User-Extensible On-Line Lexicons for Language Learning

Marjorie McShane and Ron Zacharski

Working Paper 05-05

March 8, 2005

Institute for Language and Information Technologies

University of Maryland Baltimore County

Abstract. This article describes a lexicon development and editing tool that we believe could profitably be applied to the teaching and learning of any language. Although originally developed as part of a machine translation system, the tool fills the same desiderata as one might posit for a language-learning aid: it has a fast, convenient interface, anticipates the needs of users of various profiles, incorporates means of expediting work, can monitor new submissions for completeness, is indefinitely extensible, and can be used for any language with no language-specific modifications necessary. Although the most economical application of this research and development effort would be to modify this particular tool to cater to language-learning needs, one could also exploit the research findings, design aspects and general motivation to the creation of other language-learning tools and technologies. In short, we present a proposal for incorporating cutting-edge language-processing technologies into the classroom at relatively low cost.

1. The Need for User-Extensible Lexicons

The teaching and learning of foreign languages is a fruitful avenue for the application of educational technologies. For the dozen or so more commonly-studied languages, much is already available, like sophisticated CD-ROMs to supplement textbooks, interactive exercises for learning vocabulary and word forms, and, of course, computer-based bilingual lexicons (CBBLs). CBBLs vary both in size and in the type and extent of information provided for each entry. Accordingly, like any static resource, they often do not fully meet the particular needs of a given user at a given time. Moreover, computerized lexicons do not even exist for most languages, with high development costs not being supported by the relatively small circle of interested parties – teachers, students, and those who may wish to use machine-tractable resources in natural language processing applications.[1] We believe that an environment that permits untrained users both to extend existing lexicons and to create new lexicons from scratch would be a significant contribution to language-learning technologies. As concerns the availability of lexical resources for a given language, L, three common scenarios are:

1. No CBBL exists for L, and a paper lexicon might also be lacking or minimal – as for languages that are only now being described by field linguists, have been described to some degree but are rarely or never taught, etc.
2. A sufficient paper lexicon exists but no CBBL. This case is common because, due to high development costs, technological support has not yet reached the less widely taught languages.

[1] Of course, some resources for the teaching of less-commonly-taught languages do exist. For example, the Less Commonly Taught Languages Project offers a Virtual Picture Album, Virtual Audio-Video Archives, and sample exercises for a number of LCTLs (http://carla.acad.umn.edu/lctl/lctl.html). Some clearinghouses for language-oriented materials are also available, including: a database of materials for over 100 LCTLs compiled by the UCLA Language Materials Project (http://www.lmp.ucla.edu); databases of software from the Language Center at the University of Minnesota (http://languagecenter.cla.umn.edu/lc/searchdb.html) and from the Foreign Language Multimedia Software Center, University of Hawaii (http://www.nflrc.hawaii.edu/ithompson/flmedia/); and the ERIC Clearinghouse on Languages and Linguistics (http://www.cal.org/ericcll/).

3. A CBBL exists but is insufficient in some way: e.g., a first-year language course might include a CD-ROM containing a lexicon that covers all the necessary basal items but not those necessary for intermediate or advanced study; or a sufficiently large on-line lexicon might exist but lack phrases, irregular inflectional forms, usage notes, or domain-specific terminology that a given user finds particularly important.

Scenarios 1 and 2 speak to the widely recognized problem of the paucity of instructional materials for lesser-studied languages. Brecht and Walton (no date), for example, propose that we need a "Language Learning Framework" that can "guide the design and management of instructional programs, materials development, teacher training, standards and assessment systems, and the whole range of infrastructure components […] upon which individual teachers and programs depend." We suggest that one way to speed the development of teaching and learning resources, at relatively minimal cost, is to adapt available, parametrizable resources to one's own needs – with our lexicon-development tool being a case in point.[2]

2. The Seed System: Boas

The lexicon-building tool in question is one module of a larger natural language processing system called Expedition. Expedition is a Web-based environment that promotes the fast creation of machine translation systems from any language (L) into English.[3] In a nutshell, upon delivery to a bilingual language informant, the system contains all the resources necessary for an L-to-English translation system except information about L.[4] The knowledge-elicitation system – which we call Boas, after renowned field linguist and anthropologist Franz Boas – guides the informant through the process of providing information about the morphology, syntax, and lexical stock of L in a stepwise, pedagogically supported manner.

