Time to change the "D" in "DEL"

Stephen Beale
Linguist's Assistant
Baltimore, MD
stephenbeale42@gmail.com

Abstract

The "D" in "DEL" stands for "documenting" – a code word for linguists that means the collection of linguistic data in audio and written form. The DEL (Documenting Endangered Languages) program run by the NSF and NEH is thus centered around building and archiving data resources for endangered languages. This paper is an argument for extending the 'D' to include "describing" languages in terms of lexical, semantic, morphological and grammatical knowledge. We present an overview of descriptive computational tools aimed at endangered languages along with a longer summary of two particular computer programs: Linguist's Assistant and Boas. These two programs, respectively, represent research in the areas of: A) computational systems capable of representing lexical, morphological and grammatical structures and using the resulting computational models for translation in a minority language context, and B) tools for efficiently and accurately acquiring linguistic knowledge. A hoped-for side effect of this paper is to promote cooperation between these areas of research in order to provide a total solution to describing endangered languages.

1 Introduction

The "D" in "DEL" stands for "documenting" – a code word for linguists that means the collection of linguistic data in audio and written form. The DEL (Documenting Endangered Languages) program run by the NSF and NEH is thus centered around building and archiving data resources for endangered languages. Furthermore, the recent change in the program to include computational tools hasn't changed the central focus on documentation, with one notable exception: the research headed by Emily Bender (Bender, et al. 2013) to automatically extract grammatical information from interlinear text. This paper is an argument for extending the 'D' to include "describing" languages in terms of lexical, semantic, morphological and grammatical knowledge. We present an overview of descriptive computational tools aimed at endangered languages along with a longer summary of two particular computer programs: Linguist's Assistant and Boas. These two programs, respectively, represent research in the areas of A) computational systems capable of representing and translating minority languages, and B) tools for efficiently and accurately acquiring linguistic knowledge. A hoped-for side effect of this paper is to promote cooperation between these areas of research in order to provide a total solution to describing endangered languages.

2 Documenting versus Describing

The code word "documenting" implies data. The DEL program is primarily interested in procuring data about languages that are disappearing. The rationale behind this is obvious: we need to quickly gather data from languages before they become extinct. Data in the form of transcribed audio recordings and texts is certainly invaluable. However, consider the impact of such data in two areas: 1) future analysis by linguists, and 2) revitalization and language promotion today.

Think ahead 50 or 100 years. By all accounts, a majority of the world's languages will be extinct. What resources will be available to the 22nd century linguist? The DEL program seeks to archive audio and textual data for use in the future. While this data is certainly valuable, how useful will it be? Without a living speaker of the language, extracting a useful, accurate and broad-coverage description of the language from archived data will be extremely time consuming

and probably impossible in most cases.[1] Although such data could be used for other purposes, Gippert et al. (2006) agree with the general premise that "without theoretical grounding language documentation is in danger of producing 'data graveyards', i.e. large heaps of data with little or no use to anyone." This is a shame, and quite possibly a non-optimal use of our current linguistic talent pool. On the other hand, if a linguist working today with a living informant and using appropriate computational tools and programs could efficiently and accurately describe these languages at a lexical, semantic, morphological and grammatical level, then the usefulness of such research 100 years from now would be considerably greater.

That is looking ahead. What about now? What kind of work could help revitalize endangered languages so that they will not become extinct in the first place? My experience in language projects in the South Pacific leads me to the conclusion that descriptive work - and the resulting computational and non-computational projects that are enabled by it - have a much greater impact on current language populations than documentary efforts. The community I worked with for three years were the recipients of dictionaries and story books that documented linguistic research. These efforts bore fruit: there was initially quite a bit of interest about them. However, this kind of work quickly lost appeal. On the other hand, descriptive work quickly led to the production of educational materials and interest in translation. Automatic and manual translations followed, especially of songs, religious and health-related materials. A knowledge of how the language works leads to an empowerment with the language.

