From Knowledge Elicitation System to Teaching Tool

Marjorie McShane, Ron Zacharski and Sergei Nirenburg

Working Paper 04-05

March 8, 2005

Institute for Language and Information Technologies
University of Maryland Baltimore County

Abstract

This paper discusses the potential for an extant interactive Web-based linguistic knowledge elicitation system to be exploited in the teaching of linguistic content and practices. This tool would support the learning of linguistics and the development of real-world linguistic skills by guiding students through the process of describing a language of their choice from scratch. This method of teaching linguistic content and discovery skills differs from traditional methods in a similar way that an apprenticeship differs from formal education: students will learn by doing, being generously supported with help materials, and minimal hard restrictions will be placed on their methods and course of work.

1. Introduction

Learning through supervised doing, in the spirit of age-old apprenticeships, is not an option widely offered by our educational system, at least until the advanced graduate level. The infeasibility of such a potentially fruitful path of discovery derives primarily from the ratio of students to teachers, which is far higher than what is required for traditional apprenticeship. But the computer age offers the opportunity of redefining apprenticeship, with the computer providing both the freedom of independent work and, as natural language processing capabilities evolve, the potential to ever more effectively guide the learning process.

In this paper we suggest that a Web-based linguistic knowledge-elicitation system called Boas could, with modest amendments, be effectively applied to the teaching and learning of linguistics, in effect providing students with a computer-assisted apprenticeship in linguistics.

Linguistics textbooks, which form the basis of almost all undergraduate courses in linguistics, are mostly variations on a theme. The theme works quite well: describe the basics of different branches of linguistics and ask the students to do exercises—often engaging cross-linguistic ones—that require them to apply what they have learned (e.g., Fromkin and Rodman 1998; Napoli 1996; O'Grady, Dobrovolsky and Aronoff 1997). This approach to pedagogy, common in college teaching, starts with definitions and generalizations and works toward specific instances, data sets, etc. However, practical linguistic work, like practical work in any field, tends to be much less prescriptive. When a linguist goes out into the field to describe a new language, or sets about some task in natural language processing, or even tries to develop some portion of a theory, work seldom follows a known, straight path. Creativity and flexibility in thinking are necessary from beginning to end. We suggest that it would be a step forward in the pedagogy of linguistics to foster the exploratory side of linguistics from the introductory levels.

Boas is an expert system that guides users through the process of describing any language (L), resulting in a structured profile of that language suitable for natural language processing (NLP) applications. Unlike most knowledge elicitation environments, which are designed for use by experts, Boas was designed to be accessible to linguistic experts and novices alike. This reason for this broad orientation derives from practical considerations. The original goal of Boas was to elicit sufficient information about L to support an L-to-English machine translation system.[1] Languages of particular interest were those less "popular" ones for which MT capabilities have not yet been developed. But since there is no guarantee that one can find a native speaker of all such languages who is also a trained linguist, the KE process was designed to cater to novices as well.

To serve this end, Boas contains an extensive suite of training materials that amount to a stepwise, interactive introduction to descriptive linguistics that is provided on a need-to-know basis. Thus, users of Boas not only create language profiles that can be used in many applications, they learn concepts and methods of linguistic investigation along the way. It is these latter aspects of the system that make it particularly well suited to serving pedagogical ends.

Teaching linguistic content and discovery skills through a system like Boas would not only bring cutting-edge technology into the classroom, it would be a wise means of reusing government-supported R&D. Moreover, a practical, descriptive approach to language is one of the things currently missing in many linguistics programs, having been unseated by formalisms of theory, politically and sociologically motivated topics (e.g., feminism and language), and/or interdisciplinary studies. This is unfortunate since one of the fastest-growing areas of work for linguists is natural language processing, where the ability to describe and analyze language is most valued.

[1] The larger system that houses both Boas and the other resources necessary for ramping up an MT system is called Expedition. The Expedition project was funded by Department of Defense Contract MDA904-92-C-5189.

In the next sections we briefly describe Boas (for a more detailed description see, e.g., McShane, Nirenburg, Cowie and Zacharski 2003; McShane and Nirenburg 2003a,b) and discuss the relatively minor "repackaging" that we believe could render it a useful system for teaching linguistics. We will call this system Boas-L: Boas for Learning Linguistics. The paper will conclude with a discussion of further implications of the potential collaboration between practical NLP systems and computer-aided pedagogy.

