Parallel Reverse Treebanks for the Discovery of Morpho-Syntactic Markings

Lori Levin, †Jeff Good, Alison Alvarez, Robert Frederking

Language Technologies Institute
Carnegie Mellon University
E-mail: lsl@cs.cmu.edu

†Department of Linguistics
Max Planck Institute for Evolutionary Anthropology

1 Introduction

This paper describes a corpus of syntactic structures and associated sentences. However, it is not a traditional treebank. The syntactic structures are created first and are then associated with sentences in a human language. We therefore call it a reverse treebank (RTB).[1]

The RTB has been created for elicitation of sentences in low resource languages. First, a corpus of feature structures is created using a tool suite built by the authors. The second step is to add sentences in a widely spoken language like English or Spanish that express the meanings of each feature structure. We will call this language the Elicitation Language. The third step is to have a bilingual informant translate the sentences into a low resource language. Using an elicitation tool, the informant can also graphically align the words of the Elicitation Language to the words of the low resource language. The result is a high quality parallel, word aligned corpus annotated with feature structures, which we will call a parallel RTB.

RTB sentences may have multiple clauses, but they are generally short in comparison to naturally occurring sentences in treebanks. The reason is that parallel RTBs provide small, but highly structured corpora for machine learning with small amounts of resources. Corpora such as these have been used for automatic learning of transfer rules for machine translation [8].

[1] This work is supported by the US National Science Foundation and the US Government's Reflex Program.

Hajič, J. and Nivre, J. (eds.): Proceedings of the TLT 2006, pp. 103–114.

Several aspects of RTBs may be of interest to the treebanking community. First, RTBs are created using a tool suite that is theory independent in the sense that it allows the creation of non-re-entrant feature structures with any kind of content. The theory-independence of the tool suite means that it can be used broadly in the research community to create feature structures that represent surface syntax, deep semantics, or domain-specific semantics. However, it is important to note that although the tools themselves are theory independent, they are used to construct theory-dependent representations. For example, we have chosen to base our RTBs on a theory of grammatical meanings, similar in approach to the Tectogrammatical layer of the Prague treebanks ([5]). The tool suite is described in Section 3. Our partially developed theory of grammatical meanings is described in Section 2.

The second aspect of our work that may be of interest to the treebanking community is that we have experimented with elicitation of several languages. Using Spanish as an elicitation language, we have elicited corpora of Mapudungun (Chile), Aymara (Bolivia), and Quechua (Peru). Using English as an elicitation language, we have elicited corpora in Hindi, Hebrew, and Dutch. Our current work includes Portuguese as an elicitation language for a Brazilian language and English as an Elicitation language for Inupiaq (Alaska). RTB's are also being translated into Thai, Bengali, and other languages at another site. Section 5 relates our experiences in collecting RTBs for various languages.

2 An RTB based on Morphological and Grammatical Meaning

Although our tool suite is not based on a linguistic theory, we have developed a particular RTB based on a theory of morphological and grammatical meaning, similar in feel to the Tectogrammatical layer of the Prague treebanks. We call this RTB the MILE (MInor Language Elicitation) Corpus. The MILE Corpus currently contains about 5000 feature structures corresponding to short sentences. The feature structures have been reverse annotated with English sentences and are currently being reverse annotated with Spanish sentences as well. The English RTB is about 20,000 words in length. Following are some sample sentences from the English RTB.

```
Mary is writing a book for John.
Who let him eat the sandwich?
Who had the machine crush the car?
They did not make the policeman run.
Mary had not blinked.
The policewoman was willing to chase the boy.
```

```
Our brothers did not destroy files.
He said that there is not a manual.
The teacher who wrote a textbook left.
The policeman chased the man who was a thief.
Mary began to work.
The man quit in November.
The man works in the afternoon.
The balloon floated over the library.
The man walked over the platform.
The man came out from among the group of boys.
The long weekly meeting ended.
The large bus to the post office broke down.
The second man laughed.
All five boys laughed.
```

The purpose of the MILE corpus is to support automatic learning of machine translation rules from small amounts of highly structured data. One of the areas targeted for learning is the moprho-syntactic system of the low resource language. For this reason, the features and values of the MILE Corpus are morphological and grammatical meanings. Each feature structure represents a set of such meanings that we want to elicit in the low resource language. However, since informants cannot be expected to understand the feature structures, the meanings are rendered as well as possible in sentences of the elicitation language which are then translated into the low resource language.

