Machine Translation **18:** 129–165, 2003.

Parameterizing and Eliciting Text Elements across Languages for Use in Natural Language Processing Systems

MARJORIE MCSHANE and SERGEI NIRENBURG
Institute of Language and Information Technologies, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA., E-mail: {marge,sergei}@umbc.edu USA

Abstract. This paper analyzes the structure and meaning of text elements cross-linguistically and discusses how that information can be elicited from people in a way that is directly useful for NLP applications. We describe a recently developed computer-based linguistic knowledge elicitation system that initiates a new paradigm of knowledge acquisition methodologies for NLP. In particular, we describe the natural language phenomena the system seeks to cover, the approach to knowledge elicitation and its rationale, the elicitation modules themselves, and broader implications of this work.

Key words: knowledge elicitation tools, under-resourced languages, field linguistics, morphology, lexicon

1. Introduction

Most natural language processing (NLP) applications seek to analyze every text element as a combination of lexical meaning and grammatical features, as applicable.[1] Cross-linguistically, many types of entities – stems, inflectional affixes, derivational affixes, etc. – can singularly or in combination form a text element, and any given language uses some subset of these. Creating inventories of such entities is more typical of descriptive, typological and, to a lesser degree, theoretical linguistics than of NLP: after all, most NLP systems are built to cover some specific language(s) to whatever extent is required by the given application. However, if one's goal is eliciting knowledge about any natural language for use in an NLP application, creating a comprehensive cross-lingual inventory of types of text elements and their composite entities becomes an essential preliminary stage of work. Once an inventory of this kind is established, one must develop a practice-oriented approach to organizing linguistic reality, a methodology of knowledge elicitation, and a scheme for turning elicited knowledge into processing rules. All of these challenges were met in development of the linguistic knowledge-elicitation (KE) system called Boas.[2]

We named our KE system "Boas" after renowned field linguist and anthropologist Franz Boas, whose late 19th- early 20th-century taste for innovation we try to

match in a 21st-century environment. Our work started with the formulation of a specific task that responded to the project specification: build a KE system to guide a linguistically naïve speaker of any alphabetic language (L) through the process of providing sufficient information about L to support the automatic ramping up of an L-to-English machine translation (MT) system.[3] This KE system must elicit from the user information about the ecology (writing system, orthographic conventions, punctuation, etc.), morphology and syntax of L, as well as a large bilingual lexicon. The entire elicitation environment, training materials, and means of converting the elicited information into operational static knowledge resources for the MT system must be specified and developed from the outset, with no language-specific adjustments or retrofitting. In other words, all phenomena from all natural languages must (to the extent feasible) be covered, the collected information must be automatically convertible into processing resources, and the elicitation process must be understandable to an untrained informant. Given an initially untrained user, the methodological initiative and a large degree of the responsibility for coverage must rest with the system itself. As the technological solution to the above puzzle should be practical, the informant's time must be used efficiently. If time were not a factor and resources were truly unlimited, one could resort to listing many things – like inflectional and productive derivational forms of each word – rather than generalizing by rules. However, in the real world the informant's time *is* a concern, so the listing option is used judiciously in Boas. To enhance the utility of the system in practical applications, the target KE time was set at six months, which can be increased or decreased as resources allow. The common working language of the interface is English, which not only permits some degree of English-orientation in KE (e.g., using English seed lexicons to drive lexical acquisition and preparing resident transfer rules), but also facilitates the preparation of a vast apparatus of training and reference materials, which amount to an on-line introduction to descriptive 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 an MT system and in field linguistics there is a special methodology, an inventory of lexical and grammatical phenomena to be elicited (for field linguists, this is organized 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 represent the accumulated knowledge in a machine-tractable, 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 phenomena.

Moreover, Boas must target those phenomena that can, in fact, be processed by the underlying NLP system. All this is in the spirit of the goal-driven, "demand-side" (Nirenburg, 1996) approach to computational applications. As a result, in some cases the coverage of language material in Boas is narrower than that in

published grammars of particular languages (e.g., many syntactic, semantic and discourse phenomena are not elicited by Boas because they cannot be expected to be processed by the MT system); however, in other cases the coverage is broader (published grammars are notorious for listing just a few examples of specific phenomena and ending too many lists with an "etc."). Additionally, for certain phenomena Boas adopts a descriptive grain size that is finer than is typical for published grammars aimed at human users, and for certain others, a coarser grain size. For example, even though the German noun *Zentrum* has more than one sense, there is no need to split senses in lexical acquisition through Boas because all of them are translated as English *center*.

The MT orientation does not, however, imply that the resulting language profile is useful only for MT. Instead, the profile, which is stored in XML format, can be used for any application, both within and outside of NLP.[4]

Moreover, if a given application should require more or different knowledge, our modular KE process can be amended accordingly. With a view toward the broad potential applications of knowledge elicited through Boas, this paper will focus on the KE process itself and the language profile it supplies rather than on the particular MT application for which it was originally designed (which is described in McShane et al. in press, a).

1.1. AN OVERVIEW OF BOAS

Boas is used to extract knowledge about *L* from an informant with no knowledge engineer present. In this, it differs from typical expert systems that rely on a personal interview with a domain expert carried out by a knowledge engineer (see, for example, Gaines and Shaw, 1993; Motta et al., n.d.). As concerns automated KE systems, most (like AQUINAS (Boose and Bradshaw, 1987) and MOLE (Eshelman et al., 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 (Musen et al., 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 analyzing 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 by Blythe et al. (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.

In order to lead the informant through the process of supplying the necessary information in a directly usable way, Boas must be supplied with resident (meta)knowledge about language – not *L*, but 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. The inventory takes into account phenomena observed in a large number of languages. Particular languages would typically feature only a subset of parameters, values and means of realization. The parameter values employed by a particular language, and the means of realizing them, differentiate one language from another and can, in effect, act as the formal "signature" of the language. Examples of parameters, values and realizations that play a role in the Boas knowledge-elicitation process are shown in Table I. The first block illustrates inflection, the second, closed-class 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 realization options (right column) suggest the kinds of questions that must be asked to gather the relevant information.

Treating language phenomena in terms of parameters, values and means of realization brings about conceptual and practical benefits. By doing this we are saying, both to ourselves as system developers and to the language informants, that most languages have some formal way of expressing things like tense, possession, spatial relations, etc., and there is a limited inventory of expressive means that they use for doing so. All we need to do is tease out of the informant the way this is done in their language. Using static inventories of choices turns a potentially essay-style question ("How do words in *L* inflect?") into a series of much simpler multiple-choice questions ("Does *L* inflect for tense?" [if yes] "Does *L* inflect for present, past, future, timeless and/or some other tense?"). At each stage of the elicitation process, the informant may choose to add extra parameters or values, should our inventories be incomplete; thus, the guidance afforded by inventories of parameters and values does not impose undue rigidity.

This methodology of organizing linguistic phenomena into inventories of parameters, values and realizations then helping an informant to answer questions about them in *L* is what we call "expectation-driven" knowledge elicitation. This is just one of the three types of KE used in Boas, the others being "data-driven" (as for lexical acquisition, where lists of English words/phrases act as prompts for translation into *L*) and "failure-driven" (which is a repair process to supplement acquired knowledge on the basis of failures in trial runs of the underlying MT system).

In developing Boas we used all the relevant descriptive and typological information available, with no constraints due to a particular theoretical-linguistic framework. For our purposes, issues such as the definition of "word", the line

Table I. Sample parameters, values and means of their realization.

Parameter	Values	Means of realization
Case relations	Nominative, accusative, dative, instrumental, abessive, etc.	Flective morphology, agglutinating morphology, isolating morphology, 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	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

between morphology and syntax, the difference between inflectional and derivational morphology, etc., are extraneous except to the extent that they can help us, in practical terms, to organize the process of elicitation. For example, suppose *L* had a method of pluralizing nouns similar to that of English: add the suffix *-s* to some words and the suffix *-es* to other words, and note some boundary alternations and irregularities. The informant could either choose to create inflectional paradigms for nouns in *L*, in which case the morphology-learning program would learn boundary alternations, or could choose to list *-s* and *-es* as agglutinating affixes then list

all the words with boundary alternations as exceptions in the lexicon. The latter method is not the most time-efficient and is not what most linguists would do, but in the Boas environment it is a viable option.

To summarize, 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. As a result, it is quite different from most interactive knowledge acquisition tools used in NLP (e.g., Leavitt et al., 1994; Nirenburg et al., 1996).

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. 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 fulfilled. We call the acquisition paradigm exemplified by Boas "knowledge elicitation".[5]

The KE tasks in Boas are organized 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 shows an ancestor task that can be returned to at any time. Figure 1 shows an abbreviated view of the task tree when the user is about to begin work on the paradigmatic morphology of nouns.

Although this paper focuses on just one aspect of KE in Boas – gathering sufficient information to enable the full machine analysis of text elements in *L* – relevant series of questions are interspersed throughout the system's modules, making an overview of at least the highest-level subtasks important for orientation. (The fully expanded tree includes hundreds of tasks.)

- 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

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

  - paradigmatic inflectional morphology, if needed
  - non-paradigmatic inflectional morphology, if needed
  - derivational morphology
- Syntax
  - structure of the noun phrases: NP components, word order, etc.
  - realization of grammatical functions: subject, direct object, etc.
  - realization of sentence types: declarative, interrogative, etc.
  - special syntactic structures: topic fronting, affix hopping, etc.
- Closed-Class Lexical Acquisition:[6] Provide *L* translations of some 150 closed-class meanings, which can be realized as words, phrases, affixes or features (e.g., Instrumental case used to realize instrumental *with*, as in *hit with a stick*). Inflecting forms of any of the first three realizations must be provided as well, as applicable.
- Open-Class Lexical Acquisition: Build an *L*-to-English lexicon by (a) translating word and phrase meanings from an English seed lexicon, (b) importing then supplementing an on-line bilingual lexicon, (c) composing lists of words and phrases 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 of these tasks are knowledge elicitation "threads" – i.e., series of pages that combine questions with background information and instruction. If, for example, a Russian informant indicates that nouns in Russian inflect for number, the page shown in Figure 2 will be accessed. Explanatory support for

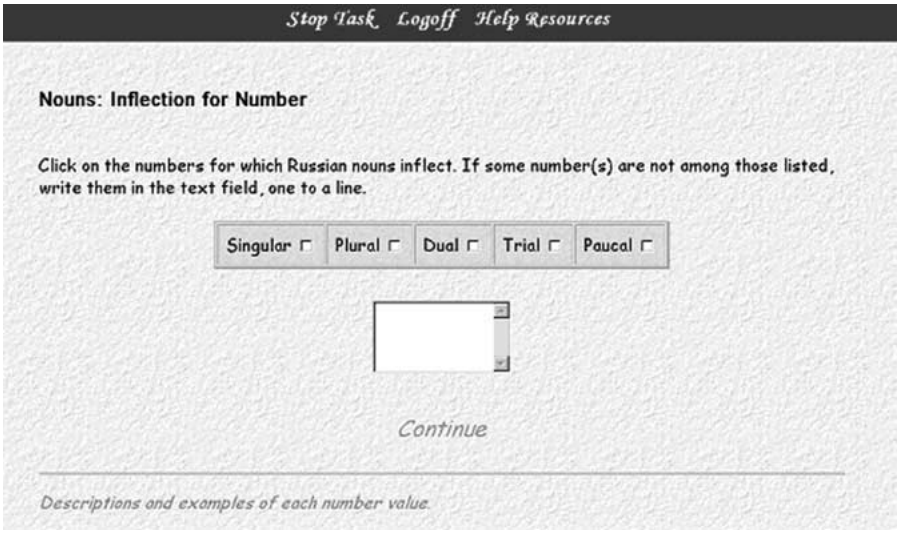

Figure 2. Some pages in Boas elicit information. Here, an informant for Russian is asked to select the values for number for which Russian nouns inflect, having indicated earlier that they do, in fact, inflect.

decision making is provided in help links at the bottom left of the page. This is one means of progressive disclosure, a method of interface design which permits a single interface to serve users with different levels of linguistic experience. Other means of progressive disclosure are hyperlinks to the resident lexicon and numerous optional tutorials and on-line reference sources available through the "Help Resources" link at the top of the page. Thus some pages, like the one in Figure 2, require user input, while others, like the one in Figure 3, are purely pedagogical.

