# Semi-Automated Elicitation Corpus Generation

**Alison Alvarez**
nosila@cs.cmu.edu

**Lori Levin**
lsl@cs.cmu.edu

**Robert Frederking**
ref+@cs.cmu.edu

**Erik Peterson**
eepeter@cs.cmu.edu

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15217

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

**Abstract**

In this document we will describe a semi-automated process for creating elicitation corpora. An elicitation corpus is translated by a bilingual consultant in order to produce high quality word aligned sentence pairs. The corpus sentences are automatically generated from detailed feature structures using the GenKit generation program. Feature structures themselves are automatically generated from information that is provided by a linguist using our corpus specification software. This helps us to build small, flexible corpora for testing and development of machine translation systems.

Keywords: corpora, elicitation, minor languages, generation

## 1 Introduction

In the field of Machine Translation fully aligned and tagged translation corpora are considered to be one of the most valuable resources for automatically training translation systems. However, among minority languages such resources are hard to find. It is possible to overcome this obstacle by using techniques inspired by field linguistics. That is, by drawing on bilingual informants to translate and align given sentences. We do this though a piece of software called the elicitation tool that presents sentences and context clues to a bilingual informant and collects translations and alignments (see Figure 1 on the following page).

Field linguists have relied on questionnaires that have remained relatively static over a number of years. We want the flexibility to change the questionnaire to reflect different semantic domains, different goals for machine translation systems, different levels of detail, etc. We also want the questionnaire to be available in multiple languages. For example, we would want a version of the questionnaire in Spanish for use by Latin American minority language speakers. We also want flexibility in lexical selection in order to avoid cultural bias and to choose appropriate lexical items for the major language. This paper will look at methods for specifying the scope and depth of an elicitation corpus as well as methods

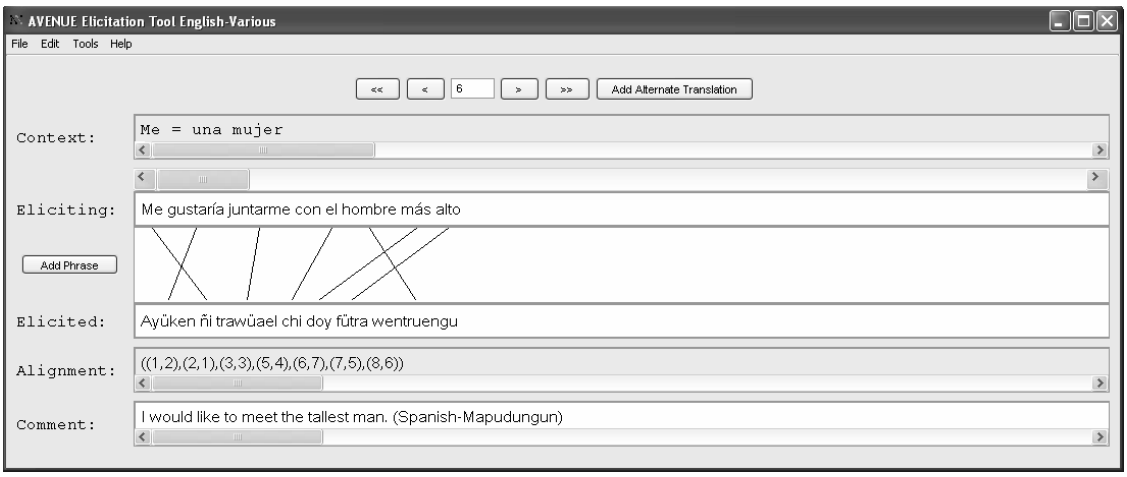


Figure 1: The elicitation tool is used by the informant for translation and alignment. Sentences are presented individually and can be annotated with context information when necessary.

for quick design and implementation of elicitation corpora.

Furthermore, we will look at a use of these methods to create a specific kind of corpus called a typological-functional corpus. This type of corpus is designed to elicit a range of language features (for example, tense, person, number) and explore the way those features are manifested in a target language.

