Natural Language Engineering **10** (1): 57–89. © 2004 Cambridge University Press
DOI: 10.1017/S1351324903003279 Printed in the United Kingdom

Mood and modality: out of theory and into the fray

MARJORIE MCSHANE, SERGEI NIRENBURG

Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA
e-mail: {marge,sergei}@umbc.edu

RON ZACHARSKI

New Mexico State University, Box 3000/MSC 3CRL, Las Cruces, NM 88003-8001, USA
e-mail: raz@crl.nmsu.edu

(*Received 10 September 2002; revised 14 March 2003*)

Abstract

The topic of mood and modality (MOD) is a difficult aspect of language description because, among other reasons, the inventory of modal meanings is not stable across languages, moods do not map neatly from one language to another, modality may be realised morphologically or by free-standing words, and modality interacts in complex ways with other modules of the grammar, like tense and aspect. Describing MOD is especially difficult if one attempts to develop a unified approach that not only provides cross-linguistic coverage, but is also useful in practical natural language processing systems. This article discusses an approach to MOD that was developed for and implemented in the Boas Knowledge-Elicitation (KE) system. Boas elicits knowledge about any language, L, from an informant who need not be a trained linguist. That knowledge then serves as the static resources for an L-to-English translation system. The KE methodology used throughout Boas is driven by a resident inventory of parameters, value sets, and means of their realisation for a wide range of language phenomena. MOD is one of those parameters, whose values are the inventory of attested and not yet attested moods (e.g. indicative, conditional, imperative), and whose realisations include flective morphology, agglutinating morphology, isolating morphology, words, phrases and constructions. Developing the MOD elicitation procedures for Boas amounted to wedding the extensive theoretical and descriptive research on MOD with practical approaches to guiding an untrained informant through this non-trivial task. We believe that our experience in building the MOD module of Boas offers insights not only into cross-linguistic aspects of MOD that have not previously been detailed in the natural language processing literature, but also into KE methodologies that could be applied more broadly.

1 Introduction

Mood and modality relate to the linguistic expression of the speaker's attitude toward an utterance – a simple enough notion at this level of abstraction. However, it is extraordinarily difficult to organise into a single, unified system not only the range of mood/modality (MOD) meanings but also their realisations in natural

language. The challenge has, of course, been taken up before, with Palmer (1986) being among the most comprehensive cross-linguistic treatments. But all cross-linguistic works, as well as those specific to a given language or language group, benefit from the opportunity to present evidence selectively – an opportunity that is not available in all linguistic applications. This paper describes a natural language processing application that underscored the cross-linguistic complexity of MOD and necessitated reframing old issues in a new way. As such, it contributes to the field of *applied theory*.

2 Overview of Boas

The need for a practical approach to MOD arose in connection with designing the Boas Knowledge Elicitation (KE) system, whose goal is to guide linguistically naïve speakers of any language (L) through the process of creating a "profile" of that language that is directly useful for natural language processing (NLP) applications.[1] The system is named after innovative descriptive field linguist Franz Boas and seeks to do for 21st century computational field linguistics what he did for 19–20th century "person-to-person" field linguistics. It is easy to perceive a similarity between the task of the Boas system and the work of a field linguist. Both in knowledge acquisition for NLP and in field linguistics there is a special methodology, an inventory of lexical and grammatical phenomena to be elicited (for field linguists, this is organised as a questionnaire of the type developed by Longacre (1964) or Comrie and Smith (1977)), and an informant. There are, however, important differences. Whereas the field linguist can describe a language using any expressive means, Boas must gather knowledge in a structured fashion; and whereas the field linguist often focuses on idiosyncratic ("linguistically interesting") properties of a language, Boas must concentrate on the most basic, most widespread, run-of-the-mill phenomena. The latter is in the spirit of the goal-driven, "demand-side" (Nirenburg 1996) approach to computational applications.

Boas contrasts in significant ways with traditional computer-oriented knowledge acquisition (KA) paradigms as well. Most KA for expert systems is carried out by developers or relies on a personal interview with a domain expert carried out by a knowledge engineer (e.g. see Gaines and Shaw (1993), Motta, Rajan and Eisenstadt (n.d.)). As for automated KE systems, most (like AQUINAS (Boose and Bradshaw 1987) and MOLE (Eshelman, Ehret, McDermott and Tar 1987)) are workbenches that help experts in any domain to decompose problems, delineate differences between possible causes and solutions, etc. Like typical knowledge engineers, such systems have no domain knowledge and therefore focus on general problem-solving methodologies. Other systems permit editing of an already existing knowledge base, with the design of the editor following from a domain model. For example, OPAL

[1] Boas is one component of the larger Expedition System, whose goal is to expedite the ramping up of translation systems from low-density languages (i.e. those lacking computational and perhaps even print resources) into English. This project, carried out at the Computing Research Laboratory of New Mexico State University, was funded by Department of Defense Contract MDA904-92-C-5189.

(Musen, Fagan, Combs and Shortliffe 1987) provides graphic forms for cancer treatment plans, which reflect how domain experts envision such plans, and these plans can be tailored by users. Boas more closely resembles the second model in that it relies heavily on a domain model; however, like the first model, it must also support not entirely predictable types of problem solving, such as analysing language data. An important aspect of Boas is that the task set to users is cognitively more complex than the tasks attempted by many KE systems. For example, the system discussed in Blythe, Kim, Ramachandran and Gil (2001) has a user provide information about travel plans. While the challenges confronting the developers of such a system are formidable (e.g. determining whether it will be less expensive for the person to rent a car or use taxis), the cognitive load on the user is minimal. In Boas, by contrast, the user plays the role of linguist which, even under close system guidance, requires natural analytical ability and much concentrated work.

To increase the practical usefulness of the system, the KE process in Boas was designed to require only about six months' work by one bilingual informant.[2] The resulting language profile, which can be supplemented at any time, is stored in XML format and, as such, can be used in any application. This underscores an important aspect of Boas: although it was originally designed to feed into an L-to-English MT system configured within the Expedition environment, and although that genesis affects some aspects of content and method, Boas is a *free-standing KE system* that can be evaluated on its own merits as well as modified for any NLP application.

Boas leads the informant through the process of supplying the necessary information in a way that is directly usable in computational applications. In order to do this, the system must be supplied with resident (meta)knowledge about language – not L, but language in general – which is organised into a typologically and cross-linguistically motivated inventory of parameters, their potential value sets, and modes of realising the latter. The inventory takes into account phenomena observed in a large number of languages. Particular languages typically feature only a subset of parameters, values and means of realisation. The parameter values employed by a particular language, and the means of realising them, differentiate one language from another and can act as the formal "signature" of the language. Examples of parameters, values and their realisations that play a role in the Boas knowledge-elicitation process are shown in Table 1. The first block illustrates inflection, the second, closed-class lexical meanings, the third, ecology, and the fourth, syntax.

In the elicitation process, the parameters (left column) represent categories of phenomena that need to be covered in the description of L, the values (middle column) represent choices that orient what might be included in the description of that phenomenon for L, and the realisation options (right column) suggest the kinds of questions that must be asked to gather the relevant information.

[2] A programmer is expected to install the system and provide limited types of system support, e.g. assistance in setting up the keyboard and in importing lexicons if they are available. The programmer need not have background in NLP. Thus, the system can be delivered to teams located anywhere in the world and they can carry out all tasks independently.

Table 1. *Sample parameters, values and means of their realisation*

Parameter	Values	Means of realisation
Case Relations	nominative, accusative, dative, instrumental, abessive, etc.	flective morphology, agglutinating morphology, isolating morphology[(a)], prepositions, postpositions, etc.
Number	singular, plural, dual, trial, paucal	flective morphology, agglutinating morphology, isolating morphology, particles, etc.
Tense	present, past, future, timeless	flective morphology, agglutinating morphology, isolating morphology, etc.
Possession	+/−	case-marking, closed-class affix, word or phrase, word order, etc.
Spatial Relations	above, below, through, etc.	word, phrase, preposition or postposition, case-marking
Expression of Numbers	integers, decimals, percentages, fractions, etc.	numerals in L, digits, punctuation marks (commas, periods, percent signs, etc.) or a lack thereof in various places
Sentence Boundary	declarative, interrogative, imperative, etc.	period, question mark(s), exclamation point(s), ellipsis, etc.
Grammatical Role	subjectness, direct-objectness, indirect-objectness, etc.	case-marking, word order, particles, etc.
Agreement (for pairs of elements)	+/− person, +/− number, +/− case, etc.	flective, agglutinating or isolating inflectional markers

[(a)] Inflection is a process used to create new forms of a word when a grammatical value (like person, number, case or tense) changes. Inflection never causes a significant change in meaning. Languages use three basic means of realizing inflectional morphology: flective affixation, agglutinating affixation and isolating words. In flective languages, words consist of one or more morphemes and each morpheme can carry more than one bit of lexical or grammatical information. E.g., the verb form *speaks* is composed of the morphemes *speak* and *s*, and *s* indicates both third person and singular number. In agglutinating languages, like Turkish, words can also be composed of one or more morphemes, but each morpheme tends to carry one bit of lexical or grammatical information. In isolating languages, each word tends to be a single morpheme and morphemes generally do not concatenated to form complex words.

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. This turns out to be a crucial aspect of knowledge elicitation for rare languages, since one must prepare for the case when available informants lack formal linguistic training. The overall KE process is driven in Boas by a combination of system guidance and user initiative. Thus, 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.[3] As a result, it is quite different from most interactive knowledge acquisition tools used in NLP (e.g. Leavitt *et al.* (1994), Nirenburg (1996)).

