Automatic Learning of Grammatical Encoding

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

November 1, 2005

**CENTER FOR THE STUDY
OF LANGUAGE
AND INFORMATION**

1.1 Introduction

Avenue (Probst et al., 2002, Monson et al., 2004, Lavie et al., 2003, Font-Llitjos et al., 2005)[1] is a machine translation system that automatically learns translation rules between two languages. In the Avenue scenario, one of the languages is a *resource rich language* like English or Spanish, for which there are many human and electronic resources (corpora, morphological analyzers, lexica, etc.). The other is a *resource poor language* with few human and electronic resources. For example, there might be no linguist available to write translation rules and there might not be large enough corpora for automatic machine learning of translation rules. This is true for the vast majority of languages.

Within the current state of the art in commercial machine translation, it is not possible to build machine translation (MT) systems for resource poor languages. However, we have met with many indigenous communities (Mapuche, Quechua, and others), who want their languages to be used in jobs, education, government, and health care. Machine translation can be a tool for maintaining functionality in their languages, because it can help them access the content of the Internet and disseminate local culture and information without having to adopt a major national language like Spanish or English.[2] The vision of the Avenue project is equal access to information for speakers of all languages.

The steps in building an Avenue system are collection of resources, automated learning of rules and morphology (Probst, 2005, Monson et al., 2004) from the collected resources, and translation correction with interactive rule refinement (Font-Llitjos et al., 2005). The Avenue rule formalism and run-time translation system are described in (Lavie et al., 2003).

1.2 Feature Detection: automatic learning of grammatical encodings

Although Avenue has many components, this paper focuses only on one aspect of resource collection, the automatic elicitation of data from bilingual speakers without the help of a human fieldworker. More specifically, this paper is about a process called *Feature Detection*, which learns the morpho-syntactic mechanisms that are used for encoding grammatical relations (Bresnan, 2001) and some aspects of meaning such as tense and evidentiality. Feature Detection guides a process of

---

[1] NSF ITR/PE 0121631

[2] Some people are opposed to interference with their language: http://www.clarin.com/diario/2005/08/05/um/m-1027754.htm.

Navigation through a typological search space by distinguishing between features that have morph-syntactic consequences, and therefore must be pursued with further questioning, and features that do not have morpho-syntactic consequences and can therefore be dropped from further questioning.

Although we have used automatic Elicitation and Feature Detection in the development of MT systems, we hope that some of the tools that we have created will be useful in other areas of linguistics. The set of tools includes the Elicitation Tool (Section 1.3), a graphical interface for translating sentences and aligning words and morphemes in two languages, which is simple enough to be used by non-linguists. The elicitation questionnaire is represented internally as a set of feature structures, not as a set of sentences. Elicitation sentences are made by generating English or Spanish sentences that correspond to each feature structure. The same set of feature structures can be used for making corpora in other resource rich languages. Furthermore, the corpus is not fixed. The Feature Specification Schema (Section 1.6) is an XML schema for defining features and values for new corpora. The Multiplier (Section 1.7) allows a linguist to specify combinations of features and values that are of interest. Using these tools, it is possible to create new corpora easily, in any resource rich language (assuming the existence of a feature structure-to-sentence generator), for specialized tasks in fieldwork or machine translation.

The following set of examples shows typical input to the Feature Detection program. In each example, the first line is a Spanish sentence that was presented to a bilingual native speaker of Mapudungun (Chile). The second line shows a translation into Mapudungun provided by the native speaker and the third line shows word alignments provided by the native speaker. The word alignments are provided graphically (see Section 1.3), but are shown here in their internal representation as indices. The index (3,1) means that the third Spanish word is aligned with the first Mapudungun word. Discontinuous, zero-to-one, many-to-many, and many-to-one alignments are allowed, although not all are illustrated here. The fourth line, in English, is for reference, and may also show some contextual information for the native speaker to take into account, such as whether the hearer is male or female. The contextual information was also presented to the native speaker in Spanish.

(1) a. La piedra cayó.
    b. Ütrünagi ti kura.
    c. ((1,2) (2,3) (3,1))
    d. The rock fell.

(2) a. Una piedra cayó.
b. Kiñe kura ütrünagi.
c. ((1,1) (2,2) (3,3))
d. A rock fell.

(3) a. Caí.
b. Iñche ütrünagün.
c. ((1,2))
d. I fell.

(4) a. Tú caíste.
b. Eymi ütrünagimi.
c. ((1,1) (2,2))
d. You fell. (Hearer is Juan.)

(5) a. Tú caíste.
b. Eymi ütrünagimi.
c. ((1,1) (2,2))
d. You fell. (Hearer is María.)

