Universal Grammar and Lexis for Quick Ramp-Up of MT Systems

Sergei Nirenburg and Victor Raskin
Computing Research Laboratory
New Mexico State University
Las Cruces, N.M. 88003, U.S.A.
{sergei, raskin}@crl.nmsu.edu

Abstract

This paper introduces Boas, a semi-automatic knowledge elicitation system that guides a team of two people through the process of developing the static knowledge sources for a moderate-quality, broad-coverage MT system from any "low-density" language into English in about six months. The paper focuses on some issues in the elicitation of descriptive knowledge in Boas and also the issue of the principled reuse of pre-existing resources, such as a lexicon, an ontology, and an English generation module, among others, made possible by the fact that the client MT system is developed for a single target language.

1. Introduction: The Boas Project

This paper presents Boas, a semi-automatic knowledge elicitation system that guides a team of two people through the process of developing static knowledge sources for a moderate-quality, broad-coverage MT system from any "low-density"[1] language into English in about six months. Boas contains knowledge about human language and means of realization of its phenomena in a number of specific languages and is, thus, a kind of a "linguist in the box" that helps non-professional acquirers with the task, whose complexity is legendary.[2]

The knowledge about language elicited by Boas from the acquirers aims to support MT output quality which is roughly commensurate with the outputs of the better commercial systems, such as Systran. These relatively modest expectations are dictated by the amount of language work which can be carried out, given the resources available. The rules of the game specifically exclude linguists and MT developers from the acquisition team. Under such conditions, the only sensible course of action is to attempt to collect as much knowledge about as many languages as possible in advance and include it in the elicitation system itself.

Section 2 below is devoted to defining the format of the descriptive language knowledge to be elicited from the acquirers through Boas. The descriptive language knowledge, which we address in this paper, is, later in the process of Boas operation, converted into operational knowledge capable of supporting the processes of source language analysis and source-target transfer. In Section 3, we discuss how work on ontological semantics in MT can contribute to Boas in a situation of a single target language, English. In Section 4, we address the procedure for descriptive language knowledge acquisition in Boas, both in terms of resources created and reused and in terms of the actual elicitation techniques, differentiating between the acquisition of grammatical and lexical parameters.

2. Defining Parameters for Boas

The descriptive knowledge about the source language is a set of statements about morphological, syntactic, and lexical properties (parameters) of a language, listed together with their values and realization options. Data about each parameter includes the language, the name of the parameter, the list of entities to which this parameter applies (its domain) and the list of parameter values (its

[1] "Density" refers roughly to the amount of effort having been previously expended in the field on computational descriptions of particular languages, resulting in the creation of a variety of machine-tractable resources—text corpora, grammars, lexicons, analyzers, etc. Thus, Spanish will most probably count as "high-density" while, say, Tagalog will not.

[2] We have introduced Boas and discussed some pertinent theoretical issues in Nirenburg (1998). In this paper, we focus on the more practical aspects of Boas implementation.

range). Moreover, parameter values have an associated set of realization options in each language. For instance, the parameter of gender in Ukrainian is described as follows:

language: Ukrainian
parameter: gender
domain: nouns, adjectives, possessives (head agreement), verbs in past tense
range (parameter values): masculine, feminine, neuter
realization: [gender markers in lexicon for nouns; inflection paradigms for adjectives, possessives and verbs in past tense]

For comparison, the Hebrew gender is described differently:

language: Hebrew
parameter: gender
domain: nouns, adjectives, possessive (non-first-person possessor agreement), finite verbs
range (parameter values): masculine, feminine
realization: [gender markers in lexicon for nouns; gender inflection paradigms for adjectives, possessives and verbs]

Instead of discovering parameters from scratch for each language, it is preferable, in order to ensure uniformity and systematicity of Boas operation, to come up with a complete list of all possible parameters in natural languages, with complete lists of their possible values attached. The attainability of such a resource becomes then a central issue.

The terms 'parameter' and 'value' are used in our task in the same sense as in the school of theoretical syntactic thought consecutively known as government and binding (Chomsky 1981), principles and parameters (Chomsky 1986) and the minimalist position (Chomsky 1995). The theory postulates a small number of general principles defining the innate human language faculty and a larger number of language parameters, which implement these principles by selecting concrete values for particular languages. The complete set of such parameters and values constitutes a universal grammar (UG)—see also Culikover (1997), Lightfoot (1991) and Webelhuth (1992).

