Modularity in Knowledge Elicitation and Language Processing

MARJORIE MCSHANE AND RON ZACHARSKI
Computing Research Laboratory
New Mexico State University

1. Introduction

This paper discusses the role of modularity in the knowledge elicitation component of a natural language processing system. The system at hand, Expedition, is intended to develop the capability for fast deployment of a machine translation (MT) system between any so-called “low-density” language (one lacking significant machine-tractable resources) and English.[1] The knowledge-elicitation component of Expedition, called Boas, guides non-expert human informants through questions about the morphology, syntax, lexical stock, and ecology (letters, symbols, punctuation, etc.) of their language. The collected source-language (SL) information provides static knowledge to fill in the blanks of the MT template. Once the informant provides all the requested information, he pushes a button and receives a moderate-quality MT system, with no need for further human intervention.

The linguistic challenges for the developers of Boas can be summarized as follows: how does one gather *all* the necessary information about *all* the phenomena that can occur in *any* natural language in a way that is both understandable to a non-expert informant and machine tractable without post-elicitation human intervention? We have chosen to start with the simplifying assumption that knowledge about a particular language is divided into modules that can be dealt with independently of each other. This modular approach covers the majority of facts about language in a pedagogically sound and computationally supportable fashion; however, it does not cover all facts, as many well-known language phenomena are cross-modular.

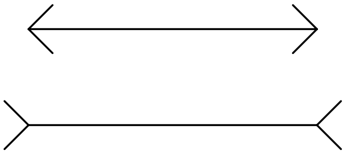
After presenting a brief overview of modularity (§2), we describe the basic modules of Boas (§3). We then present some language phenomena that cannot be handled in a strictly modular system and describe the micro-components being built to cover them (§4).

2. Modularity: Background

Modularity is the notion that complex systems are partitioned into a set of special purpose, autonomous modules. Complex systems subject to a modular architecture range from people and intelligent computational entities to the Windows Operating System. One important aspect of modularity is that the input for each module is restricted—limited to the necessary and sufficient information required for the module’s task. In ad-

[1] This project is being carried out at the Computing Research Laboratory of New Mexico State University. See *http://crl.nmsu.edu/expedition* for an overview. Also see Nirenburg 1998, Nirenburg and Raskin 1998, McShane *et al.* 2000 for more extensive coverage. The Expedition Project is supported by Department of Defense Contract MDA904-92-C-5189.

dition, any particular module is minimally affected by the operation (and output) of other modules. These two aspects together are known as information encapsulation. For example, a person's auditory system is informationally encapsulated because it has restricted input—only information from the cochlea structures—and has limited, if any, access to other modules, like the visual perception system. A particularly compelling example of encapsulation (cited in Maratsos 1992) is the Muller-Lyer optical illusion, which shows that the perceptual system cannot be 'persuaded' by the cognitive system to accept that the lines shown below are the same length, which they are.



One widely known account of modularity within cognitive science is presented in Fodor 1983, which proposes that modules are:

a. Informationally encapsulated. Modules have strictly limited input, minimal interaction with other modules, and are not driven by central cognitive processes.[2]

b. **Fast**. Fast processing is a result of encapsulation, since a given system need only consider specific information in a specific way; not all information need be interpreted by every possible cognitive system.

c. Hard wired: Fodor makes the conjecture that modules are hard-wired (not derived from induction or experience) and are localized in a particular area of the brain.[3]

d. Domain-specific. Modules are often described as either horizontal—i.e., deriving from general reasoning ability, or vertical—i.e., domain-specific. Fodor considers language ability to be accounted for by vertical modules.

The assumption of modularity has been a driving force in descriptive linguistics, theoretical linguistics (e.g., Government and Binding Theory, Minimalism) and psycholinguistics (see, for example, Chomsky 1986 and Osherson and Lasnik 1990). The general view is that there is an innate language faculty that is distinct from that part of the mind responsible for general cognitive processing. The language faculty in turn consists of specialized modules for each language subtask (syntactic processing, lexical processing, etc.); each module has a well-defined task, specific inputs, and limited types of interaction with other modules. It is assumed, for example, that the syntactic processor does not have access to the speech waveform or even to the phonetic representation of that waveform. It is further assumed that there are no input loops between modules. Syntax, for instance, cannot affect morphology since it receives input from morphology.[4]

[2] Other researchers reject encapsulation in the cognitive realm, believing that the cognitive faculty is a single, undifferentiated general processing machine. Proponents of this view include Piaget (1955) and Newell and Simon (1972).