The Expedition system was designed primarily for so-called "low density" languages, i.e., those for which few or no on-line (or, perhaps, even paper) resources are available. In pedagogical terms, these correspond to the much less commonly, least commonly, and rarely or never taught languages (Brecht and Walton). There are practical reasons for this focus: although having *some* machine translation capabilities for such languages is far better than having none—at least in the realm for which the project was contracted—the quality of a translation system generated in template form cannot compete with that of a system cater-made for a given language pair. Thus, Boas fills a specific niche in the landscape of language technologies.

The knowledge-elicitation (KE) component of the system, Boas, represents an innovative *methodology* of knowledge elicitation which 1) makes the system accessible to linguistically novice informants, 2) permits it to cover any natural language, and 3) allows its incremental extension as resources become available or the scope of interest expands. The KE process is based upon our understanding – derived of cross-linguistic research – of what phenomena occur in language and, tangentially, our view of what needs to be covered to describe a language to a reasonable degree of detail.

---

[2] Of course, user-extensible lexicons could be useful outside of the classroom as well: for example, one could expand a bilingual lexicon into specialized realms like botany or computer science, which would help specialists gain access to texts not written in their native tongue.

[3] For a broader description of the system see McShane, Nirenburg, Cowie and Zacharski 2003.

[4] Resident information includes the knowledge-elicitation system, an English lexicon, a morphological analyzer, a method of learning syntactic rules, the machine-translation engines, etc.

These expectations are organized into language parameters, their value sets, and means of realizing the latter, as shown in the examples in Table 1.[5]

Parameter	**Value Set**	**Means of Realization**
case	nominative, genitive, dative...	flective morphology, agglutinating morphology...
grammatical role	subject, direct object...	case-marking, word order, use of particles...
part-of-speech shift (a derivational process)	adjective → adverb, verb → noun...	suffixation, prefixation...

Table 1. Examples of parameters, values and realizations.

Organizing language into parameters, value sets and means of realizing the latter represents *expectation-driven* knowledge elicitation.

The KE process of Boas is organized as a series of (sub)tasks, with the order of work restricted only inasmuch as prerequisites for certain tasks obtain; apart from those restrictions, work can proceed in any order. The tasks are presented in a dynamic task tree supplemented with icons that indicate task status. Figure 1 shows an excerpt from the task tree when the user has finished work on the Introduction and Case Relations (indicated by the check mark), and has begun but not finished work on the other subtasks (indicated by the coffee cup, signaling "took a coffee break from the task").

Figure 1. An excerpt from the task tree for the Closed-Class Lexicon.

[5] More discussion of parameters and values can be found in Nirenburg 1998, McShane, Nirenburg, Cowie and Zacharski 2003, and McShane and Nirenburg 2003a,b.