3 Research in Describing Endangered Languages: knowledge acquisition methodologies

In this section we present an overview of current and past descriptive computational tools aimed at endangered languages. In general, the field can be divided into two parts: A) computational systems capable of representing and translating minority languages, and B) tools for efficiently and accurately acquiring linguistic knowledge. Up until recently, research has focused on the latter.

The most widespread line of computational research in category B can be categorized as grammatical typology questionnaires. These follow in the path of traditional, non-computational linguistic fieldwork methods characterized by Longacre (1964) and Comrie and Smith (1977). Boas (McShane, et al. 2002), the LinGO Grammar Matrix (Bender, et al. 2010) and PAWS (Black & Black 2009) all fit into this paradigm. All these systems extract salient properties of a language through typological questionnaires and then produce computational resources of varying utility. This work must be applauded, and we argue that it is indispensable for a complete solution for describing endangered languages. However, the typology questionnaire approach is limited to creating approximate grammars. Bender et al. (2010) describe the LinGO Grammar Matrix as a 'rapid prototyping' tool. Such a tool is useful, but more is needed to thoroughly describe a language and enable machine translation capabilities. Linguist's Assistant (LA, described below) promotes such a thorough description; however, it comes at a cost. LA is able to represent the kinds of knowledge that is typically extracted by the grammatical typology questionnaire approach, such as rules to represent phrase structure word ordering and phenomena such as case, agreement, nominal declensions and the like. But it is more flexible and able to describe additional linguistic phenomena that are not as easily described using a typological approach (see below for details). But the rules in LA currently must be entered manually by a computational linguist. Thus, the tradeoff: quick descriptions (using well thought-out typologies) that fall short of broad and deep coverage vs. adequate depth and breadth of coverage at a higher cost.

It is perfectly clear that some linguistic phenomena can be most efficiently described using the techniques of the typology questionnaire paradigm. However, the computational grammar and lexicon produced in an LA-type language description project are meant to be comprehensive and complete insofar as they will be able to be used in a text generator to produce accurate translations. It is exactly this completeness and the resulting usefulness of the description (especially in language revitalization) that might be a prime factor in securing research funding from organizations that are interested in endangered languages. Therefore, we argue for: 1) continued research in typology questionnaire methods for

[1] Our experience backs up this claim. We have attempted to use Linguist's Assistant to describe languages using only transcribed texts without a human informant; these experiments failed miserably.

efficiently acquiring the linguistic knowledge appropriate to that paradigm, 2) further development of complete description paradigms like LA, 3) a greater cooperation between the two paradigms, and 4) the resurrection of machine learning, example-based techniques to minimize and semi-automate the comprehensive grammatical and semantic description process needed by systems like LA.

A prime example of this latter point was the Avenue Project at Carnegie Mellon University (Probst, et al. 2003). The Avenue project was a machine translation system oriented towards low-density languages. It consisted of two central parts: 1) the pre-run-time module that handles the elicitation of data and the subsequent automatic creation of transfer rules, and 2) the actual translation engine. We are especially interested in the former:

> "The purpose of the elicitation system is to collect a high-quality, word-aligned parallel corpus. Because a human linguist may not be available to supervise the elicitation, a user interface presents sentences to the informants. The informants must be bilingual and fluent in the language of elicitation and the language being elicited, but do not need to have training in linguistics or computational linguistics. They translate phrases and sentences from the elicitation language into their language and specify word alignments graphically.
>
> The rule-learning system takes the elicited, word-aligned data as input. Based on this information, it infers syntactic transfer rules.... The system also learns the composition of simpler rules into more complicated rules, thus reducing their complexity and capturing the compositional makeup of a language (e.g., NP rules can be plugged into sentence-level rules). The output of the rule-learning system is a set of transfer rules that then serve as a transfer grammar in the run-time system." (Probst, et al. 2003:247–248)

At a high level, this is exactly the approach that LA advocates. However, LA differs from Avenue in several important features, most notably the underlying semantic representation in LA as opposed to Avenue's transfer (source surface language to target surface language) approach. LA attains a greater practicality than Avenue primarily because of this difference, because interlingual-based language description and text generation is an order of magnitude simpler and less prone to error than transfer-based approaches. But again, this benefit comes at a cost: the grammar description modules and all subsequent texts to be translated must be encoded in the semantic representation (as opposed to a natural language like English for transfer-based approaches). See the next section on Document Authoring for more details for how this limitation can be minimized.

