Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages

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

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
U.S.A.

Abstract

NICE is a machine translation project for low-density languages. We are building a tool that will elicit a controlled corpus from a bilingual speaker who is not an expert in linguistics. The corpus is intended to cover major typological phenomena, as it is designed to work for any language. Using implicational universals, we strive to minimize the number of sentences that each informant has to translate. From the elicited sentences, we learn transfer rules with a version space algorithm. Our vision for MT in the future is one in which systems can be quickly trained for new languages by native speakers, so that speakers of minor languages can participate in education, health care, government, and internet without having to give up their languages.

Keywords

Low-density languages, minor languages, feature detection, version space learning

1. Introduction and Motivation

Recently, efforts in machine translation have spread in two directions. Long-term, high-cost development cycles have given way to research on how to build MT systems for new languages quickly and cheaply (Somers, 1997; Nirenburg, 1998; Nirenburg & Raskin 1998; Sheremetyeva & Nirenburg, 2000, Jones & Havrilla 1998). Rapid deployment of MT can be useful in situations such as crises in which time and money are short, but more importantly, lowering the cost of MT has, in turn, opened the option of building MT systems for languages that do not have enough speakers to financially support a costly development process. (See Frederking, to appear; Frederking, Rudnicky, Hogan 1997; and the SALTMIL discussion group, http://193.2.100.60/SALTMIL). There is now increasing awareness of the importance of MT for low-density languages as a way of providing access to government, education, healthcare, and the internet without requiring indigenous people to give up their languages. MT additionally facilitates the design of educational programs in endangered languages, which can be a tool for their documentation and preservation.

Several low-cost or rapid deployment MT methods have been proposed, most of which are data-intensive, depending on the existence of large corpora (Somers 1997; Al-Onaizan et al., 1999). An alternative approach for low-density languages is to learn MT rules or statistics from a smaller amount of data that is systematically elicited from a native speaker (Nirenburg, 1998; Nirenburg and Raskin, 1998; Jones and Havrilla, 1998).

The NICE project (Native language Interpretation and Communication Environment) plans to combine into a multi-engine system both corpus-based MT (Al-Onaizan, et al. 1999; Brown, et al. 1990; Brown, 1996) and a new elicitation-based approach for automatic inference of transfer rules when a corpus is not available. Our vision for MT of the future includes an MT system that is omnivorous in the sense that it will use whatever resources (texts, linguists, native speakers) are most readily available and, in the extreme case, can be trained easily by a native speaker.

For corpus-based MT, we are using the EBMT engine that was developed for the Diplomat and Tongues systems (Brown, 1996)[1]. In addition, we plan to develop statistical techniques for robust MT with sparse data using exponential models and joint source-channel modeling. Both EBMT and SMT require large parallel corpora and produce statistics for associating source and target texts. The focus of this paper is the third method, an elicitation tool that will automatically learn transfer rules (not statistics) from a small, controlled corpus. We call this method Instructible Rule-Based MT (iRBMT). We ask the readers of this paper to keep in mind that this is work-in-progress. It is furthermore important to note that whereas the focus of our work is on low density languages, we have used some examples from major languages for the purpose of presentation in this paper.

2. Instructible Rule-Based MT

For iRBMT a bilingual user who is not an expert in linguistics is asked to translate a set of sentences and specify the word alignment between source and target language sentences. The goal of the learning process is to match every translation example in the elicited bilingual corpus with a transfer rule that accounts for the translation and is of an appropriate level of abstraction. For instance, after the system observes the translations of several example noun phrases of similar structure but containing different nouns and adjectives, it will automatically infer that the transfer rules of these examples can be collapsed into a single transfer rule.

A sample transfer rule that will be produced by the system can be seen in Figure 1 below[2]. Here, X

[1] For the NICE project, EBMT will be applied to new corpora being collected in Mapudungun, a language spoken in southern Chile, and Iñupiaq, spoken in Alaska.