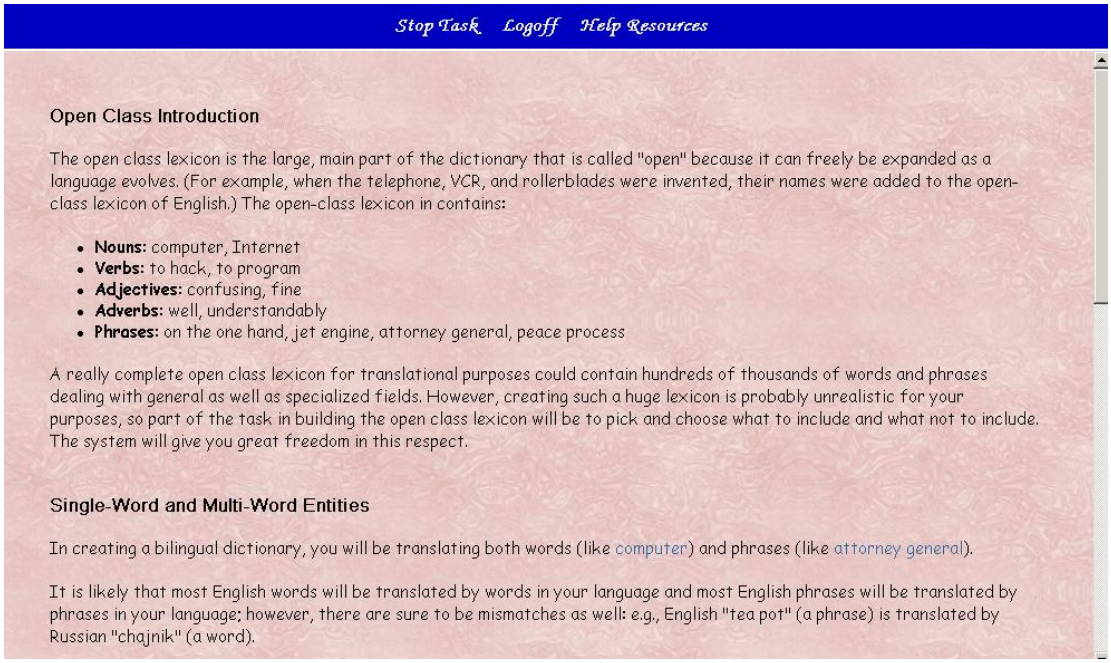
Space does not permit a comprehensive description Boas, nor is it required for a full understanding of the lexicon-oriented tools. After all, building the lexicon is conceptually the least challenging part of describing a language, reducing to primarily a listing task. (Compare this to delineating inflectional paradigms or describing details of syntax.) Moreover, since Boas was programmed in modular fashion, extracting the lexicon-building modules from the rest of the system presents no technical problems.

There is, however, some extra-lexical information that must be imported into the lexicon interfaces. This information is collected at various points in the larger Boas system but could be streamlined into a few elicitation pages in a "lexicon-only" tool. This additional information includes the following:

1. the character set used for the L; currently, Boas supports any alphabetic script but with minimal work non-alphabetic scripts could be supported as well;
2. any grammatically important inherent features, like gender for nouns, since they are largely unpredictable and must be lexically specified;
3. for which, if any, combinations of features words in L inflect; this permits irregular forms to be listed and associated with their features (e.g., the plural of *sheep* is *sheep*, not **sheeps*).
4. the cases used in L (if any), since, in case languages, prepositions and postpositions require their complements to be in a certain case.

We must emphasize that this is the *minimal* amount of pre-lexical information required for most lexicons – information that we expect to be sufficient for most users. However, language technologies have more to offer should one choose to incorporate it. For example, the full Boas system includes a module that learns productive patterns of flective morphology based on sample paradigms provided by the user. These morphological rules can then be used to generate all the regular inflectional forms of words. This facility could be very helpful, especially for beginning students but, naturally, comes at a cost: first, the lexicon-development tool would need to incorporate the morphology-learning module, which is far more complex that the lexicon-oriented components themselves; second, an informant for each L would need to provide all the necessary knowledge for the machine-learning program and test the results extensively to verify the accuracy and coverage of the rules. We will not pursue enhancements such as these in this paper (see McShane 2003 for a discussion of pedagogical aspects of Boas on the whole), however, it is worth noting that available technology can provide many enhancements to a configurable, extensible system like this one.

