Developing An Annotating Tool For Doing Formosan Language Archive

Ching-hua Yu
Institute of Linguistics
Academia Sinica, Taiwan
harryyu@gate.sinica.edu.tw

Feng-ying Chou
Lung-hwa University of
Science and Technology, Taiwan
fyc@mail.lhu.edu.tw

Abstract

One of the tasks field linguists and language documenters face is that of assigning glosses to words or morphemes, including affixes. These glosses are typically used in annotating the interlinear text at a morpheme level. But without a computer-assisted annotating tool, annotating the interlinear texts by hand is tedious and prone to errors and inconsistencies. We describe a tool which is oriented to assist field linguists and language documenters to assign glosses to grammatical morphemes and help them auto-translate the general terms based upon a mapping dictionary. Our tool has built a working model for archiving the Formosan languages and a linguistic ontology of morphosyntactic analysis from archiving practice. We illustrate this assistant from the user's point of view and from an internal perspective.

Keywords: Formosan languages, gloss, interlinear text, translation, morpheme, ontology

1. Introduction

Linguistic analysis is costly and time-consuming. If, however, we develop an adequate analysis assistant, time and effort can be saved from routine processing and more attention can be given to the thought-provoking work. One of the tasks field linguists and language documenters face is that of assigning glosses to words or morphemes, including affixes. [4] Among other applications, these glosses are typically used in annotating the interlinear text at a morpheme level, as in the following example[1]:

mani	piiʔa-ŋa	ki-ʔivaha.
就	動態.非限定:做-已經	否定-願意
then	DYN.NFIN:do-already	NEG-accept

但是長輩們還是不願意。
As earlier, the elders refused.

(DRKMn_02_013_b)

Another challenge is to maintain the bilingual version of the interlinear glossed texts in the documents.[2] As such, annotating the interlinear texts by hand is tedious and prone to errors and inconsistencies. For that reason, many field linguists usually count on interlinear text tools. [1] Pitifully, these tools are occasionally unwieldy and may produce spurious parses. More importantly, they are language-specific and therefore are not suitable for our job of archiving the Formosan languages as endangered languages. There is the need for developing a pretty handy tool for documenting the Formosan languages in the Language Archive project.

Our primary concern is with how we can ensure field linguists to annotate the transcribed data properly and efficiently in a consistent way. Initially, we believe that language documenters would benefit from access to standards for encoding the meaning of grammatical morphemes, i.e., standard glosses. (Lexical morphemes – stems and roots – are glossed with general terms.) Then, they can be made to look up the on-screen list for grammatical tags or labels while at the same time glossing the interlinear text in a word processor.

Linguists by and large conform to certain notational conventions in glossing. Initially, we developed a open set of glosses to fit the corpora in 2001. Then, we referred to the Leipzig Glossing Rules that are internationally followed. [6] The discussion of how to apply the rules for annotating the Formosan language corpora is beyond the scope of this study, but their information can be seen in [7], [8], [9], and [10].

Thus far, we have collected eleven Formosan languages and created a word list from the gloss lines appearing in the entire archive. This is a list of words (or terms) where each word is accompanied by a number indicating how many times that word occurs. [1] It is useful for building a bilingual lexicon, which will be discussed in Section 5.

The present paper describes an annotating tool for fast input of standard glosses (for grammatical morphemes) and general terms (for lexical morphemes) in the Word documents. Using it, the field linguists or linguistic documenters can produce an ontology of linguistic terminology to satisfy the need for standards for formal glossing.

The rest of this paper is organized as follows: Section 2 provides an overview of our software solution. Section 3 describes the advantage of looking up the linguistic terminology on the fly. Section 4 discusses the possibility of extending the standard glosses. Section 5 is concerned with building a mapping dictionary of general terms or words. The limitation of the tool and the expectation for a fully automatic tool is discussed in Section 6. This paper concludes with Section 7, where our annotating tool has provided a working model for doing Formosan language archive.

2. Overview of solution

A diagram giving an overview of our solution is shown in Figure 1. The software tool is abbreviated as **AnnoTool**, which is specifically designed for archiving the Formosan language corpora. It provides three great advantages: (1) looking up the linguistic terminology on the fly, (2) facilitating the annotation of interlinear texts by auto-inserting the standard glosses, i.e., the abbreviations of grammatical category labels (e.g. LOC, ASP, FP), and (3) translating the general terms (e.g. wine, cook, afraid) as well as standard glosses based upon a mapping dictionary.