[2] An alternative spelling of the Hebrew side as *ha-yeled ha-gadol* would result in a Y-side of N ADJ with enforced definiteness markers on the noun and the adjective.

represents the English side and Y represents the Hebrew side. X0 refers to the entire English phrase, X1 to the first English word, etc. The X-Y alignment specifies word-level correspondences between English and Hebrew, and X-Y constraints represent feature value projection from English onto Hebrew.

```
;; Hebrew Transfer Rule Example

English: the big boy
Hebrew: ha yeled ha gadol

NP::NP : [DET ADJ N] -> [DET N DET ADJ]
(
;;X-Y Alignment
(X1::Y1)
(X1::Y3)
(X2::Y4)
(X3::Y2)
;;X-side constraints
((X1 NUMBER) = (X3 NUMBER))
((X1 DEFINITENESS) = +)
;;Y-side constraints
((Y2 NUMBER) = (Y4 NUMBER))
((Y2 GENDER) = (Y4 GENDER))
;;X-Y constraints
((X0 NUMBER) = (Y0 NUMBER))
((X0 DEFINITENESS) = (Y0 DEFINITENESS))
)
```

Figure 1: Sample transfer rule for English to Hebrew.

Our new method to infer such transfer rules is based on Version Space (VS) hypothesis formation (Mitchell, 1982). It assumes a hypothesis space with a partial order relation between the hypotheses, as shown in Figure 2 on the following page. In this figure, each level represents a certain level of abstraction. A hypothesis can be generalized e.g. by dropping a constraint (e.g. no requirement for number), by generalizing to part of speech from a specific word (e.g. generalization from *ha* to DET), or by not enforcing a specific value for a feature (e.g. enforcing agreement in gender, but not requiring masculine gender).

We have developed a new form of version space learning called *locally constrained seeded version space* (SVS). Essentially, SVS hypothesizes an appropriate level of generality for the target concept (in this case, the transfer rule being learned). The first positive example (the "seed") for each rule is generalized to this level. Then a version space is created with up to k levels of generalization and k levels of specialization in the VS lattice around this seed and VS learning proceeds as usual or incrementally outward from the seed when new positive or negative examples arrive. Whereas worst-case complexity for VS is exponential ($O(b^d)$ where b is the branching factor of the lattice and d is its depth), SVS exhibits polynomial worst case behavior ($O(b^{2k})$ where k is typically a small constant). The risk, of course, is that if the original generalization level hypothesis is more than k lattice steps from the optimal transfer rule, SVS will converge on less than ideal transfer rules. However, SVS exhibits other desirable properties such as enabling active learning to explore the lattice around the seed.

The version space rule learner will be evaluated with three metrics. The first is the convergence rate - how many examples need to be seen before settling on a level of generalization in the version space. The second and third metrics involve false positives and false negatives.

A false positive, or overgeneralization, would be an application of a rule to an example it should not have applied to. The result will be a translation error, for example, making Swahili verbs agree with all direct objects, whereas in the corpus they agree only with definite direct objects. A false negative, or undergeneralization, will result in a rule not applying when it should. Some false negatives will be detected as sentences that are not translated, but are instances of phenomena covered in the elicitation corpus.

4. Elicitation Corpus

The input to SVS learning is a controlled corpus, which we call the elicitation corpus. The elicitation corpus contains lists of sentences in a major language (e.g., English or Spanish). During the elicitation process, the user will translate a subset of these sentences that is dynamically determined to be sufficient for learning the desired grammar rules. The Version Space learning algorithm requires the corpus to be compositional so that smaller components such as phrases can be used as building blocks for transfer rules of larger components such as clauses. The elicitation corpus is also intended to cover major typological features, using typological checklists such as (Comrie, 1977, Bouquiaux 1992). Like Boas (Nirenburg, 1998), NICE emulates the work of field linguists. However, in contrast to the Boas project, we do not expect the user to learn linguistic terminology.

As initially no bilingual dictionary is available, we start the elicitation process by asking the user to translate basic terms[3] (e.g., tree, cloud, etc.) that are likely to exist even in remote languages. This is important so that we can distinguish between features that do not exist in a language and vocabulary that does not exist. After some basic vocabulary is elicited, the system moves on to more complex structures: noun phrases with or without modifiers, transitive and intransitive sentences, and finally complex constructions such as embedded clauses, relative clauses, comparative sentences, etc.

The pilot corpus includes about 800 sentences, targeting word order (within noun phrases and clauses), grammatical features (number, gender, person, tense, definiteness), and agreement patterns (e.g., agreement of subject and verb, object and verb, head noun and possessor, head noun and adjective, head noun and article). For example, in the elicitation of basic noun phrases, we check how the language expresses definiteness and possession and whether it has numerical classifiers (as in Japanese where it is necessary to say *one volume of book* and *one stick of pencil* instead of *one book* and *one pencil*). The basic-sentences portion of the corpus identifies basic word order of subject, object, and verb, as well as agreement patterns and effects of animacy and definiteness on the expression of subject and object (for example, whether a language prefers a structure corresponding to *There is a man who left* over *A man left* but uses a more basic sentence structure corresponding to *The man left* when the subject is definite).