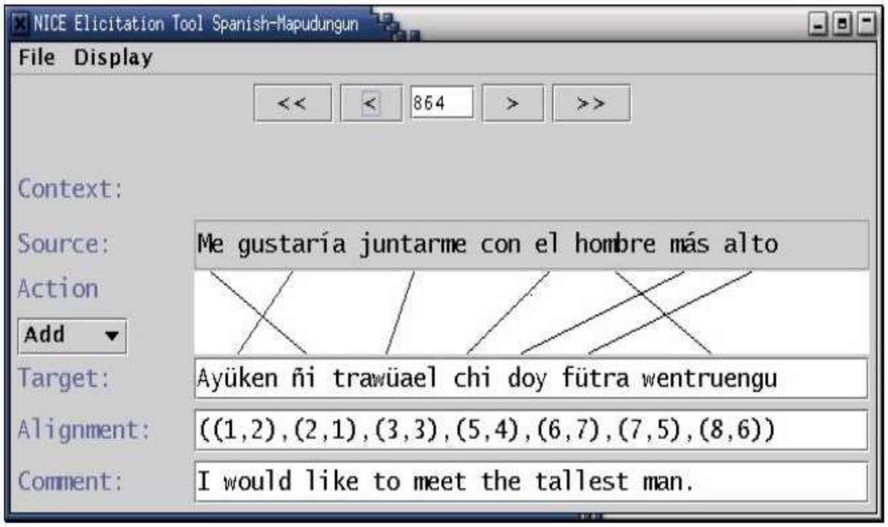
From these examples, Feature Detection would discover that the gender and animacy of the undergoer do not have morphosyntactic realizations. The person of the undergoer is realized on the word that governs it (i.e., agreement). The identifiability/specificity of the undergoer is realized by a change in word order and a change in a dependent of the undergoer (determiner). (The observation about identifiability/specificity will turn out not to hold over all Mapudungun examples.)

In carrying out our research program on Feature Detection, we face many of the difficulties faced by field linguistics. The informants might not be consistent in their translations and alignments; we may need to take into account more than one translation of each sentence; it is difficult to elicit phenomena that do not occur in the resource rich language; and the wording of sentences in the resource rich language may influence the translations in the resource poor language, so that some phenomena may be missed.

Most difficult, however, is the size of the search space of features and values. There are millions of possible combinations of feature values (four values for person, multiplied with at least three values for number, multiplied with several temporal and aspectual values, etc.) Informants cannot be expected to translate millions of sentences, and they become bored quickly when faced with irrelevant distinctions. For example, they will not be tolerant of translating masculine and feminine versions of every NP in every sentence if their language does not have morphosyntactic marking for gender. The Avenue system therefore needs to Navigate the search space in a reasonably intelligent way, as human fieldworkers do. Feature Detection (Section 1.9) tracks what has been discovered so far in the process of elicitation. We also propose to build a Navigator (Section 1.11) that makes decisions about what lines of questioning are most likely to be valuable given what has been found so far.

1.3 Elicitation

In the process of elicitation, sentences are presented to informants via the Elicitation Tool (Probst et al., 2001). The informant sees one example at a time (a sentence for elicitation and possibly a disambiguating context for the sentence) and translates it into his/her language. All Unicode fonts are supported, as are right-to-left scripts. The informant uses the mouse to align the words of his/her sentence to the words of the elicitation sentence. Alignments do not have to be one-to-one and may be discontinuous, and some words may remain unaligned. In the elicitation tool, the alignments are shown graphically as lines connecting words and as a list of pairs of numbers such as (7,5) (the seventh Spanish word aligned to fifth Mapudungun word).



864

Context:

Source: Me gustaría juntarme con el hombre más alto

Action: Add

Target: Ayüken ñi trawüael chi doy fütra wentruengu

Alignment: ((1,2),(2,1),(3,3),(5,4),(6,7),(7,5),(8,6))

Comment: I would like to meet the tallest man.

FIGURE 1 The Elicitation Tool

We have used the Elicitation Tool with various small elicitation corpora, between 850 and 2000 phrases and sentences, in several languages, including Hebrew, Mapudungun, Quechua, and Aymara. For some of these languages, several speakers have translated the corpus. We have also used the Elicitation Tool rather extensively for translating around 15,000 sentences and phrases into Hindi (Lavie et al., 2003). Learning to use the Elicitation Tool typically takes only about 20 minutes of training, including the use of the tool itself and the use of keyboard input for non-roman character sets. Informants must be instructed to translate similar sentences similarly. For example, their translations for *He is falling*, *It is falling*, and *She is falling* should be as similar as possible. Most informants can follow this requirement. Some informants like to provide multiple translations for sentences, which is allowed by the Elicitation Tool.

1.4 Underlying ssumptions and definition of the problem

In this section we will explain the basis for two assumptions that underlie our approach: that it should be possible to obtain data from a non-linguist, and that the search for morpho-syntactic features should be based on meaning and communicative function.

Eliciting data from non-linguists: Since we cannot count on the availability of a linguist who speaks the resource-poor language we are eliciting, the elicitation process must be simple enough to be carried out by a non-linguist using the Elicitation Tool. Therefore, we do not ask explicit questions about grammar and morphology (in contrast to (McShane et al., 2002)); we only ask the informants to translate sentences and align words. We do, however, assume the availability of people who are bilingual and literate in both the resource rich and resource poor language. This may require choosing an orthographic system if the resource poor language does not have a standardized orthography.

We have found some variation in the quality and consistency of translations and alignments from non-linguists. The Hindi speakers we worked with, who were students in college or graduate school, were initially inconsistent in the alignment of closed class items (postpositions and auxiliary verbs), but were able to align consistently with a small amount of supervision (from a linguist who did not know Hindi). There is occasionally a problem with an informant missing the point of a sentence, but we hope to reduce the frequency of this problem by providing richer context for each elicitation sentence.