2. A Snapshot of Boas

Elicitation of information about any L is made possible in Boas by the system's resident metaknowledge about language in general, which is organized into a typologically and cross-linguistically motivated inventory of parameters, their potential value sets and modes of realizing the latter. For example:

Parameter	Values	Realizations
case	nominative, genitive, etc.	affixes, particles, etc.
syntactic function	subject, direct object, etc.	affixes, word order, etc.
spatial relations	*in*, *on*, *at*, etc.	prepositions, affixes, etc.

The inventory (which has properties in common with traditional surveys used in field linguistics (see, e.g., Comrie and Smith 1977 and Longacre 1964)) takes into account phenomena observed in a large number of languages. Particular languages typically feature only a subset of parameters, values and means of realization. The parameter values a particular language employs, and the means of realizing them, differentiate one language from another and can, in effect, act as the formal "signature" for the language. The selection of parameters and values in Boas is made similar to a multiple-choice test, which, with the necessary pedagogical support, can be carried out even by an informant not trained in linguistics. In sum, the methodology of KE employed in Boas integrates the familiar graphical user interfaces with the (meta)knowledge about the typology and universals of human languages and a methodology of guiding the user through the acquisition process.

In addition to its methodological innovations, Boas also allows a maximum of flexibility and economy of effort. Certain decisions on the part of the user cause the system to reorganize the process of acquisition by removing some interface pages and/or reordering those that remain. This means that the system is more flexible than static acquisition interfaces that require the user to walk through the same set of pages irrespective of context and prior decisions. Moreover, a dynamic task tree graphically represents progress made and data dependencies, making it clear to the user what tasks can be carried out at any time. We call the acquisition paradigm exemplified by Boas *knowledge elicitation*, in contrast to *knowledge acquisition*, which is carried out by knowledge engineers.

The modules of Boas are as follows:

Introductory tutorials, which provide an introduction to the system, a description of resident resources, etc.

Ecology, which elicits the character set of L, how dates and numbers are expressed, punctuation conventions, common abbreviations, etc.

Inflectional morphology, where the user indicates how grammatical meanings (like tense, number, case) are realized in L; as applicable, s/he a) establishes inflectional paradigms for relevant parts of speech in L and/or b) provides affixal or word-level realizations of the inventory of grammatical meanings (this is an extensive process that includes machine-learning of productive rules of inflection for open-class lexical items).

Derivational morphology, where word-formation processes that change the meaning and/or part of speech of words in L are elicited in structured fashion (e.g., *un* + *comfortable* → *uncomfortable*).

Syntax, where information is elicited about the structure of noun phrases (inventory of elements, their ordering and coocurrence potential), the realization of grammatical functions (e.g., subject, direct object), the realization of sentence types (e.g., interrogative, negated), etc.

Closed-class lexicon, where L equivalents for an inventory of universal, closed-class semantic meanings (spatial relations, temporal relations, pronouns, numerals, etc.) are elicited along with their inflectional forms, if applicable.

Open-class lexicon, where a L-to-English lexicon is developed using many different elicitation options including translating resident English word lists into L, importing word lists in L or English and translating them, and importing then supplementing available machine-readable lexicons.

In order to carry out any of the subtasks in these modules, the user must learn the concepts and terminology applicable for that aspect of language description, and must analyze the given language in specific ways. Help to carry out each subtask consists of: 1) hundreds of alphabetically and thematically organized glossary pages of linguistic terms that were developed specifically for this system and are accessed by hyperlinks from associated acquisition

pages (see Figures 1 and 2); 2) task-specific help links presented using methods of progressive disclosure, which permit users with different levels of experience to use the same interface; 3) when applicable, novice and expert paths through a task, with the former including preparatory steps, additional instruction, etc.

Thanks to all of these features, Boas has always been a teaching tool – it simply has not been targeted to the specific needs of classroom teaching.

Figure 1. The task tree in Boas at the point when the paradigmatic morphology of nouns is being started.