The Boas lexicon-acquisition process comprises two separate modules, one devoted to open-class items and the other to closed-class items. In most languages, open-class items include nouns, verbs, adjectives and adverbs—i.e., those parts of speech whose inventory can be added to as new concepts enter a society and its language. Closed-class items, by contrast, include all of the "minor" parts of speech, like pronouns, prepositions, conjunctions, numbers, etc., whose inventory of *meanings* is quite fixed for any given language despite different realization strategies cross-linguistically. As an example of the types of tutorial materials provided in Boas, we present an excerpt from the introductions to the Open-Class and Closed-Class lexicons in Figures 2 and 3, respectively.



Stop Task Logoff Help Resources

Open Class Introduction

The open class lexicon is the large, main part of the dictionary that is called "open" because it can freely be expanded as a language evolves. (For example, when the telephone, VCR, and rollerblades were invented, their names were added to the open-class lexicon of English.) The open-class lexicon in contains:

- **Nouns:** computer, Internet
- **Verbs:** to hack, to program
- **Adjectives:** confusing, fine
- **Adverbs:** well, understandably
- **Phrases:** on the one hand, jet engine, attorney general, peace process

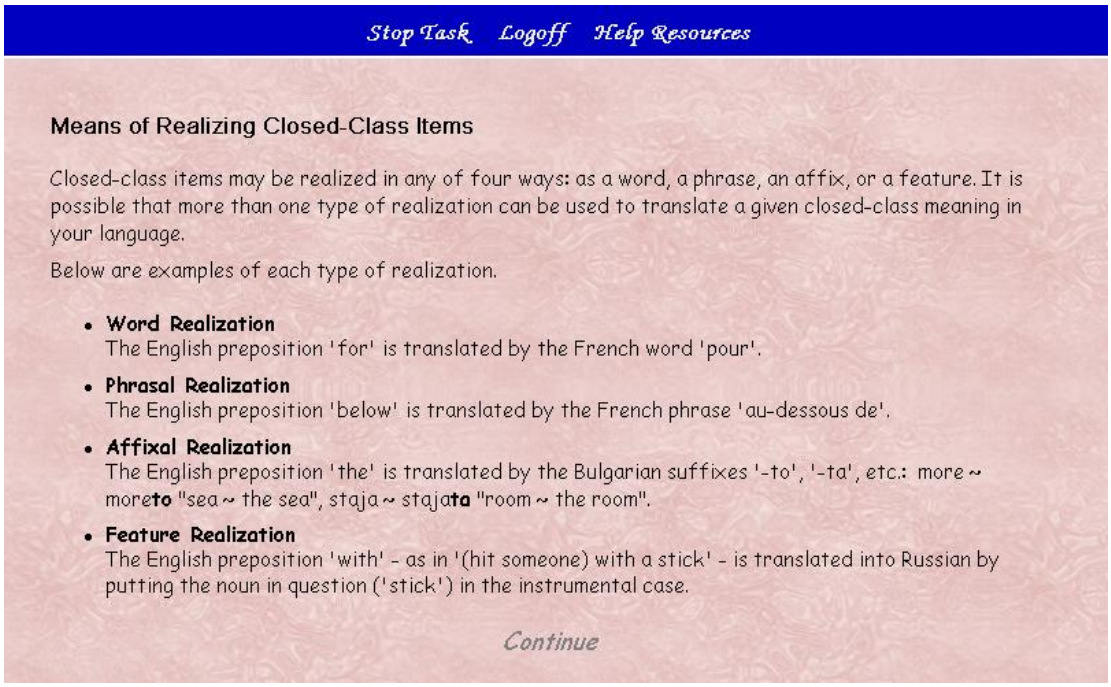
A really complete open class lexicon for translational purposes could contain hundreds of thousands of words and phrases dealing with general as well as specialized fields. However, creating such a huge lexicon is probably unrealistic for your purposes, so part of the task in building the open class lexicon will be to pick and choose what to include and what not to include. The system will give you great freedom in this respect.

Single-Word and Multi-Word Entities

In creating a bilingual dictionary, you will be translating both words (like computer) and phrases (like attorney general).