Morpho-syntactic vs Functional Approach to Elicitation: Although we are trying to discover morpho-syntactic phenomena in the resource poor language, we have organized the search by meaning and communicative function. For example, instead of looking for case markers, we look for noun phrases in different semantic roles and compare their forms. Our search for morpho-syntax is organized functionally because we cannot make assumptions about which morpho-syntactic mechanisms might be used to express a given function.

Definiteness provides a good example of the problems we are trying to avoid by adopting a functional approach in the system. Suppose we want to know how English definite NPs are expressed in another language. We first have to observe that NPs that are marked as definite in English have a variety of functions. Sometimes, determiners mark definiteness as in (6a), but sometimes they do not mark definiteness. In (6b) the definite article marks reference to a species and in (6c) the indefinite article marks a profession in a predicate nominal. Since another language may use different mechanisms for each of these three functions of English determiners, we will only get a clear picture by separating them. What we really want to know is how the semantic components of definiteness are expressed, how references to a species are made, and how predicate nominals are expressed. The portion of the elicitation corpus that covers definiteness will not include uses of *the* that are not related to definiteness, and *the* may even need to be used to mark an NP in parts of the corpus where that NP is intended to be interpreted as indefinite. This distinction between form and function is very important in our approach.

(6) a. I saw the lion.
b. The lion is a magnificent beast.
c. He is a soldier
(Compare to French *Il est soldat* (Croft, 2003).)

The mapping between form and meaning is many-to-many. One form, such as a determiner, may have many functions and one function may have many morpho-syntactic realizations. In eliciting the semantic components of definiteness (identifiability, specifiability, uniqueness, familiarity, etc.), we find changes in word order for Chinese definite direct objects, changes in marking for Hebrew definite direct objects, changes in verb agreement for Hungarian definite direct objects, and uses of existential constructions for indefinite subjects in many languages. Since translations are only good when they preserve meaning, we have to be concerned with translating meanings regardless of what forms they take. We have to find definiteness in all of its forms, and cannot focus only on determiners.

Our challenge in taking a function-based approach is to find ways of describing linguistic phenomena that are independent of morpho-syntax. Comrie et al. ((Comrie et al., 1993)) advocate a similar split between function and morpho-syntax for the production of descriptive grammars, and provide some useful insights into the possible inventory of syntax-independent functional categories.

1.5 The Process of creating an Elicitation Corpus

AVENUE is intended, in principle, to be applicable to any language—including resource poor languages which may be typologically quite divergent from the resource rich languages which serve as elicitation prompts. The elicitation corpus is a search space of all communicative functions that might have morpho-syntactic realizations in some language. Since this search space is large, and we may want to re-formulate it as we find out more about typological diversity and universals, we have designed a compact and easily modifiable way of characterizing it.

In our approach, the elicitation corpus is not represented as a set of sentences, but as a set of feature structures. This is significant because a set of elicitation sentences can be generated from the feature structures (Section 1.8) in any language, for example, English or Spanish. Maintaining consistency between the English and Spanish elicitation sentences is simply a matter of (re)generating them from the same set of feature structures. A sample of a feature structure is shown in Figure 2. Our feature structures are represented using parentheses as delimiters. Each feature-value pair is enclosed in parentheses, and each (sub)-feature-structure (list of feature-value pairs) is enclosed in parentheses.

The set of feature structures in the Elicitation Corpus is generated automatically from a compact description that is created by a linguist. There are two steps in designing the Elicitation Corpus, Feature Specification and Multiplication. Feature Specification is the process of defining the features and values that will characterize elicitation sentences. A fragment of a Feature Specification is shown in (7) in Section 1.7. Multiplication is the process of describing combinations of features and values to include in the corpus. We have formulated a control language for stating Multiplications, an example of which is shown in Figure 4.

Our approach to corpus creation allows for the corpus to be changed easily for different types of projects by changing the Feature Specification and/or Multiplications. If we decide to make a change in the

```
((actor ((np-general-type pronoun)
         (np-person person-first)
         (np-number num-sg)
         (np-biological-gender gender-male)
         (np-animacy anim-human)))
(predicate ((np-general-type common)
            (np-person person-first)
            (np-animacy anim-human)
            (np-identifiability non-identifiable)))
(c-copula-type role)
(c-secondary-type secondary-copula)
(c-v-lexical-aspect state)
(c-v-absolute-tense past)
(c-v-phase-aspect durative)
(c-imperative-degree imp-degree-n/a)
(c-ynq-type ynq-n/a))
```

FIGURE 2 An AVENUE Feature Structure

system of number or tense, it will be changed consistently and automatically in hundreds of related sentences. The system is also flexible enough to be used with any kinds of features and values. For example, even though we have chosen to organize our Feature Specification by communicative function, we could just as easily have created a Feature Specification organized by morpho-syntactic features.

1.6 The form of the Feature Specification and Multiplications

Feature Specifications are encoded using an XML-based markup system. Multiplications are partially marked up using XML but also make use of a special syntax optimized for expressing the types of feature and value combinations which need to be Multiplied (Alvarez et al., 2005).