This paper describes the elicitation of information that will permit the machine analysis of text elements in any *L*. The discussion is organized roughly parallel to the path of research and development in the project. First we will present some illustrative language examples gathered during the early period of cross-linguistic research (Section 2). Then we will categorize their morphological phenomena in general terms, without reference to the KE modules of Boas – which at the corresponding point in the development effort were only in the planning stages (Section 3). Next we will describe, by necessity, briefly, the KE modules developed to treat these and many more foreseen and unforeseen linguistic eventualities (Section 4). Finally, we will present the results of evaluation (Section 5) and suggest further implications of this R&D effort (Section 6).

2. An Inventory of Examples

The examples in (1)–(9)[7] illustrate many of the types of text elements that a linguistic KE system like Boas must treat.[8] Text elements are defined here as alphabetic strings (which may include word-level punctuation such as a hyphen

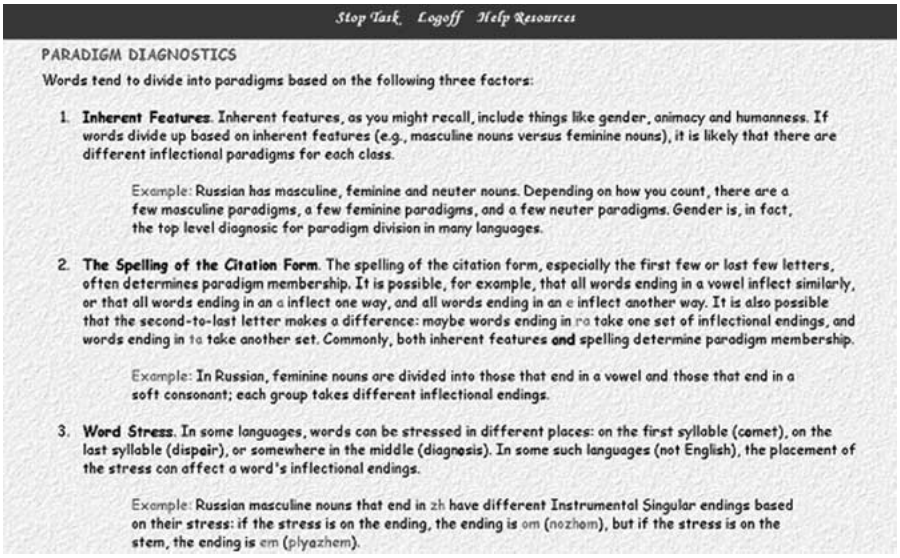

Figure 3. Some pages in Boas are pedagogical. This one explains common diagnostics for paradigm delineation.

or apostrophe) surrounded by white spaces or sentence-level punctuation.[9] The examples are shown accompanied by a literal gloss and fluent translation, and (where necessary) a transliteration for the reader's convenience; Boas accepts input in any alphabetic script (including extended Latin, Cyrillic, Hebrew, etc.). Longer examples are shown first as running text. In the examples, underscores are used to indicate agglutinating, derivational and closed-class affixes, i.e., those affixes that are not flective and can, therefore, be stripped off element by element to reveal a base form.[10] The languages illustrated are French (1), German (2), Russian (3), Ukrainian (4), Polish (5), Turkish (6), Persian (7), and Hebrew (8). Example (9) contrasts several languages as shown.

(1) *Étudie-t-elle maintenant? Non, elle m'attend à l'université.*
'Is she studying now?' 'No, she's waiting for me at the university.'
Étudie- *_t-* *_elle* *maintenant? Non, elle*
STUDY$_{3.SG.PRES}$ particle SHE$_{3.SG.NOM}$ NOW NO SHE$_{3.SG.NOM}$
m' *attend* *à* *l'* *université.*
ME$_{1.SG.OBJ}$ WAIT$_{3.SG.PRES}$ AT THE UNIVERSITY$_{MASC.SG}$

(2) *Nach Angaben der britischen Regierung schlug Blair in einem Brief an die Regierungschefs der Nato-Staaten und an den Russischen Präsidenten.*

'According to statements by the British administration, Blair, in a letter to the heads of governments of the NATO states and to Russian president Vladimir Putin, suggested the Formation of a new Russia-North-Atlantic Council.'

Nach *Angaben* *der*
ACCORDING-TO STATEMENT$_{FEM.PL.DAT}$ OF$_{FEM.SG.GEN}$
brit_isch_en *Regierung* *schlug* *Blair* *in*
BRITISH$_{FEM.SG.GEN}$ ADMINISTRATION$_{FEM.SG.GEN}$ HIT BLAIR IN
einem *Brief* *an* *die* *Regierungs_*
A$_{MASC.SG.DAT}$ LETTER$_{MASC.SG.DAT}$ TO THE$_{MASC.PL.ACC}$ GOVERNMENTS
chefs *der* *Nato-_Staaten* *und* *an*
HEADS$_{MASC.PL.ACC}$ OF$_{MASC.PL.GEN}$ NATO STATES$_{MASC.PL.GEN}$ AND TO
den *russ_isch_en* *Präsidenten*
THE$_{MASC.SG.ACC}$ RUSSIAN$_{MASC.SG.ACC}$ PRESIDENT$_{MASC.SG.ACC}$
Wladimir *Putin* *die* *Bildung*
VLADIMIR PUTIN THE$_{FEM.SG.ACC}$ FORMATION$_{FEM.SG.ACC}$
eines *neuen* *Russland-_* *Nord_atlantik_*
A$_{MASC.SG.GEN}$ NEW$_{MASC.SG.GEN}$ RUSSIA NORTH-ATLANTIC-
rats *vor.*
COUNCIL$_{MASC.SG.GEN}$ IN-FRONT-OF

(3) *Я бы ударила его палкой.*
Ja by udarila ego palkoj.
I$_{1.SG.NOM}$ cond. HIT$_{3.SG.FEM.PAST}$ HIM$_{ACC.SG.MASC}$ STICK$_{INSTR.SG.FEM}$
'I would have hit him with a stick.'

(4)a. *Я буду говорыты тыхше ниж ты.*
Ja budu govoryty tyxše, nіž ty
I$_{NOM.SG}$ WILL SPEAK$_{INFIN}$ QUIETER THAN YOU$_{NOM.SG}$

b. *Я говорытыму тыхше ниж ты.*
Ja govorytymu tyxše, nіž ty
I$_{NOM.SG}$ SPEAK$_{1.SG.FUT}$ QUIETER THAN YOU$_{NOM.SG}$
'I will speak more softly than you.'

(5)a. *My_ śmy znowu wczoraj poszli do parku.*
WE$_{1.NOM.PL}$ 1pl AGAIN YESTERDAY WENT$_{3.PL}$ TO PARK$_{GEN.SG}$

b. *My znowu_śmy wczoraj poszli do parku.*
WE$_{1.NOM.PL}$ AGAIN 1pl YESTERDAY WENT$_{3.PL}$ TO PARK$_{GEN.SG}$

c. *My znowu wczoraj_ śmy poszli do parku.*
WE$_{1.NOM.PL}$ AGAIN YESTERDAY 1PL WENT$_{3.PL}$ TO PARK$_{GEN.SG}$

d. *My znowu wczoraj poszli_ śmy do parku.*
WE$_{1.NOM.PL}$ AGAIN YESTERDAY WENT$_{3.PL}$ 1pl TO PARK$_{GEN.SG}$

'We went to the park again yesterday.'

(6) *(ben) Hasan_ a bavul_ u taşı t_ ti_ m.*
I HASAN dat SUITCASE acc-sg CARRY caus. past 1sg
'I made Hasan carry the suitcase.'

(7) سرمای شدید علی را کشت
Sarma_ ye shadid Ali ra kosht.
COLD$_{SG.}$ ezafe SEVERE ALI obj KILL$_{PAST}$
'A severe cold killed Ali.' ('Ali died of a severe cold.')

(8) כשפגשתיך
keshe_ pagash_ ti_ h_ a
WHEN MET I YOU masc
'when I met you'

(9)a. Irish: *sráid* ~ *an tsráid*
STREET THE STREET
'street' ~ 'the street'

b. Bulgarian: *море* ~ *море_то*
more ~ *more_to*
SEA$_{NEUT.SG}$ SEA THE$_{NEUT.SG}$
'sea' ~ 'the sea'

c. Czech: *ne_ znáte*
NOT KNOW$_{2.PL.PRES}$

d. Tagalog: *bulaklak* ~ *magbu_ bulaklak*
FLOWER ~ VENDOR FLOWER
'flower vendor'

3. Categorizing the Phenomena

Text elements can contain many different types and combinations of entities. Those entities could be analyzed from many perspectives, but we start from a most generic one, relying on canonical, well-known and relatively uncontroversial linguistic tenets. These include the existence of inflectional and derivational morphology (even though the split is not clean); the fact that inflectional morphology can be realized by flective affixation, agglutinating affixation or isolating words; the assumption that certain lexical items are expected to be listed as a citation form in the lexicon whereas other ones can be accounted for by applying regular rules to the citation form; the division of the lexicon into open- and closed-class (grammatical) portions, etc. Below are some descriptive observations about the structure of text elements in our examples. We will use them as a starting point for categorizing the relevant phenomena.