It is likely that most English words will be translated by words in your language and most English phrases will be translated by phrases in your language; however, there are sure to be mismatches as well: e.g., English "tea pot" (a phrase) is translated by Russian "chajnik" (a word).

Figure 2. An excerpt from the tutorial pages in the Open-Class Lexicon module of Boas.



Stop Task Logoff Help Resources

Means of Realizing Closed-Class Items

Closed-class items may be realized in any of four ways: as a word, a phrase, an affix, or a feature. It is possible that more than one type of realization can be used to translate a given closed-class meaning in your language.

Below are examples of each type of realization.

- **Word Realization**
  The English preposition 'for' is translated by the French word 'pour'.
- **Phrasal Realization**
  The English preposition 'below' is translated by the French phrase 'au-dessous de'.
- **Affixal Realization**
  The English preposition 'the' is translated by the Bulgarian suffixes '-to', '-ta', etc.: more ~ more**to** "sea ~ the sea", staja ~ staja**ta** "room ~ the room".
- **Feature Realization**
  The English preposition 'with' - as in '(hit someone) with a stick' - is translated into Russian by putting the noun in question ('stick') in the instrumental case.

Continue

Figure 3. An excerpt from the tutorial pages in the Closed-Class Lexicon module of Boas.

3. The Open-Class Lexicon

The Boas open-class lexicon helps a user to acquire the best (as defined by him or her) possible inventory of complete lexical entries in the shortest amount of time and with the least effort. For Boas's machine-translation application, "best" meant most-often met with in newspaper texts, and "complete" meant containing (i) a word or phrase in L, (ii) its English counterpart, (iii) required morphological features and (iv) irregular inflectional forms. For an educational application, "best" might mean most common in everyday speech or met with in particular coursework materials, and "complete" might include, apart from the features mentioned above, a definition, usage notes, and common collocations.

3.1 Methods of Lexical Acquisition

Boas supports four basic methods of acquiring a lexicon from scratch or supplementing an existing one. They can be carried out in any order and in any combination.

1. Translating resident English word- and phrase-senses.[6] Boas currently contains some 60,000 senses which are divided by part of speech and, within each part of speech, are grouped by frequency (e.g., group 1 of nouns contains the 50 most common nouns in our corpus, with all their senses, group 2 the next 50 most common, etc.). Our frequencies are based on newspaper corpora but, for a system catered to language-learning, could be reset to address frequency for language learners.
2. Translating word- and phrase-senses that are compiled off-line, in English or L, and are subsequently imported into the system. Importation instructions are simple and should be accessible to all users. This method streamlines the acquisition of inventories of words and phrases that are immediately needed. It is also helpful if, for example, a language uses word-formation processes widely (like compounding in German), since words thus formed will not be in the seed English lexicon.
3. Importing extant bilingual or monolingual dictionaries, which requires some limited programming skills. Boas requires files to be in XML format, which means that *relevant* tags in the source lexicon must be converted to the ones used in Boas and all others must be stripped.[7] This should

[6] We specify "senses" because a given word or phrase can have many senses.

[7] An example of the XML entry for 'account' in a Russian-English system:

```
<Entry>
  <Russian>
    <CitationForm>счет</CitationForm>      ; the main form of the word
    <Type>word</Type>                      ; word or phrase
    <PoS>noun</PoS>                        ; noun, verb, adjective, adverb
    <Gender>masculine</Gender>
    <Animacy>inanimate</Animacy>
  </Russian>
  <English>
    <CitationForm>account</CitationForm>
    <Type>word</Type>
    <PoS>noun</PoS>
    <Definition> a business or business relationship established to provide for regular services and dealings and other financial transactions</Definition>
  </English>
</Entry>
```

present no difficulties if the lexicon is amenable to reformatting; some, unfortunately, are not, which then renders their importation impossible without more extensive work on format conversion. This acquisition option permits the exploitation of all available resources in order to reduce doubling of work for a given language.

4. Adding words or phrases to the lexicon on the fly, since translating one entity can bring another relevant one to mind.