While XML-encoding of Feature Specifications is not crucial for the project at present, it is hoped that this will allow them to more easily interoperate with other relevant XML-based linguistic resources currently under development elsewhere. In particular, the Navigation system of AVENUE should be able to benefit from knowledge encoded in linguistic ontologies, like the General Ontology for Linguistic Description (Farrar and Langendoen, 2003), as they begin to encode more and more kinds of linguistic knowledge. Our XML version of (7) is shown in Figure 3.

```
<feature>
    <feature-name>c-causer-intentionality</feature-name>
    <value><value-name>intentional</value-name></value>
    <value><value-name>accidental</value-name></value>
</feature>
<feature>
    <feature-name>c-causee-control</feature-name>
    <value><value-name>in-control</value-name></value>
    <value><value-name>not-in-control</value-name></value>
</feature>
<feature>
    <feature-name>c-causee-volitionality</feature-name>
    <value><value-name>willing</value-name></value>
    <value><value-name>unwilling</value-name></value>
</feature>
<feature>
    <feature-name>c-causation-directness</feature-name>
    <value><value-name>direct</value-name></value>
    <value><value-name>indirect</value-name></value>
</feature>
```

FIGURE 3 XML Version of a Feature Specification

1.7 The Content of the Feature Specification

In this section, we discuss how we determine what features and values need to be covered in the specifications and how we encode them. Achieving broad coverage of the elicitation search space requires relying on available typological models of grammatical variation for essentially any functional feature known to have a coherent morphosyntactic realization in some language. While some functional features—e.g., features for verbal arguments—have well-established models within LFG, others are lacking such models and have required us to adapt models from the descriptive and typological literature and, in a few cases, from other formal frameworks which, for one reason or another, have worked out useful models for a relevant functional domain.

Identifying a Typological Model: As an initial strategy in ensuring that AVENUE has reasonable coverage, we have made use of the Comrie and Smith questionnaire (Comrie and Smith, 1977) for linguistic description. While dated in some respects, this questionnaire is one of the few published works which attempts to establish a detailed series of questions to guide a descriptive linguist in writing a complete grammar. Like AVENUE, Comrie and Smith is intended to be of use for any language and, therefore, has served as a useful foundation in building the feature set for the system.

The biggest divergence between work like Comrie and Smith and AVENUE results from AVENUE's reliance on basing the construction of its elicitation corpus on functional features instead of morphosyntactic features. Questionnaires like Comrie and Smith typically contain questions of two distinct types: those relating to functional distinctions which are known to be grammatically marked and those relating to *how* those distinctions are typically marked (as noted by Comrie et al., (Comrie et al., 1993)).

For example, in the domain of questions, Comrie and Smith (page 11) include an enumeration of functionally-oriented question types like *neutral yes-no*, *yes-no expected affirmative*, *yes-no expected negative*, *question-word questions*, etc., and an enumeration of how these types tend to be grammatically marked, for example, with word order changes, particles, tags, intonational patterns, etc. From the perspective of AVENUE, the former type of information is potentially quite valuable in constructing the elicitation corpus, but the latter type is not directly relevant—it is the machine's task to discover how a language marks a given functional feature grammatically.

Not surprisingly, no existing grammatical questionnaire exhaustively lists all of the grammatical features and values needed for AVENUE. Therefore, in addition to grammatical questionnaires, we make use of a wide range of other sources in determining what features need to be covered by the elicitation corpus. These include: typologically-oriented monographs covering important grammatical domains, for example, (Comrie, 1985) on tense, (Comrie, 1976) on aspect, (Palmer, 2001) on mood, and (Aikhenvald, 2004) on evidentiality; formally-oriented work with useful discussion of a relevant domain, for example, (McCawley, 1998, 692–740) on comparatives and (Foley and Van Valin, 1984, 238–320) on clause combing; and descriptions of particular languages which, for one reason or another, serve as useful guides for what distinctions need to be made during elicitation (e.g., (Haspelmath, 1993), (Good, 2003)).

Representing the typological model as a Feature Specification: Once we have established a general typological model for a given grammatical domain, we need to encode that model as a set of features and accompanying values which can be used in the construction of the elicitation corpus. This, of course, requires a fair amount of analysis, and it is not always clear that we have hit upon the ideal analysis in a given part of the Feature Specifications. They are, therefore, always subject to possible revision. For a project such as this one, the guiding principal is ensuring that functional categories that are known to behave independently of one another with respect to morphosyntactic marking are treated as values for different features so that these values can be independently varied within sets of sentences in the elicitation corpus.

To take a concrete example, consider the features and values used to model the domain of causation given in (7). The selection of these features and values was loosely based on the typology found in (Dixon, 2000).

(7) a. **Feature:** Causer intentionality
Values: intentional, unintentional

b. **Feature:** Causee control
Values: in control, not in control

c. **Feature:** Causee volitionality
Values: willing, unwilling

d. **Feature:** Causation type
Values: direct, indirect

The features and values in (7) represent a number of known types of semantic distinctions made in causative constructions. Some of these distinctions are logically dependent on each other—for example, if a causative construction specifically marks that the causer intentionally caused an action to take place then it cannot also mark that the causer acted unintentionally.[3] Other distinctions are logically independent of each other. For example, whether or not a causee performed an action volitionally is independent of whether or not the causation was direct or indirect. In some cases, it might not be completely clear whether or not two values are independent. For example, a causee who is not in control in general would not be expected to willingly do the action. However, since a causee who is in control can act willingly or unwillingly, control and volitionality were separated out as two independent features, as indicated in (7). The specifications have generally been designed to err toward allowing too many independent value combinations, rather than too few.