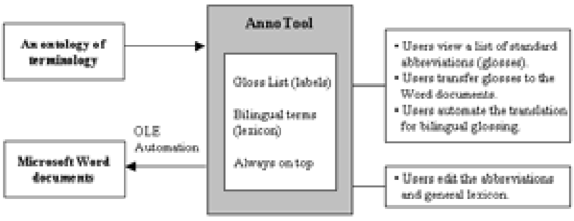

An ontology of terminology
AnnoTool
Gloss List (labels)
Bilingual terms (lexicon)
Always on top
OLE Automation
Microsoft Word documents
• Users view a list of standard abbreviations (glosses).
• Users transfer glosses to the Word documents.
• Users automate the translation for bilingual glossing.
• Users edit the abbreviations and general lexicon.

Figure 1. Overview of AnnoTool

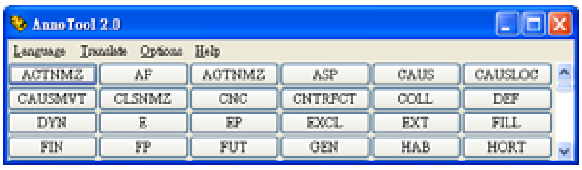


Figure 2a. AnnoTool English Interface

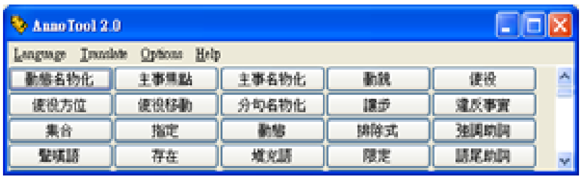
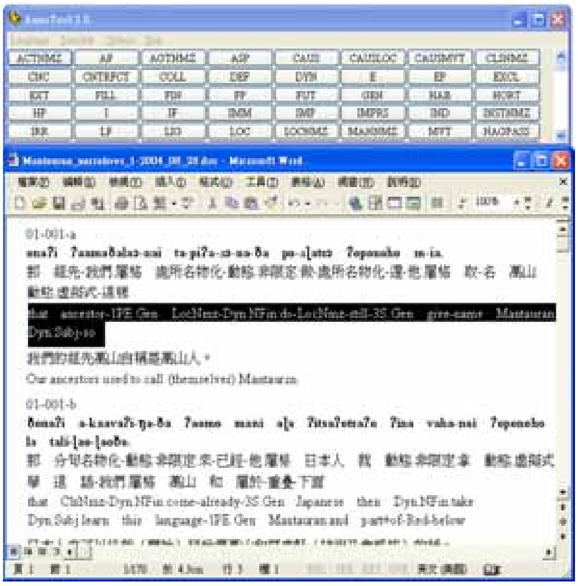


Figure 2b. AnnoTool Chinese Interface

When launched, the program pops up the list of morphosyntactic abbreviations used to annotate the interlinear texts (see Figures 2a & 2b). To respond to the potential need for more abbreviations, **AnnoTool** has been conceived to allow the expansion of abbreviations used by linguists (see Section 4). Its second major function is to translate the grammatical labels from English into Chinese – or vice versa – in order to reduce the workload in performing the bilingual glossing. The last, but not least function is the built-in mapping dictionary, which is used to transliterate the glossed words automatically.

AnnoTool uses OLE automation to communicate with Microsoft Word. The user must have both programs running conjointly. However, it is required to arrange the Desktop so that the two programs do not overlay each other. As shown in Figure 3, **AnnoTool** usually occupies one-third of the screen, and Word two-thirds. The label text can be easily exported from the **AnnoTool** window to the current Word document by clicking one of the buttons in the application window. This method makes our linguistic analysis more efficient and more accurate. It is more efficient because the linguist can view the on-screen list and stick to a pre-defined terminology. It is more accurate because the chances of misspelling the abbreviated labels are kept to a minimum.



01-001-a

Our ancestors used to call (themselves) Mantauran.

