Slavic as Testing Grounds for a Linguistic Knowledge Elicitation System

Marjorie McShane, Stephen Helmreich,
Sergei Nirenburg, Victor Raskin
New Mexico State University

1. Introduction

Among many other tasks, NLP (natural language processing) systems must be able to determine which words to list in the dictionary and how to determine citation forms in texts—a non-trivial matter for languages that have extensive inflectional and/or derivational morphology, spelling mutations, etc. This issue is central for Boas (Nirenburg and Raskin 1998), the linguistic knowledge elicitation component for the Expedition project of the Computing Research Laboratory at New Mexico State University. (See *http://crl.nmsu.edu/expedition* for an overview.) The goal of Expedition is to develop the capability for fast deployment of a machine translation system between any so-called "low-density" language (one lacking significant machine-tractable resources) and English. Boas must guide non-expert human informants through questions about the morphology, syntax, lexical stock, syntax, and ecology (letters, symbols, punctuation, etc.) of their language. Here we focus on those components of Boas associated with morphology.

Since Boas must accommodate any low-density language, and since linguistic materials are scarce or unavailable for many such languages, the system cannot be 'primed' to cover individual languages, leading to questions of coverage and efficient testing. With respect to coverage, a broad survey of languages reveals that even significantly diverse morphological phenomena fall into groups whose elements can be handled similarly by Boas. With respect to testing, since the major Slavic languages contain morphological phenomena representing each of the groups we have delineated, we hypothesize that if Boas provides sufficient

declarative knowledge for the processing engines in trials on Slavic languages (to be carried out by personnel at the lab), it should have similar success with the low-density languages for which it is being designed.

The rules of the Boas game are as follows: one language informant, who need not be a linguist, and one programmer, who need not be versed in NLP, will work for six months guided by the materials resident in the system. They may start from scratch or may incorporate existent on-line resources if the programmer can make them compatible with the relevant components of Boas. A strong informant-programmer team can incorporate extensive language-specific plug-ins, while a less experienced team can limit themselves to tasks explicitly set by Boas. At the end of six months, a moderate-quality, broad-coverage translation system should be in place.

This paper focuses on the scope of morphological phenomena presented by natural language, the distribution of these phenomena among the components of Boas, and the role that Slavic languages play in building and testing the system. The paper is organized as follows. Section 2 discusses the morphologically relevant components of Boas and the Slavic phenomena that will test each one; Section 3 discusses why most derivational morphology is not handled productively in Boas; and Section 4 concludes the paper.

2. Morphology in Boas

Morphological phenomena will be gathered in various components of Boas: inflectional paradigms, derivational affixation, and the closed- and open-class lexicons, as shown in Figure 1 (SL stands for ‘source language’). The sections below describe each component, with emphasis placed on the method of elicitation employed, the types of cross-linguistic phenomena targeted, and the Slavic examples that will serve to test the system.

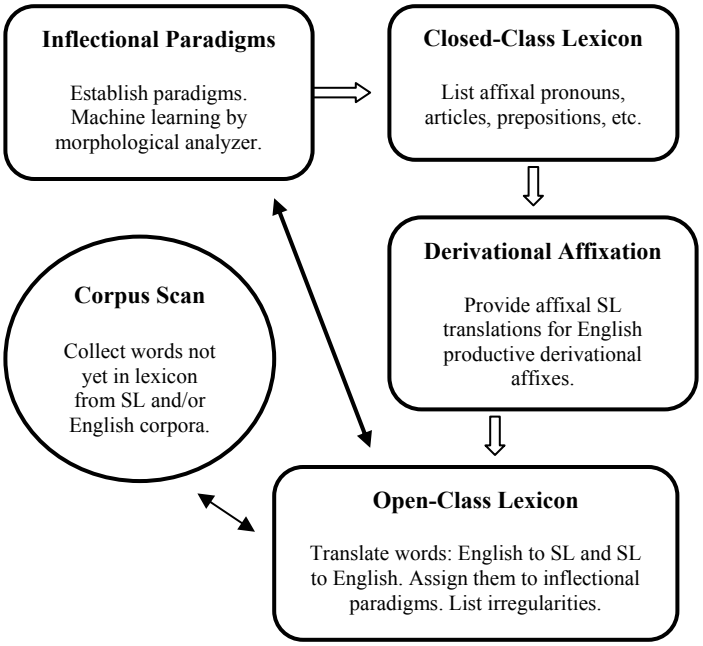

Figure 1. The Stages of Morphological Acquisition in Boas.