The Feature Specifications are intended to represent the full typological range of distinctions that can be grammatically encoded. It will therefore contain sentences that are intended to elicit causer intentionality, causee control, causee volitionality, and causation type. This does not mean that we expect every language to mark each of these distinc-

---

[3] This is not to say that a given construction could not simply be unmarked for whether or not the causer intentionally causes an action—obviously, this can also be the case.

tions. It only means that we need to check which of these distinctions are morpho-syntactically marked in each language.

The Feature Specification system also allows statements to be made about Restrictions holding among certain features. These are useful when a given feature is only relevant for a feature structure specified for a limited range of feature-value pairs. For example, specifications of values relating to the feature of pronoun inclusivity are restricted to feature structures for pronouns which are first person and plural.

The current set of Feature Specifications is not complete. However, it does have fairly wide-ranging coverage, including, for example, features relating to clause types, discourse settings, agentivity, argument roles, tense/mood/aspect, and adverbial roles—among others.

Multiplications: Choosing combinations of features and values: problem raised by decomposing functional information into features and values is the fact that logically independent functions may not be treated independently in the morphosyntactic system of a given language. In the case of causation, for example, causative marking may interact with verbal argument structure. In principle, such interactions could be detected by constructing an elicitation corpus where all possible combinations of features and values were represented by some sentence to be translated. In practice, however, this would be a very inefficient way to deal with feature interactions because the attested types of such interactions are quite limited—causative marking is known to interact with argument structure (Dixon, 2000, 43), but it would be unlikely for it to interact with, say, comparative constructions.

AVENUE deals with this problem by explicitly encoding which features and values should be Multiplied together when the feature structures of an elicitation corpus are created. One such Multiplication is given in (8). We have designed a formalism for specifying Multiplications, and have built a GUI to support their formulation. The GUI allows a linguist to browse the Feature Specification and choose feature and value names to include in the Multiplication (Alvarez et al., 2005).

(8) (lexical-aspect #all) ×
(grammatical-aspect #all) ×
(absolute-tense past, present, future)

The Multiplication schematized in (8) encodes a statement that, when feature structures for the elicitation corpus are generated, all of the features for lexical aspect (*state*, *activity*, *accomplishment*, etc.) should be cross-Multiplied with all of the features for grammatical aspect (*perfective*, *imperfective*, etc.) and three features for absolute tense, *past*, *present*, and *future*. This cross-Multiplication is designed to ensure that, if there are critical interactions between tense and aspect in a given language, the corpus will contain sentences in which those interactions can be detected.

Of course, a Multiplication like the one in (8) only specifies a small part of what a feature structure for an entire sentence might look like. Features that are not mentioned in the Multiplication will take a default value. For example, the default value for the polarity feature is *positive*—that is, affirmative, instead of negative, sentences are specified as the default type in the Feature Specification.

While Feature Specifications are intended to, ultimately, allow the generation of an elicitation corpus which can uncover most known kinds of morphosyntactically-marked functional features, it is not expected that a "complete" corpus will be generated at any one particular time. Rather, in most cases, a subset of features will be chosen as the basis a "subcorpus" optimized for the discovery of morphosyntactic marking of those features alone. For example, we have been experimenting with a "copula" corpus, which focuses only on tense, number, gender, person, and predication type (identity, role, attribute) in sentences that are expressed with copulas in English.

1.8 Making elicitation sentences from feature structures

A Multiplication is expanded into a set of feature structures. For example, the Multiplication in Figure 4 represents 288 feature structures. Each feature structure represents a set of communicative functions or meanings that we want to elicit. However, because informants cannot translate feature structures, we need to represent the feature structures in a form that the informants can understand — sentences in the language of elicitation (e.g., English).

There are two difficulties in representing the meaning of a feature structure in English (or any other language of elicitation). First, some aspects of meaning are not marked in English. So, for example, *I (masc.) fell* and *I (fem.) fell* are the same in English. Second, the feature structures in the elicitation corpus do not contain lexical items. English Lexical items therefore must be added in order to make English sentences.