01-001-b

Figure 3. Using AnnoTool with Word

Labels can be translated from English into Chinese, or vice versa. To do so, the user must first select a single term or an entire line from a document, then switch to **AnnoTool** and click English→Chinese (or Chinese→English) from the Translate menu. Accordingly, the selected sequence in Word can be translated into the target language.

If users check the “Using a built-in bilingual dictionary” option from the Options/Dictionary menu, the general terms familiar to the dictionary will be automatically replaced by their translation equivalents. As a result, most of the glossing work has already been processed and only a few strange terms are left to human translators.

3. Looking up the linguistic terminology on the fly

As a computer-assisted interlinear text tool, **AnnoTool** provides one of the great advantages: looking up the grammatical terminology on the fly. No matter if users are really transferring the labels into the Word document, the on-screen list is always on top, allowing local query. They can select a desired one from a pool of tags or labels, especially when reference material is not immediately available. It is assumed that if each language documenter uses the identical tool, the possibility of data inconsistency will be eliminated.

With the floating function, **AnnoTool** can exist outside the working document and stay where it has been dragged for its position is dynamically saved to the Windows registry . It can be moved and resized horizontally or vertically so that the users can select the best window position for their work habit.

4. Extending the standard glosses

As can be seen, there are two types of glosses used in our archive.[3] One represents grammatical morphemes, generally rendered by abbreviated grammatical category labels, the so-called “standard glosses”. The other is used for explaining the meanings of lexical morphemes or words, which are given in English or Chinese. This type of glosses are called “general terms” in this paper.

The abbreviations of standard glosses used in the Formosan language corpora are shown in Table 1.

Table 1. Abbreviations used in the Corpora

ABBREVIATION	CHINESE	ENGLISH
ACTNMZ	動態名物化	Action nominalization
AF	主事焦點	Agent Focus
ASP	時貌 (或 動貌)	Aspect
CAUS	使役	Causative
CLSNMZ	分句名物化	Clausal nominalization
CNC	讓步	Concessive
CNTRFCT	違反事實	Counterfactual
DYN	動態	Dynamic
E	排除式 （＝我們）	Exclusive
EXCL	驚嘆語	Exclamation
EP	強調助詞	Emphatic Particle
EXT.IMM	存在.近距	Existential Immediate
EXT.REM	存在.遠距	Existential Remote
FILL	填充語	Filler
FIN	限定	Finite
FP	語尾助詞	Final Particle
GEN	屬格	Genitive Case
HP	勸建助詞	Hortative Particle
I	包含式 (=咱們)	Inclusive
IF	工具焦點	Instrumental Focus
IMPRS	無人稱	Impersonal pronoun
LF	處所焦點	Locative Focus
IMP	命令	Imperative
INSTNMZ	工具名物化	Instrument nominalization
LF.HORT	處所焦點.勸建	Locative Focus Hortative
LFNMZ	處所名物化	Locative Nominalization
LIG	連繫詞	Ligature
LOC	處所格	Locative Case
LOCNMZ	處所名物化	Locative nominalization
NAGPASS	非主事被動	Non agentive passive
NEG	否定	Negation
NEGIMP	否定命令	Negative Imperative
NFIN	非限定	Non-Finite
NOM	主格	Nominative Case
OBJNMZ	受事名物化	Objective Nominalization
OBL	斜格	Oblique
PASS	被動	Passive
P, PLUR	複數	Plural
PERF	完成貌	Perfective
PF	受事焦點	Patient Focus
PF.HORT	受事焦點.勸建	Patient Focus Hortative
PRFCT	完成進行	Perfect
PROG.IMM	進行.近距	Progressive Immediate
PROG.REM	進行.遠距	Progressive Remote
QP	引述助詞	Quotative Particle
REF	反身	Reflexive
REC	相互	Reciprocal
RED	重疊	Reduplication
S	單數	Singular
STAT	狀態	Stative
STATNMZ	狀態名物化	State nominalization
SUBJ	虛擬式	Subjunctive
SUBJNMZ	主語名物化	Subjective nominalization
SUP	最高級	Superlative
TEMPNMZ	時間名物化	Temporal nominalization
TOP	主題	Topic
1	我(們)	1st Person
2	你(們)	2nd Person
3	他(們)	3rd Person
.	帶著兩種功能之詞素	Portmanteau Morpheme
:	(可區分之)詞綴	(Divisible) Affix
-	接詞	Affix or Clitic
<>	中綴	Infix
*	無法確定構詞語法功能	Morphosyntactic function undetermined