Unfortunately, work within this approach has not stressed the descriptive task of creating a comprehensive inventory of universal grammar parameters or even those for particular languages or language families. For Project Boas, it means that both the nature of the parameters it would be using and their inventory has to be developed in-house.

In order to define a set of parameters for Boas, it is essential to distinguish among the language phenomena that should be accorded the status of parameter and those that should be understood as parameter values or their realizations. Still other phenomena may remain, at least for the task at hand, outside the parameter system. We believe, with Dorr (1993), that parameters may be understood as building blocks of an interlingua in MT. We reserve judgment about whether every component of an interlingua is by definition parametric[3].

Thus, the parameter "lexical category" has a range of values {V, N, Adj, Adv, ...}. Any of these values may itself be considered a parameter. If viewed within a single language, their values are, ultimately, all words in the language which belong to the respective lexical categories. The realizations of these values are the specific forms of these words, which appear in text decorated with realizations of appropriate values of such morphological parameters as NUMBER, GENDER, CASE, etc.

An example of a syntactic parameter is HEAD-MODIFIER DEPENDENCY, whose values include such pairs as "head: noun; modifier: adjective;" "head: verb; modifier: adverb," "head: noun; modifier: relative clause" and others. Realization options for these values involve word or constituent order rules (for instance, post- or pre-posing) and agreement rules.

Lexical parameters are viewed as language-independent lexical meanings (ontological concepts), such as $\text{TABLE}_{\text{FURNITURE}}$. The values of this parameter are the word senses corresponding to this ontological concept across the inventory of languages. The realizations for these values are the words or phrases that express this meaning in each language, with a possibility of a lexical gap (a null value)

[3] Thus, for instance, a morphological analyzer for Turkish uses information that does not have to be expressed parametrically, such as data about nominal suffixes, which one needs to know in order to recognize a noun form but which do not correspond to any parametric value that needs to be expressed in English; similarly, a Russian verbal prefix may help determine the aspect value but does not realize a distinct parametric value of its own.

included.

3. Translation Environment Supported by Boas

The single-target-language (English) environment which Boas serves allows for simplification of both system implementation and the acquisition process compared to the case of multiple SLs and TLs. First, only one text synthesis module needs to be built. Second, many fewer transfer components (bilingual lexicons, transduction tables for closed-class lexical items, feature and structure transfer tables) are needed. In fact, this situation almost licenses the transfer approach, as the combinatorial argument for interlingual MT is weaker here than in the case of multiple TLs (see, however, below and fn. 3). Third, it appears that knowledge acquisition for a new SL may be aided by the presence of a number of resources already developed for the TL.

These resources include a) the vocabulary of the generation lexicon which can serve as the list of lexical parameters for compiling the bilingual dictionary; b) a world model (ontology) providing the terms in which the senses of the English words and phrases are expressed (Boas uses the ontology from the Mikrokosmos project at NMSU CRL—see Mahesh and Nirenburg 1995); c) the structure and term definitions from the text meaning representation in Mikrokosmos (see, for instance, Onyshkevych and Nirenburg 1995), to help guide parameter elicitation; d) the set of English closed-class lexical items and morphemes; e) English grammar used in text synthesis, which provides the TL side of structural transfer rules in the runtime MT system (see Figure 1 above); and f) a set of "ecological" parameters and their realizations for English. While a complete description of the use of all of the above resources is beyond the scope of this paper, we will give a few brief illustrations.

The list of English word senses seeds the acquisition of the SL lexicon. The acquirer first simply translates all the word senses into SL and then adds SL features to the corresponding entries as needed. The result is an SL-TL transfer dictionary which also serves as the lexicon for SL analysis. The acquirer gets a lemma with all its senses:

Entry: table-n1
POS: noun
Sense: furniture

Entry: table-n2
POS: noun
Sense: diagram

and produces the following SL lexicon entries (the example is in Hebrew):

Entry: shulxan-n1
POS: noun
Gender: m (plural *-ot*)
Sense: table-n1

Entry: tavla-n1
POS: noun
Gender: f
Sense: table-n2.