[3] For further discussion of the architecture of Expedition and, specifically, the methodology of Boas, see Nirenburg (1998), McShane, Nirenburg, Cowie and Zacharski (2003), McShane and Nirenburg (2003a,b), and other publications on the Expedition web site: http://crl.nmsu.edu/expedition.

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 reorganise 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. This approach holds a middle ground between rigid sequencing of tasks and a *laissez-faire* attitude of allowing the user to attempt any of the remaining tasks at any time only to be reminded later that certain prerequisites for that task have not yet been filled. We call the acquisition paradigm exemplified by Boas *knowledge elicitation*.[4]

The KE tasks in Boas are organised in a dynamic task tree, with the status of each task at any given time indicated by the associated icon. A green light means the task may be carried out, a "do not enter" icon means the task has unfilled prerequisites, a coffee cup means it was postponed mid-way through and must be finished, an X means it was deemed inapplicable by the system based on prior user responses, and an hour glass means it is an ancestor task. Figure 1 shows an abbreviated view of the task tree when an informant for some language has paused work on tense during the building of the paradigm template for verbs.

Although this paper concentrates on a specific aspect of morphology, a glimpse of the highest-level subtasks for each of the major modules in Boas will serve as useful orientation into the purview of the system (the fully expanded task tree contains hundreds of subtasks).

Ecology:

- inventory of characters
- inventory and use of punctuation marks
- proper name conventions
- transliteration
- expression of dates and numbers
- list of common abbreviations, geographical entities, etc.

Morphology:

- selecting language type: flective, agglutinating, mixed
- paradigmatic inflectional morphology, if needed
- non-paradigmatic inflectional morphology, if needed
- derivational morphology

Syntax:

- structure of the noun phrases: NP components, word order, etc.
- realisation of grammatical functions: subject, direct object, etc.

[4] There is no universal agreement about the meaning of the terms *knowledge acquisition* and *knowledge elicitation*. We do not attempt to compare and clarify terminological usage beyond stating that elicitation centrally involves system initiative and, therefore, relies on significant amounts of meta-knowledge in the system.

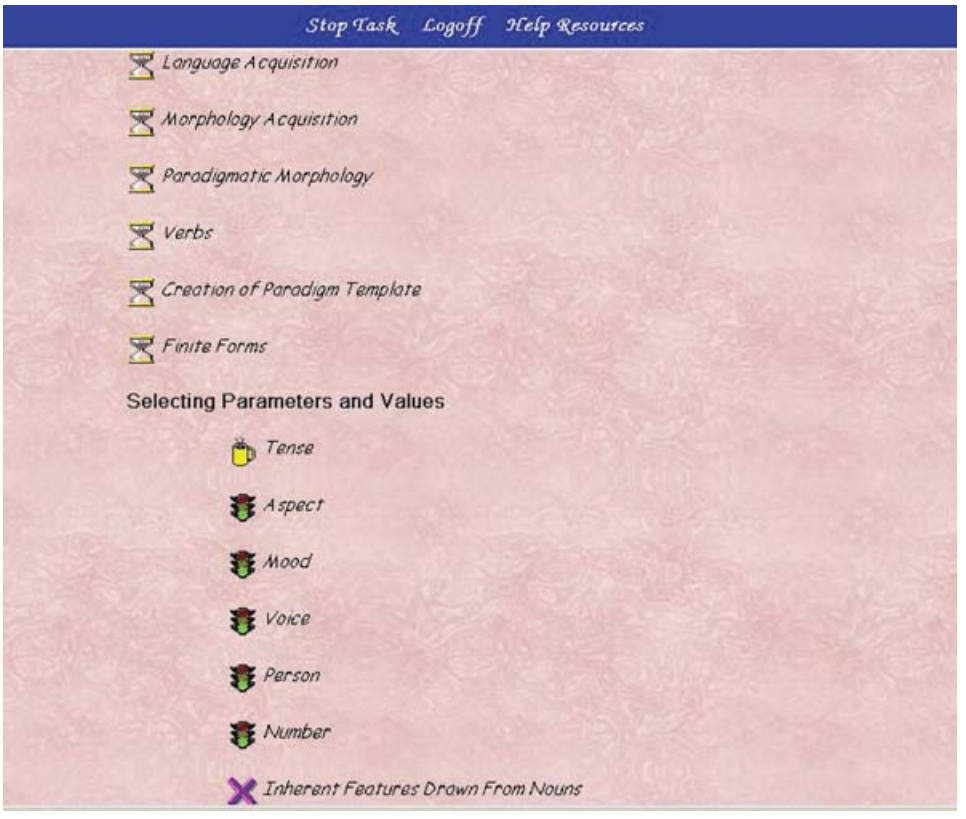

Fig. 1. The task tree in Boas during the creation of the verbal template for a sample language.

- realisation of sentence types: declarative, interrogative, etc.
- special syntactic structures: topic fronting, affix hopping, etc.

Closed-Class Lexical Acquisition:[5]

Provide L translations of some 150 closed-class meanings, which can be realised as words, phrases, affixes or features (e.g. Instrumental Case used to realise instrumental 'with', as in *hit with a stick*). Inflecting forms of any of the first three realisations must be provided, as applicable.

Open-Class Lexical Acquisition:

Build a L-to-English lexicon by (a) translating word and phrase senses from an English seed lexicon, (b) importing then supplementing an on-line bilingual lexicon, (c) composing lists of word and phrase senses in L and translating them into English, or (d) any combination of the above. Grammatically important inherent features and irregular inflectional forms must be provided.

Associated with each subtask are knowledge elicitation "threads", i.e. series of pages that combine questions with background information and instruction. For

[5] See McShane, Zacharski and Nirenburg (2003) for discussion of the lexicons in Boas.

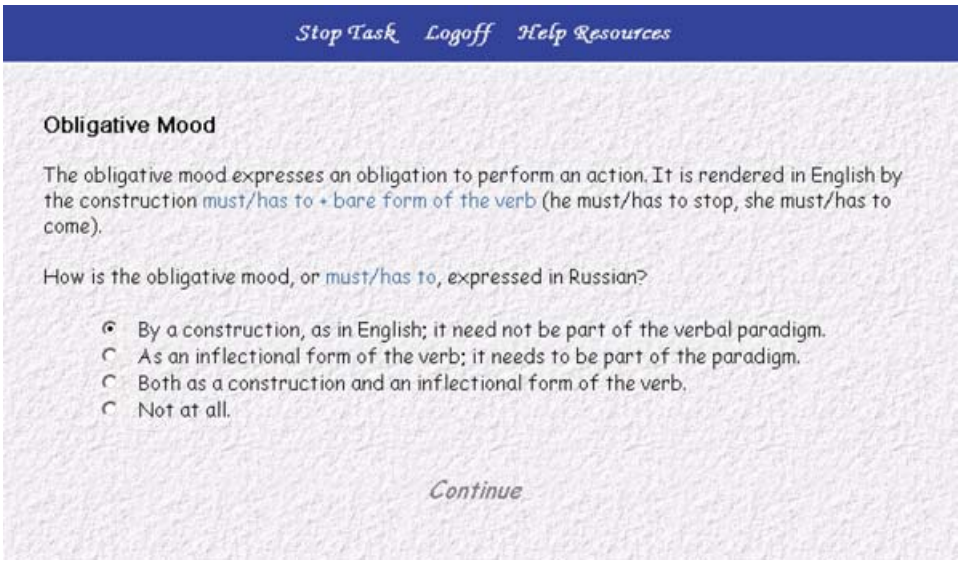

Fig. 2. Eliciting information about the obligative mood in a profile of Russian.

example, figure 2 shows the page eliciting information about the obligative mood during work on a profile of Russian.

Several methods of progressive disclosure are used in Boas to make the interface convenient for informants with different levels of experience: (1) Key terms on elicitation and explanatory pages are hyperlinked to glossary pages; (2) additional help for difficult tasks is available through hyperlinks at the bottom of the associated elicitation pages; (3) the *Help Resources* link in the toolbar provides two means of access to the full glossary – alphabetical and thematically organised. This information, taken together, amounts to an introductory course in descriptive linguistics. Figure 3 shows a view of the screen when a user is consulting the thematically organised glossary; the italicised elements indicate hyperlinks.

2.1 Inflectional morphology in Boas

The elicitation of MOD meanings must be viewed in the larger context of Boas's inflectional-morphology module.[6] Inflectional knowledge about L includes the inventory of grammatical morphemes and their features; the attachment properties of each morpheme (whether it is a prefix, a suffix, an infix or a circumfix; what parts of speech it can attach to; what class of citation forms it pertains to, etc.); and morphotactic rules (e.g. boundary alternations, like dropping English

[6] Inflection is only one of the sources of morphological processes found in natural languages. Others include derivational morphology (e.g. adding *–er* to the English verb *attack* to create the noun *attacker*); affixal realisations of closed-class lexical items (e.g. the definite article attaches to nouns in Bulgarian, so *moreto* 'the sea' is composed of *more* 'sea' + *to* 'the'); and affixal realisations of syntactic elements (e.g. the semantically vacuous French infix *-t-* as in *pleure-t-elle?* 'is she crying?').