In the following corpus fragment, the elicitation language is Spanish (English is provided for readability of this paper) and the low resource language is Mapudungun (Chile). The indices in the third line of each example indicate word alignment. The informants supply word alignments graphically as shown in Figure 2. The internal representation of word alignment is shown here. (1, 2) means that the first Spanish word corresponds to the second Mapudungun word. (Word alignments do not need to be one-to-one.) From these examples, we can discover that the gender and animacy of the thing that falls do not have morphosyntactic realizations in Mapudungun. The person of the thing that falls is realized on the word that governs it (i.e., agreement). The identifiability/specificity of the thing that falls is realized by a change in word order and a change in a dependent of the thing that falls (determiner). (The observation about identifiability/specificity will turn out not to hold over all Mapudungun examples.)

1. **Sentence 1:**
    - La piedra cayó.
    - Ütrünagi ti kura.

- ((1,2) (2,3) (3,1))
- The rock fell.

2. **Sentence 2:**
   - Una piedra cayó.
   - Kiñe kura ütrünagi.
   - ((1,1) (2,2) (3,3))
   - A rock fell.
3. **Sentence 3:**
   - Tú caíste. (Tú = Juan)
   - Eymi ütrünagimi.
   - ((1,1) (2,2))
   - You fell. (Addressee is Juan.)
4. **Sentence 4:**
   - Tú caíste. (Tú = María)
   - Eymi ütrünagimi.
   - ((1,1) (2,2))
   - You fell. (Addressee is María.)

Our goal in designing the MILE feature structures is to make as few assumptions as possible about how features might be morpho-syntactically realized. For example, identifiability (a component of definiteness) has many different morpho-syntactic realizations. In English, it is often expressed with a definite determiner (*the*) which is a syntactic dependent of the identifiable noun. There are many other realizations of identifiability in other languages. Hebrew uses a prefix (*-ha*) on both the head noun and its adjectival modifiers (*ha-tapuax ha-gadol* (the-apple the-big)) and also uses an extra particle to mark definite direct objects. In Chinese, differences in identifiability are likely to show up as differences in word order.

Because of the range of typological variation, MILE feature structures remain as neutral as possible as to the surface expression of identifiability. Identifiability is included only as a semantic property of a nominal expression. There are no determiners or agreement markers in MILE feature structures. Our job, after the corpus has been translated into a low resource language is to discover which of the many possible mechanisms were used to indicate identifiability. We do this by comparing minimal pairs of feature structures that differ only in identifiability ([6]).

Following is a MILE feature structure for a sentence like *Wasn't Mary chasing John?* Features related to causation, comparison, and filler-gap constructions have been omitted from this feature structure. Properties of this feature structure are explained in the subsections that follow.

```
((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)
   (np-person person-third)(np-number num-sg)
   (np-pronoun-exclusivity inclusivity-n/a)
   (np-distance distance-neutral)))

(undergoer
  ((np-function fn-undergoer)(np-animacy anim-human)
   (np-biological-gender bio-gender-male)
   (np-general-type proper-noun-type)
   (np-identifiability identifiable)
   (np-specificity specific)
   (np-pronoun-antecedent antecedent-n/a)
   (np-person person-third)
   (np-number num-sg)(np-pronoun-exclusivity inclusivity-n/a)
   (np-distance distance-neutral)))

(c-v-lexical-aspect activity-accomplishment)
(c-v-grammatical-aspect gram-aspect-progressive)
(c-v-absolute-tense past)
(c-v-phase-aspect phase-aspect-neutral)
(c-general-type yn-question)(c-polarity polarity-positive)

(c-assertiveness assertiveness-neutral)
(c-adjunct-clause-type adjunct-clause-type-n/a)
(c-secondary-type secondary-neutral)
(c-event-modality event-modality-none)
(c-function fn-main-clause)
(c-minor-type minor-n/a)
(c-copula-type copula-n/a)
(c-solidarity solidarity-neutral)
```

```
(c-power-relationship power-peer))
```

2.1 Argument Roles

In the MILE corpus, arguments are identified by semantic role names or by the macro-role names actor and undergoer [4], the actor being the more agent-like argument and the undergoer being more patient-like. Actor and Undergoer are similar in spirit to Dowty's proto-agent and proto-patient ([3]), and are also analogous to arg-0 and arg-1 of PropBank ([7]).