These abbreviations are universally used in the linguistic analysis of the Formosan languages. To add or remove one or more gloss strings (as items) to or from the list, the user can edit the built-in INI file, of which the internal format is shown below:

```
; Annotating the Formosan languages corpora
[Button0]
English=1200
Chinese=1400
[Lexicon]
Built-in Dictionary=1
[Font]
Name=Times New Roman
```

```
Size=9
[Configuration]
NonWordChars=.,;:!#$^&()[]{}<>+=-/\|`~"
[ItemSet]
ItemCount=80
Item0=ACTNMZ,動態名物化
Item1=AF,主事焦點
Item2=AGTNMZ,主事名物化
Item3=ASP,動貌
Item4=CAUS,使役
Item5=CAUSLOC,使役方位
Item6=CAUSMVT,使役移動
Item7=CLSNMZ,分句名物化
Item8=CNC,讓步
…
```

To annotate a particular language, users can produce a language-specific gloss list from the universal gloss list which consists of a predefined number of standard abbreviations, as shown in Table 1. Thus, a customized gloss list is created.

5. Building a mapping dictionary of general terms

Since 2002, we have built a large multilingual corpus of 11 Formosan languages spoken in Taiwan, amounting to 146,000 orthographic words. From the interlinear glosses aligned with the original transcriptions, a parse program has been designed to extract a similar number of meaningful terms.

The very frequent words of English form a large proportion of any text. [5] We found out that many a term (type) appears more than once (token). It follows that we can reach the potential terms in a new context by making good use of a lexicon. The relationship of coverage and the N most frequent words is shown in Figure 4. It may well be expected that there is a strong tendency for each text to use common words, which implies that the automatic translation of these words would be possible. This concept drives our design of automating the literal translation of lexical morphemes or words based upon a bilingual lexicon.

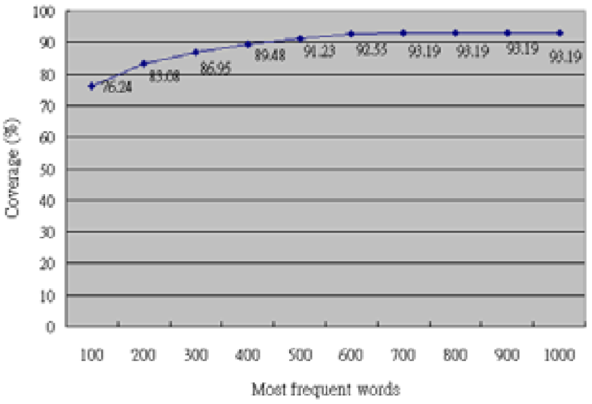


Figure 4. Relationship of coverage and top N words in the corpora

Likewise, users can modify the bilingual lexicon as in the following XML file:

```
<lexicon>
      …
  <word>
    <item lang=“English”>spring</item>
    <item lang=“Chinese”>春天</item>
  </word>
  <word>
    <item lang=“English”>sprinkle</item>
    <item lang=“Chinese”>撒</item>
  </word>
  <word>
    <item lang=“English”>sprout</item>
    <item lang=“Chinese”>長出來</item>
  </word>
  <word>
    <item lang=“English”>sputum</item>
    <item lang=“Chinese”>口水</item>
  </word>
  <word>
    <item lang=“English”>squint</item>
    <item lang=“Chinese”>斜視</term>
  </word>
      …
</lexicon>
```

It is estimated that such automation has reduced a significant amount of human translation, as far as standard glosses and general terms are concerned.

6. The limitation of AnnoTool

It must be admitted that AnnoTool was programmed to translate the familiar glosses or terms. It is not capable of processing strange words or phrases. There is the need for expanding the existing lexicon to cover a broader range of vocabulary items.

On the other hand, one source term is found, in many cases, to have different target translations. There is no one-to-one relationship between translation unit and its equivalent. At this juncture, we just choose the most frequent term translation as the typical one in the dictionary. It is instead necessary to show the users a list of all possible translations and allow them to select a desired one from the list.