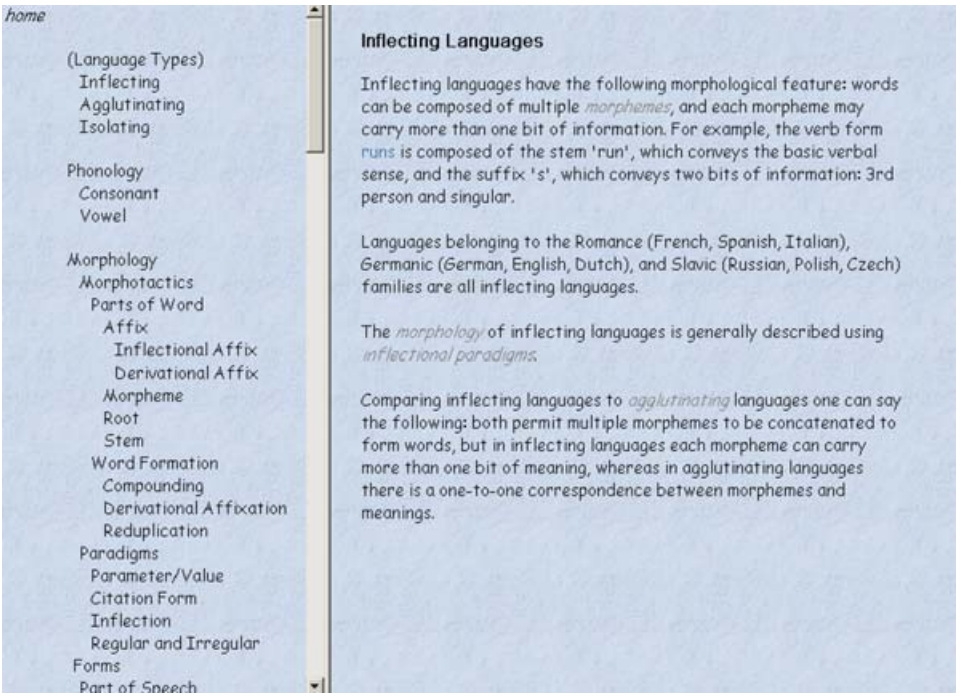

Fig. 3. A view of the thematically-organised access to instructional glossary pages.

-e to form *creating* from the citation form *create*). Inflectional morphology in Boas can be paradigmatic or non-paradigmatic, with both types freely combined in the description of L.

Paradigmatic morphology. For each part of speech in L whose inflectional patterns are best described using inflectional paradigms of well-known Latin type, Boas guides the informant through the process of providing sample paradigms from which a machine-learning program infers rules that are later applied to the whole open-class lexicon.

The process of creating inflectional paradigms involves two steps: creating a paradigm template and filling it with sample words. Creating a paradigm template involves the following major tasks (minor ones are omitted, as are interface-related details): (1) select the relevant inventory of parameters and their values; (2) order parameters (e.g. mood then tense) and values (indicative then imperative; present then past) for best mnemonic effect; (3) select a presentation format for the paradigm (a simple table or multiple tables labelled hierarchically); (4) determine valid combinations of parameter values;[7] (5) check that the template

[7] Creating a straightforward Cartesian product of parameter values is fastest, but if some feature combinations are invalid a longer process of specifying the valid ones must be carried out. This process involves answering a series of questions like: *Which values of case co-occur with the singular?,* the response to which can be 'All', 'None' or any combination of the case values selected earlier.

is correct and, in fact, practically convenient; if not, reorganise it. The idea behind this process is not only to have the informant account for the full inventory of inflectional forms, but also to organise them in a way that he or she finds convenient for later work. Once the template is built, the informant provides key examples to cover all productive patterns of inflection in L, which then act as input to a morphological learning program that generates morphological analysis rules for L.

In different implementations of Boas, different machine-learning approaches were used. The first, more sophisticated one (described thoroughly in Oflazer, Nirenburg and McShane (2001)) compiles the sample inflectional paradigms into a finite-state transducer lexicon combined with a sequence of morphographemic rewrite rules induced using transformation-based learning. The resulting morphological analyser is capable of generating as well as analysing word forms, which permits a highly effective KE methodology that we call the *learning loop*. After creating a paradigm template, the informant provides all the inflectional forms for an example – what we call the *primary example* – for Paradigm 1. The machine-learning program uses this example to hypothesize some morphological rules for this paradigm. The informant then provides the citation form for one or more secondary examples belonging to the same paradigm and the system generates what it expects to be the correct inflectional forms, based on the rules developed on the basis of the primary example. The informant corrects any errors and the program relearns the rules. This iterative process of testing and correcting the learner's rules continues until the informant has presented and tested all the slight inflectional variants s/he wants to be covered by the rules for Paradigm 1. Then s/he carries out the same process for all other inflectional patterns in L. This machine-learning program, which can tolerate virtually any degree of paradigm splitting or bunching by the informant and can both generate and analyse forms, is ideally robust. However, it engenders high distribution costs since it relies on a commercial toolkit. Therefore, for another implementation we developed a less sophisticated machine-learning program that shifts some of the work from the system to the informant. This program requires that the informant posit a stem for the each primary example and check the system's subsequent division of word forms into stem and affixes. On the basis of these manually-approved segmented forms, the program learns to associate given affixes (and, where applicable, morpheme-boundary alternations) with given combinations of parameter values. This latter approach requires that the words in a given paradigm inflect very similarly, permitting only minor bunching of variants in a given paradigm. (See figure 4 for a sample of rules learned using this second method.) Of course, between primitive and sophisticated lies a great expanse of semi-automated options that could be pursued (this is discussed in McShane and Nirenburg (2003b)). The morphology rules generated by both machine-learning programs are stored in XML format in the language profile.

Non-Paradigmatic Inflectional Morphology. This KE thread collects agglutinating affixes or independent words that convey the same grammatical meanings prompted for in paradigmatic morphology. Agglutinating and isolating inflectional units are elicited together because (a) the core inventory of grammatical meanings is the

same, and (b) the method of indicating them is the same: typing strings into text fields. The only difference is that, for affixes, the point(s) of attachment must be indicated.

The backbone of all elicitation of inflectional morphology is the inventory of parameters and values supplied to the user. It is the creation of this inventory, and ensuring that all potential means of realisation can be accounted for, that lies at the centre of the discussion to follow.

3 MOD in Boas

Of the KE methodologies employed in Boas, the one most relevant for a study of MOD is expectation-driven KE, by which the user is prompted to provide information about L based on an inventory of universal and non-universal parameters and values.[8] The notion of parameters and values must be understood in a specifically NLP-related manner, not in the manner of theoretical linguistics (e.g. Chomsky *et al.*'s "principles and parameters"), where the inventory of parameters and their values tends to be far too abstract and limited for realistic large-scale applications.

Organising the world of language into parameters and values assumes the principle of practical effability (e.g. see Raskin and Nirenburg (1998: 200)): what can be expressed in one language can *somehow* be expressed in all other languages, even if the means of realisation differ vastly. For example, even though Chinese lacks tense as an inflectional parameter, time relations *are* expressed unambiguously using other means (discourse clues, adverbials, etc.). Thus, variety in language is, at its base, "surfacy". Accounting for this surfaciness, however, is not trivial even on the descriptive level, with the challenge redoubling under the constraining factors of the Boas environment.

The method for eliciting MOD in Boas shows certain aspects of Anglo-centricity, which is motivated by two considerations. First, practically the only thing we know about the to-be language informant is that s/he must be bingual, with English as one of the languages; therefore using English both as an anchor for the extensive training materials and as a source of comparison was inevitable. Secondly, since Boas was originally developed for MT with English as the target language, some degree of reverse-engineering (with no assumption that MT is directly reversible, which it is not) was incorporated.

In designing the elicitation of MOD for Boas, we constructed an inventory of MOD meanings, and a definition of MOD itself, that takes into consideration the rather diverse work of researchers such as Croft (1990), Huddleston (1984), Jespersen (1963), Palmer (1986) and Quirk *et al.* (1971). However, our approach does not represent a synthesis of these, since achieving synthesis for such a complex

[8] The other two knowledge elicitation methodologies are data-driven (e.g. used in English-driven lexical acquisition) and failure-driven (used to incrementally improve system coverage during system testing).

Table 2. *The top level of Quirk et al.'s classification of MOD in English*

Modals	The intrinsic to extrinsic continuum covered
Can, could, may, might	Permission → Possibility
Must, have (got) to, need (nonassertive), should, ought to	Obligation → Necessity
Will, would, shall	Volition → Prediction (future)

Table 3. *Part of Huddleston's classification of MOD*

	Epistemic	Deontic
Possibility	i. *You may be under a misapprehension*	ii. *You may take as many as you like*
Necessity	iii. *You must be out of your mind*	iv. *You must work harder*

and many-faceted phenomenon would clearly be unattainable. A short comparison of two approaches should be sufficient to illustrate this point.

MOD in English is described in significant detail by, among many others, Quirk *et al.* (1971: 219–239) and Huddleston (1984). However, their descriptions have different points of departure and quite a different topography.

Quirk *et al.* start by saying that each of the major modals has both an intrinsic and an extrinsic meaning. The former involves human control over the event, whereas the latter does not, focusing on human evaluation of the likelihood of the event. However, intrinsic and extrinsic do not represent isolated points but, rather, a continuum. As such, the major modals – listed in the left hand column of Table 2 – can express the range of meanings listed in the right hand column. The details of Quirk *et al.*'s analysis builds upon this foundation.

Huddleston, by contrast, begins his inventory with a distinction between epistemic and deontic modality, whose differences include the following:[9]

1. *Apart from the modality*, the statement in a sentence with epistemic modality is a true or false proposition, whereas the statement in a sentence with deontic modality represents an action.
2. *Apart from the modality*, the time of the statement in a sentence with epistemic modality is generally present or past, whereas it is generally future for sentences with deontic modality.
3. The epistemic and deontic modalities interact differently with negation.
4. The epistemic and deontic modalities interact differently with tense.

Furthermore, both the epistemic and deontic modalities can express both possibility and necessity, as shown in Table 3, drawn from Huddleston (1984: 166).