2.1. Inflectional Paradigms

The establishment of paradigms is intended to speed up open-class lexical acquisition: rather than type in all the inflectional forms for each of the 60,000 English word senses to be translated into the source language (SL), the informant can type in just the citation form and then assign the word to one of the paradigms established earlier.

Dividing inflecting words into paradigms is anything but a trivial task for linguists, not to mention the non-expert users of Boas, since paradigms represent the epitome of regularity but natural language is often far from regular. Therefore, in addition to supplying extensive pedagogical support, Boas walks the informant through the process of establishing paradigms, first determining for which combinations of parameter values the given

part of speech (PoS) inflects,[1] then providing a paradigm template that associates each licit (as indicated by the informant) combination of parameter values with a text box, as the miniature paradigm template in Figure 2 shows.

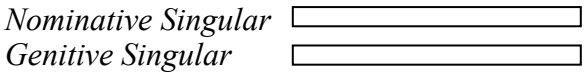

Figure 2. Templates for Inflectional Paradigms.

Then comes the hard part: the informant must select words that represent all the regular patterns of inflection in SL, interpreting 'regular' in a computationally valid sense (in contrast to the extended notion of paradigm found in many grammars). Obviously, no non-expert is expected to know intuitively what a computationally valid paradigm is, so Boas will help. First, the informant posits one citation form for what he thinks is each major inflectional pattern—a sort of first approximation of 'the truth'.[2] Let us assume he selects four patterns, instantiated by *word1, word2, word3, word4*. Each of these citation forms is generated in each of the textboxes of the paradigm template, and the informant edits the citation forms as necessary to reflect the given combination of parameter values (e.g., in Russian, the citation form *stol* would be edited to *stola* in the box labeled Genitive Singular).

After the informant establishes these paradigms, the morphological analyzer generates a preliminary set of rules for them (Oflazer and Nirenburg 1999). Then the informant provides

[1] Extensive lists of parameters and values are presented in the form of tables containing brief definitions and check boxes. Further pedagogical support can be accessed via links to a comprehensive glossary of linguistic terms, each page of which contains links to related topics such that the glossary can be employed as a free-form tutorial for novice language informants. At all stages, the informant can supplement the supplied list of parameters and values.

[2] Although the informant is free to choose what form to consider the 'citation form', it should match the citation form found in any on-line dictionaries that might be employed.

three additional citation forms for each paradigm to test these rules: if the analyzer generates all correct inflectional forms, the user can be relatively confident that the rules for the given paradigm are correct and comprehensive, and the paradigm is set. If, however, the analyzer generates some incorrect forms, the informant manually corrects the errors and tells the analyzer to relearn the rules for the paradigm. This learning loop proceeds as long as necessary until the analyzer produces correct forms for a representative sample of words belonging to the paradigm.

This learning loop is necessary not only to create a robust morphological analyzer, but also to teach the informant the strengths and limitations of computational methods. For example, the informant will be told, and may test for himself, that mutations of a similar sort can be handled within a single paradigm as long as the inflectional endings are the same and an example of each mutation is provided explicitly. For example, the analyzer has proven capable of handling Russian mutations of the type *s/O* and *z/⊙* within a single verbal paradigm. Deviations from the major paradigms will be listed explicitly in the open-class lexicon.