The GenKit generation system (Tomita and Nyberg, 1988) is used for making sentences from feature structures. So far we have only used GenKit to make English elicitation sentences, but the GenKit unification-based formalism is applicable to any language. For the purpose of AVENUE, we do not use a large, comprehensive English gen-

```
((predicatee
  ((np-general-type pronoun-type common-noun-type)
   (np-person person-first person-second person-third)
   (np-number num-sg num-pl)
   (np-biological-gender bio-gender-male bio-gender-female)))
  {[(predicate ((np-general-type common-noun-type)
                (np-person person-third)))
                (c-copula-type role)]
   [(predicate ((adj-general-type quality-type)
                (c-copula-type attributive)))]
   [(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 4 Multiplication for Copula Sentences

eration grammar, but rather several small grammars for generating sub-corpora. For example, the 288 sentences specified in Figure 4 are generated by a small English copula grammar, which covers predication of attributes (*He is happy*), identity (*He is the teacher*), and role (*He is a teacher*). (These sentences may or may not be expressed with overt copulas in other languages.) With the small grammars, we can include small lexicons with semantic features designed to ensure that selectional restrictions are met in the sentences that we generate. The grammars generate comments for features that are not expressed in English, producing strings like *I* ONE WOMAN *fell*. A post-processor prepares the English sentences for presentation in the Elicitation Tool by separating the English sentence *I fell* from the comment ONE WOMAN, which the informant then sees in different fields of the Elicitation Tool.

1.9 Initial experiment with Feature Detection

In this section we will describe a preliminary exercise with Feature Detection. For this experiment we used 48 sentences from the copula sub-corpus, which were translated into Hebrew and Japanese. The sentences included two genders for the subject (masculine and feminine), three persons for the subject (first, second, and third), three tenses (present, past, and future), and two types of predication (role and identity). Identity was only Multplied with third person subjects.

The first step in Feature Detection is identifying minimal pairs of feature structures that differ in only one value of one feature. There will be many minimal pairs for each feature value. For example, positive vs negative polarity will form a minimal pair in each of three tenses.

Each feature structure is associated with sentence from resource rich language (from the Elicitation Corpus) and a sentence in the resource-poor language (provided by the informant). Furthermore, in the process of generating the major language sentence, we produce an approximation of the LFG ϕ-inverse mapping from feature structures to constituents. In this way, we know which word of the major language is the head of each feature structure and sub-feature structure. Since we also have a word alignment, provided by the informant, from the minor language to the major language, we indirectly have an approximation of the ϕ mapping from c-structure to feature structure for the minor language.

For each minimal pair of feature structures, we want to know whether the corresponding sentences in the elicited language are the same or different. If they are the same, we will conclude that the changing the feature value has no morpho-syntactic effect in the context of that particular feature structure.

If there is a difference in the elicited sentences corresponding to a minimal pair of feature structures, we conclude that changing the feature value has a morpho-syntactic effect. In this case, we also want to know which part of the sentence is affected. For the purpose of this experiment, morpho-syntactic effects are categorized as

- **Substitutions:** different word, including different inflection of the same stem
- **Insertions/Deletions:** one member of the minimal pair contains a word that is not contained in the other.
- **Change in alignment indices:** change in word order

The locations of the morpho-syntactic effects are categorized as follows:

- **Me:** The change appears on a word that is the head of the smallest feature structure that contains the minimally different feature-value pair.
- **My Dependent:** The change appears on any dependent of Me.
- **My Governor:** The change appears on the word that is the head of the feature structure that contains Me as an argument.
- **Other:** Anything that is not one of the above, for example another dependent of My Governor, such as a secondary predicate.

Figure 5 shows two examples of elicited data for Japanese. The first minimal pair of feature structures differs in the value of tense (past or present). The sentences that correspond to those feature structures differ in one word: *desu* vs. *desita*. These Japanese words are aligned to the English words *is* and *was*. *Is* and *was* are the heads of the smallest feature structures containing the feature value pairs `(tense past)` and `(tense present)`. Therefore, the location of the change is on Me. In other words, the difference in tense is represented on the verb that is the head of the sentence.

The second Japanese example illustrates an interesting point. Japanese is usually classified as having two tenses or aspects, past (or complete) and non-past (or non-complete). However, in this exercise, the informants translated the English future tense with *naru* (become) and a change of mood marked on *desyou*. Feature Detection identifies an extra word aligned to the co-heads (*will be*) of the English sentence, and concludes that future tense in Japanese differs from present tense in an additional word being added. This may be contradicted in other parts of the corpus, for example, if *yomu* appears as both the present and future forms of *read*. We have not yet implemented the final phase of Feature Detection, which will reconcile conflicting findings by identifying the distinguishing circumstances. In this case it may (or may not) turn out that the morpho-syntactic consequences of changing tense are different in copular and non-copular sentences.

Figure 6 shows two examples from the Hebrew data. In the first example, a minimal pair of present and past tense, Feature Detection finds an extra word in the past tense, reflecting the fact that there is not an overt copula in the present tense. The second example shows variation based on the biological gender (not grammatical gender) of the subject. Two differences are found, a change on the subject pronoun itself and a change on governor of the subject. The third change, on the predicate nominal (*more* (masc.) versus *mora* (fem.)), was not detected because Feature Detection was actually run on a vowelless script different from the phonemic transcription shown here.

1.10 Remaining issues in Feature Detection

Although we have achieved some basic functionality in Feature Detection, several issues remain to be addressed.

Discovering Case and Voice: Automatic discovery of case and voice systems requires an additional definition of minimal pair. The two feature structures must have the same nominal sub-feature-structure in different roles, as in *He hit the ball*, *The ball hit him*, *I threw the ball*

```
1.
Minimal Pair:: (tense past):(tense present)
Comparing:
        He was a teacher.
        kare wa sensei desita   ((1,1 2),(2,4),(4,3))
        He is a teacher.
        kare wa sensei desu     ((1,1 2),(2,4),(4,3))
Type of difference: Substitution
Location of difference: Me