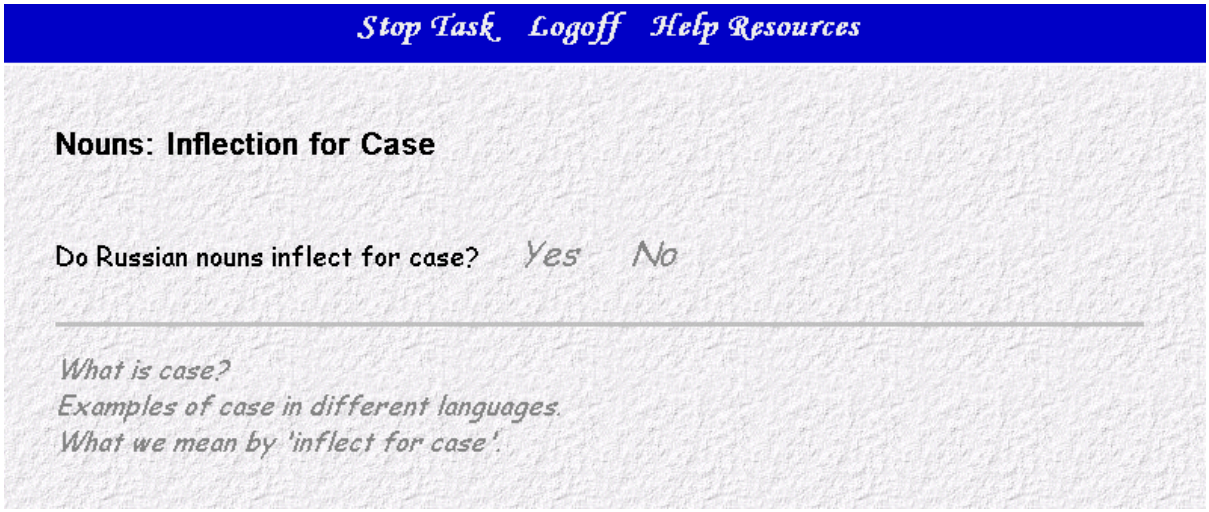

Figure 2. The first elicitation page for case, showing methods of progressive disclosure. Italics indicate action elements like links and yes/no response buttons.

3. From Boas to Boas-L: The Big Picture

In Boas-L, as in Boas, each student would be asked to select some language other than English as the topic of study and would be guided through crafting a description of it in much the same way as a field linguist would do. The major differences between Boas and Boas-L, and the system modifications necessary to produce a teaching-oriented Boas-L, are briefly described below.

User profile. Whereas the user of Boas is a bilingual speaker of L and English, the user of Boas-L would be a student who may or may not know another language to any degree of proficiency. Therefore, introductory materials and tasks will need to include information on research methods using print resources, informants (e.g., fellow students), etc. That is, this approach to learning linguistics is as applicable to students who know another language as to those who do not, it is only the means of gathering information (introspection vs. research) that will be different.

The time factor. Whereas building a basic language profile in L, including a reasonable-sized lexicon, is expected to take about six months in Boas, students will not have nearly as much time to work with Boas-L. Therefore, smaller tasks would need to be delineated: e.g., creating two inflectional paradigms rather than all productive one; providing one L equivalent for a subset of closed-class lexical items rather than all possible equivalents for all items, etc.

Use of prerequisites. Some subtasks in Boas cannot be carried out until prerequisite tasks have been accomplished. For Boas-L, means of quickly fulfilling prerequisites would need to be developed to permit the system to be used in any way a teacher chooses.

Description of motivation. As we have emphasized, Boas already contains extensive materials to teach users *how* to carry out tasks; what Boas-L would require is supplementary materials explaining *why*. For example, why is the machine learning of rules of flective morphology based upon sample paradigms rather than on rule writing by the user? What does one do with a morphological analyzer once it is built? Can a morphological analyzer generate as well? How are rules of syntax constructed based on the series of questions set to the user? Why are there so many or so few meanings of the word *x* in the lexicon? Why can't a user type in, free-form, all the information he or she knows about L and have the system learn what it needs to from that? Just as the help materials for carrying out tasks in Boas are accessed through progressive disclosure, so too could be these motivation-oriented materials.

Expected results. Whereas a full profile of L is the goal of Boas, the goal of Boas-L would be for the student user to learn linguistic concepts and practices through a real-world task and to produce a language profile of any size and any degree of coverage.

Usage scenarios. Boas-L could be used in various ways at various levels, for example as: 1) a term project for an introductory (or more advanced) linguistics course, with the individually crafted language profile being the goal; 2) the center of a course in linguistics, with class discussion revolving around language-specific and linguistic issues raised by the students' research; 3) a self-contained tool for independent study of linguistics; 4) training materials for field linguists at the graduate level; 5) a tool for carrying out actual field linguistics. For introductory courses, a more coarse-grained description of the language and/or selected use of modules of elicitation can be undertaken. For more advanced courses, or for the training or practice of field linguistics, a more fine-grained system with greater coverage and exploitation of all modules can be undertaken. Although the target audience may seem unusually diverse, the envisioned system would be suited to users of all these profiles by virtue of its configurability, modular structure, open-ended nature, and methods of progressive disclosure.