Bender et al. (2013) also provide a machine-learning component for their LinGO Grammar Matrix (Bender, et al. 2013). That is the project that is the exception to the "D" word problem. And that exceptional nature (it was funded!) should be instructional for all of us.

The missing ingredient in LA (besides the inclusion of grammar typology techniques such as LinGO and BOAS) is the sort of machine learning capability seen in the Avenue project and Bender's project. The latter system learns LinGO rules from interlinear text. Obviously, that is exciting work and has the added benefit of being able to be used directly in the DEL's data-centric context. However, it has limitations. We argue for a similar type of interlinear machine learning system, but one that is grounded in semantics and works over carefully prepared texts that will maximize the learning capabilities and allow for broad coverage of semantic phenomena. For example, assume we have the following sentences semantically represented:

John hit the tree.
John began to hit the tree.
John finished hitting the tree.
etc…

After a native speaker translates these sentences, a machine learning system could be employed to learn a grammar of inceptives, completives, etc., by comparing the semantic representations of the sentences in the module to find the differences (i.e. the addition of a "inceptive" property on the event) and then mapping those differences to the differences found in the translated texts (for example, added words, affixes or changes in word order). Example elicitation modules have been prepared (including their semantic representations) for a large variety of semantically-based phenomena. Similar techniques are also used to probe different semantic case frame realizations. Such a semantically-based "grammar discovery procedure" is the means currently employed in LA. This grammar discovery procedure can be used to quickly describe how a particular language encodes a wide range of meaning-based

communication. The resulting computational description can then be used in the embedded text generation system to enable automatic translation. A grammar discovery procedure guided by semantics will obviously not yield a complete description of a language. It will not document *everything* that can be said in the language; however, we argue that it produces a practical description that will enable future generations to answer the question, "How do you say … in this language?" The approach is also very efficient in terms of the number of man-hours of linguistic work required. Our experience is that (under the right circumstances) a field linguist will require less than a month to complete the process. We expect this timeframe to decrease further as additional techniques such as those used in BOAS and LinGO are added to LA.[2] This type of grammar discovery is also very suitable for a workshop situation where many languages within a single language family could work together.

One valid argument against such an approach comes from linguistic circles. The current trend in linguistic research discourages elicitation, relying instead on the analysis of naturally occurring texts and dialogues. For example, a respected linguist involved in and relatively supportive of LA commented that "I am, in general, a bit reluctant to use ready-made questionnaires, for all sorts of reasons - some of which you mention yourself. It so happens that my personal interest has always been on naturalistic speech... I have always paid a lot of attention to what actually shows up in everyday spoken speech…" (Alex François, personal communication). We understand and accept this inclination towards naturally occurring texts over elicited texts, and in a "normal" situation we would completely agree. However, with the extinction of thousands of languages imminent, more radical techniques are needed. Elicitation techniques are also supported in the linguistic literature, for example, Ameka et al. (2006) state that 'limiting what the grammar should account for to a corpus [of naturally occurring texts] also overlooks the fact that speakers may have quite clear and revealing judgements' and 'the view...that grammars should be answerable just to a published corpus seems an extreme position in practical terms.' And again, Gippert et al. (2006) add their warning that 'without theoretical grounding language documentation is in the danger of producing 'data graveyards', i.e. large heaps of data with little or no use to anyone.' We believe that the semantic-based grammar discovery methodology adds this theoretical grounding.

We also add the argument that "the proof is in the pudding." Allman, et al. (2012) documents that a grammar discovery procedure such as described above combined with a capable knowledge acquisition and text generation environment such as found in LA can produce translations that are as accurate and readable to native speakers as manual translations and that these results indicate that the underlying language description is accurate, natural and broad-coverage.