Although, at base, there may be nothing contradictory in these approaches, it is clear that one cannot simply extract "the best" from each in the hope of creating an

[9] See Huddleston for further explication of all points.

inventory of MOD meanings that will be (i) acceptable to all linguists, (ii) sufficient to cover the phenomena in all languages, and (iii) understandable to an untrained informant.

In the subsections below, we present seven aspects of MOD that affected the form and content of MOD elicitation in Boas, building a bridge between linguistic abstractions and our practical approach to KE for NLP.

3.1 Meanings of MOD

In general, typological categories are based on meaning, yet MOD meanings are somewhat variable across languages. As an introduction to his proposed inventory of notional moods (which excludes the indicative, imperative, and subjunctive) Jespersen (1963: 320) says: "As a tentative scheme of the purely notional ideas expressed more or less vaguely by the verbal moods and auxiliaries of various languages we might perhaps give the following list, to which I cannot, however, attach any great importance. The categories frequently overlap, and some of the terms are not quite unobjectionable."[10,11] Jespersen's pragmatism allows him to push forth to achieve an approximation of "the truth" despite the impossibility of a perfect inventory or an unassailable logic to support its compilation. It is precisely this spirit of practicality that we tried to emulate in developing Boas.

Like Jespersen, we compiled *a* (not the only possible) list of MOD meanings, shown in Table 4, to drive the KE process. This list is intended to act as a point of orientation rather than fodder for semantic hair splitting – which will be out of the depth and scope of interest of most users of Boas in any case.

[10] Jespersen's list of notional moods is as follows (1963: 320–321):

1. *Containing an element of will:*
Jussive: go (command).
Compulsive: he has to go.
Obligative: he ought to go | we should go.
Advisory: you should go.
Precative: go, please.
Hortative: let us go.
Permissive: you may go if you like.
Promissive: I will go | it shall be done.
Optative (realizable): may he be still alive!
Desirative (unrealizable): would he were still alive!
Intentional: in order that he may go.

2. *Containing no element of will:*
Apodictive: twice two must be (is necessarily) four.
Necessitative: he must be rich (or he could not spend so much).
Assertive: he is rich.
Presumptive: he is probably rich; he would (will) know.
Dubitative: he may be (is perhaps) rich.
Potential: he can speak.
Conditional: if he is rich.
Hypothetical: if he were rich.
Concessional: though he is rich.

[11] Jespersen (1963: 321) also notes, incidentally, that "the artificial languages, Esperanto and Ido, very wisely restrict their moods to the number of two besides the indicative, namely what may be called a desiderative . . . and a conditional. . . Otherwise auxiliaries or adverbs are used."

Table 4. *The MOD Meanings in Boas*[1]

Mood name	Function	Example
Indicative	Expresses actions that actually did take place, are taking place or will take place.	Cinderella *married* the prince.
Conditional	Expresses non-factual conditions upon which something else is contingent.	If I *had had* ham, I would have made ham and eggs.
Consecutive	Expresses events that result from other actions or events.	I think, therefore I *am.*
Dubitative	Expresses the probability that someone will perform the given action.	He *will arrive* tomorrow (with the nuance of uncertainty).
Hortative	Functions much like the imperative but usually includes the speaker.	*Let's go*!
Imperative	Expresses requests, orders, or commands.	*Take* my hand.
Inferential	Used when the speaker assumes (based on the available evidence) that what he is saying is true, but he is not absolutely certain.	All the test results were negative, so my headache *must have been* due to stress.
Intentional	Expresses the notion "in order to make action/event X come about".	Boris has to be back at seven *in order* for Stella *to meet* her lover at eight.
Monitory	Expresses a warning.	You *shouldn't* go outside dressed like that.
Narrative, Renarrated, Indirect Indicative	Conveys that what is being said was not personally witnessed by the speaker, and thus the speaker cannot vouch for its truth value.	I *heard that* he *demolished* his car.
Obligative, Deontic	Expresses an obligation to perform the action.	You *must call* Stella, Boris!
Optative	Expresses non-factual events, ones that might take place.	You *might tell* Boris about Stella.
Permissive	Expresses permission to perform the given action	Yes, Boris, you *may call* Stella.
Potential	Expresses someone's ability to perform the given action.	He *can speak*.
Predictive	Indicates that although the speaker is not certain that the event he is speaking of will occur, he thinks it is likely enough to make the prediction.	I *bet that* Stella *will marry* Boris.
Promissive	Expresses promises.	It *will be* done. (with the nuance "I promise")
Subjunctive	A non-factual mood used when the content of the clause is being doubted or supposed rather than definitively asserted. It often occurs in sentences containing a clause in the conditional mood.	If I *were* you, I wouldn't rob that bank.

[1] This inventory of MOD meanings was originally compiled, and many of the examples invented, by Victor Raskin.

The Boas tutorial pages from which this table was extracted explain that the English examples are only approximate and that it is the *nuances* – indicated in the Function column and in the explanatory notes – that lie at the heart of each mood; if such nuances cannot concisely be expressed in L, then L does not use that grammatical mood.[12] The order of presentation in the table is mostly alphabetical, with the exception of the indicative, which has priority since it occurs in all languages.[13]

This inventory, which represents the moods we found well attested in the cross-linguistic literature during the real-time development of the system, was sufficient to describe the approximately two dozen languages tested at various stages of R&D. Further research and testing might suggest adding more moods to the basic inventory. However, even if the inventory lacks a needed value for mood, the user can add it during the KE process. For cross-linguistic applications like the Expedition MT system, meaning can be attributed to user-added values using the translation task described in the next section.

3.2 Mapping MOD between languages

MOD features, like all other features, do not necessarily map one-to-one between languages. Moreover, the meaning of given combinations of MOD, tense, aspect, etc., are not always reached by simple concatenation. While the issue of cross-lingual mappings does not arise if the language profile created by Boas is used in a monolingual application, it does arise for MT. Since Boas originated to serve MT, certain optional transfer tasks were incorporated into the KE process; they are optional in that they are not prerequisites for any KE tasks, although they are prerequisites for configuring the MT system. We mention one such task here because it underscores an important aspect of MOD, if only for multilingual applications.

Since MOD does not ensure neat cross-lingual mappings, for languages that have inflectional paradigms for verbs, Boas replaces the traditional default transfer rules by a translation-oriented transfer method we call "bundling". (For agglutinating languages, no such approach is possible and defaults must be used.) Bundling exploits the intuitions and availability of the bilingual informant by having him or her translate all the entities in a sample paradigm using the best (or most common) English translation. For example, if a Russian informant translates the verb form *s"el* as 'ate', the system creates a correspondence between the bundle of Russian parameter values that describe *s"el* (e.g. indicative, past, masculine, singular) and the bundle of English parameter values that describes the word 'ate', which are resident in Expedition's English morphological resources. This method of deriving transfer rules circumvents potential non-compositionality of parameter-value bundles in L

[12] What is called MOD in this paper is referred to simpy as 'mood' in Boas for the benefit of the novice audience.

[13] Space does not permit the reproduction of all the explanatory materials provided for each mood. The information presented in this table is a snapshot.

and gets to the very heart of the task: teaching the system to translate from L into English. It also levels any terminological imprecision.

Of course, there are many uses of the language profile that are not translation-oriented, like monolingual knowledge extraction, summarisation, and question-answering. If the Boas profile were to be used for any of these, the challenge would shift from the inter-lingual mapping of moods to the necessity of ensuring that all moods attested in L were associated with an appropriate language-independent meaning. To clarify, if an informant building an MT system mistakenly selects the wrong name for a mood, that error will be erased during the translation task, as long as all entities in the L paradigm are translated correctly into English. Similarly, user-added values for mood can be named anything at all and need not be associated with any absolute meaning as long as they are translated correctly. If, however, the language profile is used for a monolingual application, there will be no translation task to assist in the association of meaning with names of moods, and other methods for arriving at meaning will need to be developed.

3.3 Modality vs. mood

If a distinction between modality and mood is made, then modality can be construed as a sentence feature and mood as a verbal feature. However this neat bifurcation hides a practical problem: defining the boundaries of inflection. Whereas some linguists and system developers consider only synthetic (single-word) forms to be part of inflection, others permit analytical (multi-word) forms as well. If one subsumes analytical forms under inflection, a line must then be drawn between inflectional analytical forms and analytical forms that lie outside of inflection – constructions, one might say. Within a given language such distinctions can be made for any reason or for no reason at all. However, when developing a system to cover all natural languages, principled distinctions must be reached and, in the case of Boas, clearly explained.

Consider the problem using an English example, even though English is the one language that Boas would never elicit. For the sake of conceptual simplicity, let us permit analytical inflectional forms: after all, we want a user to be able to say that English has present, past, and future tenses even though the future (e.g. *will go*) can be realised analytically. Focusing on the MOD issue and abstracting away from matters of tense, aspect, person, number, etc., if *will go* (indicative) is inflectional, then so should be *would go* (conditional), *should go* (monitory), *must go* (obligative), *might go* (optative), *may go* (permissive) and *can go* (potential) – an inventory of inflectional MOD that is far larger than one generally finds in English grammars. Moreover, mixed moods should be inflectional too, like *would have been able to go* (potential conditional). And if *that* is a mood, then so should be the hortative form *let's go* (of course, there is a linguistic reason why at least the latter analysis is not legitimate: a direct object occurs in the middle of the construction). Explaining where the line between inflection and non-inflection should be drawn, especially to an untrained informant, seems completely infeasible, which led to our collapsing mood and modality into a single semantically-oriented MOD and to our eliciting

all information about MOD realisations at the same time, as described in Section 4 and the Appendix.