- A text element may contain one stem (Fr. *elle* 'she' (1); Tur. *Hasana* 'Hasan [dat]' (6)) or multiple stems (Fr. *m'attend* 'waits for me' (1); G. *Russland-Nordatlantikrats* Russia-North-Atlantic Council [gen] (2)).[11]
- Stems may represent:
  - open-class elements: nouns (G. *Angaben* 'statements' (2); Ir. *sráid* 'street' (9a)), verbs (Tur. *taşıttim* 'I caused to carry' (6); Per. *kosht* 'killed' (7)), adjectives (G. *neuen* 'new' (2); U. *тыхше* 'quieter' (4)), adverbs (Fr. *maintenant* 'now' (1));
  - closed-class elements: pronouns (Fr. *elle* (1); Pol. *my* 'we' (5)), conjunctions (G. *und* 'and' (2); U. *ниж* 'than' (4)), prepositions (Fr. *à* 'to' (1); Ger. *der* 'of the', *in* 'in', *an* 'to', *vor* 'in front of' (2); Pol. *do* 'to' (5)), articles (Fr. *l'* (1); Ger. *den*, *die* (2)), etc.;
  - inflectional elements: auxiliaries (U *буду* 'will' (4), R. *бы* 'would' (3)), postpositions (Per. *ra* 'obj-marker' (7));
  - onomastic elements: proper nouns (Ger. *Wladimir Putin* (2)), proper adjectives (Ger. *britischen* 'British' (2)).
- Open-class stems may be inflected using synthetic flective inflection (Fr. *Étudie* 'study'-3rd-sg (1); R. *ударила* 'hit'-past-fem-sg (3); G. *Angaben* 'statement'-pl (2)), analytical inflection (U. *буду говорыты* 'will speak' (4)) or agglutinating inflection (H. *keshepagashtiha* (8)).
- Closed-class stems may also be inflected, often in suppletive paradigms (R. *ero* 'him' (3)).
- Inflection may represent syntactic information (Pol. *My* 'we' in (5) is in the nominative case, indicating that it is a subject) or lexical information (R. *палкой* 'stick' in (3) is instrumental singular, with the instrumental case reflecting the closed-class meaning 'with').
- If an element contains multiple stems, the stems may be separated by a hyphen (Fr. *Étudie-t-elle* 'does she study' (1); G. *Nato-Staaten* 'NATO states' (2)), an apostrophe (Fr. *m'attend* 'waits for me', *l'université* 'the university' (1)), or

nothing at all (G. *Nordatlantikrats* 'North-Atlantic Council', *Regierungschefs* 'heads of governments' (2)).

- Multi-stem text elements may contain: two or more open-class stems (G. *Nato-Staaten* (2)) or a combination of open-class and closed-class stems (Fr. *Étudie-t-elle* (1); H. *keshepagashtiha* (8)).
- Derivational word-formation processes that can affect a stem include compounding (G. *Nato-Staaten* (2)), affixal derivation (Cz. *neznáte* (9c); G. *britischen*, *russischen* (2)), reduplication, or some combination of the above (Tag. *magbubulaklak* (9d)).
- Syntax-level word-formation processes, which are sometimes induced by phonetic reasons, include insertion of phonetic elements (Fr. *t* in *Étudie-t-elle* (1)), affixal realizations of closed-class items (Fr. *m'attend* (1); H. *keshepagashtiha* (8)), words formed by inflectional affix hopping (Pol. *myśmy*, *znowuśmy*, *wczorajśmy* (5)), and syntactically determined spelling variants (Ir. *an tsráid* (9a)).

Many of the word-building processes described above can be carried out iteratively, as in the multiple derivations that form the English *antidisestablishmentarianism*. So the examples shown in (1)–(9) represent only a sampling of potentially highly productive processes that must be conceptualized in more general terms.

Descriptive generalizations like those above are only the first step in creating a more principled framework that derives not only from linguistic foundations but also from a reckoning of the application that the results of KE will feed into. That is, nothing is elicited in Boas that cannot be processed in the current (alpha) implementation of the system, and nothing is elicited in a way that cannot be turned into useful static knowledge resources. In the next section we will describe each of the KE modules of Boas followed by an algorithm that shows the path of processing for text elements. The modules and algorithm were, naturally, developed simultaneously.

4. The Knowledge Elicitation Modules

Knowledge about text-element structure in *L* includes: (a) the inventory of grammatical morphemes and their features; (b) the inventory of lexical morphemes and their meanings, the latter being expressed in terms of English for use in MT, although a language-independent model (e.g., one ontologically-based) could be used for other applications; (c) the attachment properties of each morpheme, whether it is a prefix, a suffix, an infix or a circumfix, and what parts of speech it can attach to; and (d) morphotactic rules like boundary alternations (e.g., dropping English *e* to form *creating* from the citation form *create*). The Boas modules that cumulatively cover the above phenomena are paradigmatic inflectional (i.e., flective) morphology, non-paradigmatic (agglutinating or isolating) inflectional morphology, derivational morphology, the closed-class lexicon, the open-class lexicon, and syntax. Developing each of these modules meant not only writing

questions that could be answered using a small inventory of expressive means, it also meant teaching the informant – be he or she an expert or a novice – how to work within this system, a necessary initiation into a mode of thinking that is designed to produce the best results with the least effort. In describing each of the KE modules below, we will indicate which of the text elements in our original list of examples should be treated by information provided in that module.

4.1. PARADIGMATIC INFLECTIONAL (I.E., FLECTIVE) MORPHOLOGY

In this module, the user establishes inflectional paradigms for open-class parts of speech in *L* (nouns, verbs, adjectives and adverbs, as applicable[12]) whose inflectional forms have any of the following properties:

1. They are finite in number (i.e., listable without necessitating thousands of forms per word).
2. They are created using affixes that carry more than one bit of meaning: e.g., for English verbs, *-s* indicates three inflectional parameter values: present tense, 3rd person, and singular.
3. They are formed by a morphological process other than affixation: e.g., Irish "slendering", as in *gas**ur***$_{\text{NOM.SG.}}$ ∼ *gas**uir***$_{\text{GEN.SG.}}$ 'child'.
4. They are marked by word-internal or boundary spelling alternations that cannot easily be generalized, for example:
   - Finnish consonant gradation as in *kau**pp**a*$_{\text{NOM.SG}}$ ∼ ***p**at*$_{\text{NOM.PL}}$ ∼ *kau**p**an*$_{\text{GEN.SG}}$ ∼ *kau**pp**ojen*$_{\text{GEN.PL}}$ 'shop(s)';[13]
   - Belorussian graphotactic vowel reduction as in *cm***о***л*$_{\text{NOM.SG}}$ ∼ *cm***а***ла*$_{\text{GEN.SG}}$ [*st**o**l* ∼ *st**a**la*] 'table';
   - Polish consonant alternations as in *wo**ż**ę*$_{\text{1.SG.PRES}}$ ∼ *wo**zi**sz*$_{\text{2.SG.PRES}}$ 'drive';
   - Blackfoot vowel shortening as in *kakk**óó**wa*$_{\text{SG.}}$ ∼ *kakk**ó**íksi*$_{\text{PL.}}$ 'pigeon(s)'.
5. They are marked by suppletive stems or forms (like English *good* ∼ *better* rather than *good* ∼ **gooder*): e.g., Comanche intransitive verbs are suppletive for singular versus plural subjects, while transitive verbs are suppletive for singular versus plural objects; Blackfoot intransitive verbs have different stems for animate and inanimate subjects: *siksinámma* 'it$_{\text{ANIMATE}}$ is black' ∼ *siksináttsiwa* 'it$_{\text{INANIMATE}}$ is black'.

Boas guides the informant through the process of providing sample paradigms from which a morphology-learning program can infer rules of inflection to be applied to the whole open-class lexicon.[14] This process includes:

– indicating which parts of speech require inflectional paradigms; selecting, for each, the relevant inflectional parameters (number, case, etc.) and their values (singular, plural; nominative, accusative, dative; etc.);
– choosing licit combinations of parameter values (e.g., nominative singular; nominative plural);
– designing a conveniently laid-out paradigm template;

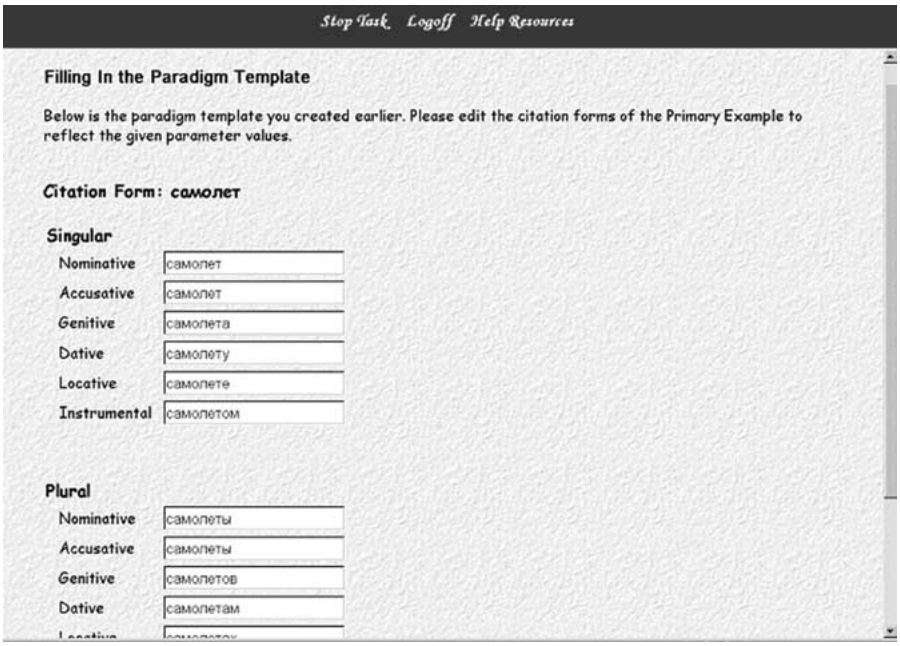

Figure 4. A screen of paradigm elicitation in Boas using a Russian example.

– filling in that template with sample words that represent all productive inflectional patterns in *L* (see the Russian example in Figure 4, which shows part of the paradigm for the noun *самолет* 'airplane').

The reason for asking the informant establish inflectional paradigms, even though this task is conceptually rather difficult and requires extensive instructional materials, is three-fold:

– to free them from having to type all inflectional forms of all inflecting open-class words,
– to have a means of associating inflectional forms with their parameter values, and
– to have rules capable of analyzing unexpected input (e.g., an unknown word ending in *-ed* in English might be assumed to be the past participle of a verb, unless syntactic evidence contradicts this hypothesis).

In Boas, inflectional paradigms can include synthetic (single-word) as well as analytical (multi-word) forms, even though from both theoretical and language-processing standpoints a case can be made for analyzing analytical forms as part of syntax rather than inflectional morphology.[15] However, when one considers the orientation of Boas – both in terms of organizing language phenomena into parameters, values and realizations, and in terms of guiding an untrained informant – there is strong motivation for permitting analytical forms in inflectional paradigms. Consider the evidence of the parameter "tense". If *L*, like English, has three tenses, the information about realizations of those tenses is most easily collected at the

same time. If an inflectional paradigm were limited to synthetic forms, then English verbs would inflect for only some forms of the past tense (e.g., *went* but not *had gone*), only some forms of the present tense (e.g., *goes* but not is *going*), and no forms of the future tense. Another module would have to be built for analytical forms, starting from the same inventory of parameters and values but limiting realizations to multiple words. This would certainly be difficult for an informant, especially if a single combination of parameter values could be realized either synthetically or analytically (like the verb in Ukrainian example (4) above).

Boas *does* have a bifurcation between eliciting synthetic and analytical inflectional forms, but only *after* the entire paradigm template has been built and the cells of the paradigm template are labeled with parameter values (e.g., present singular 3rd simple). At this point the informant is asked how many words are needed to realize each inflectional form. Single-word realizations remain in the core paradigm to serve as input to the morphology learning program, whereas multi-word realizations are postponed until later, where they are built up as combinations of auxiliaries and forms of the head word. Some words in some languages permit more that one realization of a given parameter-value combination. Both variants could be synthetic, as with the so-called second locative in Russian (*лесе*$_{\text{LOC.SG}}$ ~ *лесу*$_{\text{LOC.SG}}$ [*lese* ~ *lesu*] 'forest'), both could be analytical, or there could be a combination, as in our Ukrainian example *говорытымy* vs. *буду говорыты* (4). Boas has facilities to cover all these eventualities

Covering the Examples. The word forms or parts thereof from our examples that should be described using inflectional paradigms are listed below. It deserves note that the citation forms of words in flective languages are considered a member of the paradigm if they are also full-fledged inflectional forms (e.g., the infinitive of verbs in many languages).