[3] For an alternative view regarding hard-wiring, see Karmiloff-Smith 1994.

[4] For convincing counterevidence to the 'no loop' hypothesis, see Levelt and Maassen 1991, Dell 1986, and Bock 1987.

The notion of modularity, particularly the idea of information encapsulation, is also fundamental to computer science. The now standard technology of object-oriented programming (a primary motivation for languages such as C++ and Java) is based on this notion of modularity (see Booch 1994 among others). Modularity also plays a key role in the fields of artificial intelligence and computational linguistics. For example, Marvin Minsky's popular theory of "Society of Mind" (Minsky 1985) describes a 'mind' as a group of encapsulated, highly specialized, agents.[5] Many systems within subsymbolic artificial intelligence (e.g., neural network systems), which traditionally were constructed with homogeneous architectures, now use a modular approach (see, for example, Miikkulainen 1993, which describes a modular neural network approach to natural language processing). Modularity is important in all these areas of computer science in part because a system composed of encapsulated, special-purpose modules is more easily built and studied than a non-modular one. It would be virtually impossible to build a Microsoft Word, a Windows Operating System, or a commercial machine translation system without the application of modularity.

3. The Modules of Boas

The simplifying assumption of modularity was adopted as a first-cut approach in Boas for the following reasons. First, one needs a strict organizational principle for knowledge elicitation, especially when the expected language informant has little or no formal linguistic training. Second, one needs an anchor for cross-linguistic research and generalization, and most descriptive accounts of language phenomena either adhere to or point out deviance from generally accepted modules: morphology, syntax, lexicon, etc. Finally, programs built to process language require narrowly defined types of input and output.

There are six basic modules in the Boas System, described briefly below.

a. Ecology: This module collects information about the writing conventions of SL, including the inventory of letters and punctuation marks, the treatment of numbers, dates, etc.

b. Inflectional Morphology: This module "learns" rules of inflection based solely on sample inflectional paradigms. The informant is guided through the process of building a paradigm template for each inflecting part of speech by answering questions about the parameters and values for which words inflect. For example, nouns might inflect for the parameter case using the values nominative, genitive, dative, and for the parameter number using the values singular, plural, dual. Once a paradigm template is established for a given part of speech, the informant provides all the inflectional forms of an inventory of examples that he selects. This inventory of examples should reflect all the productive patterns of inflection in SL. Paradigms thus established can be tested and made more robust

[5] "Each mental agent by itself can only do some simple thing that needs no mind at all. Yet when we join these agents in societies—in certain very special ways—this leads to true intelligence" (Minsky 1985: 17).

using additional examples.[6] A morphological learner (Oflazer and Nirenburg 1999) creates rules of inflection based on the inflectional paradigms. The morphological learner, however, imposes one significant restriction: it permits only single-word input, thereby excluding as input inflectional forms like *would have been going*. This restriction is actually not surprising either theoretically or computationally: "inflection" is defined by many as the realm of single words. To circumvent this restriction in Boas, the micro-component Multi-Word Inflection is being developed (described in §4.1)

c. Productive Affixation: This module represents Boas's minimal treatment of derivational morphology. It collects two types of SL derivational affixes: (i) those that correspond to a small inventory of productive derivational affixes in English—e.g., affixes expressing negation (*un- non- in-; anti- counter-*) and lesser degree (*mini- sub- under-*); (ii) affixes that only change the part of speech of the word with no significant shift in meaning—e.g., English *-ly* (*joyful ~ joyfully*).

Considering that derivational morphology—particularly compounding and reduplication—is extremely widespread in natural language, the question is why does Boas not handle it productively? The answer is that the results of these word-formation processes are often semantically ambiguous or non-compositional. Therefore, even if Boas were to "understand" that a compound were composed of stem A and stem B, or that a reduplicative form were composed of prefix X plus root A with the first syllable reduplicated, how could that information automatically be translated into a natural English equivalent?