It goes without saying that the rich and complex inflectional morphology of Slavic languages provides rigorous testing grounds for Boas's morphological analyzer. More importantly, however, Slavic languages contain all the *categories* of paradigm-related complications that we have found in surveys of other languages. Consider the following sampling of phenomena:[3]

Stem-internal alternations will be handled by positing multiple stems in the lexicon for each word assigned to the given paradigm. In Slavic, listing multiple stems will account for Russian fleeting vowels (*otec* 'father.NOM.SG' ~ *otca* 'father.GEN.SG') and Belorussian graphotactic vowel reduction (*stol* 'table.NOM.SG.' ~ *stala* 'table.GEN.SG.'). Outside of Slavic, listing multiple stems

[3] For reasons of space, examples are presented sparingly; "etc." should be assumed in all instances. In addition, sources are not presented after each language example, as all the language examples were either taken from the grammars listed in the bibliography (page numbers are cited there) or gathered from in-house native speakers.

will account for suppletion in Comanche and Blackfoot verbal paradigms: in Comanche, intransitive verbs are suppletive for singular versus plural subjects, while transitive verbs are suppletive for singular versus plural objects (113); in Blackfoot, intransitive verbs have different stems for animate and inanimate subjects (*siksinámma* 'it.ANIMATE is black' / *siksináttsiwa* 'it.INANIMATE is black') (38).[4] To reduce overgeneration, the user will be asked to indicate, when possible, which stem is associated with which parameter value or combination of parameter values (e.g., stem 1: present tense, stem 2: past tense).

Boundary alternations can be incorporated into paradigms as long as they are mandatory; for example, one cannot include in a single paradigm words that do and do not have $s \rightarrow ś$ shift in the Present 1st Singular. The morphological analyzer has been positively tested for boundary alternations in Polish verbs (*wo**ż**ę* 'drive.1.SG.PRES' ~ *wo**z**isz* 'drive.2.SG.PRES') and is expected to work equally well for Finnish consonant gradation (*kau**pp**a* 'shop' ~ *kau**p**at* 'shops') and Blackfoot vowel shortening (*kakk**óó**wa* 'pigeon' ~ *kakk**ó**íksi* 'pigeons') (9).

Multiple realizations of an inflectional form can be incorporated into the paradigm, when regular. In Slavic, the future tense has regular duplicate forms in Polish (*będę robił ~ będę robić* 'will work.1.SG') and Ukrainian (*robitimu ~ budu robiti* 'will work.1.SG'). In Blackfoot, many verbs have two or three acceptable past tense forms, which are formed by rules of different paradigms: e.g., *Nitókska'si/Nitsííkska'si* 'I ran' (36). If duplicate forms are idiosyncratic or apply to a limited number of lexical items, they can be added to the paradigm at the stage of open-class acquisition. This applies to the so-called 2nd locative in Russian (*lese*. ~ *lesu*. 'forest') and to variant realizations of absolutive nouns in Nahuatl (*tochin ~ tochtli* 'rabbit') (17).

[4] Page numbers for examples that are drawn from grammars are presented in 'bare' form, since there is generally only one grammar per language listed in the references.

Slight Irregular Modifications of Paradigms will be handled by overriding one or more inflectional forms during open-class lexical acquisition. The user will assign the word to the paradigm that has the closest fit, then click on the 'show me forms' button to see what inflectional forms the morphological analyzer generates. The forms are generated in an editable field where the user can make all necessary corrections. In Slavic, this "manual override" process can be used for Russian masculine plurals in stressed *–a* (*adres* 'address' ~ *adresa* 'addresses') and for Polish masculine dative singulars in *–u* (*brat* 'brother.NOM' ~ *bratu* 'brother.DAT'). Outside of Slavic, it should be particularly useful for languages that have highly complicated, largely unpredictable inflectional patterns. For example, in describing nominal inflection in Irish, Ó'Siadhail (1989:159) notes,

> It is very difficult to predict how the plural of any given noun is formed; nevertheless the phonetic environment and the function of the plural play a certain part in determining the formation.

Thus, rather than force the user to create some ideal system of paradigm delineation, which might necessitate positing hundreds of paradigms, Boas permits him to posit far fewer paradigms and tweak them as necessary during open-class lexical acquisition.