4 Document authoring: a bridge to practical MT (and language description) in endangered languages

We have already argued that a semantically-based language description environment is superior to a transfer-based system. We will try to bolster that argument here. In terms of machine translation, the analysis of a source text will always be the bottleneck in terms of translation quality. On the other hand, an interlingual text generation process is relatively simple and accurate - assuming the presence of an accurate semantic description of the input text. Furthermore, a semantic description "language" is much simpler than natural languages since it has no ambiguity, fewer atoms (concepts vs. words), and fewer "syntactic" combinations. This leads to an economy when trying to describe how a particular language encodes it (as opposed to trying to describe how a language would encode arbitrary free text from a source language). And finally, as described above, a semantic-based description provides the framework for efficient and potentially machine learnable acquisition of grammar via an organized grammar discovery procedures.

The glue that holds this together is the concept of "document authoring." Authoring a semantic description of a text (or of the elicitation modules) can be accomplished through a semi-automatic authoring interface. Such an interface typically accepts a standardized (or "controlled") subset of a natural language as its input. The input is run through an analyzer and the results are visually presented to the user, who checks and/or assigns semantic concepts and relationships. The

[2] The discovery process itself as well as the underlying semantic representation language need to be refined and validated by our colleagues; we expect such refinements to also improve efficiency.

steps in preparing a semantic analysis of a text or set of elicitation sentences is thus: 1) manually "translate" the text into the controlled language, 2) run this through the automatic analyzer, and 3) manually check and correct the resulting semantic analysis. Although unlimited free text cannot be translated in an LA language project, a wide variety of texts can be semantically authored. This process only needs to be done once and the results can then be used for any language. See (Beale, et al. 2005) for more information on document authoring in the context of endangered languages.

We believe that a semantically-based description of a language is the key to the practical description of endangered languages. It provides an inherently efficient framework for language description in the field. The resulting description not only provides invaluable data for future linguists, but also enables present-day translation capabilities that can aid in language revitalization. A document authoring system provides the means for overcoming one of the main drawbacks to a semantically-based system in that it allows for a relatively quick, once-for-all preparation of semantic representations that can be used in a grammar discovery procedure and in machine translation of texts.

We now present longer summaries of Linguist's Assistant and BOAS.

5 Linguist's Assistant

The Linguist's Assistant (LA) is a practical computational paradigm for describing languages. LA is built on a comprehensive semantic foundation. We combine a conceptual, ontological framework with detailed semantic features that cover (or is a beginning towards the goal of covering) the range of human communication. An elicitation procedure has been built up around this central, semantic core that systematically guides the linguist through the language description process, during which the linguist builds a grammar and lexicon that 'describes' how to generate target language text from the semantic representations of the elicitation corpus. The result is a meaning-based 'how to' guide for the language: how does one encode given semantic representations in the language?

Underlying this approach to knowledge acquisition in LA is a visual, semi-automatic interface for recording grammatical rules and lexical information. Figure 1 shows an example of one kind of visual interface used for "theta-grid adjustment rules." The figure shows an English rule used to adjust the "theta grid" or "case frame" of an English verb. Grammatical rules typically describe how a given semantic structure is realized in the language. The whole gamut of linguistic phenomena is covered, from morphological alternations (Figure 2) to case frame specifications to phrase structure ordering (Figure 3) to lexical collocations – and many others. These grammatical rules interplay with a rich lexical description interface that allows for assignment of word-level features and the description of lexical forms associated with individual roots (Figure 4). As stated above, the user is currently responsible for the creation of rules, albeit with a natural, visual interface that often is able to set up the requisite input semantic structures automatically. As mentioned, we also seek to collaborate with researchers to enable semi-automatic generation of rules similar to what can be found in the Boas (McShane, et al., 2002), LinGO (Bender, et al., 2010), PAWS (Black and Black, 2009) and Avenue (Probst, et al., 2003) projects. Such extensions will make LA accessible to a larger pool of linguists and will shorten the time needed for documenting languages.