Consider in this respect the following examples. Example (1) shows a Swedish compound, *frukosten,* that could have five different interpretations, from which a machine could not be expected to choose the most logical:

(1)	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' (from Karlsson *et al.*, 1995: 28).

Example (2) shows one of the many patterns of reduplication in Tagalog: a noun undergoes a non-trivial pattern of reduplication resulting in another noun that refers to the vendor of the original noun. Both the semantics and the formal rules underlying this reduplicative pattern would be difficult to capture and convey in English in a fully automated manner: [prefix *mag*] + [first two letters of the base, reduplicated] + [base] results in "vendor of [original noun]" (Schachter 1972).

(2)	a. *mag**bu**bulaklak*	'flower vendor'	(*bulaklak* 'flower')
	b. *mag**ka**kandila*	'candle vendor'	(*kandila* 'candle')

A final complication is the theme-and-variations nature of reduplication, illustrated below by the Turkish method of showing intensity of color:

[6] This is just a sketch of the process of eliciting inflectional paradigms. Since the informant is expected to be a linguistic novice, he is lead methodically through every step, offered extensive suggestions and examples, provided with redo capabilities, etc.

(3)	a.	*siyah ~ **sim**siyah*	‘black ~ very black’
	b.	*mor ~ **mos**mor*	‘purple ~ very purple’

Derivational processes such as this present practically insurmountable problems for a template system like Boas.

Boas’s answer to the problems of ambiguity, lack of compositionality, and formal variation in derivational word-formation processes is to have the informant enter all common words thus formed in the open-class lexicon. SL corpus scans will assist the informant in selecting the most common words of this type.

Although Boas’s module Productive Affixation unquestionably does not cover the majority of derivational word formation processes, neither do systems built explicitly for individual languages (i.e., in situations when the investigators are trained linguists with a full understanding of the patterns in question). For example, Dura (1998) suggests that the best way to deal with compounding in Swedish 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. Boas’s Productive Affixation module is, however, expected to permit significant time savings for informants of some languages: e.g., in Czech one negates verbs by adding the suffix *ne-*; this can be recorded and exploited to great effect by Boas.

d. Syntax: This module, still under construction, learns phrase structure rules based primarily on SL translations of a graduated set of English phrases and clauses.[7] This set has been designed such that the values of basic syntactic parameters like word order and agreement constraints can be determined. The syntactic learner works as follows. Suppose the informant translates *a new book* into Japanese as *atarasii hon-ga* (‘new book-NOM’). The learner will generate a set of proposed rules to account for this word sequence. One proposed rule will be the very specific one that occurrences of *atarasii* precede occurrences of *hon-ga.* The learner will also posit increasingly generalized rules: adjectives precede nominative case nouns, adjectives precede nouns. Once the inventory of candidate rules has been established, the learner selects the most general rule that is consistent with the entire set of translated examples. The learned grammar produces fairly flat syntactic analyses; the goal is to perform basic ‘chunking’ of a sequence of words into noun phrases and clauses, not to generate a linguistically sophisticated detailed grammar of the language. Some preliminary empirical work suggests that ‘chunking’ grammars are more robust than typical computational grammars (see Beale *et al.* 1999).

e. Closed-Class Lexicon: The closed-class lexicon contains an inventory of English closed-class items (pronouns, prepositions, conjunctions, etc.), organized semantically. Informants are asked to provide as many equivalents for each English sense as are employed in SL. The equivalents can take the form of a word, a phrase, an affix or a feature.

[7] See Sheremetyeva and Nirenburg 2000 for details about this module.

Examples (4) and (5) show cross-linguistic examples of affixal and feature realizations of closed-class items, respectively.[8]

(4)	a. BULGARIAN definite article:	*more* ~ *more**to*** ‘sea ~ the sea’
	b. RUSSIAN reflexive/reciprocal affix:	*myt’* ~ *myt’**sja*** ‘wash ~ wash oneself’
	c. PERSIAN possessive pronoun:	*kt|b* ~ *kt|b**t*** ‘ book ~ your book’
	d. ARABIC preposition:	*byt* ~ ***b**byt* ‘house ~ in a house’
	e. CREE possessive pronoun:	*astotin* ~ ***nit**astotin* ‘cap ~ my cap’

(5) RUSSIAN

a.