In the examples, the senses are conveniently explained not in any specially designed lexicon/ ontology notation, but rather through translation into English. Because each English translation is the entry head for a sense which is already explained in an ontology-based semantic metalanguage in the already existing Mikrokosmos lexicon, Expedition can benefit from richer semantic information than that acquired using Boas.We use the Mikrokosmos ontology as a search space to support word sense disambiguation. The method (suggested by Jim Cowie) depends on the bilingual dictionary of the kind illustrated above. Coarse grain-size lexical mappings of TL word senses to ontological concepts are established (for instance, *chihuahua* and *poodle* may be both linked to the ontological concept DOG). The system, thus, knows that both chihuahuas and poodles have four legs, are carnivorous, domesticated, etc.

The disambiguation method uses such ontological constraints by computing a distance in the ontological space between ambiguous word senses on the one hand and the senses of other words in their context. SL syntactic information helps to guide the disambiguation process by providing additional constraints. Thus, closeness between senses of words belonging to the same syntactic unit is weighed more heavily than that across unit boundaries.

The acquisition of the complete list of parameters in the single-TL environment is facilitated not only by the availability of the initial set of lexical parameters but also by the prominence of the syn-

tactic and morphological parameters activated in English. Thus, for morphology and syntax, the existence of such comprehensive grammars of English as Quirk *et al.* (1985) allows a quick round-up of the major parameters. One cannot always limit oneself, however, to TL-induced acquisition as we have demonstrated in the previous section on the example of GENDER in English.

4. Source Language Knowledge Acquisition

Acquisition of descriptive knowledge about a language consists in Boas of a set of elicitation "episodes." The episodes have been clustered, very unevenly, into six large classes, namely, morphology, closed-class items, open-class items, syntax, transfer features, and ecology. Each episode is an HTML document, accessible through the standard Web browsers. Each page deals with one parameter and elicits information on its values present in the source language as well as the realizations of these values. It is morphology which seems to require the greatest number of parametric episodes, though the total is not very high: verbs, around 30 episodes for the finite forms, and about 40 for the non-finite forms; nouns, around 20; adverbs and adjectives, under 5. Morphology does include these four sections.

Closed-class items are pronouns, temporal relations, spatial relations, and case-like relations, e.g., prepositional phrases (the morphological case is, of course, handled in the noun section of the morphology class). Each closed-class page deals with one English closed-class item in one appropriate sense and elicits all the possible translations of that item into the source language (or, more accurately, all possible expressions in the source language which may be translated into English with this item in this sense), with the complete morphological and syntactic information on each such translation. Because there are, roughly, 200 closed-items in English (and many other languages), this class requires the greatest number of Web pages but they are mot parametric and quite straightforward.

Open-class items are acquired lexically, with the help of, essentially, one huge standard elicitation episode/Web page. Lexical acquisition proceeds as described in Section 3 and further aided by a special resource created for Boas/Expedition: continuing our work on significantly reducing the number of different senses in a lexicon entry by combining related senses in MRDs (see Nirenburg *et al.* 1995) and, more rarely, deleting the marginal ones, we have manually reduced a combined (Mikrokosmos and other sources) English lexicon of about 28,000 words to about 40,000 word senses, each of which serves as a lexical parameter for SL acquisition. In addition, frequency analyses of SL corpora will provide the requirements for adding lexical parameters from SL, not just TL.

Syntax will have rather few elicitation episodes since much of it will be collected automatically from a large corpus, pre-tagged morphologically by Boas--automatically.

The elicitation pages in transfer features class will deal with non-standard transfer correpondences, and ecology with proper names, punctuation, standard acronyms for numbers, and other print conventions of the source language. It is unlikely to be a very numerous class and it is, of course, largely non-parametric.

It is the morphology class which has necessitated the heaviest use of and most remedial effort on parameter inventories. We have largely expanded the inventory of parameters, previously acquired in the PROPERTY branch of the Mikrokosmos ontology: most of the "grammatical" meanings, realized in any one of the Mikrokosmos languages, such as English, Spanish, and Chinese, are already recorded and systematized there. We have also had to compile what we hope to turn out to be the most complete list of both parameters and their values, such as noun case (around 30 values), verb mood (about a dozen), verb aspect (about two dozen), etc.