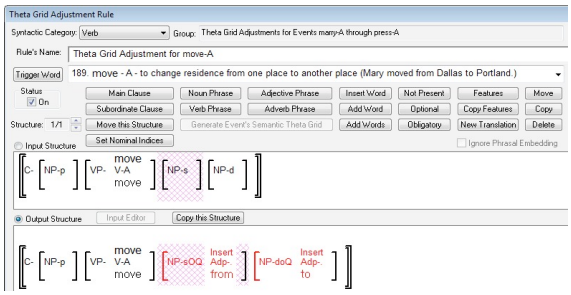

Figure 1. Visual interface for grammatical rules

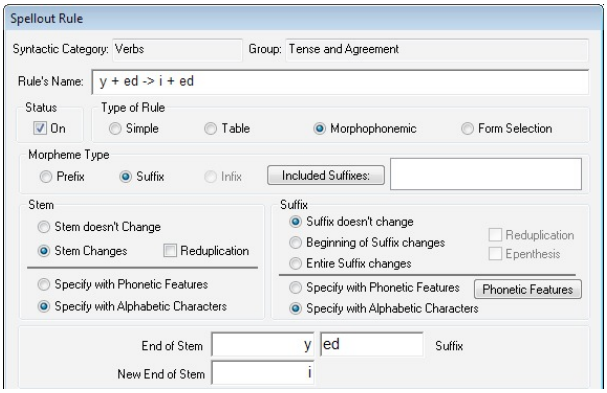

Figure 2. Morphological alternation rule

Integrated with these elicitation and description tools is a text generator that allows for immediate confirmation of the validity of grammatical rules and lexical information. We also

provide an interface for tracking the scope and examples of grammatical rules. This minimizes the possibility of conflicting or duplicate rules while providing the linguist a convenient index into the work already accomplished. And finally, we provide a utility for producing a written description of the language - after all, a computational description of a language is of no practical use (outside of translation applications) unless it can be conveniently referenced. Refer to Beale (2012) for a comprehensive description of Linguist's Assistant.

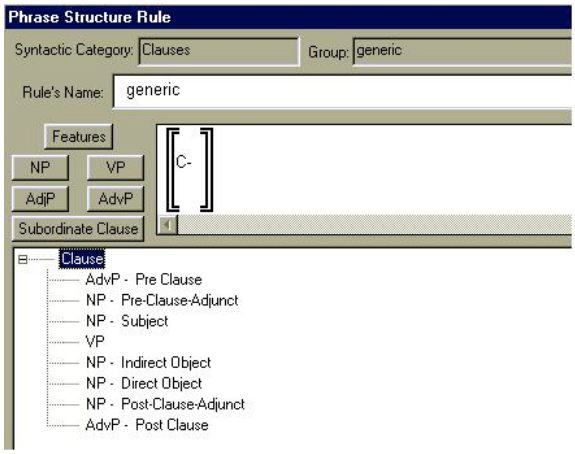

Figure 3. Phrase structure ordering rule

	Stems	Glosses	infinitive	present indic 1st sing
1	aprend	learn	aprender	aprendo
2	habl	speak	hablar	hablo
3	ten	have	tener	tengo
4	viv	live	vivir	vivo

present indic 2nd sing	present indic 3rd sing	present indic 1st pl	present indic 3rd pl
aprendes	aprende	aprendemos	aprenden
hablas	habla	hablamos	hablan
tienes	tiene	tenemos	tienen
vives	vive	vivimos	viven

Figure 4. Lexical forms for Spanish

LA has been used to produce extensive grammars and lexicons for Jula (a Niger-Congo language), Kewa (Papua New Guinea), North Tanna (Vanuatu), Korean and English. Work continues in two languages of Vanuatu (and a new avenue of research has recently opened as a result of a partnership with De La Salle University in the Philippines). The resulting computational language descriptions have been used in LA's embedded text generation system to produce a significant amount of high-quality translations. Figures 5 and 6 present translations of a section of a medical text on AIDS into English and Korean. Please reference Beale et al. (2005) and Allman and Beale (2004; 2006) and Allman et al. (2012) for more information on using LA in translation projects and for documentation on the evaluations of the translations produced. We argue that the high quality achieved in translation projects demonstrate the quality and coverage of the underlying language description that LA produces.