## 2 The AVENUE Project

The work described in this paper takes place in the context of the AVENUE machine translation project[1].

AVENUE is focused on the development of machine translation systems for low-resource languages. Application of AVENUE to a new language involves three stages: elicitation, automatic learning, and rule refinement. This paper concerns only the elicitation stage of the project. Automatic learning of transfer rules from elicited data is described in (Probst et al. 2002). Automatic learning of morpheme boundaries and morphological paradigms is described in (Monson et al. 2004). Rule refinement via interaction with a consultant is described in (Font-Llitjos et al. 2005). At run time, the AVENUE system consists of a transfer engine and a decoder. The transfer engine encompases analysis, transfer, and generation and produces a large lattice of possible translations. The decoder uses statistical techniques to zero in on the best scoring hypothesis. (Lavie et al. 2003)

## 3 Elicitation Corpora

Our elicitation corpus will be sentence-based and have two main components: firstly, a feature structure (an example can be found in Figure 5) and secondly, the accompanying major language surface text with its optional context information. The informant will only see the major language sentence and its context comments, but the feature structure will be used to specify the coverage of the elicitation corpus and provide annotation to the sentences. These feature structures are designed to be as language neutral as possible; that is, they can be used as a guide to generate elicitation sentences in any language where a grammar

[1] AVENUE is supported by the US National Science Foundation,
NSF grant number IIS-0121-631

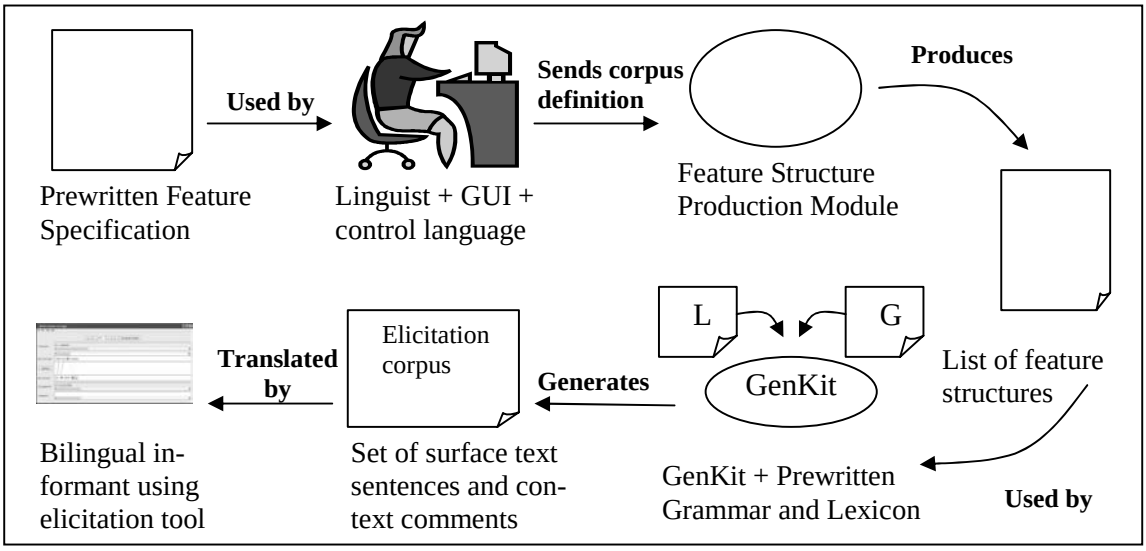


Figure 2: The start-to-finish process of the corpus generation system. Ovals indicate software components. The page-boxes indicate human or computer generated documents

and a lexicon can be built. For our purposes, feature structures will be used to generate the elicitation corpus in the major language. This corpus will then be translated and aligned into a minor language by an informant.