---

[3] This follows the tradition of the Swadesh List (named after the linguist Morris Swadesh).

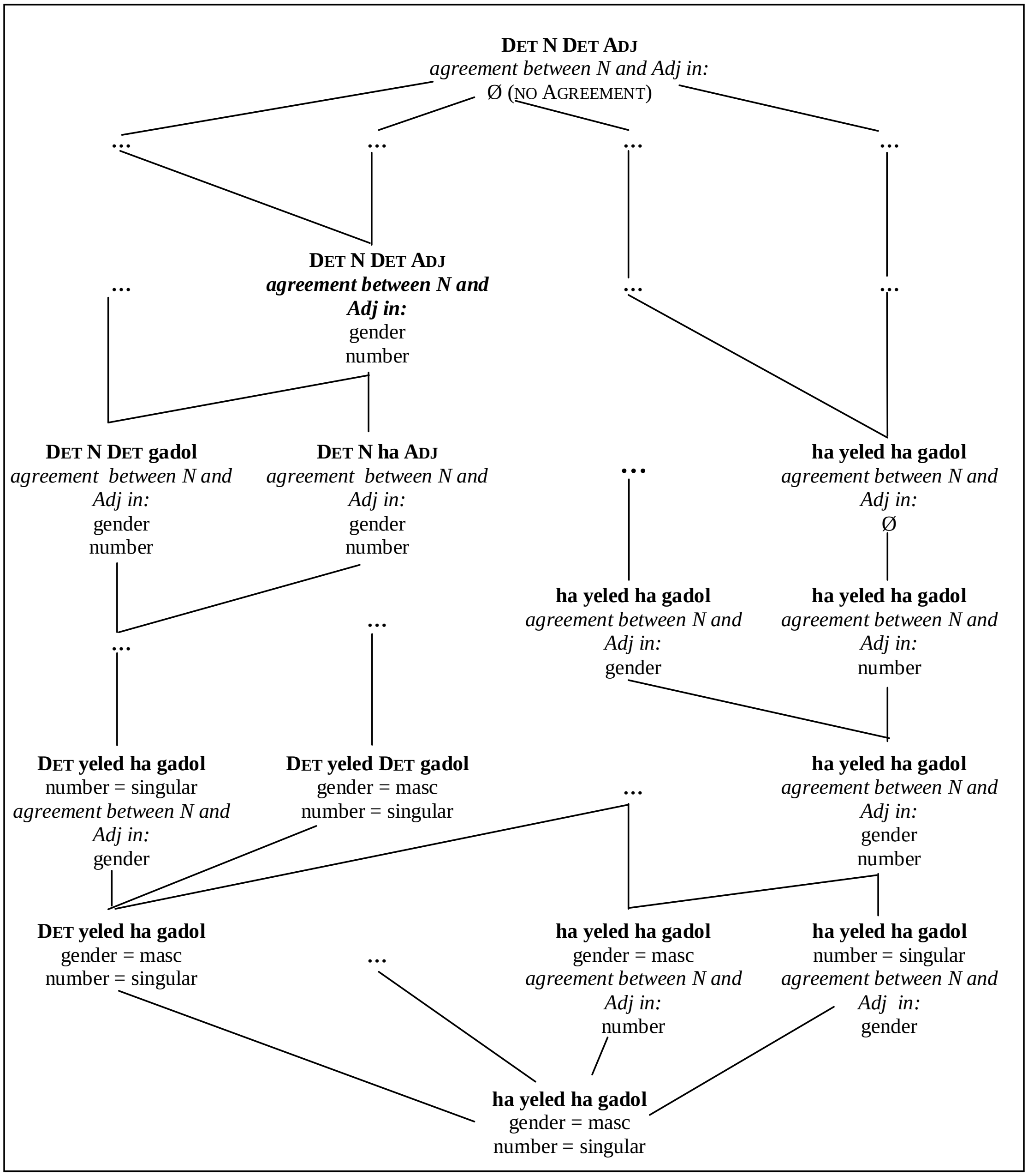

Figure 2: Partial representation of the version space for the example given in figure 1.