An alternative approach to eliciting MOD meanings would be to elicit only narrowly-defined inflectional realisations (say, only affixes) in the paradigm-building section, then elicit word-level, construction-like meanings separately. At first, this is in fact what we did, placing modals like *must, may, should* etc., in the English-driven closed-class lexical acquisition. But for the pedagogical reasons described above, this option was ultimately rejected.

3.4 Universal vs. non-universal MOD

There is an inventory of MOD meanings that most languages can express in some "concise" way: either inflectionally, using modal verbs, or using some isolated particles or clitics. There are also certain MOD meanings that are inflectional in some languages but require creative means of expression in other languages (following the principle of practical effability, we assume that all meanings can somehow be expressed in all languages). The case where the source language holds more information than can easily be expressed in the target language has particular import for MT systems, including the one configured through Expedition.

Consider, for example, Bulgarian verbs, whose flective moods include the narrative. The narrative mood indicates the speaker's lack of certainty regarding the content of the information conveyed. Since a complete profile of Bulgarian must include morphological rules for the analysis of narrative-mood forms, this information must be elicited in Boas. However, there is no concise way to transfer this mood into English – it must be done using well-placed, contextually appropriate references to the speaker's lack of certainty. In the Expedition environment, we did not develop transfer rules of this nature, opting instead to omit this and certain other semantic nuances in translation. However, the fact that this particular NLP application does not *fully* exploit some knowledge does not render it useless in the language profile: in the Expedition application, information about all moods can be crucial to support parsing of the source text, and in other applications, the semantic nuances of all moods could be treated more fundamentally.

3.5 MOD and tense

There is a complex relationship between MOD and tense: e.g. some MOD, like the imperative, tend not to have tense at all. In addition, the line between tense and MOD can be indistinct. Palmer (1986) says, "... it can be argued that WILL and SHALL in English are markers of modality rather than tense... because they are members of a clearly defined system of modal verbs." However, such language-specific analyses cannot be supported by a template-based system like Boas: informants are asked to provide tense-oriented information in the tense section and mood-oriented information in the mood section, with this bifurcation based on generalised principles of language description. Then, if inflection is paradigmatic as

opposed to agglutinating or isolating, the necessary restrictions on parameter-value compatibilities (e.g. Imperative has no tense) can be indicated when constructing the paradigm template.

3.6 MOD and sentence type or discourse function

There is also a non-trivial relationship between MOD and sentence type or discourse function. For example, declarative statements are often in the indicative mood and commands in the imperative mood, but there is no mood for interrogatives, and even the first two correlations represent a simplification of reality since a sentence like *You are coming tomorrow* can be a statement, a question, or a command (Palmer 1986: 23–24, 32). Similarly, when used in subordinate clauses, MOD can shift its meaning or lose meaning entirely, functioning as a grammaticalised relic. Such discourse functions of MOD lie beyond the scope of the current implementation of Boas, the main idea behind which is to enable the analysis of tokens in text (i.e. labelling each token with a lexical meaning and grammatical features), and the building of a feature structure.[14] Choosing not to incorporate particularly difficult MOD-oriented problems into this alpha version of Boas reflects a judgment about the best use of limited R&D resources rather than an opinion that these problems are either unsolvable or unimportant.

3.7 Theory vs. practice

After discussing many of the issues mentioned in the preceding subsections, Palmer (1986: 6) says, "It follows from all that has been said that it will often be very difficult to decide what to include and what to exclude from a grammatical study of modality." For Boas, it would be counterproductive to use as a rudder theoretically oriented distinctions like "grammatical study", as contrasted with semantic, pragmatic, or mixed type studies. Instead, theory must defer to the practical task at hand: teasing out of the informant information that will ultimately benefit NLP systems.

4 The Boas approach to MOD

All of the considerations listed above led to the decision to take a maximally semantic, minimally surfacy approach to the elicitation of MOD information. The elicitation process runs approximately as follows.[15] The actual elicitation text, which exemplifies both method and content, is presented in the Appendix.

Part 1: Eliciting MOD meanings that can be concisely expressed in English. In Part I of MOD elicitation, the user is presented with a selected subset of the

[14] Even current systems catered to a given language do not include high-quality discourse-oriented interpretative processors.

[15] I say "approximately" because some interface matters are glossed over in order to avoid extraneous detail.

cross-linguistic inventory of MOD meanings one by one (with indicative being assumed for every language): imperative, conditional, hortative, deontic, inferential, obligative, optative, permissive, potential and monitory. This subset is presented first because English expresses all of these using inflectional forms or common modals, permitting direct transfer. The priority of English is due to the following practical considerations: (a) English is the common language of the interface and, thus, the anchor for explaining phenomena, and (b) if an MT system is built using Expedition, English is the fixed target language. Each elicitation page contains a description of the given mood, an English example, a description of how it would be best to analyse the English realisation (for purposes of orientation) and a list of choices for how L might realise it, as shown in Figure 2.

If the "construction" choice is selected, the information is stored for later use in the constructions subtask, where the user is asked to build up the construction as a concatenation of modals or other words in combination with some form(s) of the main verb. If the inflection choice is selected, the next step depends upon the nature of L. If L is an agglutinating or isolating affix, the user need only list all realisations of the given MOD meaning in a text field. If L is a flective language, all relevant MOD meanings are included in the inventory of parameter values to be used in creating the inflectional paradigm for verbs. The "Both" option activates both subtasks in the KE process. The "Not at all" option should be used if the given meaning cannot be expressed in any of the concise ways we have been discussing.[16]

Part 2: Eliciting MOD meanings that cannot be concisely be expressed in English. Part 2 of MOD elicitation seeks only *inflectional* realisations of the MOD meanings for which English has no convenient means of realisation: the consecutive, dubitative, intentional, narrative, predictive, promissive and, of course, any other values for MOD an informant may want to add. The reason for eliciting this information is to ensure that the user teaches the system to recognise the morphological forms that convey them.

Let us reiterate the factors in favor of this semantically driven elicitation, which starts from an inventory of meanings and elicits any and all realisations of them:

- All MOD meanings are presented at the same place in the system and informants are asked to think about the issue of modality as a whole, abstracting away from the many possible means of realisation cross-linguistically. No principled division between inflectional, modal-verb, and other realisations of MOD meanings is required and the system keeps track of which subtasks a user must complete based on his or her answers to the initial questions about MOD realisations in L.

[16] The user is instructed *not* to provide free-form, prose-like translations for mood meanings. In monolingual applications, this is not required, and in multi-lingual applications it could have negative repercussions. For example, if one decided to realise the narrative mood by a phrase like "I heard that...", every single clause containing a verb in that mood would contain that phrase. Automatically filtering out such repetition is a language-specific matter that extends beyond the current state of the art in NLP.

- The system's known orientation toward English can be exploited. Since we know what expressive means English (as the target language for MT) offers, we use them to reverse-engineer the elicitation of information about L.
- There is no reason why an informant cannot mix and match realisations of MOD meanings: some via constructions, other inflectionally. The matter of right and wrong does not exist in the Boas environment as long as the necessary information is somehow collected and turned into processing rules.
- By having a separate constructions section, rather than just a listing of modal words, particles, etc., we capture the co-occurrence patterns of complex verbal entities: e.g. it is helpful for the system to know patterns like *might have been going* rather than just *might* + some unknown form of the verb *go* (as would occur if modals were simply listed in the closed-class lexicon).

5 The resulting language profile

As we have tried to emphasise, Boas is a free-standing KE system that produces a language profile in XML format that can be applied to any natural language processing application. Its original role of serving an MT system, while affecting certain decisions regarding its content and methods, does not in any way restrict the range of application for that profile.

Since the wide-scale usefulness of the profile depends crucially upon its format, below we present a number of excerpts from files that represent the output of KE modules of Boas devoted to inflectional morphology, syntax, the open-class lexicon, and the closed-class lexicon. Although details of the latter three types of elicitation were not presented in this paper, it is worth illustrating the structural similarity of all types of system output. All of the examples below were drawn from a small profile of Polish produced as part of system testing.

Figure 4 shows an excerpt from the morphological rules learned by the second morphology learning program used in Boas (the simpler one that does not rely on outside toolsets; the rules from the more sophisticated learning program are presented in Oflazer *et al.* (2001)). The example is from a verbal paradigm in Polish, *ruszać* 'to throw'. For the sake of readability, characters with diacritics, which are represented in the file as numbers and/or symbols, have been restored. The illustrative forms, which are the 27th and 29th in the large inflectional paradigm, are: (1) the feminine 3rd plural past indicative form, *ruszały*; and (2) the masculine 1st singular conditional form, *ruszałbym*. Figure 5 shows an excerpt from the XML file that stores information about this inflectional paradigm. Forms shown are the base form (which is also the infinitive), *ruszać* and the third plural past indicative, *ruszały*.

Figure 6 is an excerpt from the XML file containing information about NP structure in L. All possible structures of the NP are elicited by asking the user to select: (a) which categories can exist as free-standing elements in an NP (article, quantifier, adjective, etc.); (b) the ordering of each with respect to the head noun (before, after or either); and (c) which other categories can intervene between each such pairing. These results are then computed to yield the full, much larger inventory

```
// PARADIGM ruszać; "x." indicates parameters
ruszać < GeneralRule;
ruszać =
<  <ruszać-1 x.verb[]> >;

ruszać-1 < GeneralRule;
ruszać-1 =
// An indication of word forms and their associated parameter-values
< ...
  <ruszać-1-27 x.verb[gender: x.feminine, person: x.third, number: x.plural, tense: x.past, mood:
  x.indicative]> |
  ...
  <ruszać-1-29 x.verb[gender: x.masculine, person: x.first, number: x.singular, mood:
  x.conditional]> |
... >
// The morphological alternations. The structure of rules is determined by the framework of our
// morphological analyzer.
...
ruszać-1-27 < GeneralRule;
ruszać-1-27 = <
 <$1=String  "ły">
 x.verb[exp: "$1$ć"]
>;
ruszać-1-29 < GeneralRule;
ruszać-1-29 = <
 <$1=String  "łbym">
 x.verb[exp: "$1$ć"]
>;
```

Fig. 4. An example of morphological rules learned in Boas.

of NP structures in L (some over-generation might occur, but for language analysis, in contrast to generation, this is not expected to introduce undue noise).