### 3.1 Feature Specification

```
srcsent:  I was the teacher.
context:  I = one_man

srcsent:  I was interesting.
context:  I = one_man

srcsent:  I was a teacher.
context:  I = one_man

srcsent:  I am a teacher.
context:  I = one_man

srcsent:  I was a teacher.
context:  I = one_woman
```

Figure 3: These sentences and context comments are part of an elicitation corpus used to elicit copula sentences. The feature structure for the first sentence can be found in Figure 5

How do we determine the range and number features that we would like to cover in our elicitation corpus? It may be important for us to look at plural and singular noun phrases, but it also might be important to determine whether a language delineates between plural, dual and paucal. The purpose of the feature specification is to define the list of features and corresponding values that are available for producing feature structures. Depending on our elicitation goals, the specification might include just singular and plural or all four possible values of number for a noun phrase. Choosing to have all four would insure that all possible properties of number are addressed in the feature corpus, but it might also cause combinatoric bloat.

Additionally, the feature specification determines what kind of phrases can use what kinds of features. For example, the polarity feature carries the value of positive and negative but can only be applied at the clause level.

Many features are also assigned a default value. This attribute will be detailed

```
<feature>
  <feature-name>np-my-number
  </feature-name>

  <value>
     <value-name>num-sg
     </value-name>
  </value>
  <value>
    <value-name>num-pl
    </value-name>
  </value>
  <value>
    <value-name>num-dual
    </value-name>
  </value>

  <note>
    Additional values of num
    ber: trial, quadral, pau-
    cal.  We will ignore
    these for now.
    (Notes for analysis of
    data: CS, 2.1.2.4.1 page
    38, seem to imply that-
    some combinations of
    numbers are more expected
    than others.)
   </note>
</feature>
```

Figure 4: A sample feature entry from the current feature specification

more closely when we look at corpus design in Section 3.3.

Additionally, the feature specification defines what feature-values cannot be combined. For instance, we may not want to apply first or second person to common nouns. We call these 'exclusions'.

We have written the feature specification with XML markup language. The specification itself is realized as a hierarchical structure of values contained within features. Each level also contains markup listing exclusions and further source notes.

### 3.2 Feature Structures

Our feature structures draw inspiration from Lexical Functional Grammar (see Bresnan (ed.) 1982). They are multi-level sets of feature-value pairs that are used to reflect the grammatical structures intended for elicitation. They can be designed to specify lexical items, but in order to reuse our set of feature structures with multiple languages we keep lexical items out of our feature structures and enable their specification in a specially designed GenKit grammar and lexicon (Tomita et al. 1988).

A feature structure is made of feature-value pairs that correspond to each phrase. Within the feature structure noun phrases may be labeled as subjects, objects, possessors or predicate nominatives (such as 'He is the teacher.'). Verbs generally correspond to the top level of the feature structure and there is no specific verb phrase node in the feature structure. Language specific headings such as 'subject' or 'predicate' can be dismantled and reconfigured for languages that do not have such syntactic phenomena.

Feature names and feature values must come from the feature specification.

```
((subj ((np-my-general-type pronoun)
       (np-my-person person-first)
       (np-my-number num-sg)
       (np-my-biological-gender gender-male)
       (np-my-function fn-predicatee)
       (np-my-animacy anim-human)…))
(predicate ((np-my-general-type common)
         (np-my-person person-first)
         (np-my-function predicate)
         (np-my-animacy anim-human)
        (np-my-definiteness indefinite)…))
(c-my-copula-type role)
(c-my-secondary-type secondary-copula)
(c-v-my-lexical-aspect state)
(c-v-my-absolute-tense past)
(c-v-my-phase-aspect durative)
(c-my-imperative-degree imp-degree-n/a)
(c-my-ynq-type ynq-n/a)…)
```

Figure 5: An abridged feature structure for the sentence "I was the teacher"

### 3.3 Corpus Design Control Language

The feature specification defines the allowable feature-value pairs in the elicitation corpus. However, it is not feasible to elicit every possible combination of features and values. Our current feature specification results in tens of millions of combinations of features and values. The corpus control language is used by a linguist to delimit a subset of features and values to include in the elicitation corpus.