- The verb forms in French (*étudie*, *attend*) (1), German (*vorschlug*) (2), Russian (*ударила*) (3), Ukrainian (*буду говорыты*, *говорытымy*) (4), Czech (*znáte*) (9c) and Polish (*poszliśmy*) (5d).
- The nominal forms in German (*Angaben, Regierung, Blair, Brief, Regierungschefs, Nato-Staaten, Präsidenten, Wladimir, Putin, Bildung, Russland-Nordatlantikrats*) (2), Russian (*палкой*) (3), Polish (*parku*) (5) and Irish (*sráid*, but not necessarily *tsráid* – see section 4.5) (9a), which will be entered in the regular lexicon or in the onomasticon (lexicon of proper names), as applicable.
- The adjectives in German (*britischen*, *neuen*) (2) and Ukrainian (*тыхщe*) (4), which will be entered in the regular lexicon or in the onomasticon, as applicable.

4.2. NON-PARADIGMATIC INFLECTIONAL MORPHOLOGY

Non-paradigmatic inflectional units are (agglutinating) affixes or free-standing (isolating) words that relatively freely combine with each other and with stems

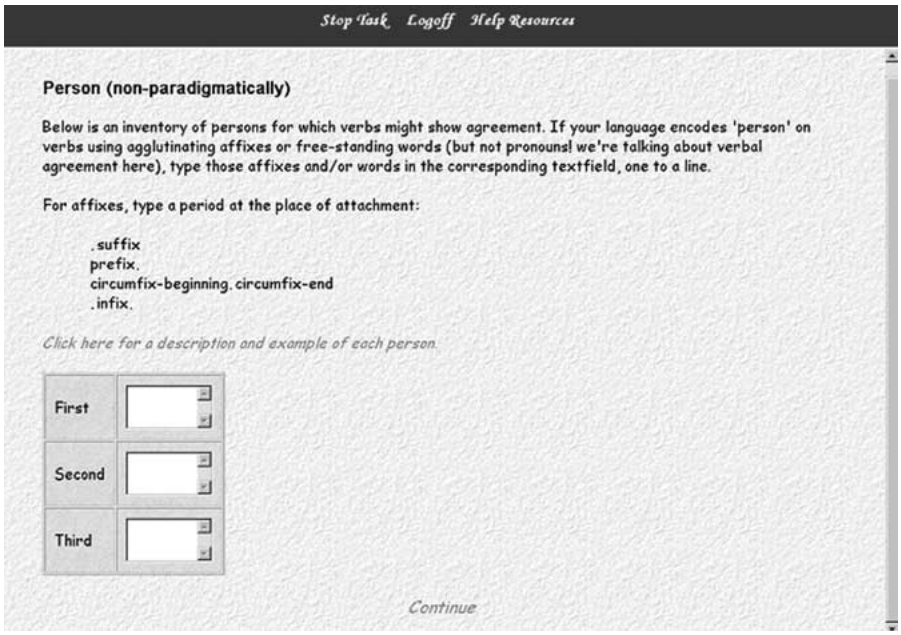

Figure 5. Eliciting non-paradigmatic realizations of the inflectional parameter "person".

to create inflectional forms. In Boas, the abovementioned inventory of inflectional parameters and their values is presented to the informant in tabular form and associated with text fields in which one or more affixes or free-standing words can be entered as realizations. Figure 5 shows the KE page for eliciting agglutinating and isolating realizations of grammatical "person".[16]

Agglutinating and isolating inflectional units are elicited together because the parameter-value prompts are the same and the method of recording realizations of them is the same: typing one or more strings into a text field. The only difference is that for affixes the point of attachment must be indicated. During processing, non-paradigmatic affixes are stripped off in sequence to ultimately yield a base form that is listed in the lexicon.

Covering the Examples. A Turkish informant should provide here the nominal and verbal inflectional affixes shown in (6). A linguistically insightful Polish informant might include the "hopping" affix *śmy* as well (5), since it has agglutinating properties; however, affix hopping will be elicited separately in the syntax module as well. (Some redundancy in the recording of knowledge may occur and will not affect processing, its main shortcoming being non-optimal use of the informant's time.) A Hebrew informant should enter at least the affix *a* (for masculine) here (8), and may choose to enter *ti* '1.sg.' and *h* '2.sg.' as well, since the elicitation process provides for this common bunching of features among agglutinating affixes. Alternatively, the informant may enter *ti* and *h* as affixal realizations of 'I' and 'you', respectively, in the closed-class lexicon. A Persian informant trained

in linguistics might see the similarity between the particle *ra* and the accusative case (7) and thus choose to list that particle here; however, this affixal means of indicating object status will be elicited in syntax as well.

4.3. DERIVATIONAL MORPHOLOGY

Derivational morphology is difficult for machine processing because, both in terms of form and of meaning, simple concantenation often does not obtain. Formwise, adding derivational affixes to words often causes boundary and/or word-internal spelling changes. For example, inexact reduplication in Turkish is used to form the superlative of adjectives that convey intensity of color, as in *siyah* ∼ ***sim****siyah* 'black' ∼ 'very black' and *mor* ∼ ***mos****mor* 'purple' ∼ 'very purple'. Ponapean shows similar formal variations, as evidenced by the following reduplicative forms (leaving the meanings aside): *pa* ∼ *pahpa*, *it* ∼ *itiht*, *alu* ∼ *alialu*. Even if the rules for such spelling changes could be listed, which is possible for some processes in some languages, the semantics of the resulting entity are often not predictable, as derivational affixes are often ambiguous. For example, *-er* in English is typically taken to be an affix that, when attached to a verb, V, produces a noun whose meaning is 'the agent of V-ing'. However, this analysis certainly does not apply to the English word *cooker*.

A common challenge in analyzing derived word forms is ambiguity. Consider, for example, the Swedish surface form *frukosten*, which can have the five analyses in (10) (from Karlsson, 1995: 28).

(10)a.	*frukost* + *en*	'the breakfast'
b.	*frukost_en*	'breakfast juniper'
c.	*fru_kost_en*	'wife nutrition juniper'
d.	*fru_kost* + *en*	'the wife nutrition'
e.	*fru_ko_sten*	'wife cow stone'

Such compounding ambiguities abound in Swedish, and Dura (1998) suggests that the best approach to them is to list the most common compounds explicitly in the lexicon then use these ready-made chunks as set units for the further analysis of compounding forms.

Another complexity of compounding, also well illustrated by Swedish, is that some morphemes are spelled the same in their free-standing and compounding forms, whereas others are not. Compare the examples of saga which in its free-standing form can mean 'Icelandic saga' or 'fairy-tale', which in the first meaning forms compounds as *saga-*, but in the second as *sago-* (from Dura, 1998: 78). In order to prepare a morphological analysis program to trace the compounding form *sago-* back to the citation form *saga*, one would need either to supply the compounding form overtly in the lexicon, to write rules (if they could be formulated)

for common boundary alternations, or to rely on fuzzy matching that will likely, however, produce much noise in analysis.

Another problem inherent in compounding is the opaque semantics of many compounds. For example, a Comanche grammar calls the word for 'Mexican restaurant' a compound composed of the elements 'fat-white-man-*possessive*-eat-house'. Even if Boas could decompose the components of such a compound, it would be unrealistic to expect the analysis engine to arrive at the correct meaning or the English generator to produce a reasonable equivalent. Such semantic non-compositionality affects practically all derivational word-formation processes at least to some extent. As such, Boas trains the informant to use corpus tools, failure-driven methods, and their own insights to create a large enough open-class lexicon to include the most common words in *L* that are created by non-compositional word-formation processes.

However, listing derived words in the lexicon is not a perfect solution since it does not guarantee adequate coverage. For this reason, some derivational morphological phenomena are elicited in Boas, but only those for which there is a realistic expectation of semantic regularity.

The elicitation of derivational affixes is driven by an inventory of some 100 productive derivational affixes found in English, which are grouped into the subclasses negation (*un*, *non*), lesser degree (*mini*), numerical relations (*bi*, *tri*), similarity (*quasi*), temporal relations (*pre*, *post*), etc.[17] This bit of Anglocentricity is justified, we believe, in a KE system that feeds into an *L*-to-English translation system. Affixes like these may attach to one or many parts of speech and may or may not change the part of speech of the word to which they attach – information that is elicited from the informant. A sample elicitation screen is shown in Figure 6.

Some derivational affixes are semantically empty or impoverished and function primarily to change the part of speech. Here, Boas uses English prompts primarily for pedagogical purposes since such processes are rather limited and idiosyncratic in English (e.g., the noun-to-verb change can be realized by any of the affixes marked here in bold, among others: *referr**al***, *polish**ing***, *abdicat**ion***). Each part-of-speech pair is elicited: noun-to-verb, verb-to-noun, noun-to-adjective, etc. Affixes that change the part of speech are rare enough in some languages to suggest lexical listing as a better option, but for truly agglutinating languages, productive analysis of such derivations is essential.

The final KE section of the derivational module of Boas permits any other semantically full affixes in *L* to be listed along with their English translations. The kinds of affixes we expect to be provided here have meanings like: '[when added to a verb] the place where that type of action typically takes place'; '[when added to a noun meaning a good] the seller of that good'; '[when added to a verb] a person typically associated with that action, not necessarily as an agent'. Obviously, in order for the system to translate such affixes, a generic translation must be supplied. We ask for translations using the variable X, like *the place where X typically occurs, the vendor of X, the person typically associated with X*. Translation equivalents

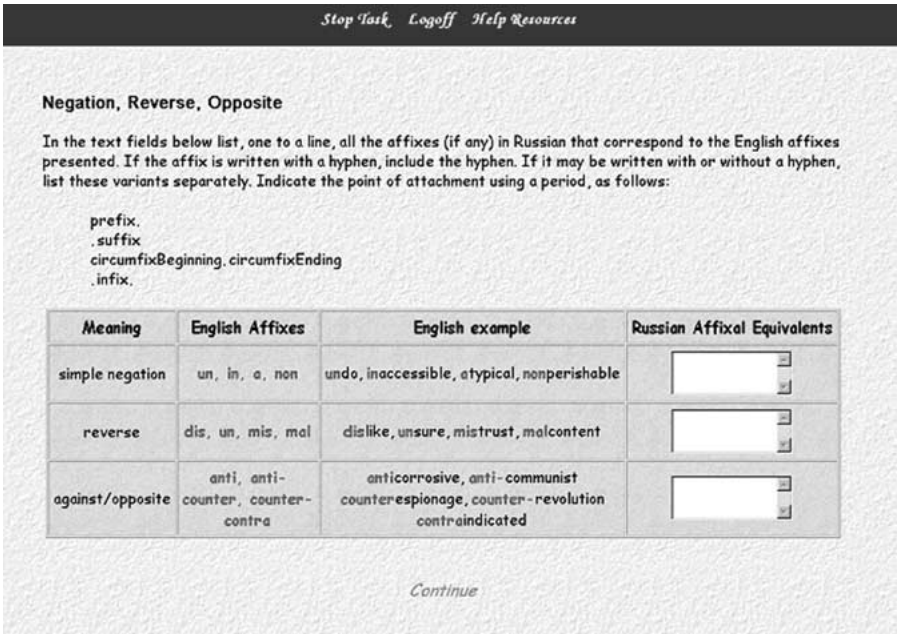

Figure 6. Eliciting productive derivational affixes in Boas.

like this will not produce refined English but they will produce a comprehensible rendering of the meaning that is preferable to no equivalent at all.