On	šel	**lesom.**
he$_{NOM}$	walked	**woods$_{INSTR}$**

‘He walked **through the woods**.’

b.

On	ubil	vora	**nožom.**
he$_{NOM}$	killed	thief$_{ACC}$	**knife$_{INSTR}$**

‘He killed the thief **with a knife**.’

The affixal and feature realizations of closed-class items actually represent cross-modular phenomena: a morphological process is required to convey a full-fledged semantic meaning. However, this bit of cross-modularity does not pose problems for Boas because the module Closed-Class Lexicon was originally developed with this functionality in mind.

f. Open-Class Lexicon: The open-class lexicon collects SL translations of nouns, verbs, adjectives, adverbs, phrases, collocations, and idioms. Translations may be words or phrases; multiple translations may be posited (e.g., English *blue* would be translated by Russian *sinii* ‘dark blue’ and *goluboj* ‘light blue’; Russian has no generic word for ‘blue’); allomorphs can be listed, as can irregular inflectional patterns; lexical items carrying grammatically relevant inherent features (e.g., gender) can be so tagged. Acquisition is primarily English-driven, but SL-driven acquisition is also possible, especially if large SL corpora are available to generate wordlists.

Modules (a)-(f) of Boas cover most language phenomena. The adherence to modularity allows an informant to focus on providing just the knowledge associated with a particular aspect of language (inflectional morphology, syntax, etc.) rather than face the daunting task of interacting with an undifferentiated knowledge elicitation system that places the burden of organization on the informant. As concerns processing the information acquired from the informant, modularity allows us to create efficient, specialized programs to handle different aspects of language. For example, finite state machines can handle morphological analysis and chart-based parsing algorithms can handle syntactic analysis. In sum, the simplifying assumption of modularity provides numerous advantages for the architecture of the Boas system. However, not all facts about language fall neatly into the abovementioned modules; some crucial language phenomena fall between

[8] The Cree example is from (Wolfart 1981). All examples, here an elsewhere, that are not attributed to a source were elicited from informants or created by the authors.

the cracks. These cross-modular phenomena, and Boas's treatment of them, are the subject of the next section.

4. Micro-Components for Cross-Modular Phenomena

For language phenomena that do not neatly fall into one of Boas's major modules, we are developing tailor-made micro-components. A sample of these, described in terms of their expected functionality, is presented below.

4.1. The Micro-Component for Multi-Word Inflection

Multi-word inflectional forms, like *would have been going*, straddle the line between morphology and syntax and are not acceptable input for Boas's morphological learner (cf. §3b). Therefore, the task of establishing inflectional paradigms must be split into single-word and multi-word subtasks. Once the informant establishes a paradigm template, he is presented with that template and asked to indicate whether each combination of feature values is realized as a single word, multiple words, or either.[9] All single-word and "either" entities remain in the main paradigm and are processed as described in §3b. All multi-word and "either" entries are extracted and sent to the Multi-Word micro-component.

The Multi-Word micro-component asks the informant to describe multi-word inflectional forms as the combination of auxiliaries and head words. The inventory of auxiliaries will be collected as a prerequisite task. All the necessary forms of the head word (e.g., infinitive, participles) should have already been collected in the single-word task and need only be pointed to in this module.

As concerns processing, multi-word inflectional forms present the same complexities as phrasals and idioms: often they can be scrambled and/or split by intervening words (*I would* ***definitely*** *have gone*). Processing of multi-word inflectional forms is done via pre-syntactic analysis. First, the basic morphological analyzer tags every individual word. Then the auxiliaries used as component parts of the inflected forms are deleted and their features transferred to the head word. For example, in the word sequence *will have been going*, the auxiliaries *will*, *have*, and *been* will be deleted and *going* will be assigned the features 'future', 'passive', and 'perfect'.