The formalized description of the desired set of feature structures is called a "multiply". This description is written using a GUI to select features and set their range of values. Features can be specified with just one set value, a list of values that will be alternated throughout the set of generated feature structures, or they can have the string '#all' written next to them. The '#all' notation indicates that all values are to be multiplied out into the set of feature structures. Furthermore, disjoint statements can be used to create structures where one or more features vary in tandem. See Figure 6 for an illustration of the details.

The Corpus Control Language allows a linguist to summarize a set of feature structures. After then linguist writes a multiply, the feature structures are automatically created. The feature structures represent a cross product of all the feature-values that were contained in the multiply, minus those ruled out by exclusions. For example, if we want to create an elicitation corpus that looks at three values for tense (past, present and future) along with all combinations of polarity (positive and negative) we would end up with six sentences. If we also wanted to look at those features along with the values for the subject as first, second and third person we would end up with 18 sentences (= 2 values * 3 values * 3 values).

Not all features need to be specified in a multiply. All features carry a default neutral value that is automatically invoked when that feature is not used in a particular multiply. For example, if the feature for *polarity* is left unspecified, then the value is automatically set at *positive*. If a default setting for a particular feature is unacceptable, then an alternative can be specified within the multiply. This keeps our control language from being too cumbersome and tedious.

## 4 Generation

Generation is performed using GenKit generation software (Tomita et al. 1988). It takes the feature structures along with a corresponding grammar and lexicon

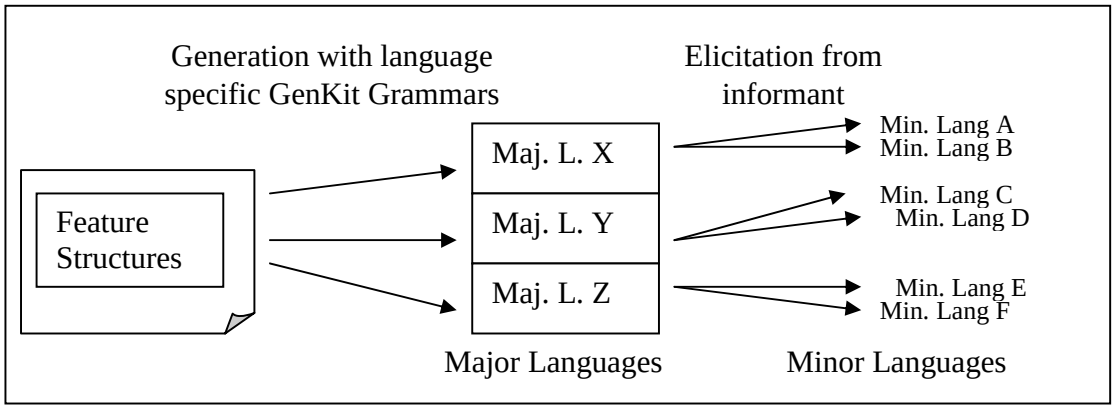


Figure 6: An illustration of the flexibility of the feature structures used for elicitation

and generates a surface string along with a comment. Generated comments are used to show pieces of meaning that might not be evident in the major/source language but may be found in the target/minority language. For example, the first person singular pronoun in English does not carry gender, so a comment will be generated indicating that "I = gender-female" or "I = gender-male". When using the elicitation tool this information is presented to the bilingual informant using the context field.

## 5 Work Related to the Functional-Typological Corpus

AVENUE is a system for learning translation rules from a word aligned bilingual corpus. One phase of rule learning is feature detection, which uses the elicitation corpus to discover morpho-syntactic properties of a minority language. For this we drew our inspiration from Robert Longacre's *Principles of Grammar Discovery* (1964). Thus, we expect to generate sentences with high degrees of uniformity that can easily be compared in order to discover typological properties such as whether the verb agrees with the subject, whether nouns have singulars and plurals, etc. In order to discover these properties, we compare sentences like "The child read a book" and "The children read a book" in order to see if the translation of "child" or "read" changes when "child" is understood as plural.