Covering the Examples. The instances of derivation from our original inventory of examples fall into two groups, those that Boas elicits for productive analysis and those that Boas does not. Among the first group are the Czech form *neznát* (9c) and the German forms *britischen* and *russischen* (2). The Czech prefix *ne* (*znát* ∼ *neznát* 'know$_{\text{INFIN}}$' ∼ 'not know$_{\text{INFIN}}$') is an example of a productive derivational affix that has a direct English counterpart.[18] During translation, the word-level translation *not* will be selected instead of the affixal translations *non-* or *un-* when the word forms **unknow* and **nonknow* are not found in the resident English lexicon or available corpora. The German forms *britischen* and *russischen* could either be entered in the lexicon explicitly (due to their very common usage) or could be analyzed as noun-to-adjective word formation using the productive suffixes *isch* + *en*.

Among text entities that Boas would not elicit and its associated programs would not analyze are the Tagalog word *magbubulaklak* (9d) and the German derivational compounds *Regierungschefs*, *Nato-Staaten* and *Russland-Nordatlantikrats* (2). Tagalog *magbubulaklak* 'flower vendor' is derived by a combination of reduplicating the first syllable of the base word *bulaklak* 'flower' and adding the prefix *mag-*. It is not trivial to elicit or process (i.e., learn and then automatically analyze at runtime) all the possible variations of exact and inexact reduplication.[19] Moreover, arriving at a translation for such entities can be as difficult as for other derivational

processes, since, for example, a person can be a *flower vendor* but a *car salesman* and a *fishmonger*.

4.4. THE CLOSED-CLASS LEXICON

The closed-class lexicon elicits *L* realizations for a relatively universal inventory of semantic meanings including spatial and temporal relations, conjunctions, numerals, pronouns, etc. Closed-class meanings may be realized in *L* by words or phrases, like open-class meanings, but they may also be realized by affixes or inflectional parameter values. For example,

- the definite article is realized by Bulgarian suffixes, as in (9b);
- the reciprocal 'oneself' can be realized by the Russian suffix -*ся* [-*sja*] as in *мыть* [*myt'*] ∼ *мыть***ся** [*myt'**sja***] 'to wash' ∼ 'to wash oneself', and by the Comanche affix *na-*;
- the demonstrative 'this' can be translated by the Ponapean suffix -*et* as in *wahr* ∼ *wahr**et*** 'canoe' ∼ 'this canoe'.

Feature realizations of closed-class meanings include the well-known use of the instrumental case to indicate instrumental-*with*: e.g., Polish *rewolwerem*, the instrumental singular of *rewolwer* 'pistol', can mean '(shoot, kill, etc.) with a pistol'.

If closed-class items inflect, they often require different paradigms than the ones used for open-class parts of speech. For example, whereas English nouns do not inflect for case, English pronouns do (e.g., I vs. *me*). Moreover, inflectional forms of closed-class items are often idiosyncratic and not subsumed under the same types of broad-coverage rules as open-class items. Because of these special properties of closed-class items, they are elicited using a separate interface in Boas. The elicitation strategy for closed-class items requires the informant to provide the equivalents in *L* of a variety of grammatical meanings presented using English words, phrases and examples. Figure 7 shows a portion of the temporal relations page in a system devoted to Russian. Russian equivalents have already been acquired.

Several features of closed-class elicitation are particularly important for purposes of analyzing text elements:

- There are special means of indicating affixal realizations in the text field, so the single text field can accept word-level, phrasal and affixal realizations of meanings.
- If the entity requires that its complement be in a certain case, which is typical for propositions and postpositions in case languages, that case must be indicated. The inventory of cases presented to the informant is drawn from information provided in the morphology module of the system.
- If some meaning is realized by case-marking alone (e.g., instrumental case to mean 'with'), the text field is left empty and only a value for case is selected.

Figure 7. The closed-class lexicon interface.

– If the entity has inflectional forms, they are collected in a separate elicitation thread accessed by clicking the "Add" button.[20]

Covering the Examples. The closed-class elicitation thread should elicit sufficient information to permit analysis of the following of our examples:

– all pronominal meanings, whether realized as full words (e.g., Fr. *elle* 'she' (1), R. *я* 'I' and *его* 'him' (3), U. *я* 'I' and *мы* 'you(sg.)' (4), Pol. *my* 'we' (5), Tur. *ben* 'I' (6)) or affixes (Fr. *m'* 'me' (1));
– articles, realized as words (Ger. *der*, *einem die*, *eines* (2); Irish *an* (9a)) or affixes (Fr. *l'* (1), Bul. *-to* (9b));
– so-called case relations, like "instrument" (R. *палкой*$_{\text{INSTR.SG}}$ 'with a stick' (3)), "recipient" (Ger. preposition *an*) and *of/by* (Ger. preposition *der*) (2);
– spatial relations, realized in all of our examples as prepositions (Fr. *à* (1); Ger. *in*, *vor* (2); Pol. *Do* (5)) although postpositional, affixal and parameter-value realizations are also possible.

Another point of analysis that will be supported by closed-class information is the fact that the genitive case marking of *parku* in Polish (5) does not represent a semantic meaning, like partitive, but rather is an instance of lexical case marking imposed by the preposition *do*.

4.5. THE OPEN-CLASS LEXICON

The open-class lexicon in Boas is the repository for pairs of *L* and English words and phrases from the major parts of speech – nouns, verbs, adjectives and adverbs – plus proper nouns, adjectives derived from proper nouns, acronyms and abbreviations.[21]

The goal of open-class elicitation is to help the language informant to acquire the best (in NLP terms) possible inventory of complete entries in the shortest time and with the least effort.

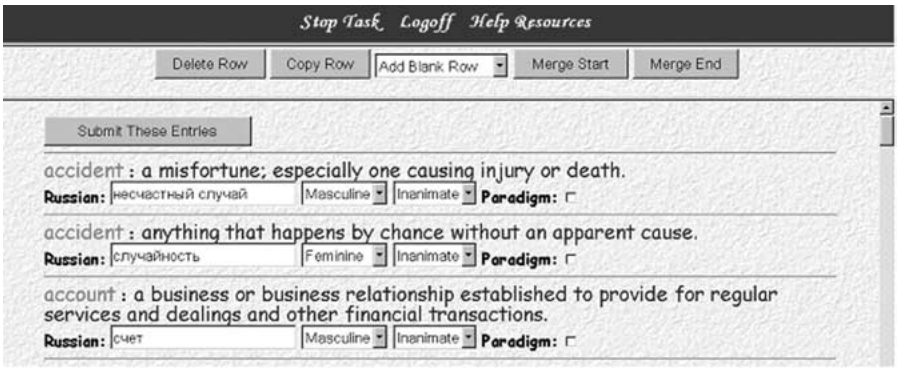

Figure 8. The open-class lexicon interface.

Since Boas is intended for languages for which few or no NLP resources are available, the method of translating lists of word meanings (hereafter simplified to "word lists") is expected to dominate the acquisition process. English-driven acquisition using resident word lists is one option, with the word senses being distinguished using modified Wordnet definitions.[22] Another option is for the informant to translate word lists that they and the programmer have compiled off-line. Such lists can be in *L* or in English, can cover a specific subject area or be generalized, and can be gathered using Boas's corpus tools or any other means. Importation instructions are provided. Working from externally generated lists is highly recommended, at least as a supplement, for languages with widespread derivational word-formation processes like compounding and reduplication since most such forms will not have correlates in the English seed lexicon. Listing *L*-English pairs of common phrasals is also recommended because a large inventory of phrasals considerably improves the performance of MT systems. The goal of presenting all of these options is to cater the acquisition process to the envisioned needs, resources, and preferences of the user.

L entries should be entered in one or more base forms, otherwise known as citation forms, upon which word-formation processes occur. The citation form (or its head, for phrasals) may be a root, a stem or a word, the choice depending on (a) the tradition in *L*, (b) informant preference and/or (c) the convention used in any lexicons or portions thereof that are imported.[23] In addition, the informant must: (a) supply relevant inherent features (e.g. gender), as indicated in the morphology module; (b) list any irregular inflectional forms; (c) for phrasals, mark the head; (d) for entries produced from external word lists, indicate the part of speech.[24] An example of the interface, shown during creation of a Russian language profile, is shown in Figure 8.[25]

Covering the Examples. All of the words in our examples except for those belonging to the closed class and those with purely syntactic function (see below for the latter) must have a corresponding entry in the open-class lexicon. Of course, the lexicon will contain only a citation form and any irregular forms, all other pos-

sible forms being analyzed based on learned rules. There are four types of entries based on their usual means of elicitation in Boas. The examples reflect the English variants of the corresponding words from our original inventory:

– common (not proper) nouns, verbs, adjectives and adverbs, including: *study*, *wait*, *university*, *statement*, *administration*, *hit*, *letter*, *president*, *formation*, *new*, *stick*, *speak*, *quiet(er)*, *again*, *go*, *park*, *suitcase*, *carry*, *cold*, *kill*, *meet*, *street*, *sea*, *know*, *flower*;
– "famous" proper nouns, adjectives and adverbs, some of which are in the English seed onomasticon and others of which can be added as needed: *British*, *Blair*, *Russian*, *Vladimir Putin*;
– "non-famous" proper nouns, adjectives and adverbs, which – if not in the onomasticon – will be transliterated using the transliteration conventions provided by the user: *Hasan*, *Ali*;
– words formed by derivational word-formation processes, which are not included in the English seed lexicon and must be incorporated judiciously based on frequency in *L*: *government heads*, *NATO states*, *Russia North Atlantic Council*, *flower vendor*.

One phenomenon from our inventory requires special comment. The Irish spelling variant *sráid* ∼ ***t****sráid* (9a) exemplifies a productive, language-wide series of word-initial alternations called "eclipsis", which is most often induced by the phonetic form of the preceding word. A similar process with different graphotactic reflexes is called "lenition". While the rules for *generating* eclipsis and lenition in appropriate contexts in Irish are complex, teaching a system to recognize variant forms is not since the problem reduces to a list of predictable alternations (11)–(12).

(11)

Lenition: $c \rightarrow ch$, g $\rightarrow gh$, $t \rightarrow th$, $d \rightarrow dh$, $p \rightarrow ph$, $b \rightarrow bh$, $s \rightarrow sh$, $m \rightarrow$, mh, $f \rightarrow fh$
e.g., ***b****ad* 'boat' → *ar* ***bh****ad* 'on (the) boat'

(12)

Eclipsis: $c \rightarrow gc$, $g \rightarrow ng$, $t \rightarrow dt$, $d \rightarrow nd$, $p \rightarrow bp$, $b \rightarrow mb$, $f \rightarrow bhf$
e.g.,

The most efficient way of preparing to analyze such variant forms would be to write a global lexical rule; however, in the Boas environment this is not expected of the informant-programmer team (it is also not prohibited, should the team be particularly skilled in NLP). Alternatively, one could develop a KE thread in which the informant were asked to list letter variants and the place in the word in which they occur – word-initially or word-finally (word-internal variations would introduce undue complexity). These rules could then be used to supplement the lexicon either prior to or at run-time.