4.2 Movement of Inflectional Affixes

Another phenomenon spanning morphology and syntax is the movement of inflectional affixes from their head words to another place in the sentence. Sometimes a moved affix cliticizes onto another word, sometimes not. A case in point is certain Polish person markers, which can move from their head verb to virtually any pre-verbal position. For

[9] An example of "either" is the Ukrainian future tense: *robitimu* and *budu robiti* are both valid ways of expressing the 1st person singular 'will work'.

example, the 1st person plural suffix *śmy* has the legal placements shown in (6a-d) (hyphens are included only for emphasis).

(6) a. My–**śmy** znowu wczoraj poszli do parku.
we-1PL again yesterday went to park
b. My znowu–**śmy** wczoraj poszli do parku.
c. My znowu wczoraj–**śmy** poszli do parku.
d. My znowu wczoraj poszli–**śmy** do parku.
e. * My znowu wczoraj poszli do-**śmy** parku.
f. * My znowu wczoraj poszli do parku-**śmy**.
'We went to the park again yesterday.' (Franks and Bański 1999: 125)

The processing problems for sentences like (6a-d) are obvious: the morphological analyzer will not find lexical matches for words like *myśmy* 'we-1PL' *znowuśmy* 'again-1PL' or *wczorajśmy* 'yesterday-1PL'. In addition, the left-over verb forms in (6a)-(6c) will be incorrectly analyzed as 3rd person plural (plural verb forms have no person suffix).

The movement of inflectional affixes is handled in Boas using the micro-component Affix Movement. After the inflectional paradigms for a given part of speech are created, the informant is asked if affix movement occurs in SL. If so, he/she selects one paradigm to serve as a test case and highlights all affixes that can move. If different affixes from different paradigms can move, the process is repeated for more paradigms. In the end, Boas will contain an inventory of mobile affixes similar to the inventories of affixes collected through the Productive Affixation and Closed-Class modules.

For each inflectional affix that can move the system generates a set of morphological rules. One rule recognizes the affixless form of words in the source paradigm (i.e., wordforms the affix can hop from). For example, *poszli* in (6a) will be recognized as a verb that is missing inflection for person and number (*poszli* will *also* be recognized as the 3rd person plural form of the verb; this bit of ambiguity will be resolved at a later stage). A second rule strips the hopped affix off the target word, revealing its underlying form. For example in (6a), *śmy* will be stripped off of *myśmy* and *my* will be recognized as a pronoun in the regular way. In post-morphological analysis, the features associated with the hopped affix (1st person plural for *śmy* in (6a)) are unified their source stem (*poszlli*).

4.3 Spelling Changes Induced Word-Externally

Yet another phenomenon that straddles morphology and syntax is spelling changes induced by word-external factors. For example, lenition and eclipsis in Irish are word-initial mutations triggered by certain types of preceding words. Table 1 presents a small sampling of such alternations:[10]

[10] For a full description of lenition and eclipsis in Irish, see Ó'Sé and Sheils 1993, and Ó'Siadhail 1989, 1995.

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...?' (These processes occur in many other contexts as well and affect many other letters.)

In order to avoid the acceptable yet labor-intensive approach of having the informant list two variants of each affected word in the lexicon, Boas employs the micro-component Lenition Etc. If alternations induced word-externally occur in SL, the informant is asked to indicate the basic letter or cluster, the resulting letter or cluster, and where the mutation occurs: word-initially, word-finally, or both. Boas converts this information into a lexical redundancy rule covering the entire open- and closed-class lexica. We do not need to elicit in what contexts such alternations occur since we are not producing, only decoding, SL.

4.4. Noun Incorporation

Noun incorporation is a subset of compounding, namely, noun-verb compounding. In incorporating structures, the verb and one of its arguments (usually the subject or object, or just the head noun of the subject or object) either occur as a single word or occur in series with certain morphosyntactic modifications that indicate that incorporation has occurred.[11] Incorporation presents all the elicitation and processing problems of noun-noun compounding plus a host of others. This would suggest that Boas should handle incorporation like it handles noun-noun compounding and most other derivational word-formation processes—lexically. For some languages this approach seems feasible, as incorporation is lexically restricted and/or semantically non-compositional anyway. However, for other languages a productive approach to compounding appears necessary. Below are some cross-linguistic properties of incorporation that pose particular challenges to the Boas system.