In order to remain neutral about the the voice system of the low resource language, MILE does not use grammatical relations such as subject and object. For example, Mapudungun [9] has an inverse voice system. When a third person is acting on a first or second person (*He hit me/you*), the verb acquires an inverse marker and the undergoer takes on subjecthood properties. The elicitation language, English or Spanish, may use active or passive voice in this case. MILE feature structures do not represent the voice system of the elicitation language. They represent a meaning such as third person actor acting on first person undergoer with the goal of discovering the morpho-syntactic realization of that meaning in the low resource language.

Other semantic role names used in MILE are recipient, beneficiary, cognizer, perceiver, emoter [4], predicate, and many role names used in the Lingua checklist [2] including around 80 locative relations.

2.2 Features of Nominal Arguments

Because there was a size limit on the MILE corpus, the set of nominal features and values is limited. There are three numbers (singular, plural, and dual); three persons (first, second, and third) with an option to specify exclusivity on first person plurals; two biological genders; two components of definiteness (identifiability and specificity); three reference types (common noun, proper noun, and pronoun); and a distance feature for deictics. Some features related to anaphora, quantification, and filler-gap constructions have been omitted from the feature structure above.

As mentioned above, there are no determiners in MILE feature structures. Grammatical gender systems are also not represented in MILE feature structures. Some semantic features like biological gender and animacy are included with the goal of discovering whether or not these have morpho-syntactic realizations in a grammatical gender system.

The MILE Corpus does not mark information structure (old and new information) on nominal arguments or clauses. This is because the sentences are presented to the informants with very limited discourse context. (See Section 4.) We hope in

the future to have a more elaborate elicitation system so that this important typological feature can be taken into account.

2.3 Clause Level Features

Again, because the MILE Corpus was limited in size, a small selection of clause level features was used. Clauses are labeled with three general types: declarative, open question, and yes no question. Clauses are also labeled with functions: main, complement, quoted, relative, argument (e.g., actor), or adjunct (with a semantic role like temporal). Clauses can have secondary types: impersonal, existential, and copula. Secondary clause types can occur in embedded clauses. There are also minor clause types that only occur in main clauses. These include commands (*Wash the dishes*), surprise (e.g., *Why do such a thing?*), laments (e.g., *If only I had apologized*), and a few others. The minor types are communicative functions, not syntactic forms. Their syntactic forms will undoubtedly be quite different from English in other languages.

Clauses are also marked for tense, aspect, mood, and assertedness. Tense is currently limited to present, past, future, and recent past. Aspect is divided into lexical aspect, phase aspect, and grammatical aspect with a limited number of values for each. Evidentiality is elicited using the feature `source` with several values including sensory, hearsay, inferred, assumed, etc. Clauses also have values for assertedness including asserted, presupposed, and wanted (desiderative). Modality includes permission and obligation, both internally and externally imposed. There are also special sets of features covering typological variation in causative and comparative sentences.

3 Tool Suite

Our tools for corpus creation are described elsewhere ([1]), and so are only briefly summarized here. However, we would like to emphasize that our priority is flexibility in creation of new RTBs. We do not expect the features and values of MILE to be carved in stone, and fully expect to revise and extend the feature set and also to restrict the feature set for other applications. The tools described in this section, are very simple, and are not theory-dependent. They have been used for the theory of grammatical and morphological meanings behind MILE but they could also be used with syntactic categories such as NP and VP or with domain specific semantic categories such as body-part and symptom in order to describe sentences like *I have a pain in my arm*.

The first step in creating a corpus of feature structures is to list the features

and values that they will contain. This list is called a *feature specification*. The partial feature specification shown here includes four features related to causation and their values. Feature specifications are written in XML, but are shown here in human-readable form.

- **Feature:** Causer intentionality
  Values: intentional, unintentional
- **Feature:** Causee control
  Values: in control, not in control
- **Feature:** Causee volitionality
  Values: willing, unwilling
- **Feature:** Causation type
  Values: direct, indirect

Feature specifications may also include restrictions on co-occurrences of features and values. For example, the feature exclusivity (including or excluding the addressee) is only defined for plural pronouns, usually first person; biological gender is only defined for animate nouns, and so on.