Figure 7, also from the realm of syntax, shows the realisations of some syntactic functions in L. These data say that a subject in Polish has Nominative case-marking, and that the direct object can have various types of case-marking: Accusative as the default in positive clauses, Genitive in negated clauses, and Locative, Dative, Genitive or Instrumental if these cases are required by the selecting predicate (so-called lexical or "quirky" case-marking).

Figures 8 and 9 show lexicon entries from the closed- and open-class lexicons, respectively. In Figure 8, the English preposition 'about (circa)' is realised by Polish *okolo,* which takes a genitive-case complement. The English preposition 'of (related to)' is realised by the Genitive case; that is, there is no lexical equivalent in Polish and, instead, the given word (in our example, *England*) is placed in the Genitive case. The closed-class elicitation interface makes it convenient to express such non-lexical realisation options. Figure 9 shows the open-class lexical entry for the verb **rzucać** 'to throw' (discussed above). It has no irregular inflectional forms, which is indicated by the fact that the <Paradigm></Paradigm> tags are empty. This signals that the rules learned during flective morphological elicitation should be applied to this word.

Since Boas was developed without a bias toward any of the current competing grammar formalisms, the data stored in the output files can readily be converted

```
<Lang-desc><Lang-name>Polish</Lang-name><Lang-type>morphology</Lang-type>
<Paradigms>
<Paradigm affixSimilarity="0">
<POS>verb</POS>
<Name>ruszać</Name>          ; citation form for this paradigm
<Table>                      ; paradigms are presented to the user in tabular form
<Row cit="yes" member="">
<Form>ruszać</Form>          ; the infinitive is both the citation form and an
<Seg>                                ; inflectional form that carries features
<Stem frequency="172" name="0" type="regular" weight="4">
rusza</Stem>                 ; the stem is determined by comparing what is common
<Affixes>ć </Affixes>        ; to all inflectional forms
<Sequence>$1+ @1</Sequence>
</Seg>
<Rules>                      ; rules created by the simpler learning program
<Rule type="&">
<LHS> $1+ "ć"</LHS>
<RHS> $1+ "ć"</RHS>
</Rule>
</Rules>
</Row>
<Row cit="no" member="">
<gender>feminine</gender>
<person>third</person>
<number>plural</number>
<tense>past</tense>
<mood>indicative</mood>
<feature>verb</feature>
<Form>ruszały</Form>
<Seg>
<Stem frequency="172" name="0" type="regular" weight="4">rusza</Stem>
<Affixes>ły</Affixes>
<Sequence>$1+ @1+</Sequence>
</Seg>
<Rules>
<Rule type="&">
<LHS> $1+ "ły"</LHS>
<RHS> $1+  "ć"</RHS>
</Rule></Rules></Row></Table></Paradigm></Paradigms></Lang-desc>
```

Fig. 5. Stored data about the inflectional paradigm for the Polish verb *rusza_* 'to throw'.

```
<entry><pos>Positive_Adjective</pos>
<ordering><order>preceding</order>     ; i.e., a Pos. Adj. can precede the Head Noun in a NP
<intervening>Ordinal Numeral</intervening> ; Pos. Adj. + Ord. Num. + Head Noun is valid
<intervening>Positive Adjective</intervening>
<intervening>Comparative Adjective</intervening>
<intervening>Superlative Adjective</intervening>
</ordering></entry>
```

Fig. 6. Stored data about the structure of the NP in L.

to rule formalisms used in any approach (e.g., HPSG, Dependency Grammar) in at least four ways:

1. Grammar writers can read the output files and write appropriate rules based on Boas's structured description with no need to develop their own KE methodologies or spend time working with informants.
2. One can use the results of Boas as input to machine learning techniques, using this structured information to supplement corpus-based techniques.

```
<caseMarked>
<gramCat>Subject</gramCat>
<cases>Nominative </cases>
</caseMarked>
<caseMarked>
<gramCat>DirectObject</gramCat>
<cases>
<Base>Accusative</Base>
<Negative>Genitive</Negative>
<Quirky>Locative Dative Genitive Instrumental</Quirky>
</cases>
</caseMarked>
```

Fig. 7. Stored data about the realisation of syntactic functions in L.

```
<entry>
<EnglishForm>about (circa)</EnglishForm>
<Example>He was born circa 1060 and died about 1118. </Example>
<Realization>
      <wordForm>okolo</wordForm>
      <case>genitive</case>
</Realization>
</entry>
<entry>
<EnglishForm>of (related to)</EnglishForm>
<Example>the king of England</Example>
<Realization>
      <case>genitive</case>
</Realization>
</entry>
```

Fig. 8. Closed-class lexical entries.

```
<Entry><Polish><CitationForm>rzucać</CitationForm>
<Type>word</Type><PoS>verb</PoS>
<Paradigm></Paradigm>                    ; this word has no irregular inflectional forms
</Polish>
<English><CitationForm>throw</CitationForm>
<Type>word</Type><PoS>verb</PoS></English>
<Description>to hurl through the air: He threw the ball</Description></Entry>
```

Fig. 9. An open-class lexical entry.

3. One can write an automatic conversion program from the results of Boas to the desired formalism, which we have, in fact, done as part of the Expedition project.
4. One can use information elicited through Boas to supplement the left-hand and/or right-hand sides of existing rules written in any formalism.

We know that automated conversion of Boas results to a grammar formalism is possible because we developed these capabilities in the Expedition project. The underlying machine translation system for Expedition is based on MEAT, a Multilingual Environment for Advanced Translation that represents linguistic knowledge in typed feature structures (Amtrup and Zajac (2000), Amtrup, Megerdoomian and Zajac (2000), Zajac (1992)). When a user finishes creating a profile of L to

the desired level of detail, he invokes an MT-Build program, which automatically converts the XML files recorded in Boas into the MEAT formalism (for syntactic information, there is an intermediate step of producing PATR-style rules). Since the MEAT formalism is comparable to any of the other well-known formalisms the possibility of automatic conversion of Boas results should apply similarly (for further discussion of MT-Build, see McShane, Nirenburg, Cowie and Zacharski (2003)). While this article focused on the acquisition of mood, the Boas system is used to describe a much larger number of language parameters, values and their realisations (see Section 2, Table 1 and Figures 4–9), making mood only an example of our approach.

6 Evaluation

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.[17] 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 the system to be used in the way it was intended – over a 6-month period of time.

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;

[17] The student is Katherine Triezenberg, working under Victor Raskin.

- 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).

It must be said, however, that the most demanding users were the developers themselves, so no revolutionary changes were made on the basis of outside input.

The most conceptually difficult task for users has been in making generalisations about what constitutes an inflectional paradigm, despite the many iterations of instructional materials developed in response to user feedback. Perhaps the most important addition to those training materials has been the repeated emphasis that making good generalisations from the outset is not obligatory. A user can provide a few, even random, samples of paradigms and then, when the KE system is plugged into an application, see what inflectional forms were unknown and gradually add the necessary information. Practically all users have done better than choosing random samples, but doing so would not be a roadblock for the KE process: after all, making inflectional generalisations is primarily a time-saving measure that circumvents typing out inflectional forms for all words.

Funding limitations have not permitted a full-scale, multi-user field test, nor have there been volunteers willing to spend the requisite time to ramp up a full language profile. However developers have made small profiles, including representative samples of all grammatical phenomena and a limited lexicon, of Polish, Russian and French. All of the necessary inflectional morphological phenomena in those languages were covered, include all needed moods. Phenomena as yet not covered by the system on the whole include: (a) certain less common syntactic phenomena; (b) reduplication as a productive word-formation process (such forms need to be listed explicitly); and (c) certain types of morphotactic rules of word formation, especially for agglutinating languages. Such lacunae result only from restrictions in development time and cost.

In Boas, practical R&D considerations necessitated certain tradeoffs. One that we already mentioned (Section 2.2.) was replacing a very robust but expensive-to-distribute morphology-learning program by a simpler one that required more user input. Further development of this system should pursue alternative options to provide the functionality and ease-of-use of the first machine-learning module with the portability of the second.

Another trade-off that affects flective morphology on the whole was deciding to fix the means of eliciting paradigms by starting from abstract features and moving toward actual inflectional forms of words. An alternative – which could replace or accompany this method – would be to start by having the informant list all the forms of words s/he could think of and then backtracking to the features each represents. With the pros and cons of each method being numerous, and the development costs of producing both being too high, we selected the former but do not exclude the potential benefits of the latter, especially for the most inexperienced of users.