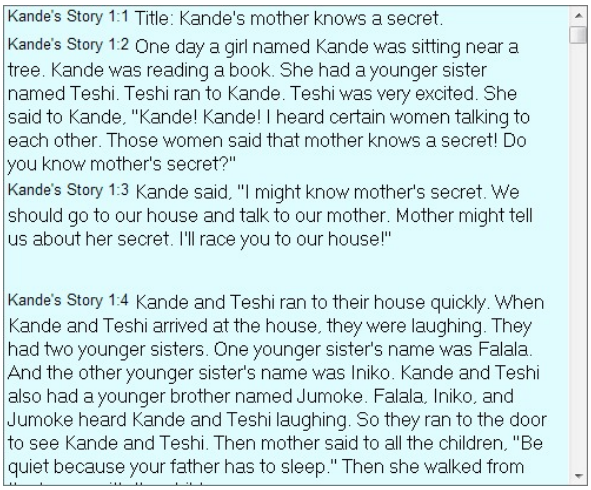

Figure 5. English translation of a medical text

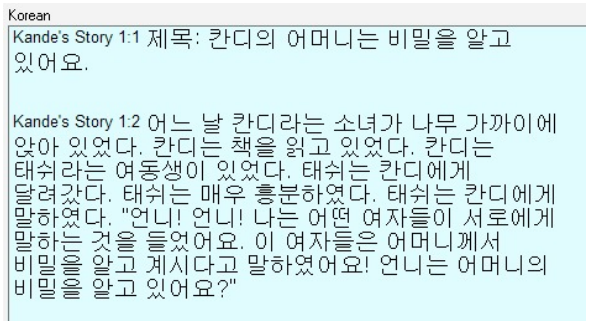

Figure 6. Korean translation of a medical text

6 BOAS

Boas (McShane et al. 2002) is an example of a typology-based questionnaire approach that can be useful for quickly eliciting certain properties of a language. This section is meant as an overview that is representative of this class of programs. The author has no direct connection with the Boas system; permission was given to use the following description.

Boas is used to extract knowledge about a language, L, from an informant with no knowledge engineer present. Boas itself leads the informant through the process of supplying the necessary information in a directly usable way. In order to do this, the system must be supplied with 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" for the language. Examples of parameters, values and their realizations that play a role in the Boas knowledge-elicitation process are shown in Table 1. 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.

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, etc.	numerals in L, digits, punctuation marks (commas, periods, percent signs, etc.) or a lack thereof in various places
Sentence Boundary	declarative, interrogative, imperative, etc.	period, question mark(s), exclamation point(s), ellipsis, etc.
Grammatical Role	subjectness, direct-objecness, indirect-objectness, etc.	case-marking, word order, particles, etc.
Agreement (for pairs of elements)	+/- person, +/-number, +/- case, etc.	flective, agglutinating or isolating inflectional markers

Table 1: Sample parameters, values and means of their realization

The selection of parameters and values in Boas is made similar to a multiple choice test which, with the necessary pedagogical support, can be carried out even by an informant not trained in linguistics. This turns out to be a crucial aspect of knowledge elicitation for rare languages, since one must prepare for the case when available informants lack formal linguistic training. 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.

The five major modules of the Boas system are:

Ecology:
- inventory of characters
- inventory and use of punctuation marks
- proper name conventions
- transliteration
- dates and numbers
- list of common abbreviations, geographical entities, famous people, etc. (which can be expanded indefinitely)

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

Syntax:
- structure of the noun phrases: NP components, word order, etc.
- 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:

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 a L-to-English lexicon by a) translating entries from an English seed lexicon, b) importing then supplementing an on-line bilingual lexicon, c) composing lists of words 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 user indicates that nouns in L inflect for number, the page shown in Figure 7 will be accessed. Explanatory support for decision-making is provided in help links at the bottom of the page.