The second step in creating a corpus of feature structures is to specify which combinations of features and values are desired. For example, a corpus that is intended for the study of verb paradigms might include all values of person, number, and gender for nouns combined with all values of tense and aspect on the verb. The desired combinations are defined in a feature structure template. The template shown in Figure 1 represents 288 feature structures related to copular sentences (which may or may not be expressed with overt copulas in some languages). It specifies that the subjects (predicatee) of the feature structures should cycle through all combinations of pronouns and common nouns; first, second, and third person; singular and plural; and male and female. These should be combined with predicates expressing role (*He is a teacher*), attributes (*He is happy*), and identities (*He is the teacher*).

Templates such as the one shown in Figure 1 are automatically expanded into sets of feature structures. The remaining step in the creation of an RTB is reverse annotation. Reverse annotation is the creation of a sentence that expresses whatever is meant to be expressed by a feature structure. For example, the sentence *He is a teacher* expresses a feature structure with a third person singular masculine predicatee, a predicate that expresses the role of the predicatee, present time, and durative aspect.

```
   [(predicate ((np-general-type common-noun-type)
                (np-person person-third)
                (c-copula-type identity)))]}
(c-secondary-type secondary-copula) (c-polarity #all)
(c-general-type declarative)
(c-speech-act sp-act-state)
(c-v-grammatical-aspect gram-aspect-neutral)
(c-v-lexical-aspect state)
(c-v-absolute-tense past present future)
(c-v-phase-aspect durative))
```

Figure 1: Multiplication for Copula Sentences

4 Reverse Annotation

Reverse annotation is not an easy task, especially with the approach to grammatical and morphological meanings that we have chosen. The annotators have to be aware of the meanings contained in the feature structure and how to express those meanings in a natural way in the elicitation language. We have done a small amount of checking of inter-annotator agreement to make sure that the reverse annotators have the same understanding of the meanings of the feature structures.

Some difficulties in reverse annotation stem from cross-linguistic differences in grammaticalized and periphrastic expressions of meanings. For example, evidentiality is grammaticalized as a set of verbal inflections in Quechua, but is doesn't have such a fixed method of expression in English, resulting in various paraphrases like *I heard that . . .* , *Apparently, . . .* , *They say that . . .* .

One difficulty arises in eliciting things that are not grammaticalized in the elicitation language. For example, consider eliciting Quechua evidentials using English as the elicitation language. There are a few options. One is to present sentences like *They say that he stole the tapes* and hope that the Quechua speaker does not translate the words *They say that . . .* but rather uses an evidential morpheme. Another

option is to present the sentence *He stole the tapes* with instructions to translate this as if it is hearsay. A third option is to present an extensive discourse context that favors the use of an evidential. This option is currently beyond our capability. We have not yet determined which of the first two elicitation methods is the most effective.

Perhaps the most difficult issue in reverse annotation is that the annotators must remember that the morphological and grammatical meanings in the feature structures are not in one-to-one correspondence with English words and constructions. For example, identifiability is not isomorphic with the use of *the* in English. Also, it is necessary to choose from multiple possible expressions of each meaning. Impersonal sentences, for example can be expressed with *one* (*One doesn't chew tobacco at an elite country club*), or with an impersonal use of *you* or *they* or with an agentless passive. The choice depends on which is more likely to be understood and translated as an impersonal by the informant. Also, because we are learning translation rules from the MILE corpus, we really want the most common and natural English sentence, which might not be the most unambiguously translatable sentence. We hope that in our future work, we will have a larger search space of elicitation sentences that will be navigated automatically (see below) including several possible variants of each sentence.

5 Experiences with various languages

The MILE corpus has only recently been completed and is currently being translated into Thai and Bengali. Several other languages are scheduled for the next few years. We do, however, have extensive experience with a pilot elicitation corpus of about 850 short sentences (around 5 words each). The pilot elicitation corpus has been translated into Mapudungun, Quechua, Aymara, Hebrew, Dutch, and Hindi. In general, after a short period of instruction most informants translate and align words consistently. In one small experiment, three Japanese speakers translated a corpus of around 40 copula sentences (e.g., *He is a teacher*). The three informants translated most, but not all, of the sentences identically. Using the elicitation tool (Figure 2), informants can provide multiple translations for each sentence of the elicitation language.