It is not trivial to predict what coverage such a corpus will provide, but it is known that some linguistic phenomena are more common among languages than others, i.e. they are used in more languages and within those languages they are more commonly used. We started by covering these prominent features. This will provide us maximal coverage for minimal work. The corpus can then be expanded to cover more obscure phenomena that do not occur in most languages, or are very rare. In the extreme, coverage of this part of our system will depend only on the number of sentences a user is willing to translate.

We envision that the elicitation corpus should gradually expand to between 10,000 and 25,000 sentence pairs. However, the more common phenomena will be covered by the earlier part of the corpus, and thus translations can be produced (albeit not perfect ones) without utilizing the full elicitation corpus. Moreover, the dependencies among the features elicited imply that only a fraction of the corpus will be employed for each particular language, as discussed below.

5. The Elicitation Interface

The tool that orders sentences for elicitation and presents them to the user is called *GLAD* for Generalized Linguistic Acquisition Device. A picture of the interface can be seen in Figure 3. The users are presented with a source language sentence and translate the sentence into their native language. Then they specify how the words in

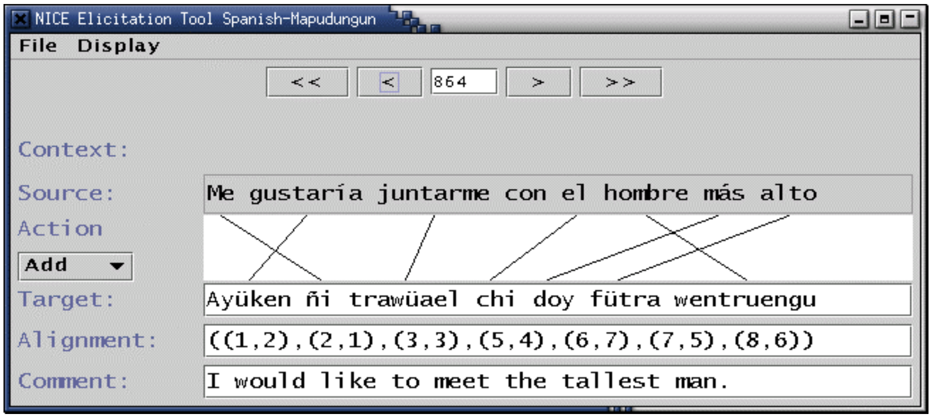

Figure 3: Elicitation Interface. The elicited language here is Mapudungun, the source language Spanish. The English translation is given as a comment for the developer.

the two sentences (source and target) align. The tool allows for the specification of one-to-one, one-to-many, many-to-one, and many-to-many alignments. It also allows for words to have no correspondence in the other sentence at all. This flexibility is necessary as all of the above alignments are possible between languages.

As can be seen in the GLAD output, the word-by-word alignments are stored in a parenthetical notation. When the user clicks at two corresponding words, this alignment is stored internally as a pair consisting of the indices of the words in their respective sentences. This is done not only to compress the representation to a minimum; storing the alignments as indices also ensures that no ambiguities are caused when a word occurs more than once in a sentence.

6. Detection of Grammatical Features

In addition to SVS learning of transfer rules, another type of learning takes place during elicitation – detection of grammatical features (such as number, person, gender, definiteness, and animacy) that play a role in agreement and well-formedness constraints.

Detection of typological features also allows us to navigate through only the appropriate subset of the elicitation corpus, thereby minimizing the required number of translations. Research on implicational universals has found that certain features are guaranteed not to exist in a language if certain other features are not present. For instance, if a language does not mark plural, then it will also not mark dual or paucal (on Implicational Universals see e.g. Comrie, 1998 and Greenberg, 1966). The elicitation process will thus first inquire about the existence of a plural, and only ask about dual and paucal if plural was found to exist.

In other cases, tests cannot be performed properly without knowing the results of other tests. For example, the form of adjectives often depends on the noun class of their head noun. If we have not yet performed tests to determine the noun classes in a language, it would be difficult to learn the morphology of adjectives.

The sequence of sentences elicited from the native speaker is controlled by a hierarchical organization of features. The system is similar to a tree, in that the logical flow of the system can be thought of as a depth-first search, where the states contain lists of tests. Figure 4 below illustrates the basic design of the system.