The downfall of global lexical rules is, however, that they are often not truly global. Consider in this respect Ukrainian, which puts *у* and *в* in free alternation word-initially for many words: e.g., *учитель/вчитель* [*učitel/včitel*] 'teacher'. Some words, however, lack the *в-* variant, like place names (*Урал* [*Ural*] 'Urals') and foreign words (*уран* [*uran*] 'uranium'). On the one hand, since the language profile created by Boas is meant to support analysis not generation, this allowance of never-to-be-attested forms might seem irrelevant. On the other hand, it would occasionally introduce spurious ambiguity: e.g., Ukrainian *уклад* [*uklad*] means 'regime' while *вклад* [*vklad*] means 'contribution', so a lexicon-wide rule that put *у-* and *в-* in free variation word-initially will cause each instance of *уклад* and *вклад* to be incorrectly tagged with two meanings. The risks associated with instantiating global lexical rules, and the difficulties in accurately eliciting their restrictions, led us to exclude such a facility in the alpha version of Boas. However, future development could include a routine that would generate all potential variants then ask the user to remove non-existent forms.

4.6. SYNTAX

A KE module devoted to syntax might seem like the least likely place to find information about word structure but, in fact, some languages contain words and/or affixes that have only grammatical (not lexical) meaning, making their elicitation among other syntactic phenomena natural. These include the noun-phrase markers found in languages like Persian and Hebrew, the subject and topic markers found in Japanese, the basic interrogative particle in Polish (*czy*), the interrogative affix in Malay (*-kah*), etc. In addition, case marking often carries grammatical meaning, like indicating subject or object status, which contributes to a full analysis of word meaning. Although the inventory of such entities in any language is frozen, they should not be considered part of the closed-class lexicon because it reflects an inventory of universal *semantic meanings*, whereas grammatical sentence elements are neither universal nor semantically full.

In Boas, syntactic elements in *L* – which can be free-standing words or affixes – are elicited using the same types of expectation-driven methodologies we have been describing thus far. We compiled an inventory of syntactic parameters that include things like subject status, object status, possession, sentence type (e.g., interrogative, declarative), components of an NP, the ordering of components within an NP, etc., and present the user with options regarding how each might be realized in *L*. For example, the syntactic function of an NP might be indicated by case, a particle or word order; possession might be indicated by an affix on the possessor, an affix on the thing possessed, a particle or word order; and so on. Figure 9 shows a screen on which the diagnostics for direct objecthood in Russian are being elicited.

The output of the syntactic elicitation in Boas supports the analysis of text elements inasmuch as it provides an inventory of grammatical words and affixes

and their associated meanings as well as attributes grammatical meaning to the case-marked forms elicited in the inflectional morphology module.

Covering the Examples. The French affixal particle *t* (1), the Persian postposition *ra* and the affixal "ezafe" (*-ye* in *sarmaye* 'cold') (7) will be elicited in the abovementioned types of elicitation threads. In addition, the potential syntactic function of all case marking will be indicated, like the fact that the dative case can be used as the "direct" object of causative verbs in Turkish.

4.7. PHENOMENA STRADDLING MORPHOLOGY AND SYNTAX

Many linguistic phenomena straddle the traditional branches of linguistics, with the morphology-syntax overlap being particularly common. One such phenomenon that we already discussed is analytical inflection, by which multiple words are used to convey a lexical meaning and its features. Another interplay between morphology and syntax is realized by ambulant inflectional affixes, as found in the Polish example (5). Sometimes the ambulant affix cliticizes onto another word, sometimes it stands alone. The processing challenges are obvious, with various outcomes possible: (a) the morphological analyzer does not recognize the "source" word without its inflection; (b) the morphological analyzer does not recognize the "target" word with its unexpected inflection; (c) the morphological analyzer does not recognize the bare inflection realized as a word (not reflected in our examples but possible in some languages); (d) the morphological analyzer recognizes the source word and/or target word in the given form but the analysis is incorrect: e.g., Polish *poszli* 'went' is a valid word with 3rd person plural features, but the verb form intended in this sentence has 1st person plural features (*poszli* + *śmy*).

Information about ambulant inflectional affixes is elicited in Boas in a separate thread that follows the establishment of inflectional paradigms. If inflectional affixes in *L* can move, the user selects a paradigm to serve as a sample case and highlights all ambulant affixes. If different affixes from different paradigms have movement potential, the process is repeated for as many paradigms as necessary.

As a result of this process, Boas will contain an inventory of ambulant affixes similar to the inventories of affixes conveying agglutinating inflectional morphology, derivational morphology, and affixal realizations of closed-class meanings. For each inflectional affix that can move, the system will generate a set of morphological rules. One rule will recognize the affixless form of words in the source paradigm (i.e., word forms from which the affix can move): e.g., *poszli* in (5a) will be recognized as a verb that is missing inflection for person and number (*poszli* will also be recognized as the 3rd person plural form of the verb; this bit of ambiguity will be resolved at a later stage). A second rule will strip the hopped affix off the target word, revealing its underlying form. For example, in (5a), *śmy* will be stripped off of *myśmy* and *my* will be recognized as a pronoun in the regular way. In postmorphological analysis, the features associated with the hopped affix (1st person plural for *śmy* in (5a)) will be unified with their source

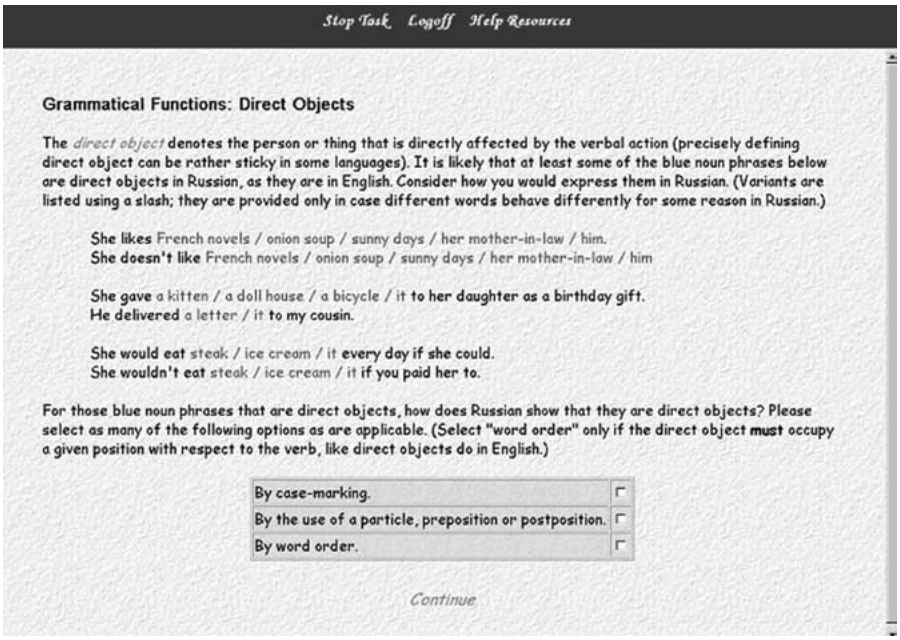

Figure 9. Eliciting indicators of the direct-object function.

stem (*poszli*). Elicitation threads for other morphosyntactic phenomena are planned for later implementations of the system.

Covering the Examples. Information about Polish mobile affixes will be elicited here.

4.8. THE PROCESSING ALGORITHM

The lexical analysis algorithm that Boas must support takes a text as input and outputs a set of candidate lexical readings for each input text element. Each reading consists of a lexical item (i.e., the citation form listed in the lexicon) plus the parameter values represented by the particular form of the word used in the text. The algorithm for this process is illustrated in Figure 10.

We will use English examples to explicate the algorithm even though English will not be analyzed as a source language by Boas.

1. If the given string directly matches one or more citation forms in the lexica, the given analyses are added to the Output Candidate Set. For example, English *move* is a citation form with several nominal and several verbal meanings, all of which will be added to the Output Candidate Set.
2. Analysis is continued since there might be homography between a listed word form and a derived one. For example, the English word *speaker* might be listed in the lexicon as a noun with the meaning ‘loudspeaker’ but might not be listed as a noun with the meaning ‘person who speaks’, since the latter

is predictable based on productive derivational processes. Following the algorithm above, after the Output Candidate Set has been augmented by the 'loudspeaker' analysis, it will be determined that English does, in fact, use derivational morphology. The list of derivational affixes will be checked, *-er* will be identified and stripped off, the citation form *speak* will be located in the lexicon, and the analysis *speak + er* will be added to the Output Candidate Set.

3. Analysis is continued in case of additional homography. For example, if the text element were English *hit*, it would be analyzed as a nominal and verbal base form (in several meanings each) in the original lexical lookup. Then, since English uses derivational affixes, it would be checked for affixes, of which there are none. Next, since English uses flective morphology, morphological analysis will be carried out and the verbal paradigm for *hit* – depending on how the user decided to organize it – will show that several parameter-value combinations are homographic with the base form: the infinitive (minus *to*), all simple-present-tense forms, the simple past tense and the past participle. All of these analyses will be added to the Output Candidate Set.
4. Processing continues until no new sets of citation forms and parameter-value combinations are found.

This algorithm shows why, in presenting the original inventory of examples, we distinguished between affixes that can be stripped off in turn (indicated by underscores) and those that cannot: they are treated by different procedures in the analysis algorithm.

4.9. COVERAGE

Boas is a largely expectation-driven system that does not rely upon language-specific rule-writing by trained computational linguists, and cannot productively use a free-form presentation of information (e.g., a prose description of grammatical processes). For these reasons, and because system development proceeded under the usual constraints of time and manpower, certain types of language phenomena are currently not treated. The most compelling reasons for excluding a phenomenon were its incomplete description in the literature and/or our inability to formulate sufficiently structured elicitation threads to permit useful computer-based generalizations. An example of a phenomenon that "fails" on both of these points is incorporation.