The design of our elicitation corpora are also modeled after questionnaires used by field linguists. The two most pertinent are the Comrie-Smith Questionnaire (1977) and *Studying and Describing Unwritten Languages* by Bouquiaux and Thomas (1992). We use these field linguistics guides to make an assessment of the type of morpho-syntactic phenomena that exist not only in our major language, but also those that have the potential to exist in almost any natural human language. We used these checklists as inspiration for the basic format of our feature structures and to determine scope and breadth of language features.

## 6 Functional-Typological Corpora

An elicitation corpus is the untranslated major language corpus that will be presented to the language informant using the elicitation tool. Although our corpus

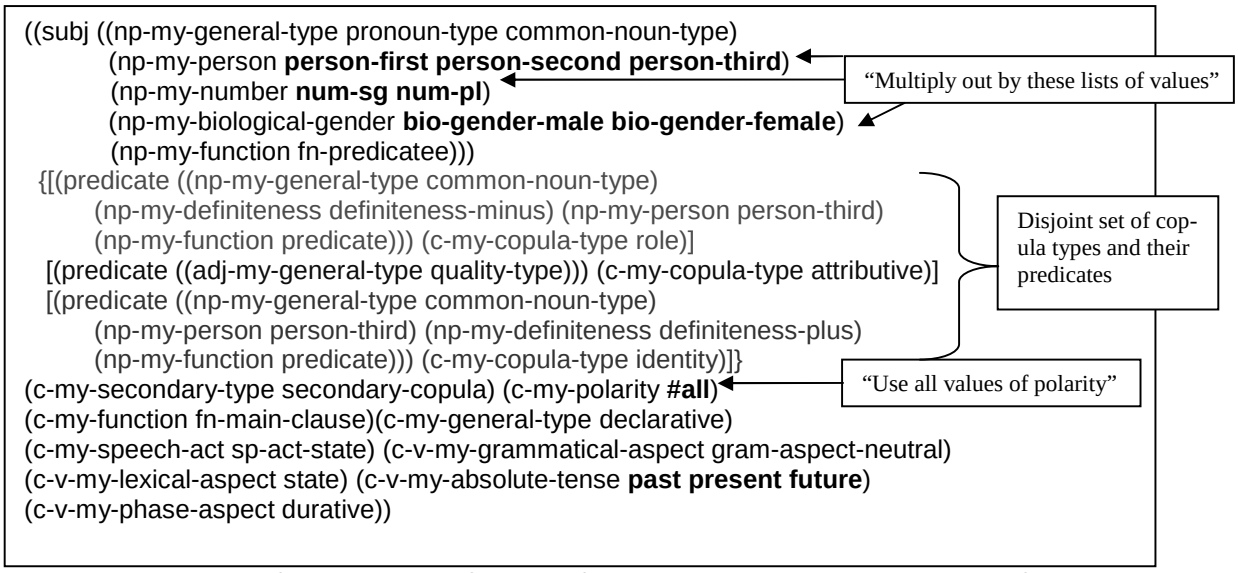

Figure 7: A multiply specification used to define a set of copula sentences with all combinations of tense, and subject person, number and gender

creation tools allow the creation of any kind of corpus, we have focused on linguistic functions such as cardinality and identifiability rather than on linguistic forms such as suffixes and determiners. Our focus on function is a consequence of the AVENUE rule learning scenario, in which it is possible that nothing is known about the form of the minor language. Our goal is to vary the functions and observe changes in the forms.

Our complete elicitation corpus is not generated as a whole; it is actually made of many smaller elicitation corpora. Each "mini-corpus" focuses on one general typological category and concentrates on building sentences with similar lexical items and structure for ease of grammatical discovery. Incremental generation lowers the development time of each mini-corpus and reduces the possibility of generating nonsensical sentences. Step-by-step development also increases the ease of testing, updating and expanding corpora.