The central dispatching unit is the state labeled *Diagnostic Tests*. It contains tests that serve merely to provide an initial idea of whether a linguistic feature exists in a language or not. The mechanism for detecting

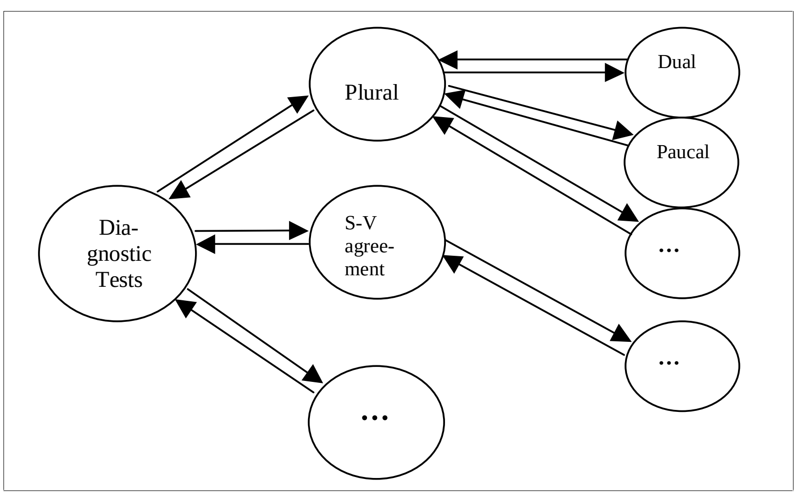

Figure 4: Logical flow of data elicitation

features is comparison of minimal pairs – sentences that differ in only one feature, for example, whether a noun is singular or plural. For instance, the minimal pair of sentences *The rock fell* and *The rocks fell* is designed to detect whether singular and plural nouns are marked differently in the target language (and also whether the verb is marked differently depending on the number of the noun).

After a minimal pair of sentences is elicited, feature detection proceeds in the following way. The system first identifies which words of the target language have been aligned with the minimally different elements of the source language sentence. For example, which words have been aligned with *rock* and *rocks.* The target language words are then examined to see if they are the same or different from each other. In the first version of the system we have made the simplifying assumption that the feature that is tested for is expressed as an affix rather than in a change of word order or a separate word. Also, as a cushion in cases where the user decides to vary the vocabulary in a minimal pair (e.g., using near synonyms like *rock* and *stone*) we actually elicit several minimal pairs for each feature we are examining.

If the diagnostic tests detect a difference between singular and plural nouns, control is passed to the state *Plural*, which will then perform tests that are related to plural, including dependent features like dual and paucal. The system returns to the node *Diagnostic Tests* after the list of plural tests is exhausted.

The tests in each node are sequenced based on linguistic knowledge. For instance, the state *Diagnostic Tests* first performs tests regarding noun classes, before it starts inquiring about adjectival forms.

In order for previously elicited information to be useful to later tests, it is necessary that the states pass along information to other states. This is achieved by having a central *Results* unit. The results of all tests are written to this unit. Before prompting the user for the translation of a minimal pair of sentences, this unit is then consulted to see whether the minimal pair still presents an open question. For instance, if it is found that a language has no dual, the sentences *My two cats are brown* and *My three cats are brown.* would not be expected to show a difference in the possessor (*my*). Accordingly, the system will prune the number of sentences that the user is asked to translate.

It is also important that the sentences and tests are stored separately, so that in the test we only store the index of each sentence. The actual sentence is then retrieved from the database by this index. This ensures that sentences can be re-used for a number of different tests. For example, the sentence "The men danced" is used in the test for number, but also in a further test for definiteness without having to be presented to the user again.

7. Future work

To reach our ultimate goal of building a rapid-deployment tool that automatically learns transfer rules, we have to expand and improve our prototype. The corpus will have to be expanded to cover more typological features, and tests for these features will have to be integrated in the tool hierarchy. Furthermore we will continue to work on how to navigate through the corpus in the most efficient way for each language.

Another task for the coming months is to tag each sentence of the elicitation corpus (source language) with a feature vector showing what it exemplifies (e.g., definite inanimate subject, relative clause with a gap in object position, possessive noun phrase with first person singual possessor, etc.). The SVS learning mechanism will use the feature vector to find sentences that bear on its hypotheses. For instance, to be able to produce a transfer rule that applies only to adjectives of definite nouns, we need to know the part-of-speech tags of a sentence in addition to information about the definiteness of a sentence.

8. Conclusion

We believe that this work is significant to the future of machine translation for several reasons. In recent years, the research community has recognized the importance of tools that can build MT systems quickly. Especially in the case of low-density languages, the communities as well as the governments have limited financial resources that can be spent on the development of a translation system. We believe that NICE can provide a tool that will make such extreme costs unnecessary. In addition, NICE does not require the existence of a large bilingual dictionary, while its performance can be improved if one is available or can be collected over time. It will be able to build an MT engine using very little data compared to other systems, especially rapid-deployment systems. Again, this is a crucial element when working with low-density languages.