In order to organize work on the lexicon and ensure that all entries that enter The Dictionary (i.e., the inventory of compete entries) are complete, Boas has a three-tier division of the lexicon: 1) word lists (in either language) that have not yet been worked on; 2) entries that have been begun but not completed; this holding bin, which we call "Purgatory", permits users to postpone things must be verified, things that require extra time (e.g., listing irregular forms), etc.; and 3) The Dictionary, in which all entries are complete, although they may be edited at any time.

Work on the Open-Class lexicon is organized into a similar hierarchy of tasks as is used throughout Boas. An abbreviated task tree is presented in Figure 4.

Introduction
Main Gate
- Nouns
  - Work on new word lists
    - List 1
    - List 2
    - ...
  - Complete entries in "Purgatory" (holding place for entries started but not finished)
  - View/edit The Dictionary (the inventory of completed entries)
  - View/edit list of words deleted or skipped
- Verbs
- Adjectives
- Adverbs
- Phrases

Create Word Lists
Import Existing Lexicon

Figure 4. A representation of the Open-Class task tree.

3.2 Interface Functions

The basic Open-Class interface in Boas is illustrated in Figure 5 using an example from Russian. It reflects information collected through pre-lexicon knowledge elicitation: 1) the informant posited two inherent features for Russian nouns: one with at least the values masculine and feminine, and the other with at least the value inanimate; 2) the informant has created inflectional paradigms for Russian, otherwise the "Paradigm" checkbox would not be present; 3) the informant does not think that any of the entries in L has irregular inflectional forms, since no checkboxes are checked. All of this information was imported into the lexicon interface from the morphology section of Boas; for a pedagogical system, it

would be elicited using a number of interactive Web pages that use the 'help-supported' KE methodologies of Boas. [8]

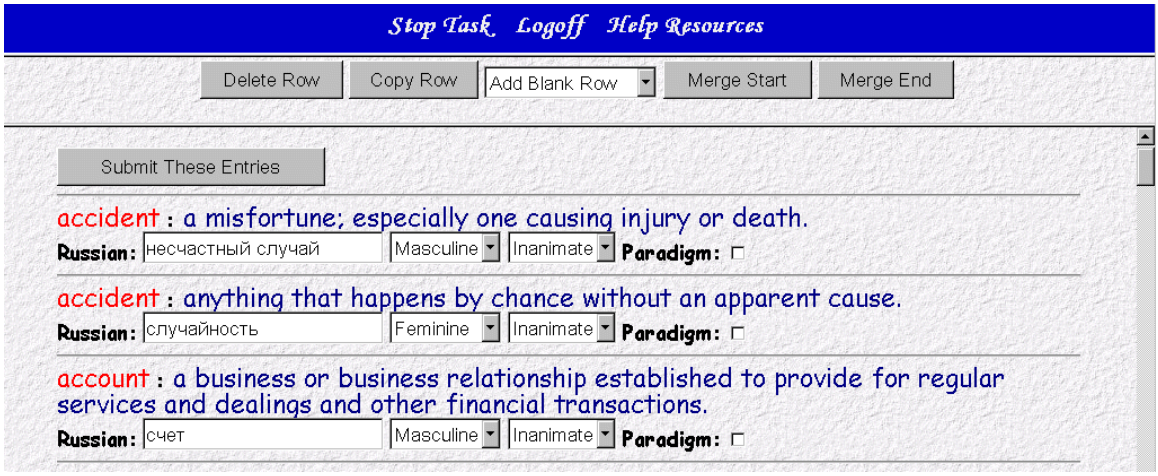


Figure 5. The basic open-class lexicon interface in Boas.

The interface functions that speed acquisition are as follows:

Delete Row is used to remove a word from the list and put it into a trash bin. This is for words that cannot be translated or are not important enough in L to be bothered with. The cursor must be in the text field of the given row before clicking on Delete Row. After clicking on it, the screen refreshes with that row missing. (These cursor and refresh comments apply to most functionalities and will not be repeated.)