"Incorporation" describes a situation in which lexical elements with different syntactic functions (often a verb and one of its arguments) combine to form a single word. Incorporation presents many complexities including boundary mutations between incorporated elements, the loss of inflectional morphology on the nominal element, the splitting of verbal morphology from the stem leaving the incorporated noun in the middle, a change in verbal transitivity, and unpredictable semantic nuances of the incorporated structure (see for example, Allen et al. 1984;

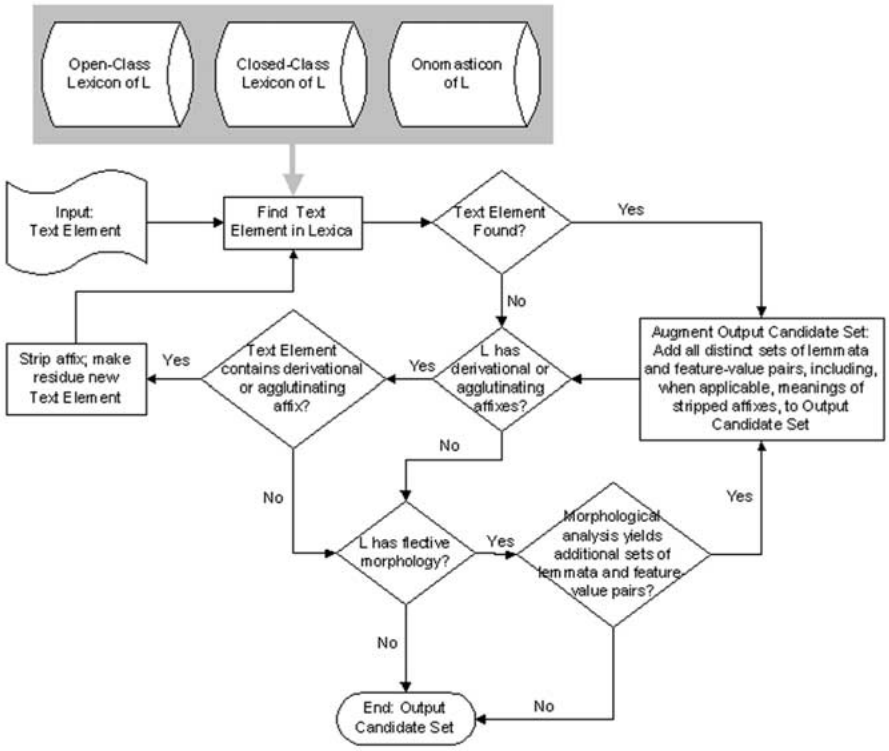

Figure 10. The algorithm for text analysis. Lexica contain lemmata, not inflected forms; the output candidate set can, according to this algorithm, be empty. Means of recovery from such failure are beyond the scope of Expedition.

Baker 1988; Bok-Bennema and Groos 1988; Fortescue 1984; Mithun 1984; Payne 1995 and Weggelaar 1986).[26] Because of these challenges, incorporation would be difficult to treat adequately even in a system designed for a particular language by computational linguists. Further research is required to determine to what extent the methodologies of Boas could be effective in treating the most descriptively complex of linguistic phenomena.

5. Experience with and Feedback on the System

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 Shore 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.[27] 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 sufficiently understood by most users, the work would have been easier and fewer questions would have arisen if time permitted the system to be employed in the way it was intended, that is, 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 (usually we had, in fact, planned for 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 permitting them to occupy valuable screen space;
- devoting more attention to the elicitation of agglutinative morphology;
- augmenting the inventory of parameters and values;
- fundamentally redesigning the open- and closed-class interfaces to increase speed of acquisition.

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 results of Boas have not yet been used to ramp-up full-scale MT systems, although the XML files that store all data generated using Boas are available and can be applied to MT or any other task. An excerpt from the XML file from the open-class lexicon of a profile of Polish is shown in Figure 11; similar XML files are produced for all other types of information elicited in the system.

6. Broader Implications

Boas offers a good example of an advanced KE system by combining, for the first time in a single system, extensive and parameterized descriptive material about language, a rich set of expressive means in the user interface, and extensive pedagogical resources. While there may be potential for Boas to serve as a blueprint for other similar systems, we believe that it should instead be considered an implemented example of an entire class of computer systems.

The KE methodology developed for Boas proceeds from the non-trivial assumption that untrained informants *can* be valuable sources of knowledge without the mediation of a domain expert ("knowledge engineer" in the parlance of the expert-system efforts of two decades ago) as long as metaknowledge about the subject

```
<Entry>
  <L>
    <CitationForm>drzewo</CitationForm>
    <Type>word</Type>
    <gender>neuter</gender>
    <OtherInherentFeature>
       <Name>virility</Name>
       <Value>inanimate</Value>
    </OtherInherentFeature>
    <PoS>noun</PoS>
    <Paradigm></Paradigm>     ;; there are no irregular forms
  </L>
  <English>
    <CitationForm>tree</CitationForm>
    <Type>word</Type>
    <PoS>noun</PoS>
  </English>
  </Entry>
```

Figure 11. Excerpt from the XML file from the open-class lexicon of a profile of Polish.

area in question is incorporated into the elicitation process. Of course, this incorporation can hardly be carried out without domain experts, but the idea is that their time is better spent working on metaknowledge than on carrying out broadscale acquisition.

It is clear that in some types of KE applications it will be difficult to develop an interface that obviates the need for the user to learn the metalanguage in which the knowledge they impart to the system is encoded. Boas did not require users to know the metalanguage (XML), since developers provided rules that generated metalanguage expressions from HTML forms filled out by the user. Some other application may require users not only to know the content of some subject domain but also to be well-versed in expressing their knowledge through the system's metalanguage.

It is not at all a trivial task for experts to be able to express their knowledge in *any* language – how many times did we hear the opinion that "I'd rather do it myself; it's too much trouble explaining things to others"? It is not only the perceived inability of people to learn that underlies this state of affairs. To use another popular simile, remember what happened to the centipede, arguably, an expert in many-legged locomotion, when somebody asked him how he manages to operate so many legs at once? So, systems that extend the capabilities of Boas must help users both to understand how best to formulate their knowledge and, if necessary, to express it in the metalanguage used by the system.

A good example of an area where such capabilities would be beneficial is in the acquisition of ontologies, including ontologies to support NLP in specialized domains (e.g., bioterrorism, nuclear physics). This task requires domain knowledge available only to experts. But since such experts are usually not trained ontologists, recording the relevant knowledge using the expressive means available in the given

ontological system is a logjam, usually necessitating the guidance of an ontologist who asks the expert the right questions in the right order.

We believe, however, that a KE system of the Boas class can be designed such that it facilitates ontology acquisition in both its content and metalanguage aspects, turning the task of the domain expert into traversing a series of well-defined questions and choices. So, whereas in the current version of Boas the parameters, values and realizations are of a linguistic nature, in ontological acquisition they could be oriented toward procedures for organizing and encoding knowledge in an ontology, supported by the same types of progressive-disclosure assistance as were developed for Boas.

Linguistically-related lessons of Boas involve achieving a better understanding of the very nature of language description and "airing out" issues that have become stagnant. For example, although we have not discovered any hitherto unknown types of word structure, the picture we paint is quite different than existing treatments. In an environment where established schools, theories and perspectives dominate, such novelty may provide a springboard to greater descriptive coverage and a finer grain size of description.

We believe that Boas could be readily applied to various realms, including, for example, education. With relatively minor augmentations, Boas could support training in general linguistics, computational linguistics and field linguistics, since working through the process of providing information about a language in a structured manner would be a hands-on means of learning linguistic content and developing discovery skills. When modified for this purpose, the Boas system would: prepare students to work creatively and independently as linguists; permit a customized, user-modeled approach to problem solving; offer a truly empirical basis for learning; promote a flexible definition of "success" since the language chosen and the user's knowledge of it would need to be taken into consideration for purposes of evaluation; encourage students to think globally, since rare languages will be more interesting research candidates than better studied languages; and facilitate the interaction between NLP and linguistics, since the content covered and means of covering it are largely driven by the ultimate processing needs.

Acknowledgments

Thanks to Jim Cowie, Igor Drugov, Stephen Helmreich, Wanying Jin, Denis Elkanov, Denis Kamotsky, Denis Loginov, Kemal Oflazer, Victor Raskin, Ron Zacharski and Rémi Zajac for their contributions to various aspects of the work.

Notes

[1] We use "text element" as a shorthand for "lexical text element", which excludes so-called "ecological" phenomena like numbers, dates and sentence-level punctuation.

[2] Boas is one component of the 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, recently carried out at the Computing Research Laboratory of New Mexico State University, was funded by Department of Defense Contract MDA904-92-C-5189. Descriptions of other aspects of the system can be found in McShane and Nirenburg (2003), McShane et al. (in press a,b), and other articles on the Expedition Web site: `http://crl.nmsu.edu/expedition`.

[3] Restricting the system to alphabetic languages that have distinct word boundaries was a programmatic decision. This approach to KE could, however, be extended to non-alphabetic languages as well.

[4] We believe that profiles of low-density languages could, for example, promote the teaching and learning of low-density languages.

[5] 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 a significant amount of metaknowledge in the system.

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

[7] Many of the examples throughout the paper were compiled from informants, others were drawn from grammars or other print sources. When examples from grammars are accompanied by original analysis of the author, the citation is provided explicitly. Otherwise, the examples from the following languages are due to the following sources: Albanian: Newmark et al. (1982); Blackfoot: Frantz (1991); Comanche: Charney (1993); German: a newspaper article; Irish: Ó'Sé and Sheils (1993), Ó' Siadhail (1989, 1995); Malay: Lewis (1954), Trask (1993); Nahuatl: Sullivan (1988); Polish: Franks and Bański (1999); Ponapean: Regh (1981); Tagalog: Schachter (1972); Ukrainian: Medushevsky and Zyatkovska (1963) (but example (4) was provided by a native speaker).

[8] Although Boas is intended primarily for less common languages for which MT capabilities have not been developed, we use more common languages for illustration since examples from them will be more transparent to readers.

[9] The decision to consider all word-level punctuation, including apostrophes, to be within a text element rather than to represent a word boundary has no special implications for this system.

[10] 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. For example, the English verb form *speaks* is composed of the morphemes *speak* and *-s*, where *-s* indicates both 3rd person and singular number. In agglutinating languages, words can also be composed of one or more morphemes but each morpheme tends to carry exactly one bit of lexical or grammatical information. For example Turkish *tas'ttim* 'I caused to carry' in example (6). In isolating languages, each word is generally a single morpheme and morphemes are not concatenated to form complex words. Inflection may be realized by synthetic (single-word) or analytical (multi-word) forms. Derivational affixation, as contrasted with inflectional affixation, contributes substantial new meaning to a word: for example, when *-er* is added to the stem *garden*, the meaning shifts from a place where flowers are located to the person whose takes care of them (*gardener*).

[11] A "root" is the simplest form of a morpheme, e.g., Polish *czyt-* 'read'. A "stem" is a form of the root upon which word-formation processes occur, e.g., Polish *czyta-* is the present-tense stem from which forms like $czytam_{1.SG}$ and $czytasz_{2.SG}$ are created via suffixation. A "citation form" is whatever form is listed in the dictionary; it is most commonly either a root or an inflected form, like the infinitive.

[12] No inventory of parts of speech is acceptable to all linguists for all languages. We fix the open-class inventory as noun, verb, adjective and adverb for purposes of English-driven lexical acquisition, but users are never required to describe morphological or grammatical properties of any part of speech

that they do not attribute to *L*. We circumvent the need to specify closed-class parts of speech by using a meaning-oriented elicitation procedure. None of the analysis programs in Boas require the explicit naming of closed-class parts of speech in *L*.

[13] The full paradigm is in Bright (1992: 15) with the note: "Finnish is a suffixing, relatively agglutinative language. However, since there are several dozen morphophonological alternations like gradation and vowel mutation, Finnish is by no means typically agglutinative".

[14] For a discussion of the morphological learning programs used in Boas, see Oflazer et al. (2001) and McShane and Nirenburg (2003).

[15] The morphology-learning program in Boas does not, however, productively treat inflectional reduplication. Some cross-linguistic examples are:

- in Malay full-word reduplication creates the indefinite plural: *bunga* 'flower' ∼ *bungabunga* 'flowers';
- in Ponapean full-word reduplication signals a change in aspect (tense is conveyed pragmatically – assume past tense in this example): *kang* '(I)-ate' ∼ *kangkang* '(I)-was-eating$_{\text{DURATIVE}}$';
- in Nahuatl partial reduplication (reduplication of the first syllable) in combination with suffixation (addition of the suffix *-tin*) is used to form plurals: *teuctli* 'lord' ∼ *teteuctin* 'lords'.