Multiple Paradigm Templates can exist for a given part of speech. For example, *pluralia tantum* nouns have no singular forms (e.g., Russian *časy* 'watch'), and certain classes of verbs lack a given mood or aspect in some languages (e.g., intransitive verbs commonly have no passive). Multiple templates can also exist for reasons idiosyncratic to a given language: for example, Nahuatl permits plurals only for animate nouns, so inanimate paradigms will be half the size of animate ones (16). Similarly, Blackfoot has no singular-plural distinction for non-particular nouns (11). The user can create as many different templates as necessary when establishing paradigms in the morphological component of Boas.

The only paradigm-related process for which there is no obvious test material in Slavic is inflectional reduplication, which is used, for example, to create different verbal aspects in Ponapean (tense is conveyed pragmatically; assume past tense in this example): *kang* '(I)-ate' ~ *kangkang* '(I)-was-eating.DURATIVE' (74). Reduplication is also used to form plurals in Nahuatl according to the following rule: reduplicate the first syllable and add the suffix *–tin: teuctli* 'lord' ~ ***te****teuctin* 'lords' (17).[5]

Since the morphological analyzer in Boas is a finite state machine, it interprets only strings of characters, not patterns. Therefore, reduplication must be handled by scripts unconnected to the morphological analyzer. When the informant is establishing inflectional paradigms, he will indicate what, if any, forms are generated via reduplication. These forms will be excluded from the learning loop of the morphological analyzer. Then, in a separate task, the informant will select from a list the appropriate pattern of reduplication and provide a handful of examples. Based on this information, a script will be automatically generated, which the informant and programmer can modify, as necessary. If the patterns of reduplication are too complex to be captured in explicit rules, the reduplicative forms will have to be listed individually for every lexical item.

2.2. Closed-Class Lexicon

In some languages, closed-class items regularly attach to stems and must be stripped off to reveal the citation form. The affixal realization of closed-class items will be captured during closed-class lexical elicitation, in which English closed-class senses can be translated into SL as a *word, phrase, affix*, or *feature* (e.g., case). Below are some Slavic and non-Slavic examples.

(1) BULGARIAN articles: *more* ~ *more****to*** 'sea ~ the sea'

[5] Derivational reduplication will not be handled productively, for reasons explained in Section 3.

RUSSIAN reflexive/reciprocal affix: *myt'* ~ *myt'**sja*** 'wash ~ wash oneself'
PERSIAN possessive pronouns: *kt|b* ~ *kt|b**t*** ' book ~ your book'
ARABIC prepositions: *byt* ~ ***b**byt* 'house ~ in a house'
Cree possessive pronouns: *astotin* ~ ***nit**astotin* 'cap ~ my cap' (44)
NAHUATL possessive pronouns: *michin* ~ ***no**mich*[6] 'fish ~ my fish' (26)
PONAPEAN demonstratives: *wahr* ~ *wahr**et*** 'canoe ~ this canoe' (86)
MALAY interrogative particle: ***-kah*** attaches to the word questioned (123)
COMANCHE reflexive/reciprocal affix: ***na-*** attaches to the verb (103)
MALAY prepositions: *rumah sakit* ~ ***di**-rumah sakit* 'hospital ~ in the hospital' (79)

When applicable, allomorphs of attached affixes and/or their inflectional forms will be listed explicitly in the closed-class lexicon. In Bulgarian, e.g., allomorphs will be listed for the masculine singular definite article: *-ъt, -jat, -a, -ja*; in Cree, they will be listed for the possessive pronoun 'my': *ni-, nit-, n-* (44-46); and in Comanche, they will be listed for the locative postpositions *on, under, in, at, from,* and *beside*, each of which has between two and five allomorphs (73).

For each affixal realization of a closed-class sense, the user will indicate what part(s) of speech it can attach to for purposes of disambiguation. For example, in Russian and Polish, instrumental *with* is reflected on noun phrases by the feature 'Instrumental case', while in Ponapean it is reflected on verbs by the suffix *–ki*. Compare the translations of *I will write with this pen.*

(2) I pahn nting**ki** pehnet. [Ponapean: 224]
I will write-**with** pen-this

[6] The absolutive suffix is removed when the possessive prefix is added.