Copy Row is used when there is more than one translation for a given prompt. For example, there are two Russian words for English *blue* – one meaning *light blue* and the other meaning *dark blue* (there is no neutral word for *blue*). Multiple translations must be typed in separate rows because they might have different inherent features, or one might be a word whereas another is a phrase, or one or both might have irregular inflectional forms.

Add Blank Row is used to add a completely new entry for which variants in both English and L must be provided. Add Blank Row is actually not a button but a pull-down menu requiring the informant to indicate which part of speech the new item will belong to, since L might require different kinds of information for different parts of speech (e.g., nouns might have inherent features whereas verbs do not); therefore, it is important that a new row of the right profile be added. This function permits the informant to add, on the spot, entities that occur to him or her during work on the open class—like idioms, phrases, or compounding forms based on a word just translated.

Merge Start and Merge End are a pair of functions that permit the informant to bunch word senses that have the same translation, thus reducing acquisition time, especially if a given entity in L requires additional work, like listing irregular inflectional forms.

[8] As mentioned earlier, it would not be necessary to incorporate the full paradigm-development module of Boas into a lexicon-building system. The "Paradigm" checkbox could be renamed "Irregular Forms", and checking it would lead to a page in which irregular forms could be listed in text fields and associated with features that are selected from pull-down menus.

Since speed is at the center of the interface design, keyboard-centered methods of working with the interface are encouraged. For example, tabbing takes the user from one action point to the next and if some variety of a Latin keyboard is being used, typing in the first letter of a given word in a drop-down menu will pull up that word.

4. The Closed-Class Lexicon

The closed-class lexicon in Boas contains a finite inventory of cross-linguistically prevalent semantic meanings that include: spatial relations; temporal relations; case relations; personal, reflexive, relative, interrogative, indefinite, predicative, demonstrative and possessive pronouns; conjunctions; articles; quantifiers; cardinal and ordinal numbers; and interrogative adjectives and adverbs.[9] From the cross-linguistic perspective, it is important to conceptualize the closed-class lexicon as *meaning oriented* rather than part-of-speech oriented because methods of realizing this collection of meanings reach beyond the familiar word and phrase options of the open-class lexicon—that is, closed-class meanings are also regularly realized as an affix or an inflectional feature. Affixal realizations of closed-class items include the definite article in Bulgarian realized by the suffixes *-to*, *-ta,* etc.: *more* 'sea' ~ *more*to 'the sea'; and the English reciprocal *oneself* realized by the Russian suffix *-sja*: *myt'* 'to wash' ~ *myt'*sja 'to wash oneself'. 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* 'revolver', can mean '(shoot, kill, etc.) <u>with</u> a revolver'.

Apart from expanded realization options, there are other features that distinguish closed-class elements from open-class ones. For example, if closed-class items inflect, they often require different paradigm templates than those found for the open-class parts of speech (e.g., pronouns tend to be singular only or plural only), and their inflectional patterns are often idiosyncratic.

The closed-class interface in Boas was designed to speed acquisition while providing for all possible realizations of the English word senses. The look and feel of the interface is illustrated in Figure 6 using a portion of the Temporal Relations page, with Russian equivalents listed.

[9] A given sense is elicited separately when used as different parts of speech in order to avoid presenting syntactic apples and oranges to the informant. For example, "before" functions as a conjunction in *he turned the lights out before leaving* and as a preposition in *before November 17*. In many languages, these uses of "before" will have different translations.

Figure 6. The Closed-Class lexical acquisition interface.