Some trade-offs specific to MOD elicitation were also necessary. As noted in Section 3.6, although we elicited inflectional realisations of all MOD meanings, when we used the information in the Boas profile as a resource for the Expedition

MT system, we did not attempt to generate transfer rules for those that require context-sensitive translations, like the narrative mood. Not translating the narrative mood means a loss of the nuance that the speaker himself/herself cannot vouch for the information. However, in the media-style texts that the Expedition system is intended to process, at the rather coarse grain size of processing realistically expected by a system built in half a year by a non-expert, the loss of this nuance is quite acceptable. We also did not explicitly elicit mood-related information that can be detected by other means. For example, the interrogative mood, which is posited for some languages in some grammars, is not among our inventory of MOD because research showed that all languages have special punctuation for interrogative sentences. Therefore, this system – which is intended to process text, not speech – has an alternative means of knowing that a sentence is interrogative. (Of course, the interrogative mood, if inflectional, could be posited by a user to improve token analysis, but no special transfer rules would be elicited.)

As mentioned above, the task of establishing inflectional paradigms for flective languages is arguably the most conceptually difficult one in Boas. With the ever-developing state of the art of corpus-based machine-learning techniques, one could seek means of simplifying this task for the user. For example, words with a similar stem could be automatically collected and presented to the user, he or she could group them into parts of speech, and a machine learning program could then attempt to learn rules of inflection for each group. The user would still need to attach parameter value descriptions to the forms, as well as provide any forms that were not attested in the corpus, but having such additional options for carrying out language description tasks (at least for languages for which a sizeable enough corpus were available) would be valuable.

The machine learning of syntactic structures is another area in which Boas could be improved (see McShane *et al.* (2003) for discussion). Currently Boas learns rules of syntax by being told: the user answers questions that generate pre-defined types of phrase structure rules. A more ambitious method of learning syntactic rules is being pursued, for example, in the knowledge elicitation system called AVENUE (Probst and Levin 2002). In AVENUE, language informants are asked to translate a large inventory of sentences (currently 850, expected to grow to 10,000) then align the elements in the source and target variants; machine learning then takes over to infer transfer rules. This approach, in contrast to the one being developed in Expedition, shifts a larger proportion of the work from the language informant to the machine-learning engines (Carbonell *et al.* 2002). Finding a golden mean between these two methodologies – exploiting the user's knowledge without overloading him, and exploiting machine learning capabilities without expecting too much of them – should be a fruitful line of further research.

The standard for NLP system evaluation has been dictated by application-oriented systems. For example, when one builds an information retrieval system it is evaluated in terms of precision and recall, and although correct assignment of blame for failures is central for system improvement, it does not affect the score assigned to the results of the system on any given run. Corpus-based and machine-learning methods are

virtually always application-oriented and their current prominence in the field has nurtured our collective expectation for numerical evaluation. Boas, by contrast, is not bound by any specific application and it is therefore awkward to use such metrics to evaluate it. But while evaluating the linguistic foundation and user-friendliness of Boas distinctly resists typical NLP-style methods, evaluating the coverage of the profile is more readily attainable – but only when the profile is plugged into an application. Here again arises the familiar challenge of assigning blame for failures, since deficiencies of the application engines and deficiencies of a particular type of data must be distinguished.

If necessary one *could* develop independent evaluation metrics for submodules of Boas, like its coverage of mood. For example, one could have speakers of various languages mark up texts for mood to determine if the inventory of moods and realisation options presented in Boas were sufficient to cover all examples. The problem with any such task, however, is that text mark-up is fraught with inconsistency even for a given informant, as Mitkov *et al.* (2000) have amply shown for the task of pronoun resolution. In fact, it is arguably more difficult to assign moods to a text than reference relations because the former is theoretically based whereas the latter links real antecedents with their coreferential categories.

Boas approaches language not as a set of surface strings, but in terms of mental representations presented using the vocabulary of parameters, values and realisations. As further research teaches us more about these in their cross-linguistic diversity, that knowledge can and must be incorporated into "smarter", more streamlined and comprehensive elicitation threads.

Mood and modality is a complex linguistic issue, MT is a difficult NLP application, and eliciting knowledge from naïve informants is a delicate matter. Put together, they present a considerable challenge to the development of a system like Boas. This paper has shown one approach to meeting that challenge, which was driven in part by knowledge, in part by creativity, and in part by the necessity to implement a broad-coverage working system, with a view to later, failure-driven improvements.

Appendix

Below is the text of the KE pages devoted to mood for a user whose language requires inflectional paradigms for verbs. Presenting the actual KE threads is the most direct way of showing not only what is elicited but also the methodology employed. For reasons of space, formatting has been modified, functional elements (action buttons, text fields, etc.) are omitted or rendered in simplified form, dynamically generated tables and lists are briefly described, and the alternate paths of elicitation for languages without inflectional paradigms and/or with agglutinating realisations of some inflectional forms are not followed. L is the variable that would be replaced by the language name in a given elicitation process. Tutorial hyperlinks are indicated selectively. The first part of KE about mood occurs when the informant is creating inflectional paradigms for verbs; at this point, inflectional values of tense and aspect have already been selected.

Introduction to Mood

As you'll recall from the section on aspect, there were three possible statuses for any aspect in your language:

- the aspect is used and is part of the inflectional paradigm,
- the aspect is used but is best handled outside of the paradigm, as a type of construction,
- the aspect is not used.

The same three options will be available for moods, but whereas the 'construction' option was available only twice for aspect (for inceptive and cessive), it will be available much more often for mood.

Mood is a feature of the verb that reflects the speaker's attitude toward what he is saying. Of course, in all languages there are lots of ways that a speaker can reflect his attitude toward what he is saying: through intonation, by overtly commenting upon what he says, by raising his voice or swinging his fist... However, here we're interested in linguistic ways of conveying mood. As with aspect, we'll move step by step through the moods.

English has only two full-fledged moods and one historical holdover:

1. The *indicative* mood, which expresses actions that actually did take place, are taking place, or will take place: The hamster *spilled* its food.
2. The *imperative* mood, which expresses requests, orders, and commands: *Don't spill* your food, *eat* it!
3. The remnants of the *subjunctive* mood: If I *were* a hamster, I wouldn't spill my food. Though she *be* the queen, she does not have full freedom of action.

The *indicative* mood exists in all languages, and virtually any language will translate sentences like the following using it: The Earth *revolves* around the Sun. Napoleon *conquered* most of Europe. It *will be* much warmer in the summer. So, we will assume that L has the indicative mood.

Does L have special verb forms for the imperative mood? Yes/No
Does L have special verb forms for the subjunctive mood? Yes/No

Hortative Mood

Now we move on to moods that English can express, but that are not part of an English verb's paradigm because they are construction-like rather than inflectional. In L, these moods may be construction-like, they may be inflectional, or they may not be expressed at all.

The first mood of this type, the hortative mood, is sort of like the imperative but it includes the speaker. It is rendered in English by the construction Let's + bare form of the verb (let's go).

How is the hortative mood, or let's, expressed in L?

- By a construction, as in English; it need not be part of the verbal paradigm.
- As an inflectional form of the verb; it needs to be part of the paradigm.

- Both as a construction and an inflectional form of the verb.
- Not at all.

[*There are similar elicitation pages for the conditional, deontic, inferential, obligative, optative, permissive, potential and monitory moods – all of which are conveyed in English by constructions.*]

More Moods

Below is a list of moods that English does not have special verb forms to convey. That is, although there are ways of expressing such nuances in English, there are no special verb forms or constructions used exclusively for these purposes.

[*Here is a table that is a subset of the one in Table 4; it includes the consecutive, dubitative, intentional, narrative, predictive and promissive moods.*]

If L has special verb forms that are used precisely and exclusively to convey any of these meanings, you'll have to include those verb forms in your paradigm. To do so, write the names of those moods (if any) in the text field below, one to a line. As with aspect, we won't require you to use opaque terminology: use any names you'd like as long as they are memorable, written in English letters, and are each a single word (which may contain capitalisation and/or hyphens). If you write nothing in the text field, we'll just forget about these verbal nuances, as we would if we were working on English. [text field]

Naming Moods

Below is a list of the moods you have chosen to include in your verbal paradigm. The ones that you haven't already named yourself are provided with text fields where you can type in whatever you want to call these moods in the future. [table]

If there are more moods that you need to include that have not been covered here, type their names (using English characters) in the text field below, one to a line. But beware! Not every verbal nuance reflects mood (or aspect) – we haven't even gotten to the voices yet. So, before you type anything, look at the overview of voices [a hyperlink]. [text field]

[*At this point, information for two types of further work has been collected: a) what moods need to be included in the inflectional paradigm, a template for which is being built at this stage of the KE process; b) which moods need to be handled in the later subtask of multi-word inflection. We jump to the latter, since paradigm building has no further special implications for the description of mood.*]

Multi-word inflectional forms: Introduction

Multi-word inflectional forms are composed of some form of the main word plus one or more auxiliaries or other so-called helping words, which will be referred to

collectively from now on as **Aux**. Consider, for example, the following multi-word inflectional forms in English. (Subjects are included just for reference; they are not part of the inflectional forms.)

Example 1
The future tense in English is formed by the fixed-form **will** plus the base form of the main verb (i.e. the infinitive minus **to**).

(Subject)	**Aux.**	**Main verb**
(I, you, he, she, it, we, they)	will	go

Example 2
The present perfect in English is formed by the present simple of **have** plus the participial form of the main verb.

(Subject)	**Aux.**	**Main verb**
(I)	have	gone
(you)	have	gone
(he, she, it)	has	gone
(we)	have	gone
(they)	have	gone

Example 3
The future passive perfect progressive in English is formed by four fixed auxiliaries in a row (**will, have, been, being**) plus the participial form of the main verb.

