The MILE Corpus for Less Commonly Taught Languages

Alison Alvarez, Lori Levin, Robert Frederking, Simon Fung, Donna Gates
Language Technologies Institute
5000 Forbes Avenue
Pittsburgh, PA 15213
[nosila, lsl, ref+, sfung, dmg]
@cs.cmu.edu

Jeff Good
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig
good@eva.mpg.de

Abstract

This paper describes a small, structured English corpus that is designed for translation into Less Commonly Taught Languages (LCTLs), and a set of re-usable tools for creation of similar corpora. [1] The corpus systematically explores meanings that are known to affect morphology or syntax in the world's languages. Each sentence is associated with a feature structure showing the elements of meaning that are represented in the sentence. The corpus is highly structured so that it can support machine learning with only a small amount of data. As part of the REFLEX program, the corpus will be translated into multiple LCTLs, resulting in parallel corpora can be used for training of MT and other language technologies. Only the untranslated English corpus is described in this paper.

1 Introduction

Of the 6,000 living languages in the world only a handful have the necessary monolingual or bilingual resources to build a working statistical or example-based machine translation system. Currently, there are efforts to build *language packs* for Less Commonly Taught Languages (LCTLs). Each language pack includes parallel corpora consisting of naturally occurring text translated from English into the LCTL or vice versa.

This paper describes a small corpus that supplements naturally occurring text with highly systematic enumeration of meanings that are known to affect morphology and syntax in the world's languages. The supplemental corpus will enable the exploration of constructions that are sparse or obscured in natural data. The corpus consists of 12,875 English sentences, totaling 76,202 word tokens.

This paper describes the construction of the corpus, including tools and resources that can be used for the construction of similar corpora.

2 Structure of the corpus

- 247: John said "The woman is a teacher."
- 248: John said the woman is not a teacher.
- 249: John said "The woman is not a teacher."
- 250: John asked if the woman is a teacher.
- 251: John asked "Is the woman a teacher?"
- 252: John asked if the woman is not a teacher.
- …
- 1488: Men are not baking cookies.
- 1489: The women are baking cookies.
- …
- 1537: The ladies' waiter brought appetizers.
- 1538: The ladies' waiter will bring appetizers.

Figure 1: A sampling of sentences from the complete elicitation corpus

[1] AVENUE/MILE is supported by the US National Science Foundation NSF grant number IIS-0121-631 and the US Government's REFLEX Program.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 5–8,
New York, June 2006.

srcsent: Mary was not a doctor.
context: Translate this as though it were spoken to a peer co-worker;

((actor ((np-function fn-actor)(np-animacy anim-human)(np-biological-gender bio-gender-female)
(np-general-type proper-noun-type)(np-identifiability identifiable)
(np-specificity specific)…))
(pred ((np-function fn-predicate-nominal)(np-animacy anim-human)(np-biological-gender bio-gender-female) (np-general-type common-noun-type)(np-specificity specificity-neutral)…))
(c-v-lexical-aspect state)(c-copula-type copula-role)(c-secondary-type secondary-copula)(c-solidarity solidarity-neutral) (c-power-relationship power-peer) (c-v-grammatical-aspect gram-aspect-neutral)(c-v-absolute-tense past) (c-v-phase-aspect phase-aspect-neutral) (c-general-type declarative-clause)(c-polarity polarity-negative)(c-my-causer-intentionality intentionality-n/a)(c-comparison-type comparison-n/a)(c-relative-tense relative-n/a)(c-our-boundary boundary-n/a)…)

Figure 2: An abridged feature structure, sentence and context field

The MILE (Minor Language Elicitation) corpus is a highly structured set of English sentences. Each sentence represents a meaning or combination of meanings that we want to elicit from a speaker of an LCTL. For example, the corpus excerpts in Figure 1 explore quoted and non quoted sentential complements, embedded questions, negation, definiteness, biological gender, and possessive noun phrases.

Underlying each sentence is a feature structure that serves to codify its meaning. Additionally, sentences are accompanied by a context field that provides information that may be present in the feature structure, but not inherent in the English sentence. For example, in Figure 2, the feature structure specifies solidarity with the hearer and power relationship of the speaker and hearer, as evidenced by the features-value pairs *(c-solidarity solidarity-neutral)* and *(c-power-relationship power-peer)*. Because this is not an inherent part of English grammar, this aspect of meaning is conveyed in the context field.