Ja	budu	pisat'	ètoj	ručkoj.	[Russian]
I	will	write	this.INSTR	pen.INSTR	

2.3. Derivational Affixation

For reasons explained in Section 3.1, Boas will not attempt to generate rules for all derivational processes in SL. There is, however, one subset of derivational forms that will be handled productively: forms created by affixes that have a direct English counterpart. Productive affixes in English have been divided into several dozen semantic classes, like 'simple negation' (*un-, in-, im-, non-*), 'very' (*super-, extra-*), 'against' (*anti-*). The informant will supply corresponding affixes in SL, if they exist. Although this elicitation might yield no results for some languages, for others it will prove fruitful: for example, it will catch negated verbs in Czech, which are formed by the prefix *ne-* (*mluvím* ~ ***ne****mluvím* '(I) speak ~ (I) don't speak') and negated adjectives in Ponapean, which are formed by the prefix *sa-* (*peik* ~ ***sa****peik* 'obedient ~ disobedient'). Likewise, it will cover Blackfoot words modified by affixes meaning 'very' (*iik-*) and 'extraordinarily' (*sska'-*) (92). The English generator will be responsible for mapping, 'not + proper' to *improper* and 'not + obedient' to *disobedient* (avoiding invalid formulations like **inproper, *inobedient* and **non-proper, *non-obedient*).

This elicitation also gathers affixes whose main function is to change the part of speech with little or no accompanying semantic shift (e.g., noun → adjective in *success* → *successful*). This subtype of derivational morphology was singled out because it permits (relatively) direct transfer from SL to English, something impossible for many other derivational processes.

2.4. Open-Class Lexicon

The open-class lexicon in Boas, as everywhere, is the seat of things unpredictable: translations of words, their paradigm membership (if not fully predictable by the spelling of the citation form),

irregular inflectional forms, allomorphs, etc. As is well known, the bigger the lexicon, the better the machine translation system. The challenge for Boas lies in maximizing the effectiveness of limited lexicon-building resources—namely, having just one language informant devoting less than six months to the task. Due to these time constraints, it is unlikely that the informant will be able to translate all of the 60,000 English word senses resident in Boas, as well as all the words in the SL corpus not covered by these word senses. Therefore, the informant is encouraged to organize open-class lexical acquisition in the way most efficient for his language and most in keeping with his goals (e.g., coverage of articles dealing with medicine or nuclear proliferation). Below we consider the most morphologically salient aspects of open-class lexical acquisition, again focusing on classes of phenomena and the Slavic examples that will serve as testing material.

Non-Inflectional Spelling Variants. Some languages have non-inflectional spelling variants of lexical items, like Russian *značen'e/značenie* 'meaning' and *predstavlen'e/predstavlenie* 'presentation'. Mokilese, for example, has so-called cluster metathesis, by which clusters composed of a labial and a velar stop can occur in either order: *apkas* ~ *akpas* 'now'. Since these alternations are rather idiosyncratic, they are best listed in the open-class lexicon as allomorphs.

More problematic are widespread alternations, like Irish lenition and eclipsis, which are phonologically driven processes that modify word-initial consonants based on the preceding lexical item. Table 1 presents a sample of such alternations:

Table 1. Lenition and Eclipsis in Irish

basic consonant	lenited consonant	eclipsed consonant
c	ch	gc
b	bh	mb
g	gh	ng

Lenition can occur, for example, after the preposition *ar* 'on': ***b**ad* 'boat' → *ar* ***bh**ad* 'on (the) boat'. Eclipsis can occur after the

positive interrogative particle *an*: ***b****ris* 'break' → *An* ***mb****riseann se...?* 'Does he break...?'