(Subject)	**Aux.**	**Aux.**	**Aux.**	**Aux.**	**Main verb**
(I, you, he, she, it, we, they)	will	have	been	being	seen

In the pages to follow, you will be asked to describe multi-word inflectional forms based on their components. The following two things, illustrated by the examples above, should be kept in mind:

1. Some Aux. words are fixed in form whereas others inflect based on things like the subject they are used with. In analysing example 1, we would simply say that the future tense is composed of **will** plus the base form of the main verb. It is irrelevant that **will** is, officially, the future tense of **be**; it is much

simpler to write the rule in terms of a fixed-form word since this word does not inflect when used to create the future tense. The same is true of all the Aux. in example 3. In example 2, by contrast, the Aux. does inflect: it is the present simple form of the verb **have**; therefore, this rule should be described as "present simple of **have** plus past participle of the main verb".

2. When dividing up inflectional forms for purposes of rule creation, think in terms of columns. That is, forms containing two Aux. cannot be collapsed into a single rule with forms containing three Aux.; likewise, forms using different Aux. words cannot be collapsed into a single group.

Collecting Aux. Words

In this task, you will create a preliminary list of Aux. words that are used in multi-word inflectional forms. This list should cover the Aux. used for all parts of speech. This list is preliminary because if you should forget to list some Aux. now, you will have a chance to list them later.

Below is a list of the inflectional forms that you said were realised in L by multiple words. In the table, you should list all the Aux. needed to cover all of these forms.

[*Here is a generated lists of inflectional forms of nouns, verbs, adjectives and adverbs that the informant said could be realised by multiple words. They are presented using their parameter-value descriptions, e.g. future indicative first singular*]

In the left-hand column of the table below, list the citation forms of all Aux. used in L. If the form is fixed when used as Aux., click the box in column 2. If the citation form inflects when used as Aux., click the box in column 3. If you need extra rows for more auxiliaries, click on the 'Add a row to the table' button.

Citation form	**It is fixed**	**It inflects**
[text field]	[radio button]	[radio button]

Inflecting Aux.

You will now create inflectional paradigms for all inflecting Aux. The elicitation pages are basically the same ones as found in the closed-class lexicon (i.e., they provide a streamlined way of creating paradigms). When establishing paradigms for Aux. keep in mind that each paradigm should only contain single-word Aux. forms. That is, think in terms of columns, as described earlier: if there are two or more Aux. words, each should be handled separately.

[*We omit the paradigm-creation process for Aux.*]

Grouping multi-word inflectional forms

In the left column below are the combinations of parameter values for {nouns/verbs/adjectives/adverbs, as applicable} that can or must be realised by multiple words in L. In the right column is a series of text fields. The idea is to group the entities on the left according to (i) which Aux. they require, (ii) the form of Aux. and (iii) the form of the main word. [*Interface instructions and more explanation are omitted; examples include the directive to put 'have gone/has gone' in one group, 'has been going/have been going' in another, etc.*] Why this task should not be difficult [a tutorial hyperlink].

[generated list of all multi-word forms, represented by their parameter-values]	[10 text fields, labeled Group 1-10, with the option to add more]

Describing groups

In the bottom frame below are the combinations of parameter values that you assigned to Group 1 and a text field provided for your use as a scratch pad (the information will not be processed). We suggest you write down a couple of examples of this inflectional group for reference as you work. This bottom frame, including your scratch pad, will be available for reference throughout this elicitation.

How many Aux. words are required for Group 1? [pull-down menu with #s 1-10]

[*In the bottom pane is a generated list of, e.g., Indicative Future 1st Singular, Indicative Future 2nd Singular… A Russian informant then might type in, for his own use,* _____ _____, _____ _____ *('will sleep' for 1st and 2nd persons sg., respectively).*]

Describing groups, continued

In the table below, please make a template for the multi-word inflectional forms in Group 1. This means selecting which Aux. in which order they are are used, and where the main word is located with respect to them. (If any of the Aux. inflect, you'll be able to indicate which inflectional forms are used on the next page.) Note that you may only select one position for the main word: if it can occupy various positions, choose the most common one. Then, in the far right column of the table, select the form(s) of the main word relevant for Group 1.

[*For the sake of clarity, we fill this table with an English example rather than present it in abstract. Assume that we have already indicated that Group 1 has one Aux. word and that we have written in our scratch pad, for reference, the forms have gone, has gone. The emboldened words are those that are then selected by the user.*]

Group 1

Main word here	Aux1	Main word here	Main word
○	be **have** must . . .	•	infinitive **past participle** present participle . . .

Group 1: The Full Inventory of Inflectional Forms

The idea of this task is to finalise the full inventory of Group 1 forms. In an attempt to save you some time, the system has guessed which forms of Aux. go with which forms of the head word (this guessing is based on parameter values they share). The system's guesses are shown in the right-hand side of the table, with each component in a pull-down menu. If a guess is correct (i.e. the given forms of the given words create a good multi-word inflectional form), accept it; if any of the forms are wrong, correct them.

[*We omit the interface instructions for the rather complex generated table. The English results would be* have gone {1st sg., 2nd sg., 1st pl., 2nd pl., 3rd pl.}, has gone {3rd sg.}. *When Group 1 is finished, the system cycles through the rest of the groups established by the user, which represents the full inventory of multi-word inflectional forms in L.*]

Acknowledgements

Sincere thanks to Stephen Helmreich, Victor Raskin, Jim Cowie, Igor Drugov, Wanying Jin, Denis Elkanov, Denis Kamotsky, Denis Loginov, Natasha Elkanova and Rémi Zajac for their various contributions. We are grateful also to the anonymous reviewers at JNLE for their many constructive suggestions.

References

Amtrup, J. W., Megerdoomian, K. and Zajac, R. (2000) Rapid development of translation tools: Application to Persian and Turkish. *COLING-2000*, Saarbrücken, Germany.

Amtrup, J. W. and Zajac, R. (2000) A modular toolkit for machine translation based on layered charts. *COLING-2000*, Saarbrücken, Germany.

Blythe, J., Kim, J., Ramachandran, S. and Gil, Y. (2001) An integrated environment for knowledge acquisition. *International Conference on Intelligent User Interfaces*, Santa Fe, NM.

Boose, J. H. and Bradshaw, J. M. (1987) Expertise transfer and complex problems: using AQUINAS as a knowledge acquisition workbench for knowledge-based systems. *Int. J. Man-Machine Stud.* **26**(1): 3–28.

Carbonell, J., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R. and Levin, L. (2002) Automatic rule learning for resource-limited MT. *Proceedings of AMTA*.

Comrie, B. and Smith, N. (1977) Lingua Descriptive Questionnaire. *Lingua* **42**.

Croft, W. (1990) *Typology and Universals*. Cambridge University Press.
Eshelman, L., Ehret, D., McDermott, J. and Tan, M. (1987) MOLE: A tenacious knowledge acquisition tool. *Int. J. Man-Machine Stud.* **26**(1): 41–54.
Gaines, B. R. and Shaw, M. L. G. (1993) Eliciting knowledge and transferring it effectively to a knowledge-based system. *IEEE Trans. Know. & Data Eng.* **5**(1): 4–14.
Huddleston, R. (1984) *Introduction to the Grammar of English*. Cambridge Textbooks in Linguistics.
Jespersen, O. (1963) *The Philosophy of Grammar*. Allen & Unwin.
Leavitt, J. R. R., Lonsdale, D. W., Keck, K. and Nyberg, E. H. (1994) Tooling the lexicon acquisition process for large-scale KBMT. *Proceedings 5th International IEEE Conference on Tools for Artificial Intelligence*, New Orleans, LA.
Longacre, R. E. (1964) *Grammar Discovery Procedures*. Mouton: The Hague.
McShane, M. and Nirenburg, S. (2003a) Blasting open a choice space: Learning inflectional morphology for NLP. *Computational Intelligence* (forthcoming).
McShane, M. and Nirenburg, S. (2003b) Parameterizing, eliciting and processing text elements across languages. *Machine Translation* (submitted).
McShane, M., Nirenburg, S., Cowie, J. and Zacharski, R. (2003) Nesting MT in a linguistic knowledge elicitation system. *Machine Translation* (forthcoming).
McShane, M., Zacharski, R. and Nirenburg, S. MS. User-extensible on-line lexicons for language learning.
Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L. and Sotirova, V. (2000) Coreference and anaphora: Developing annotating tools, annotated resources and annotation strategies. *Proceedings Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2000)*, pp. 49–58. Lancaster, UK.
Motta, E., Rajan, T. and Eisenstadt, M. (n.d.) A methodology and tool for knowledge acquisition. Available at http://citeseer.com.
Musen, M. A., Fagan, L. M., Combs, D. M. and Shortliffe, E. H. (1987) Use of a domain model to drive an interactive knowledge editing tool. *Int. J. Man-Machine Stud.* **26**(1): 105–121.
Nirenburg, S. (1996) The inflexible fickleness of fashion. *IEEE Expert Intell. Syst. & Applic.* **4**:15–16.
Nirenburg, S. (1998) Project Boas: "A linguist in the box" as a multi-purpose language resource. *Proceedings First International Conference on Language Resources and Evaluation*, Granada, Spain.
Oflazer, K., Nirenburg, S. and McShane, M. (2001) Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics* **27**(1).
Palmer, F. R. (1986) *Mood and Modality*. Cambridge University Press.
Probst, K. and Levin, L. (2002) Challenges in automated elicitation of a controlled bilingual corpus. *Proceedings TMI*.
Quirk, R., Greenbaum, S., Luch, G. and Svartvik, J. (1972) *A Grammar of Contemporary English*. Longman.
Raskin, V. and Nirenburg, S. (1998) An applied ontological semantic microtheory of adjective meaning for natural language. *Machine Translation* **13**: 135–227.
Zajac, R. (1992) Inheritance and constraint-based grammar formalisms. *Computational Linguistics I* **18**(2).