The table is labeled with the name of the semantic group being worked on (here: Temporal Relations), which can be clicked on to lead to a help page describing the properties of the group. To the right of the group name are two links with the following functions: the **Add row** link adds a row to the table should there be more than one realization of a meaning in L (as for *before* in the figure), and the **Interface help** link leads to a page explaining interface functions. Each row of the table contains:

- The English prompt.
- An example of its use, for disambiguation. Clicking on this example pops up a help page with a full definition and more examples.
- A text field in which to type the translation, if the translation is a word, phrase, or affix. If the translation is simply some case form (e.g., Instrumental case to convey *with*), the text field is left empty.
- A pull-down menu of cases, included only if nouns in L inflect for case and only in some elicitation tables. That is, in tables that elicit case relations, selecting a case implies that a noun in the given case can have the given meaning: e.g., instrumental *with* realized by a noun in the Instrumental case, and recipient *to* realized by a noun in the Dative case. In tables that elicit prepositional/postpositional meanings, choosing a case implies that, in the given meaning, the preposition/postposition takes a complement in that case.
- An **Add** button, which indicates that the (head) word or affix has inflectional forms. Clicking on this button leads to the related paradigm elicitation pages.

5. The Benefits of a Lexicon Authoring Tool

A lexicon-authoring tool for pedagogical use must foresee the needs and preferences of the teachers and students who will use it as well as demand no special skills or training; otherwise, the tool has little chance of being adopted by the target community. The tool we describe caters namely to novices, and, moreover, can be used for any language by users with mainstream or idiosyncratic needs.

The primary advantages of this system (and, by extension, the approach of creating configurable, reusable systems in general) are as follows:

1. It allows anyone to build or expand a bilingual lexicon for any language.

2. As a Web-based tool, it supports collaborative work on a given lexicon, even by people in different locations.
3. Since the data is saved in simple XML format, it is readily available to be incorporated into any computerized language-learning exercises.
4. The technology already exists, having been developed for a US-government-sponsored project, so customizing the resources for distribution as a language-learning resource would be relatively inexpensive.
5. Such a tool supports non-traditional teaching and learning of languages, which includes teaching less common languages as well as learning such languages independently, even perhaps as a field linguist.

There are no technological or conceptual challenges inherent in converting the current tool to one for language-learning purposes, but some development time would be required. Tasks include: extracting the lexicon-elicitation modules from the machine-translation architecture; reordering the frequency-based word lists to focus on learners' needs; reducing the tutorials to contain only those materials relevant for lexicon building; writing the extra-lexical knowledge-elicitation pages; including an inventory of optional text fields, as for comments or notes; incorporating more viewing options of The Dictionary (e.g., strictly alphabetically rather than by part of speech); and field testing the system before widespread distribution. Considering the broad potential use of such a system, and the fact that it would represent a significant contribution especially for the less-commonly-taught languages, we believe that such development costs seem well justified.

Acknowledgments

The Expedition project was carried out at the Computing Research Laboratory of New Mexico State University and was funded by Department of Defense Contract MDA904-92-C-5189. Our thanks to Jim Cowie, Denis Elkanov, Natasha Elkanova, Stephen Helmreich, Wanying Jin, Denis Loginov and Remi Zajac.

References

Brecht, R. D. and Walton, A. R. (No Date). National Strategic Planning In The Less Commonly Taught Languages. Available at http://www.councilnet.org/pages/CNet_VLib_Pubs.html#Council papers.

McShane, Marjorie. 2003. Applying Tools and Techniques of Natural Language Processing to the Creation of Resources for Less Commonly Taught Languages. Under review at *IALL Journal*.

McShane, Marjorie and Nirenburg, Sergei. 2003a. Blasting Open a Choice Space: Learning Inflectional Morphology for NLP. Forthcoming in *Computational Intelligence*.

McShane, Marjorie and Nirenburg, Sergei. 2003b. Parametrizing, Eliciting and Processing Text Elements Across Languages. Under review at *Machine Translation.*

McShane, Marjorie, Sergei Nirenburg, James Cowie and Ron Zacharski. 2003. Nesting MT in a Linguistic Knowledge Elicitation System. Forthcoming in *Machine Translation.*

Nirenburg, Sergei. 1998. Project Boas: "A Linguist in the Box" As A Multi-Purpose Language Resource Paper. *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.