3 Building the Corpus

Figure 3 shows the steps in creating the corpus. Corpus creation is driven by a Feature Specification. The Feature Specification defines features such as tense, person, and number, and values for each feature such past, present, future, remote past, recent past, for tense. Additionally, the feature specification defines illegal combinations of features, such as the use of a singular number with an inclusive or exclusive pronoun (*We = you* and *me* vs *we = me* and *other people*). The inventory of features and values is informed by typological studies of which elements of meaning are known to affect syntax and morphology in some of the world's languages. The feature specification currently contains 42 features and 340 values and covers. In order to select the most relevant features we drew guidance from Comrie and Smith (1977) and Bouquiaux and Thomas (1992). We also used the *World Atlas of Language Structures* (Haspelmath et al. 2005) as a catalog of existing language features and their prevalence.

In the process of corpus creation, feature structures are created before their corresponding English sentences. There are three reasons for this. First, as mentioned above, the feature structure may contain elements of meaning that are not explicitly represented in the English sentence. Second, multiple elicitation languages can be generated from the same set of feature structures. For example, when we elicit South American languages we use Spanish instead of English sentences. Third, what we want to know about each LCTL is not how it translates the structural elements of English such as determiners and auxiliary verbs, but how it renders certain meanings such as

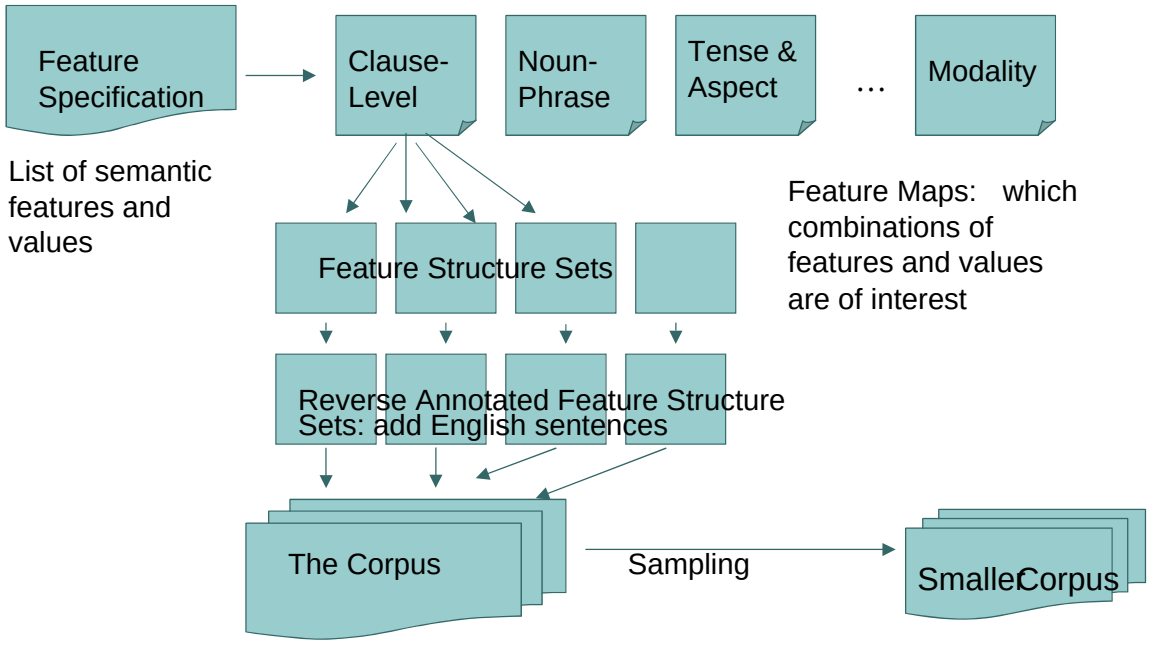

Figure 3: An overview of the elicitation corpus production process

definiteness, tense, and modality, which are not in one-to-one correspondence with English words.