The more primitive but fool-proof way to handle such alternations would be to list the variant spellings as allomorphs in each relevant lexical entry, but this method would carry high time costs for the informant. Alternatively, the linguist/programmer team could write lexicon-wide rules for all such alternations, but this would require a type of unguided rule writing potentially out of the reach of less experienced linguist/programmer teams (much depends on their respective knowledge of NLP). We are currently working to develop rule templates to assist in this process, and will test them on the Ukrainian word-initial alternations *u-/v-* and *i-/j-* as in: *učitel/včitel* 'teacher' and *idu/jdu* '(I) go'.[7]

Rule writing carries its own complications, as evidenced by the Ukrainian phenomena mentioned above. For example, imposing lexicon-wide rules can lead to overgeneration: in Ukrainian, place names like *Ural* 'Urals' and foreign words like *uran* 'uranium' do not have a *v-* variant (27-8). In most instances, such overgeneration is irrelevant (and simply adds a bit of dead weight to the lexicon), since the translation system will be primarily interpreting, not generating, SL.[8] However, in some instances overgeneration will lead to ambiguity. For example, Ukrainian *uklad* means 'regime' while *vklad* means 'contribution'. A lexicon-wide rule that puts *u-* and *v-* in free variation word initially will cause each instance of *uklad* and *vklad* to be incorrectly tagged with two meanings. But considering the amount of lexical and other ambiguity that all translation systems face, these additional sources of ambiguity are relatively insignificant.

3. Derivational Morphology

[7] These letters also alternate as freestanding prepositions, but as prepositions they will simply be listed as allomorphs in the closed-class lexicon.

[8] 'Primarily' interpreting because the morphological analyzer will be used to generate forms in the morphological learning loop and in the open-class lexicon during paradigm selection/modification.

Words formed by derivational morphological processes present significant problems to MT systems even if the source language is known and can be prepared for individually. These problems are compounded when source language is unknown. The sections below detail the problems inherent in derivational morphology and Boas' rather unconventional approach to dealing with this aspect of the grammar.

3.1. Derivational Morphology: The Problem

Source language words formed by productive derivational processes (like the German *Donaudampfschiffahrtskapitän* 'Donau steam ship driver captain') will, in large part, not be captured by Boas' English-driven open-class lexical acquisition since such words are equivalent to multi-word English phrases. This presents a considerable problem for languages with widespread compounding (German, Swedish) and/or reduplication (Tagalog, Ponapean). One obvious way to handle derivational word formation would be to prepare the system to analyze such forms based on knowledge elicited from the informant. While creating a series of questions would be trivial ('How many roots can typically be joined in a compound?' 'What, if any, letters can be added between compounding roots?' 'Do compounding forms use different roots than non-compounding forms?'), processing the results presents significant complications. The problem lies in the fact that derivational forms are often semantically ambiguous and/or non-compositional. Thus, even correct formal analysis of derived words would often be of little help in SL-to-English lexical transfer.

Consider, for example, the Swedish surface form *frukosten*, which can have the following five parses (from Karlsson *et al*, 1995:28).

(3) 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. Dura (1998) suggests that the best way to deal with them is to list the most common compounds explicitly in the lexicon, then use these ready-made chunks as set units for further analysis of compounding forms.

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 never generate a correct English equivalent.
Derivational reduplication poses even more fatal problems, both in formal and in semantic terms. For example, in Tagalog the meaning ‘a vendor of the product indicated by the base’ is created as follows (103): [prefix *mag*] + [first two letters of the base, reduplicated] + [base].

(4)	*mag**bu**bulaklak*	‘flower vendor’	(*bulaklak* ‘flower’)
	*mag**ka**kandila*	‘candle vendor’	(*kandila* ‘candle’)

Clearly, such word formation processes can only be captured by language-specific rules that are (i) difficult, if at all possible, to elicit in a generalized way, (ii) limited to certain semantic classes of lexical items, and (iii) not always strictly compositional in meaning.

Further complications arise from the ‘theme and variations’ nature of reduplication. For example, Turkish color terms can be intensified by reduplication that includes various consonant additions/mutations: *siyah ~ simsiyah* ‘black ~ very black’, *mor ~ mosmor* ‘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.*

3.2. Derivational Morphology: Boas’ Answer

Because of the complications associated with derivational word formation, Boas will treat it lexically, assisted by the on-line corpus.[9] More specifically, the informant will begin open-class lexical acquisition by providing SL equivalents for some minimum number of high-frequency English words. Open-class acquisition can then proceed in a number of ways, as deemed best by the informant. He can: (i) continue to translate English word senses resident in Boas; (ii) scan the SL corpus and translate the most frequent SL words into English; or (iii) scan an English corpus devoted to some special topic and translate the most frequent words therein. These methods of lexical acquisition can be carried out in loops, as dictated by the informant (e.g., 200 English words, then 300 words from the SL corpus, then another 200 words from the SL corpus...).