2.
Minimal Pair:: (tense present):(tense future)
Comparing:
        He is a teacher.
        kare wa sensei desu.   ((1,1 2),(2,4),(4,3))
        He will be a teacher.
        kare wa sensei ni naru desyou   ((1,1 2),(2 3,5 6),(5,3))
Type of difference: Insertion/Deletion
Location of difference: Me
```

FIGURE 5 Elicited data from Japanese

```
1.
Minimal Pair:: (tense past):(tense present)
Comparing:
        He was a teacher.
        Hu haya more.    ((1,1),(2,2),(4,3))
        He is a teacher.
        Hu more.        ((1,1),(4,2))
Type of difference: Insertion/Deletion
Location of difference: Me

2.
Minimal Pair::(biological-gender male):(biological-gender female)
Comparing:
        He will be a teacher.
        Hu yihye more.   ((1,1),(2,2),(3,2),(5,3))
        She will be a teacher.
        Hi tihye mora.   ((1,1),(2,2),(3,2),(5,3))
Type of difference: Substitution
Location of difference: Me
Type of difference: Substitution
Location of difference: My governor
```

FIGURE 6 Elicited data from Hebrew

at him.

Incomparable Target Language Sentences: As shown in 9 Hebrew predicative sentences do not need an overt copula in the present tense, but they can have a pronominal copula (Doron, 1983). In order to detect the morpho-syntactic realization of gender, we want to compare(9a) and (9b), or (9c) and (9d). Comparing (9a) and (9d) would give the misleading result that words are added or deleted to express gender. Therefore, when there are multiple translations for the same feature structure, we need to identify the most comparable sentences for Feature Detection.

(9) a. Lea mora.
Leah teacher.FEM
Leah is a teacher.

b. Avi more.
Avi teacher.MASC
Avi is a teacher

c. Lea hi mora.
Leah 3SG.FEM teacher.FEM
Leah is a teacher.

d. Avi hu more.
Avi 3SG.MASC teacher.MASC
Avi is a teacher.

Incomparable Source and Target Language Sentences: In Machine Translation, the term *translation divergence* (Dorr, 1994) refers to source and target language sentences that do not translate literally. A typical example of a divergence between English and German is *Ich esse gern* (literally: I eat gladly) and *I like to eat*. These sentences are problematic because the head of the German sentence is *esse* (eat) whereas the head of the English sentence is *like*. Because the head of the German sentence is aligned to the XCOMP of the English sentence, Feature Detection as described above will not find the correct location of morpho-syntactic change for this pair of sentences. Divergence detection is therefore a pre-requisite for Feature Detection.

Near Minimal Pairs: Since we may not be able to use the same lexical items throughout the corpus — informants have complained about repetitious vocabulary — we might need to compare sentences such as *He is a teacher* and *She is a firefighter* whose feature structures are a minimal pair except for the lexical items. We must be careful about concluding that gender is marked on the predicate nominal only because *teacher* and *firefighter* are different words. In this case, it would be helpful to combine Feature Detection with the Morphology learning component of AVENUE (Monson et al., 2004) so that we can compare the morphology of the two words instead of just comparing the two strings of characters.

1.11 Navigation

Currently, the size of the Elicitation Corpus is controlled with Multiplications, Restrictions, and Defaults. Multiplications specify exactly which combinations of features and values we are interested in, Restrictions disallow incompatible combinations of features, and Default values of features are used in place of fully multiplying out all values. Eventually we hope to replace these mechanisms with a more intelligent Navigation component.

In the Navigation component we envision, the search for features with morpho-syntactic consequences can be directed by explicitly modeling the cost/benefit tradeoff of having the informant translate and align each possible next sentence, using a marginal utility calculation. Marginal Utility (MU) is a concept originally from the field of economics/operations research, with applicability to any process of rational decision making under uncertainty. One example of the use of Marginal Utilities in decision making is Carbonell and Goldstein (Carbonell and Goldstein, 1998), where it is used in text summarization. The fundamental idea is to try to decide what to do next, if every choice has possible rewards and costs, and you have imperfect knowledge of what will actually result from your specific action. The MU of a particular choice is defined as:

MU = Expected Value - Expected Cost

The Expected Value is in turn defined as:

EV = Sum (over all outcomes f) P(f)*Value(f)

which is to say it is the sum over all possible outcomes of the value of each of the possible outcomes multiplied by the probability of that outcome. In this system, the EV of a sentence will be the sum of the EVs of all potential facts that might be discovered by Feature Detection from translating/aligning this sentence at this point (i.e., given what has already been translated/aligned), each multiplied by the probability that the result will actually lead to the deduction of a fact. If a feature structure contains values for tense, aspect, and polarity it will have an expected value based on whether or not minimally different feature structures have been observed – feature structures that differs only on the value of polarity, tense, or aspect. If such feature structures have been observed, then it will be possible to make a comparison that might yield a fact. The expected value might also depend on whether tense, aspect, and polarity have been found in previous examples to have morpho-syntactic effects.

The Expected Cost (EC) represents the expected human cost of translating/aligning the sentence, and may initially be approximated by the number of words in the English sentence plus a term for the number of sentences translated so far; later we may try to empirically determine a more sophisticated formula. For both EV and EC, one essentially gets a weighted average of all possible outcomes from this action, and the MU thus simply reflects the value of the action minus its cost.