4. Evaluation of Boas

Boas has undergone continuous informal testing by the authors as well as by students and colleagues at various stages of its development. Students at the 1999 CRL Language Technologies Summer School at New Mexico State University, most of whom knew a second language natively or well, created a short profile of that language as a laboratory exercise. Students of the African Languages Center of the University of Maryland Eastern Shores used the system to develop profiles of Yoruba and Ibu, and a student at Purdue University used the system as part of a linguistically-oriented introduction to Swahili. The drawback of most of these tests is that time did not permit students to read and absorb all of the instructional materials; so although most tasks were understood by most users, the work would have been easier and fewer questions would have arisen if time had permitted unhurried use of the materials.

The student comments, in conjunction with comments from colleagues who have viewed and tested the system, led to changes including:

- improving the look and feel of the interface;
- developing a map of the system that previews what types of information are elicited at what points in the process; this was a point of

concern for many users, who would think of a phenomenon and would either want to provide information about it immediately or would fear that the system would never get to it;
- extending explanatory materials to target particularly difficult issues; for example, in some cases it is possible to provide the same information in more than one place, in which case the user can choose to provide it in one module, the other module, or both;
- demoting some explanatory materials to links rather than permit them to occupy valuable screen space;
- expanding the elicitation of agglutinative morphology in specific ways;
- augmenting the inventory of parameters and values;
- fundamentally redesigning the open- and closed-class interfaces to increase speed of acquisition (see McShane, Zacharski and Nirenburg 2003 for a description of lexical acquisition in Boas).

5. Challenges for Conversion

The work involved in converting Boas to Boas-L is of two types: content-oriented and programming-oriented. The content side of the work is well understood, with modifications including: a) adding materials to explain the motivation behind tasks, b) adding explanations of NLP-oriented biases and limitations of the systems (e.g., why the morphology-learning program operates on single-word but not multi-word entities), c) incorporating links to selected resources on the Web, d) adding elicitation pages for phenomena that, although not currently automatically processable by Boas (and therefore not elicited there), will be useful for a description of L – e.g., a larger inventory of syntactic constructions, e) incorporating more summary functions to permit users to print out various views of their nascent profile, and f) incorporating options for each type of task depending upon student level, time allocated, etc.

The programming side of the work poses more significant challenges. Most importantly, the system must be further tested and debugged to make it sufficiently stable for classroom use. In addition, we would like to replace our current morphology learning program with one that has generation capabilities in order to provide students with more useful feedback about the validity of their paradigm-oriented generalizations. One version of Boas actually had such capabilities (Oflazer, Nirenburg and McShane 2001), but it engendered prohibitive distribution costs. Yet another technological challenge involves the "smartness" of redo capabilities. The task tree is a hierarchy whose tasks have predefined prerequisites. If an early task is redone, and the results of that task affect the later path of elicitation, planning for the smartest possible behavior the system in each potential scenario is both a research and an implementational challenge. Finally, the system must be prepared to accept any type of writing system—currently, only alphabetic scripts (using any character set) are permitted.

As Boas continues to develop as a KE system to support NLP, all enhancements could be imported into Boas-L since both systems are modular and extensible.

6. Benefits of this Approach

We see many potential benefits of using Boas-turned-Boas-L to teach linguistic concepts and practices:

Prepare students to work as linguists. Scientific work is creative, largely exploratory, and rarely of the concrete-sequential type found in traditional methods of teaching linguistics. Boas permits students to explore linguistics on the whole by carrying out a task of the type that might encountered in real-world linguistics, a task requiring research, creativity and self-motivation.

Permit a customized, user-modeled approach to problem solving. The problem posed by Boas is to describe some language to some degree of detail. The language chosen, the means of gaining information about it, the grain-size and completeness of description, the use (or not) of help materials available in the system—all of this is up to the individual student.

Promote a flexible definition of success. Some tasks are harder than others. For example, describing French using a combination of available grammars, textbooks, native-speaker informants, and 12 years of prior study is easier than describing Swahili, with its very limited resources, the difficulty in finding a native speaker (perhaps only over e-mail), etc. Success, therefore, must be judged relative to the difficulty of the problem.

Promote creative, integrated research techniques. Knowledge about L may be elicited from native-speaker informants living on campus or in the community, from guides written by field linguists, from formal grammars, from Web resources, from e-mail correspondence with students overseas, or from any number of other resources. The specific learning experience of each student will be unclear from the outset, as is always the case with science. For example, some informants may be more insightful,

some less so; for some languages much information may be available, for others less; in some instances one or two sources of information may suffice, in others creative research methodologies might have to be used. The search is, in large part, the point.