For languages with extensive derivational morphology, the SL corpus scan should be used broadly, since English-driven lexical acquisition will miss many common words (cf. 'flower vendor' and 'very black' above). For languages with less extensive derivational morphology, English-driven lexical acquisition should provide relatively good coverage.

When the corpus scan is employed, it will generate a list of unknown words in order of frequency for the informant to potentially translate. 'Potentially' is an important notion, as a given corpus might include a large number of rare words.

While Slavic languages do not show extensive compounding or reduplication, they have other derivational word formation processes on which the corpus-scan method of lexical supplementation can be tested. For example, English-driven lexical acquisition will not capture quasi-productive prefixation (like Russian *dopet'* 'sing to the end' and *pereutomit'sja* 'get overtired') or diminutive/endearing forms (like Russian *košečka* < *koška* 'cat'). These forms, like compounds and reduplicative forms, often do not have entirely compositional semantics.

---

[9] The primary goal of Boas is to translate on-line resources into English; thus we assume the existence of on-line texts that can be compiled into a corpus by the programmer.

4. Conclusions

It is virtually impossible to build a knowledge elicitation system that specifically caters to every linguistic eventuality encountered in every natural language, since it is virtually impossible even to list all such eventualities. Therefore, when facing the task of creating a knowledge elicitation system with maximal coverage, strategy plays a crucial role. Chance also plays some role. In the case of Boas, chance dictated that the linguist developers had more knowledge of Slavic languages than, say, of African languages, making testing of the former language group more realistic than testing of the latter. However, strategy proves no less important: by dividing language phenomena into typologically valid classes whose members can be handled similarly, we can test a given elicitation process on Slavic languages with relative confidence that equally good results will be achieved in the more 'exotic' languages for which testing lies beyond our reach.

REFERENCES

Charney, J.O. 1993. *A Grammar of Comanche*. Lincoln: University of Nebraska Press.

deBray, R.G.A. 1980. *Guide to the East Slavonic Languages*. Columbus, Ohio: Slavica Publishers.

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

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

Harrison, S.P. 1976. *Mokilese Reference Grammar*. Honolulu: The University Press of Hawaii.

Heim, M. 1982. *Contemporary Czech*. Columbus, Ohio: Slavica Publishers, Inc.

Hubenova, M. et al. 1968. *A Course in Modern Bulgarian*. Columbus, Ohio: Slavica Publishers, Inc.

Karlsson, F. 1995. "Designing a Parser for Unrestricted Text." In Karlsson, F., A. Voutilainen, J. Heikkilä, A. Anttila (eds.) *Constraint Grammar*. New York: Mouton de Gruyter.

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

Medushevsky, A. and R. Zyatkovska. 1963. *Ukrainian Grammar*. Kiev: Radyanska shkola.

Nirenburg, S. and V. Raskin. 1998. "Universal Grammar and Lexis for Quick Ramp-Up of MT Systems," in *COLING-ACL '98* (36th Annual Meeting of the Association for Computational Linguistics), vol. II, 975-979.

Oflazer, K. and S. Nirenburg. 1999. "Practical Bootstrapping of Morphological Analyzers." In *Proceedings of the Workshop on Computational Natural Language Learning at EACC '99*, Bergen, Norway.

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

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

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

Rehg, K.L. 1981. *Ponapean Reference Grammar*. Honolulu: The University Press of Hawaii.

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

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

Computing Research Laboratory
Box 30001, 3CRL
New Mexico State University
Las Cruces, NM 88003
marge@crl.nmsu.edu
http://crl.nmsu.edu/Staff.pages/Technical/marge.htm