It is also important to note here that our paradigm of conducting MT research is a novelty in the field. We have formed partnerships with native communities who will determine their own MT needs and take responsibility for collecting and archiving their own data.

Finally, it is our belief that MT research has not touched all language families, and that current techniques may therefore not be fully general or applicable. For example, whereas there are MT systems for morphology-rich, word-order poor languages, we do not know of an MT system for morphologically extreme languages such as polysynthetic languages. For our research on corpus-based methods, we are collecting corpora in two polysynthetic languages – Mapudungun (Chile) and Inupiaq (Alaska) – which we encourage other researchers to use as test data for their own systems.

To summarize, our vision for the next 10 years is that MT can be learned quickly from native speakers without linguistic expertise, which will allow MT to branch out to lower-density languages and thus to more language families.

9. Bibliography

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical Machine Translation, Final Report, JHU Workshop 1999. Technical Report, CLSP/JHU.

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

Brown, Peter F., J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S.Roossin. (1990). A Statistical Approach to Machine Translation. In Computational Linguistics, 16(2):79-85.

Brown, Ralf. (1996). Example-Based Machine Translation in the Pangloss System. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96).

Comrie, Bernard. (²1989 [¹1981]). Language Universals & Linguistic Typology. The University of Chicago Press.

Comrie, Bernard; N. Smith. (1977). Lingua Descriptive Series: Questionnaire. In: Lingua, 42:1-72.

Frederking, Robert; A.Rudnicky; C. Hogan; K. Lenzo. (To appear). Interactive Speech Translation in the DIPLOMAT Project. In Machine Translation Journal, Special issue on Spoken Language Translation.

Frederking, Robert; A.Rudnicky; C. Hogan (1997). Interactive Speech Translation in the DIPLOMAT Project. In: Steven Krauwer et al. (eds.). Spoken Language Translation: Proceedings of a Workshop. Association of Computational Linguistics and European Network in Language and Speech.

Greenberg, Joseph H. (1966). Universals of language. 2nd. ed. Cambridge, MA: MIT Press.

Jones, Douglas; R. Havrilla. (1998). Twisted Pair Grammar: Support for Rapid Development of Machine Translation for Low Density Languages. In Proceedings of AMTA.

Mitchell, Tom. (1982). Generalization as Search. In Artificial Intelligence, 18:203-226.

Nirenburg, Sergei. (1998). Project Boas: “A Linguist in the Box” as a Multi-Purpose Language. In Proceedings of LREC.

Nirenburg, Sergei; V. Raskin. (1998). Universal Grammar and Lexis for Quick Ramp-Up of MT Systems. In Proceedings of COLING-ACL.

Sherematyeva, Svetlana; S. Nirenburg. (2000).Towards a Universal Tool For NLP Resource Acquisition. In Proceedings of The Second International Conference on Language Resources and Evaluation (Greece, Athens, May 31 -June 3, 2000).

Somers, Harold L. (1997). Machine Translation and Minority Languages. In Translating and the Computer 19: Papers from the Aslib conference (London, November 1997).

```
Category:     Plural
Test:         Dual
       First sentence of minimal pair:
              Source:            Two cats ran across the street
              Target:            Paka wawili walivuka barabara
              Alignment:   ((1,2),(2,1),(3,3),(4,3),(5,0),(6,4))
       Target word1:       =>Paka
       Second sentence of minimal pair:
              Source:            Four cats ran across the street
              Target:            Paka wanne walikimbia na kuvuka barabara
              Alignment:   ((1,2),(2,1),(3,3),(4,4),(4,5),(5,0),(6,6))
       Target word2: =>Paka
       =>The two words are equal.
       First sentence of minimal pair:
              Source:            Two men danced
              Target:            wanaume wawili walicheza dansi
              Alignment:   ((1,2),(2,1),(3,3),(3,4))
       Target word1:       =>wanaume
       Second sentence of minimal pair:
              Source:            Many men danced
              Target:            wanaume wengi walicheza dansi
              Alignment:   ((1,2),(2,1),(3,3),(3,4))
       Target word2: =>wanaume
       =>The two words are equal.
Feature Dual not detected.
```

Appendix: Sample output of elicitation tool. The target language here is Swahili.