Foster organizational skills. A crucial aspect of research work is organizing and evaluating data. Gathering information from diverse sources and presenting it in the structured form that the system like Boas-L can accept requires analysis and organizational skills.

Promote collaborative work. Creating a language profile is, of course, only part of the goal of this system. It should also lead to discussions of the material (in class or outside of class), cooperative work among students who may be working on similar types of languages, interaction with faculty (e.g., a professor of the language on which a student is working or a professor of some other discipline who comes from the given country), establishing contacts with native speakers of the language over e-mail, etc. Most of science is, after all, collaborative, if not directly then indirectly.

Encourage students to think globally. The teaching of linguistics always includes cross-linguistic aspects, e.g., textbook exercises with data suites from diverse languages. Boas will push the global angle further by encouraging students to look at linguistics from the perspective of a language other than English.

Promote conversation about linguistics as used in information technology. Information technology and the science of linguistics are developing simultaneously, informing each other at every step. The challenges of natural language processing are pushing linguists to find solutions to extremely complex problems, problems that are driving linguistic inquiry in a different direction than it might otherwise have taken. For example, ontological semantics (Nirenburg and Raskin 2003) has developed in an attempt to answer the need for textual "understanding" in natural language processing (NLP), and Lexical-Functional Grammar is a theory that has always had practical application as a goal. Rather than hide the application-specific aspects of Boas, we will highlight them in the context of introducing students to thinking about linguistics suited for practical applications.

Advance knowledge about linguistics. Boas not only permits students to learn what is known, it encourages them to discover what is not known (e.g., when describing a little-studied language), to frame known facts in a different way (e.g., deciding on new paradigm delineations for French verbs), and to seek out new evidence of known facts from wide-ranging resources (e.g., if a student knows L quite well, he or she can read texts written in L and collect from them examples of syntactic structures). Students' findings may be non-trivial, in fact, they could be the beginning of descriptive work that has bearing on the field.

Introduce students to the challenges and limitations of informational technology. We live in a world quickly being changed by information technology. Many if not most students use the Web regularly and, no doubt, find themselves frustrated with the fact that search engines, for example, do not work ideally well. The NLP aspects of Boas-L will provide insights into the challenges and limitations of things like search engines, machine translation systems, voice-activated systems, etc.

7. Further Implications

One goal in writing this paper is to describe how a particular cutting-edge technology that was developed for a natural-language-processing application could be transformed into a teaching tool. However, even more important, perhaps, are the broader implications of this type of cross-cutting work.

The Boas approach to knowledge elicitation is predicated upon several beliefs: a) people can have valuable knowledge despite lack of expertise in expressing that knowledge using formal means; b) people can be taught to provide that knowledge, without the intervention of a knowledge engineer, by a system that embeds elicitation tasks in pedagogical materials; c) such a system must itself contain extensive information about the realm for which it is eliciting information (in this case, cross-linguistic phenomena and their means of realization). Users of knowledge elicitation systems of this profile not only support practically useful work – in our example, the building of NLP systems – they necessarily learn about formal aspects of the content in the process. This means that such a KE system *is* a teaching tool, whether or not teaching is its explicit goal.

One question that might arise is, since Boas was originally developed to support L-to-English machine translation, why not have students build a machine translation system and test their language profile using machine translation? The answer is that it would be practically impossible to impart to students a sufficiently acute understanding of the state of the art in MT for them to accurately evaluate success versus failure. That is, MT systems produced by the larger system that house Boas are not expected to produce perfect results – the challenges of MT are too great and the field too young. It would also be impossible to craft the perfect "problem set" for

students such that good input would yield good MT results regardless of which language were selected.

The matter of "why not MT?" leads to two broader issues: what can be expected of NLP in the educational realm both in the short term and the in long term, and what is the best use of limited resources for developing realistic and reliable teaching aids?

It is interesting to note that the sections on artificial intelligence and computational linguistics in Chapelle 2001 (pp. 32-36) focus in a large part on developing dialogue systems, with the computer being expected to understand the faulty input of the student and generate a flawless response or correction using natural language. While research in this area is fascinating and both draws from and contributes to other natural language processing tasks (e.g., question answering, machine translation, text summarization), carrying on a free-form conversation between man and machine, especially with imperfect input, is one of the most daunting of all natural language processing tasks and is realistic today and in the foreseeable future only in a severely limited domain of discourse.