The current version of the functional-typological feature specification was written in XML markup and then converted to a machine-readable format. The feature set is functional in the sense that it describes functions like actor and undergoer rather than morpho-syntactic realizations such as nominative and accusative. A process of feature detection (not included in this paper) determines which functions have morpho-syntactic realizations in the minor language that is elicited.

Currently, the feature specification contains about 50 features and a few hundred values and their corresponding exclusions. In addition, existing mini-corpora include open questions, copula sentences and declarative sentences. Each comes complete with its own GenKit grammar and lexicon.

So far we have used our functional-typological corpora in conjunction with a Hebrew-English translation system.

## 7 Conclusions and Future Work

It is possible to produce resources for minor languages using elicitation corpora and bilingual informants. It is also possible to produce these elicitation corpora in a way that will work across any major language-minor language pair and in a timely manner. We can use trained linguists to design the feature scope of a corpus, and produce a set of feature structures to cover that scope. Furthermore, each individual feature structure can produce a surface string and context cues with just a major language grammar and lexicon.

Our goal is to continue to produce a complete functional-typological corpus and use it to automatically discover the typological features of a language. We will do this by drawing comparisons between aligned and translated sentences with similar feature structures. In addition, we are exploring ways to control the size of a corpus while still ensuring that the features of a language have been fully explored.

We have also generated several narrow domain corpora, namely one for medical situations. We would also like to further test elicitation corpus generation for this purpose.

## 8 Acknowledgements

We would like to thank those who designed GenKit and those who maintain its current incarnation. Namely, we would like to thank Ben Han and Eric Nyberg for their help.

An earlier version of the elicitation corpus was written by Katharina Probst.

## 9 References

Bird, Steven, Gary Simons. 2003. Seven. Dimensions of Portability for Language Documentation and Description. Proceedings of the Workshop on Portability Issues in Human Language Technologies, Third International Conference on Lan-

guage Resources and Evaluation. Last Accessed: April 30, 2005. Website: http://www.ldc.upenn.edu/sb/home/papers/0204020/0204020-revised.pdf.

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

Bresnan J. editor. The Mental Representations of Grammatical Relations. MIT Press, Cambridge, Massachusetts, 1982.

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

E-Meld School of Best Practices in Digital Language Documentation. (2005) Last Accessed: May 2, 2005. Website: http://emeld.org/school/index.html

Font-Llitjos, Ariadna , Jaime Carbonell, Alon Lavie. 2005. "A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation" EAMT 10th Annual Conference 30-31 May 2005, Budapest, Hungary.

Harris, Alice C. 1981. Georgian Syntax: Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Alon Lavie, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjos, Rachel Reynolds, Jaime Carbonell, Richard Cohen. 2003. Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario. Transactions on Asian Language Processing (TALIP).

Levin, L., R. Vega, J. Carbonell, R. Brown, A. Lavie, E. Canulef and C. Huenchullan. June 2002. "Data Collection and Language Technologies for Mapudungun". In Proceedings of International Workshop on Resources and Tools in Field Linguistics at the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Canary Islands, Spain.

Longacre, Robert. (1964) Grammar Discovery Procedures. Mouton & Company, the Hague

Monson, Christian, Alon Lavie, Jaime Carbonell, and Lori Levin. 2004. Unsupervised Induction of Natural Language Morphology Inflection Classes. In Proceedings of the Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON).

Probst, K., R. Brown, J. Carbonell, A. Lavie, L. Levin, and E. Peterson. September 2001. "Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages". In Proceedings of the MT-2010 Workshop at MT-Summit VIII, Santiago de Compostela, Spain.

Katharina Probst, Lori Levin, Erik Peterson, Alon Lavie, Jaime Carbonell. 2002. MT for Minority Languages Using Elicitation-Based Learning of Syntactic Transfer Rules. Machine Translation, 17/4.

Tomita, Masaru and Eric Nyberg III. Generation and Transformation Kit, Version 3.2 User's Manual. CMU-CMT-88-MEMO, 1988.