In Boas, reduplicative inflectional forms must be listed explicitly in the open-class lexicon for each relevant word, the same way as exceptions would be listed. If, for example, one form in an inflectional paradigm is created using reduplication but the other forms are created using affixation and other such rules, the morphology-learning program will learn the latter, leaving only the reduplicative form to be listed.

[16] In some agglutinating languages, person and number values are combined in a single set of affixes (e.g., one affix might indicate 1st person singular, another 1st person plural, etc.). Combined values are elicited on the Web page preceding the one shown in Figure 5.

[17] Some of these are reminiscent of Mel'čuk's lexical functions (Mel'čuk et al. 1984, 1988), a similarity that underscores the necessity to organize linguistic reality in terms of language universals in a system like Boas.

[18] Not only European languages have productive correlates for the words in our English inventory: e.g., adjectives in Ponapean can be negated using a productive affix *-sa* as well: *peik* ∼ ***sa****peik* 'obedient' ∼ 'disobedient'.

[19] The program would have to learn to identify syllables in *L*, which can be quite complex, then use the abstract notion of syllables as a basis for rule creation.

[20] In the current version of Boas, machine learning is not applied to closed-class items for three reasons: (a) inflectional patterns are commonly idiosyncratic, making machine learning infeasible; (b) in most languages, there are not very many closed-class forms, so typing them out should not be prohibitively time-consuming; (c) circumventing machine learning allowed us to streamline the paradigm-creation process, making the preliminary stages (i.e., establishing the template) much quicker and therefore speeding up work for most users in most cases. Given more development time, we could include more options regarding the best balance of typing out forms in a streamlined paradigm-creation process and having the machine learn rules in a more lengthy one.

[21] This version of Boas does not elicit affixal realizations of open-class items since they occur most commonly in incorporating languages which, for reasons described below, are not in our current purview.

[22] The inventory of word senses in the open-class lexicon elicitation thread of Boas is significantly smaller than that for the corresponding words in Wordnet. This "bunching" of senses reflects realistic expectations for word-sense disambiguation capabilities of the underlying MT system.

[23] In theoretical terms, the citation form of words with inflectional paradigms might be considered just one of the forms of the paradigm (unless the tradition for that language is to use a stem as the citation form). However, many NLP applications – Boas included – use the citation form as a base form upon which rules of inflectional morphology act, thus giving it special status. Different

languages use different conventions for listing citation forms, and even within a given language what is used as the citation form can be variable. For example, in Albanian the citation form of the verb is generally 1st person singular active indicative present common aspect. But some verbs do not have an active voice, so they are cited in the non-active; and some verbs do not have a 1st singular, so they are cited in the 3rd singular. Moreover, in some languages – Albanian is, again, a good example – there is more than one equally basic root: e.g., for verbs the root morpheme is actually a set of allomorphs, as in *djeg-/digj-/dogj-* 'burn' – with the choice of root depending on tense.

[24] Since we did not find convincing examples of instances in which eliciting inherent features for verbs, adjectives and adverbs would enhance analysis, we do not currently elicit them in Boas.

[25] See McShane and Zacharski (2003) for further description of interface functions.

[26] In many languages incorporation occurs either exclusively or primarily with nouns indicating body parts (Weggelaar 1986: 301f). This is true, for example, of Panare, in which "most incorporated nouns are body parts, and the verbs that allow incorporation are verbs of 'removal' or 'destruction', e.g., 'cut' (of various kinds), 'break', 'hit', 'pluck', etc. (Payne 1995: 300).

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

References

Allen, B. J., D. B. Gardiner, and D. G. Frantz: 1984, 'Noun Incorporation in Southern Tiwa', *International Journal of American Linguistics* **50**, 292–311.

Baker, M.C.: 1988, 'Morphology and Syntax: An Interlocking Independence', in M. Everaet et al. (1988), pp. 9–32.

Blythe, J., J. Kim, S. Ramachandran, and Y. Gil: 2001, 'An Integrated Environment for Knowledge Acquisition', in *International Conference on Intelligent User Interfaces*, Santa Fe, New Mexico, pp. 13–20.

Bok-Bennema, R. and A. Groos: 1988, 'Adjacency and Incorporation', in M. Everaet et al. (1988), pp. 33–56.

Boose, J. H. and J. M. Bradshaw: 1987, 'Expertise Transfer and Complex Problems: Using AQUINAS as a Knowledge Acquisition Workbench for Knowledge-Based Systems', *International Journal of Man-Machine Studies* **26**, 3–28.

Charney, Jean Ormsbee: 1993, *A Grammar of Comanche*, University of Nebraska Press, Lincoln, NB.

Comrie, Bernard and Norval Smith: 1977, 'Lingua Descriptive Studies: Questionnaire', *Lingua* **42**, 1–72.

Dura, E.: 1998, *Parsing Words*, Göteborg, University, Göteborg, Sweden.

Eshelman, L., D. Ehret, J. McDermott, and M. Tan: 1987, 'MOLE: A Tenacious Knowledge Acquisition Tool', *International Journal of Man-Machine Studies* **26**, 41–54.

Everaet, M., A. Evers, R. Huybregts, and M. Trommelen (eds): 1988, *Morphology and Modularity*, Foris Publications, Dordrecht.

Fortescue, M.: 1984. *West Greenlandic*, Croom Helm, London.

Franks, S. and P. Bański: 1999, 'Approaches to "Schizophrenic" Polish Person Agreement', in K. Dziwirek and C. M. Vakareliyska, (eds), *Annual Workshop on Formal Approaches to Slavic Linguistics: The Seattle Meeting, 1998*, Michigan Slavic Publications, Ann Arbor, pp. 123–143.

Frantz, D. G.: 1991, *Blackfoot Grammar*, University of Toronto Press, Toronto, Ontario.

Gaines, B. R. and M. L. G. Shaw: 1993, 'Eliciting Knowledge and Transferring it Effectively to a Knowledge-Based System', *IEEE Transactions on Knowledge and Data Engineering* **5**, 4–14.

Karlsson, F.: 1995, 'Designing a Parser for Unrestricted Text', in F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila (eds), *Constraint Grammar*, Mouton de Gruyter, New York, pp. 1–40.

Leavitt, John R. R., Deryle W. Lonsdale, Kevin Keck, and Eric H. Nyberg: 1994, 'Tooling the Lexicon Acquisition Process for Large-Scale KBMT', in *Proceedings of the 5th International IEEE Conference on Tools for Artificial Intelligence*, New Orleans, pp. 283–289.

Lewis, M. B.: 1954, *Teach Yourself Malay*, English Universities Press.

Longacre, R. E.: 1964, *Grammar Discovery Procedures*, Mouton, The Hague.

McShane, Marjorie and Sergei Nirenburg: 2003, 'Blasting Open a Choice Space: Learning Inflectional Morphology for NLP', *Computational Intelligence* **19**, 111–135.

McShane, Marjorie, Sergei Nirenburg, James Cowie, and Ron Zacharski: in press, a, 'Embedding Knowledge Elicitation and MT Systems within a Single Architecture', to appear in *Machine Translation*.

McShane, Marjorie, Sergei Nirenburg, and Ron Zacharski.: in press, b, 'Mood and Modality: Out of Theory and into the Fray', to appear in *Journal of Natural Language Engineering*.

McShane, Marjorie and Ron Zacharski: 2003, 'Preparing for Eventualities in User-Extensible On-Line Lexicons', manuscript, Institute of Language and Information Technologies, University of Maryland Baltimore County.

Medushevsky, A. and R. Zyatkovska [Медушевський А. и Р. Зятьковська]: 1963, *Українська грамамика* [Ukrainian Grammar]. Київ: Радянська школа.

Mel'čuk, I. A., N. Arbatchewsky-Jumarie, L. Elnitsky, L. Iordanskaja and A. Lessard: 1984, *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I* [Explanatory and combinatorial dictionary of contemporary French: Lexico-semantique research I]. Les Presses de l'Université de Montréal, Montréal.

Mel'čuk, I. A., N. Arbatchewsky-Jumarie, L. Dagenais, L. Elnitsky, L. Iordanskaja, M.-N. Lefebvre, and S. Mantha: 1988, *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques II* [Explanatory and combinatorial dictionary of contemporary French: Lexico-semantique research II]. Les Presses de l'Université de Montréal, Montréal.

Mithun, M.: 1984. 'The Evolution of Noun Incorporation', *Language* **60**, 847–895.

Motta, Enrico, Tim Rajan, and Marc Eisenstadt: n.d., 'A Methodology and Tool for Knowledge Acquisition', Technical Report TR-32, Human Condition Research Laboratory, Open University, Milton Keynes, UK; available at
`http://citeseer.nj.nec.com/cache/papers/cs/319/ftp:zSzzSzhcrl.open.ac.ukzSzwebzSztechreportszSzpaperszSztr32.pdf/a-methodology-and-tool.pdf`.

Musen, M. A., L. M. Fagan, D. M. Combs, and E. H. Shortliffe: 1987, 'Use of a Domain Model to Drive an Interactive Knowledge Editing Tool', *International Journal of Man-Machine Studies* **26**, 105–121.

Newmark, L., P. Hubbard, and P. Prifti: 1982, *Standard Albanian: A Reference Grammar for Students*, Stanford University Press, Stanford, CA.

Nirenburg, S.: 1996, 'On Supply-Side vs. Demand-Side Lexical Semantics', in *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, Santa Cruz, CA.

Nirenburg, Sergei, Stephen Beale, Kavi Mahesh, Boyan Onyshkevych, Victor Raskin, Evelyne Viegas, Yorick Wilks, and Rémi Zajac: 1996, 'Lexicons in the Mikrokosmos Project', in *Proceedings of the AISB Workshop on Multilinguality in the Lexicon*, Brighton.

Oflazer, Kemal, Sergei Nirenburg, and Marjorie McShane: 2001, 'Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning', *Computational Linguistics* **27**, 59–85.

Ó'Sé, D. and J. Sheils: 1993, *Irish*, NTC Publishing Group, Lincolnwood, IL.

Ó'Siadhail, M.: 1989, *Modern Irish*, Cambridge University Press, Cambridge.

Ó'Siadhail, M.: 1995, *Learning Irish*, Yale University Press, New Haven, CT.

Payne, T. E.: 1995, 'Object Incorporation in Panare', *International Journal of American Linguistics* **61**, 295–311.

Regh, K. L.: 1981, *Ponapean Reference Grammar*, University Press of Hawaii, Honolulu, HI.

Schachter, P.: 1972, *Tagalog Reference Grammar*, University of California Press, Berkeley, CA.

Sullivan, T. D.: 1988, *Compendium of Nahuatl Grammar*, translated from the Spanish by T. D. Sullivan and N. Stiles., University of Utah Press, Salt Lake City, UT.

Trask, R. L.: 1993, *A Dictionary of Grammatical Terms in Linguistics*, Routledge, London.

Weggelaar, C.: 1986. 'Noun Incorporation in Dutch', *International Journal of American Linguistics* **52**, 301–305.