction
E: Predicator in Predicator-Complement Construction
F: Complement in Predicator-Complement Construction
G: Attributive in Attributive-Head Construction
H: Head in Attributive-Head Construction
I: Adverbial in Adverbial-Head Construction
J: Head in Adverbial-Head Construction
K: Appendage

Major Grammatical Functions of English Constituents

P: Predicate

PR: Predicator

C: Complement

including:

- S: Subject
- O: Object
- PDC: Predicative Complement
- LCC: Locative Complement
- PPC: Prepositional Complement
- CTC: Catenative Complement
- CLC: Clausal complement
- PC: Complement of Preposition

A: Adjunct

including but not limited to

- ATM: Attributive Modifier
- ETM: External Modifier
- APM: Appositive modifier

SP: Supplement

GP: Gapping

SL: Embedded Constituents

Major Semantic Roles

CS: Causer

AG: Agent

PT: Patient

EX: Experiencer

ST: Stimulus

TH: Theme

RL: Relevant

RG: Range

LC: Location

Types of Complements

PT: Potential

DG: Degree

RS: Result

DR: Direction