The action with the largest MU is thus simply the one that will probably be best. Based on the MU calculation for all remaining Elicitation Corpus sentences, the sentence with the best expected payoff in the current stage of exploring the current language is explored next. The process then repeats until the MU drops to zero, at which point we believe we have learned everything that is reasonable to try to learn from this process: everything likely to be of value that can be learned for a reasonable cost.

As we have just seen, the MU of acquiring a new translation depends in part on the expected value (EV) of any new language facts that might be inferred from the new translation; this inference depends on combining the new facts with previously learned related facts about this language. The EV calculation thus requires a significant amount of linguistic and logical knowledge. This knowledge will be represented in the form of a set of decision graphs that declaratively represent the inference process for each kind of inference that might be made by the Feature Detection system. Describing the detailed design of these decision graphs is beyond the scope of this paper; but in brief, they must be constructed (just once) by computational linguists to encode some of the knowledge used by field linguists, and are used to remember the status of many different partially-completed inferences at run-time, in order to compute EV estimates. A decision graph might for example direct Feature Detection to look at plural number before looking at dual because a language that does not mark plural will also not mark dual.

Each time the informant translates and aligns a new elicitation corpus sentence, the system may learn new facts, and these facts in turn will inform the next MU calculation for all remaining Elicitation Corpus sentences. This is a very large update calculation; it is necessary to update all the affected Decision Graphs in the system. And it must be done quickly, while the informant is waiting, since it determines the next sentence that the informant will translate/align. The design for achieving this is again beyond the scope of this paper, but is related to earlier work by Forgy (Forgy, 1982).

1.12 Conclusion

We have presented a framework that uses pairs of translated sentences to discover which meanings and communicative functions are morphosyntactically marked in a language. Although the system has been developed in the context of a Machine Translation project, it has consists of tools for elicitation and corpus creation that may be useful in other areas of linguistics. The true promise of the system, however, is that the development of automated Feature Detection will lead to discoveries about form-function relationships that will give us insight into better methods for the description and documentation of languages, and possibly also insights on which to base linguistic theories.

Bibliography

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University.

Alvarez, Alison, Lori Levin, Robert Frederking, Erik Peterson, and Jeff Good. 2005. Semi-automated elicitation corpus generation. In *Proceedings of MT-Summit X*. Phuket, Thailand.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Carbonell, Jaime and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR, Special Interest Group in Information Retrieval*. Melbourne, Australia.

Comrie, Bernard. 1976. *Aspect*. Cambridge: Cambridge University.

Comrie, Bernard. 1985. *Tense*. Cambridge: Cambridge University.

Comrie, Bernard, W. Croft, C. Lehmann, and D. Zefferer. 1993. A framework for descriptive grammars. In *Proceedings of the XVth International Congress of Linguists*, pages 159–70. Sainte-Foy, Canada.

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

Croft, William. 2003. *Typology and Universals*. Cambridge University Press.

Dixon, R. M. W. 2000. A typology of causatives: Form, syntax, and meaning. In R. M. W. Dixon and A. Y. Aikhenvald, eds., *Changing valency: Case studies in transitivity*, pages 30–83. Cambridge: Cambridge University.

Doron, Edit. 1983. *Verbless Predicates in Hebrew*. Ph.D. thesis, University of Texas at Austin.

Dorr, Bonnie J. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics* 20(4):597–633.

Farrar, Scott and Terry Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International* 7:97–100.

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

Font-Llitjos, Ariadna, Jaime Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In *Proceedings of EAMT 10th Annual Conference*. Budapest, Hungary.

Forgy, Charles L. 1982. Rete: A fast algorithm for the many pattern/many object pattern match problem. *Artificial Intelligence* 19(1):17–37.

Good, Jeff. 2003. Clause combining in Chechen. *Studies in Language* 27:113–170.

Haspelmath, Martin. 1993. *A grammar of Lezgian*. Berlin: Mouton.

Lavie, Alon, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, and Jaime Carbonell. 2003. Experiments with a hindi-to-english transfer-based mt system under a miserly data scenario. *TALIP (ACM Transactions on Asian Language Information Processing)* 2:143–163.

McCawley, James D. 1998. *The Syntactic Phenomena of English,* (Second Edition). Chicago: University of Chicago.

McShane, M., S. Nirenburg, J. Cowie, and R. Zacharski. 2002. Embedding knowledge elicitation and mt systems within a single architecture. *Machine Translation* 17.

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 (S)pecial (I)nterest (G)roup in Computational (Phon)ology*, pages 52–61. Barcelona, Spain.

Palmer, Frank Robert. 2001. *Mood and modality (second edition)*. Cambridge: Cambridge University.

Probst, Katharina. 2005. *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.

Probst, Katharina, Ralf Brown, Jaime Carbonell, Alon Lavie, Lori Levin, and Erik Peterson. 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.

Probst, Katharina, Lori Levin, Erik Peterson, Alon Lavie, and Jaime Carbonell. 2002. Mt for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation* 17.

Tomita, Masaru and Eric Nyberg. 1988. *Generation and Transformation Kit, Version 3.2 Users Manual*. CMU-CMT-88-MEMO.