6 Summary and Future Work

We have presented a framework for creating reverse treebanks (RTBs) and a specific application of the framework in the MILE Corpus. The RTB framework itself

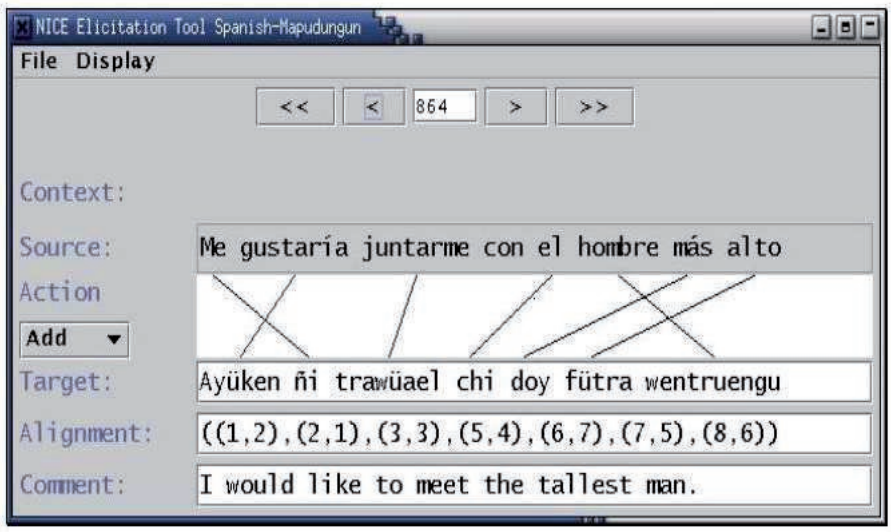

Figure 2: The Elicitation Tool

is not theory dependent, but the MILE Corpus is based on a theory of morphological and grammatical meanings and is specifically crafted for machine learning of translation rules from small, highly structured data sets.

Currently we are working on a process of Feature Detection from parallel RTBs. The purpose of feature detection is to determine which morphological and grammatical meanings correspond to morpho-syntactic markings or changes in word order in the low resource language. Feature Detection feeds into a Navigation module. The Navigation module allows us to start with a much larger search space of feature structures and present to the informant only the ones that are relevant to his/her language. Relevance is determined by the discoveries of the Feature Detector and knowledge of language typology. For example, if the Feature Detector has failed to find any morpho-syntactic markings expressing plural number, it will advise the Navigator to stop investigating the effects of plural number in various paradigms. It will also advise the Navigator not to investigate dual number because this is typologically unlikely in the absence of plural number.

References

[1] Alison Alvarez, Lori Levin, Robert Frederking, Simon Fung, Donna Gates, and Jeff Good. The mile corpus for less commonly taught languages. In *Pro-*

ceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 5–8, New York City, USA, June 2006. Association for Computational Linguistics.

[2] Bernard Comrie and Norval Smith. Lingua descriptive studies: Questionnaire. *Lingua*, 42:1–72, 1977.

[3] David Dowty. Thematic proto-roles and argument selection. *Language*, 67:547–619, 1991.

[4] William A. Foley and Robert D. Jr. Van Valin. *Functional Syntax and Universal Grammar*. Cambridge: Cambridge University, 1984.

[5] Eva Hajičová. Dependency-based underlying-structure tagging of a very large czech corpus. *T.A.L.*, 41(1):47–66, 2000.

[6] Lori Levin, Alison Alvarez, Jeff Good, and Robert Frederking. Automatic learning of grammatical encoding. In *Jane Grimshaw, Joan Maling, Chris Manning, Joan Simpson and Annie Zaenen (eds) Architectures, Rules and Preferences: A Festschrift for Joan Bresnan*, CSLI Publiations, in press.

[7] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105, 2005.

[8] Katharina Probst. *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. Unpublished Ph.D. Thesis, Carnegie-Mellon University, School of Computer Science, 2005.

[9] Fernando Zúñiga. *Mapudungun*. Lincom Europa, 2000.