Morphological Complexities. Frequently, spelling changes occur as part of incorporation (just as they often occur in noun-noun compounding and reduplication). For example:

1. Incorporated nouns generally lose their inflectional morphology (Baker 1988: 26), making their grammatical role (subject, object, etc.) opaque to an MT system.
2. An epenthetic vowel can sometimes be inserted between the V and N, as in Tuscarora: [a] is inserted when a consonant-final N and a consonant-initial V are joined (described by Williams 1976; cited in Baker 1988: 23). Insertions, deletions and mutations at morpheme boundaries are typically difficult to describe and they are difficult to prepare for in a template system like Boas.

[11] In some languages adjuncts can also be incorporated. See, for example, Spencer 1995, which gives examples of adjunct incorporation in Chukchi.

3. The incorporated noun can occur between the verb stem and its inflectional affixes. This means that the verb forms collected via paradigms in the Inflectional Morphology section of Boas will have to be understood as splittable, with possible mutations occurring at the new morpheme boundaries created during incorporation. That is, whereas one type of mutation might occurred at the boundary of the verb stem and its inflectional ending, another mutation might occur at either edge of the inserted noun.

Syntactic Complexities. The syntax of incorporating structures can differ in significant ways from the syntax of non-incorporating structures. For example:

1. When the direct object is incorporated, the verb might become intransitive or it might remain transitive. In the latter instance, the oblique object or possessor is often promoted to the direct object role, as in Panare (Payne: 299). For Boas, this means that basic source-language-to-English transfer rules will fail in incorporating structures. Assume, for example, that possessors in the source language are normally in the Genitive case, such that Genitive case maps to *'s* in English; in incorporating structures, possessors could bear Accusative case, which normally maps onto direct-object status in English. Thus, a special set of transfer rules would have to be invoked in Boas for incorporating structures.
2. Generally only the head of the incorporated NP is incorporated, leaving modifiers as separate words, as in the following West Greenlandic example (from Fortescue 1984; cited in Bok-Bennema *et al*. 1988):

(7)	kissartu	-mik	**kavvi-sur**	-put
	hot	-instr	**coffee-drink**	-3Pl.Ind.

(they hot coffee-drank)

This means that, formally, the adjective appears with an elided head noun whose antecedent occurs as a bound morpheme attached to the verb.

3. In languages in which the incorporated noun remains a separate word, numerous syntactic changes may take place. Mithun (1984: 850-851) notes the following examples: in Samoan, particles that generally cliticize to the right of the verb cliticize to the right of the verb-noun complex under incorporation; the same applies to aspect suffixes in Micronesian languages; subjects in ergative languages are case-marked absolutive in intransitive incorporating structures as opposed to ergative in the transitive non-incorporating counterparts.

Semantic Complexities. Incorporation is at once a lexical and a syntactic process. Its lexical aspect can give rise to to the same types of semantic shifts as other lexical processes. Linguistic descriptions of incorporating languages tend to be less rigorous regarding semantic compositionality than is necessary for an MT system. For example, Mithun (1984: 853) describes the following incorporating structures as "somewhat idiomatic", although in the world of MT they would be considered completely idiomatic (an MT system would, at best, be able to produce the literal glosses): *heart+be.numerous* = 'to be fickle'; *rump+be.heavy*= 'to be sluggish'. Similarly, in describing Dutch, Weggelaar (1986: 302) groups together truly compositional incorporating structures with the following, which lack strict semantic compositionality: *to-buzz+(child's)head* = 'to be dizzy'; *to-roll+(child's) head* = to tumble; *to lick + beard* = 'to lick one's lips'.

Another semantic complexity is shown in Panare: unincorporated "head cut" describes a person getting a cut on the head, whereas incorporated "head-cut" asserts that the head was cut off (Payne: 300). Thus, incorporating and non-incorporating 'cut' in Panare are actually separate lexical items, at least in combination with body parts.