Boas-L would represent a suite of expert-level tools that provide significant natural language processing functionality that is available immediately and reliable in its current state of development. As we have said, this system, like Boas, would not immediately be able to do everything that one can envision for it: the machine learning of morphology and syntax could be made more sophisticated, the redo capabilities could be improved, new language phenomena are being discovered by linguists daily and need to be covered, and perfection in interface design is an always elusive goal. However, the NLP capabilities that the system provides are both cutting-edge and feasible.

We have described elsewhere how Boas could be applied to developing resources for teaching languages, especially those less common ones for which few resources are available (McShane 2003a,b, McShane, Zacharski and Nirenburg 2003). The potential for teaching both language and linguistics using a single system (modified in certain ways) derives from its theoretical foundation: the resident formal representation of linguistic reality in terms of parameters, value sets and means of realizing the latter. We hypothesize that any field of study that can be represented using such a schema could support both a KE system and teaching aids of the type we have described here. In other words, we suggest that our means of representing metaknowledge, our knowledge elicitation methodology, and our system design could all be replicated in other realms, giving rise to systems that have both practical and pedagogical use.

The natural question is, are there any realms other than language in which non-experts have special knowledge whose elicitation could serve a practical purpose, thus justifying the building of a KE system that could then inexpensively be converted into a teaching tool? One realm that comes to mind is the ever-growing field of ontology building—that is, the building of formal world models used to support NLP. If, for example, a semantically oriented NLP system needs to process texts about microbiology, its ontology will need to achieve a greater level of detail than a typical knowledge engineer can provide. Accordingly, either the knowledge engineer must work cooperatively with a microbiologist to formalize the knowledge of the latter, or an expert system must be designed to guide the microbiologist through the process of expressing his/her knowledge. If an architecture like the one used in Boas were applied to ontology building, the ontology builder would naturally learn about the structure and function of ontologies while carrying out the practical task of supplying structured world knowledge. And here would be another teaching tool in the making.

It would seem that practically any computer system that is used by experts for practical purposes could be wrapped in pedagogical materials that explain the task, its methods and justification. Such reuse of technologies would not only make financial sense, it would expose students to the real-world work engaged in by scientists and researchers.

References

Chapelle, Carol A. 2001. *Computer Applications in Second Language Acquisition.* Cambridge: Cambridge University Press.

Comrie, B. and N. Smith. 1977. Lingua descriptive questionnaire. *Lingua 42*.

Fromkin, Victoria and Robert Rodman. 1998. *An Introduction to Language.* Sixth Edition. Fort Worth: Harcourt Brace College Publishers.

R.E. Longacre. 1964. *Grammar discovery procedures* .The Hague: Mouton.

Marjorie McShane. 2003a. Applying tools and techniques of natural language processing to the creation of resources for less commonly taught languages. Forthcoming in *IALL Journal*.

Marjorie McShane. 2003b. Redefining 'paradigm' for (computer-aided) language instruction. Forthcoming in *Foreign Language Annals*.

Marjorie McShane and Sergei Nirenburg. 2003a. Blasting open a choice space: Learning inflectional

morphology for NLP. Forthcoming in *Computational Intelligence*.

Marjorie McShane and Sergei Nirenburg. 2003b. Parameterizing, eliciting and processing text elements across languages. Under review at *Machine Translation.*

Marjorie McShane, Sergei Nirenburg, Jim Cowie and Ron Zacharski. 2003. Nesting MT in a linguistic knowledge elicitation system. Forthcoming in *Machine Translation.*

Marjorie McShane, Ron Zacharski and Sergei Nirenburg. 2003. User-extensible on-line lexicons for language learning. Under review at *CALICO*.

Napoli, Donna Jo. 1996. *Linguistics: An Introduction.* New York: Oxford University Press.

O'Grady, William D., Michael Dobrovolsky and Mark Aronoff. 1997. *Contemporary Linguistics: An Introduction.* Third Edition. New York: St. Martin's.

Oflazer, Kemal, Sergei Nirenburg and Marjorie McShane. 2001. Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics* 27(1).

Sergei Nirenburg. 1998. Project Boas: A linguist in the box as a multi-purpose language resource, *Proc. 1st International Conf. on Language Resources and Evaluation,* Grenada, Spain, 1998.

Sergei Nirenburg and Victor Raskin. 2003. *Ontological semantics*. MIT Press. Forthcoming.