Boas offers a good example of an advanced elicitation system by combining extensive and parameterized descriptive material about language, a rich set of expressive means in the user interface, and extensive pedagogical resources.

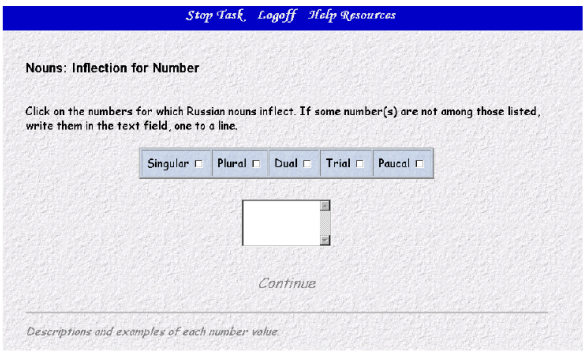

Figure 7: Selecting the values for number for which nouns inflect

7 Conclusion

A quick perusal of the grants awarded by NSF/NEH in the DEL program over the last five years confirms the underlying assumption of this paper: the DEL program funds projects that produce or aid audio and textual documentation (i.e. data) on endangered languages. We argued that descriptive work might return a higher payback as regards to potential linguistic utilization in the future. We also argued that the value of descriptive work in revitalizing languages today exceeds that of purely documentary work. Furthermore, we described several lines of research that would allow such descriptive work to proceed, along with a rationale for continued research to improve the computational tools employed in such work. Linguist's Assistant and Boas represent two sides of the same coin for descriptive work in minority languages. Cooperation between the various research programs that represent each side of that coin is critical to attaining a total solution to describing endangered languages.

References

Tod Allman, Stephen Beale and Richard Denton. 2012. Linguist's Assistant: A Multi-Lingual Natural Language Generator based on Linguistic Universals, Typologies, and Primitives. In Proceedings of 7th International Natural Language Generation Conference (INLG-12), Utica, IL.

Tod Allman and Stephen Beale. 2006. A natural language generator for minority languages. In Proceedings of SALTMIL, Genoa, Italy.

Tod Allman and Stephen Beale. 2004. An environment for quick ramp-up multi-lingual authoring. *International Journal of Translation* 16(1).

Felix Ameka, Alan Dench & Nicholas Evans. 2006. Catching language: the standard challenge of grammar writing. Berlin: Mouton de Gruyter.

Stephen Beale. 2012. Documenting endangered languages with Linguist's Assistant. *Language Documentation and Conservation* 6(1), pp. 104-134.

Stephen Beale, S. Nirenburg, M. McShane, and Tod Allman. 2005. Document authoring the Bible for minority language translation. In Proceedings of MT-Summit, Phuket, Thailand.

Emily Bender, Michael Wayne Goodman, Joshua Crowgey and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: inferring large-scale typological properties. In Proceedings of the ACL 2013 workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities.

Emily Bender, S. Drellishak, A. Fokkens, M. Goodman, D. Mills, L. Poulson, and S. Saleem. 2010. Grammar prototyping and testing with the LinGO grammar matrix customization system. In Proceedings of the ACL 2010 System Demonstrations.

Sheryl Black and Andrew Black. 2009. PAWS: parser and writer for syntax: drafting syntactic grammars in the third wave. http://www.sil.org/silepubs/PUBS/51432/SILForum2009-002.pdf.

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

Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel. 2006. *Essentials of language documentation.* Berlin: Mouton de Gruyter.

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

Marjorie McShane, Sergei Nirenburg, Jim Cowie, and Ron Zacharski. 2002. Embedding knowledge elicitation and MT systems within a single architecture. *Machine Translation* 17(4), pp. 271-305.

Katharina Probst, Lori Levin, Erik Petersen, Alon Lavie and Jaime Carbonell. 2003. MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation* 17(4), pp. 245-270.