A standard morphological episode elicits the values for a parameter which the user has already marked as present in the source language on the previous Web page. The moment the box for that parameter was checked there, the user is taken to the values page, where Boas offers a complete list of existing values for that parameter and requests that the user select all that apply.

Two additional factors deserve a special mention. First, each elicitation episode is supported with context-sensitive online help, which can be also accessed as a complete morphological, syntactic, closed-class, etc. tutorial. This tutorial, as far as we

know, is the only available sketch of universal grammar. Secondly, each parameter and value choice provides for the selection of "other" unlisted values, and great care is taken to assist the user in naming the parameter or value as well as determining the appropriate values for each user-introduced parameter with the appropriate realizations.

At the conclusion of each elicitation cycle, such as nouns or verb finite forms, all the elicited information is presented to the user for checking, correcting, and editing in the form of a paradigm table, which id the Cartesian product of all the established parameters and values. The user is also guided through the parts of the source language grammar which deals with exceptional paradigms. It should be also noted that, in open-class acquisition, the paradigms for each acquired source language item will be assigned to one of the already established types or, alternatively, a new exceptional type will be added, if necessary.

The most difficult issues in acquisition involve the transcategorial realization of values, such as the signalling of a noun case in the verb or non-standard clitics, or the lexical realizations in SL of grammatical parameters in TL, such as the possible absence of continuous tenses in a SL and the choice of a grammatical realization of such lexical values as "right now" in the SL as the present continuous marker on the corresponding verb. Interestingly also, clitics and similar morphological "complications" of source languages are unlikely to present a significant problem either in elicitation or in transfer, primarily because of the single target language environment in Boas and ensuing lack of necessity to generate (rather than just to analyze) much morphological complexity.

5. Conclusion: Computational Field Linguistics?

Boas exemplifies the broad-coverage descriptive approach to NLP (see, for instance, Nirenburg and Raskin 1996) and adds to it a complementary new commitment to developing and using automated field-linguistic methodology (cf. Nirenburg 1998). This goes hand in hand with the evolving reorientation of theoretical linguistics from selective theorizing, in terms of prevalent atomistic rule postulation and testing, back to the primary goal of linguistics, which is a theory-based language description.

A full evaluation of Boas, that is, the development of the first actual SL to English MT system over a six-month time interval, will take place within the next two years.

Acknowledgments

The research reported in this paper was supported by Contract MDA904-92-C-5189 with the U.S. Department of Defense. Victor Raskin is grateful to Purdue University for permitting him to consult CRL/NMSU.

References

Chomsky, N. 1981. **Lectures on Government and Binding**. Dordrecht: Foris.

Chomsky, N. 1986. **Knowledge of Language: Its Nature, Origin, and Use**. New York: Praeger.

Chomsky, N. 1995. **The Minimalist Program**. Cambridge, MA: MIT Press.

Comrie, B., and N. Smith 1977. Lingua Descriptive Studies: Questionnaire. Lingua 42:1, pp. 1-72.

Culikover, P. W. 1997. **Principles and Parameters. An Introduction to Syntactic Theory.** Oxford University Press.

Dorr, B. 1993. Interlingual Machine Translation: A Parametrized Approach. *Artificial Intelligence* 63, 429-92.

Lightfoot, D. 1991. **How to Set Parameters**. Cambridge, MA: MIT Press

Mahesh, K., and S. Nirenburg 1995. Semantic Classification for Practical Natural Language Processing. In: **Proceedings of the Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting**. Chicago, IL.

Nirenburg, S. 1998. Project Boas: "A Linguist in the Box" as a Multi-Purpose Language Resource. In: **Proceedings of The First Lexical Resources and Evaluation Conference**. Granada, Spain.

Nirenburg, S., and V. Raskin 1996. Ten Choices in Lexical Semantics. MCCS-96-304, Las Cruces, N.M.: NMSU CRL.

Nirenburg, S., V. Raskin, and B. Onyshkevych 1995. Apologiae Ontologiae. **TMI '95**, Leuven.

Onyshkevych, B., and S. Nirenburg. 1995. "A Lexicon for Knowledge-Based MT." *Machine Translation*, 10:1-2, pp. 5-57.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik 1985. **A Comprehensive Grammar of the English Language**. London: Longman.

Webelhuth, G. 1992. **Principles and Parameters of Syntactic Saturation**. New York and Oxford: Oxford University Press.