Finally, the statistical-based gloss tool is needed to parse the gloss information without the need for human intervention. Nevertheless, the use of this tool makes our linguistic analysis easier than it used to be.

7. Conclusion

We have described an annotating tool, **AnnoTool**, which is specifically designed to assist our language documenters to assign glosses to grammatical morphemes and help them translate the general terms automatically. From the user’s point of view, this is a glossing tool, but **AnnoTool** builds a working model for the Formosan languages and assigns morphosyntactic features to the glossed morphemes.

AnnoTool allows the user to choose glosses for the Formosan languages from a linguistic ontology of morphosyntactic properties. In addition, **AnnoTool** performs the mapping between the general terms and their translation equivalents by way of a bilingual dictionary.

We have specified the functional design of this tool and the possible contributions to the archive of Formosan languages as endangered languages. Though the tool is far from perfect, it has indeed automated linguistic processing in a positive way.

Acknowledgements: This research was supported by the Taiwan National Science Council under the Language Archive project of the NDAP program (2002-2006). We thank Elizabeth Zeitoun, the leader of the Formosan language archive project, for her helpful comments on this paper.

Notes

1 This example is taken from the Formosan language archive project website (http://formosan.sinica.edu.tw).

2 The most common form of interlinear text in linguistic description is the three-line format: a line of transcribed data, often broken down by morpheme, a line of grammatical and gloss information aligned with the text in the first line, and a line representing some form of free translation. [2] Variations to this basic form, however, are often seen in literature. [3] In the case of the Formosan language corpora, for example, the five-line format is used, because bilingual information has to be considered.

3 Based upon the distinction of Maxwell et al. (2002), we divide glosses into two types: "standard glosses" and "general terms". The former represents the abbreviations of grammatical category labels, and the latter the general lexicon used for literal translation.

References

[1] Antworth, E. and Valentine, J. "Software for doing field linguistics", in Lawler, John and Dry, Helen Aristar (eds) *Using Computer in Linguistics: A Practical Guide*. London and New York: Ithaca: Routledge, 1998.

[2] Catherine Bow, Baden Hughes and Steven Bird, 2003. Towards a General Model of Interlinear Text, in "Proceedings of EMELD Workshop 2003: Digitizing & Annotating Texts & Field Recordings". LSA Institute: Lansing MI, USA. July 11-13, 2003.

[3] Lewis, W. D. Mining and Migrating Interlinear Glossed Text, in "Proceedings of the EMELD Workshop on Digitizing and Annotating Texts and Field Recordings", East Lansing, MI, 2003.

[4] Maxwell, M., Simons, G., and Hayashi, L. "A Morphological Glossing Assistant." In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain, 2002. [www.mpi.nl/lrec/papers/lrec-pap-25-MorphologicalGlossingAssistant.pdf]

[5] Sinclair, John. A Way with Common Words, in Hasselgård, Hilde and Oksefjell, Signe (Eds.), Out of Corpora: Studies in Honour of Stig Johansson. Amsterdam – Atlanta, GA: Rodopi, pp. 157-179, 1999.

[6] The Leipzig Glossing Rules: http://www.eva.mpg.de/lingua/files/morpheme.html

[7] Zeitoun, Elizabeth. "Discussion on the digitization of the Formosan Language Archive: The Formosan language archive: a linguistic perspective." In Proceedings of The First Workshop on Digital Archives Technologies. Taipei: Institute of Information Science, Academia Sinica, 2002.

[8] Zeitoun, Elizabeth, Yu, Ching-hua, and Weng, Cui-xia. "The Formosan Language Archive: Development of a Multimedia Tool to Salvage the Languages and Oral Traditions of the Indigenous Tribes of Taiwan." Oceanic Linguistics. 42.1: 218-232, 2003.

[9] Zeitoun, Elizabeth and Yu, Ching-hua. "The Formosan Language Archive: Language Processing and Linguistic Analysis." International Journal of Computational Linguistics & Chinese Language Processing. 10.2. June 2005.

[10] Zeitoun, Elizabeth. and Yu, Ching-hua. "The Formosan Language Archive: Language Processing and Linguistic Analysis," In Proceeding of 1st International Joint Conference on Natural Language Processing (IJCNLP-04) Fourth Workshop on Asian Language Resources, March 25, 2004, Sanya, Hainan Island, China.