Creation of feature structures takes place in two steps. First, we define which combinations of features and values are of interest. Then the feature structures are automatically created from the feature specification.

Combinations of features are specified in Feature Maps (Figure 3). These maps identify features that are known to interact syntactically or morphologically in some languages. For example, tense in English is partially expressed using the auxiliary verb system. An unrelated aspect of meaning, whether a sentence is declarative or interrogative, interacts with the tense system in that it affects the word order of auxiliary verbs (*He was running*, *Was he running*), Thus there is an interaction of tense with interrogativity. We use studies of language typology to identify combinations of features that are known to interact.

Feature Maps are written in a concise formalism that is automatically expanded into a set of feature structures. For example, we can formally specify that we want three values of tense combined with three values of person, and nine feature structures will be produced. These are shown as Feature Structure Sets in Figure 3.

4 Sentence Writing

As stated previously, our corpus consists of feature structures that have been human annotated with a sentence and context field. Our feature structures contain functional-typological information, but do not contain specific lexical items. This means that our set of feature structures can be interpreted into any language using appropriate word choices and used for elicitation. Additionally, this leaves the human annotator with some freedom when selecting vocabulary items. Due to feedback from previous elicitation subjects we chose basic vocabulary words while steering clear of overly primitive subject matter that may be seen as insulting. Moreover, we did our best to avoid lexical gaps; for example, many languages do not have a single word that means *winner*.

Translator accuracy was also an important objective and we took pains to construct natural sounding, unambiguous sentences. The context field is used to clarify the sentence meaning and spell out features that may not manifest themselves in English.

5 Tools

In conjunction with this project we created several tools that can be reused to make new corpora with other purposes.

- An XML schema and XSLT can be used to make different feature specifications
- A feature structure generator that can be used as a guide to specify and design feature maps
- A feature structure browser can be used to make complicated feature structures easier to read and annotate

6 Conclusion

The basic steps for creating a functional-typological corpus are:

1. Combinations of features are selected
2. Sets of feature structures representing all feature combinations are generated
3. Humans write sentences with basic vocabulary that represent the meaning in the feature structure
4. If the corpus is too large, some or all of the corpus can be sampled

We used sampling and assessments of the most crucial features in order to compile our corpus and restrict it to a size small enough to be translatable by humans. As a result it is possible that this corpus will miss important feature combinations in some languages. However, a corpus containing all possible combinations of features would produce hundreds of billions of feature structures.

Our future research includes building a Corpus Navigation System to dynamically explore the full feature space. Using machine learning we will use information detected from translated sentences in order to decide what parts of the feature space are redundant and what parts must be explored and translated next. A further description of this process can be read in Levin et al. (2006).

Additionally, we will change from using humans to write sentences and context fields to having them generated by using a natural language generation system (Alvarez et al. 2005).

We also ran small scale experiments to measure translator accuracy and consistency and encountered positive results. Hebrew and Japanese translators provided consistent, accurate translations. Large scale experiments will be conducted in the near future to see if the success of the smaller experiments will carry over to a larger scale.

7 References

Alvarez, Alison, and Lori Levin, Robert Frederking, Jeff Good, Erik Peterson September 2005, Semi-Automated Elicitation Corpus Generation. In Proceedings of MT Summit X, Phuket: Thailand.

Bouquiaux, Luc and J.M.C. Thomas. 1992. *Studying and Describing Unwritten Languages.* Dallas, TX: The Summer Institute of Linguistcs.

Comrie, Bernard and N. Smith. 1977. *Lingua descriptive series: Questionnaire.* In: Lingua, 42:1-72.

Haspelmath, Martin and Matthew S. Dryer, David Gil, Bernard Comrie, editors. 2005 *World Atlas of Language Strucutures. Oxford University Press.*

Lori Levin, Alison Alvarez, Jeff Good, and Robert Frederking. 2006 "Automatic Learning of Grammatical Encoding." To appear in Jane Grimshaw, Joan Maling, Chris Manning, Joan Simpson and Annie Zaenen (eds) *Architectures, Rules and Preferences: A Festschrift for Joan Bresnan* , CSLI Publica tions. In Press.