Within Boas, semantic non-compositionality and/or unpredictability of the types described above must be handled by explicit lexical listing.

Lexical Restrictions. Incorporation is highly lexically restricted in some languages, such that lexical specification of relevant word complexes would be both feasible and preferable in Boas. For example, in many languages incorporation occurs either exclusively or primarily with nouns indicating body parts (Weggelaar: 301-2). This is true of Panare, in which "most incorporated nouns are body parts, and the verbs that allow incorporation are verbs of 'removal' or 'destruction', e.g., 'cut' (of various kinds), 'break', 'hit', 'pluck', etc." (Payne: 300). It is also true of Dutch: only about thirty verbs support incorporation, and the nouns that incorporate must refer to body parts (incorporation can be used somewhat more productively, but not to an extent that would be crucial for Boas, it appears) (Weggelaar: 301).

As the above evidence makes clear, productively handling incorporation in an MT system would be extremely difficult even if one were dealing with a single language for which extensive data were available and a highly trained linguist provided analysis.[12] The challenge grows exponentially under the constraints of the Boas environment.

In the current implementation of Boas, no attempt will be made to elicit specific information about patterns of incorporation in SL. Only two questions will be asked: Is incorporation employed in SL? If so, is it employed in a highly productive manner with the resulting NV complexes having compositional semantics? (Of course, all of these notions will be explained.) If incorporation is used only in a limited or non-compositional manner, the informant will be asked to enter the most common incorporating structures in the lexicon, using source-language corpus scans to help compose this list. If incorporation is used highly productively and compositionally, we will create a last resort program to deal with it: all unknown words will be submitted to a fuzzy match algorithm which will assume that: (i) all nouns and all verbs in the lexicon can potentially have affixal status; (ii) inflectional affixes on the verb are mobile, and (iii) morpheme boundaries can show mutations. This algorithm will assume that the form of the noun and the form of the verb are basically the same in incorporating and non-incorporating structures and that at least some semantic compositionality obtains. This should provide at least some degree of coverage of this complex and largely idiosyncratic linguistic process.

[12] For example, a trained linguist working with Eskimo might include a class of "noun-verb postbases" in the closed-class lexicon; these are verbal elements that can never stand alone but, rather, must participate in incorporating structures (Baker 1988: 16). Boas would analyze these the same way as other affixal elements gathered in the Productive Affixation and Closed-Class Lexicon modules.

5. Conclusions

Modularity in complex systems is assumed by researchers in a variety of fields. There has been considerable debate within theoretical linguistics, neurolinguistics, developmental linguistics and other areas regarding the exact nature of modularity and whether modularity is needed at all (see Karmiloff-Smith 1994, Mueller 1996, Bates 1994, among others). For example, in the area of cognitive science, several contemporary theories (modern cognitivism, cognitive linguistics, associationistic empiricism) reject modularity and argue that all mental processes are interconnected and exchange data freely.[13] However, when it comes to building large practical systems that deal with natural language, modularity is indisputably a sound architectural principle.

In this paper we have described a system that acquires information about any natural language from untrained human informants and uses that information to ramp up a SL-to-English machine translation system. We have shown that adopting the simplifying assumption of modularity helps to organize the acquisition of knowledge and to structure the resulting machine translation system. Strict modularity must, however, fail because some language phenomena (e.g., multi-word inflection, movement of affixes, noun incorporation) span modules. The solution we propose is to develop a highly specialized micro-component for each cross-modular phenomenon we identify. We have sketched out a few of these micro-components in this paper. Continued work on Boas will concentrate on identifying still more cross-modular phenomena that occur in natural languages and developing micro-components to elicit and process the varied instances of those phenomena found in the world's languages.

REFERENCES

Baker, M.C. 1988. "Morphology and Syntax: an Interlocking Independence". Everaet, M., A. Evers, R. Huybregts and M. Trommelen, eds., *Morphology and Modularity*. Dordrecht: Foris Publications.

Bates, E. 1994. "Modularity, Domain Specificity, and the Development of Language". *Discussions in Neuroscience* 10:136-149.

Beale, S., S. Nirenburg, J. Cowie, and K. Oflazer. Ms. "Quick Ramp-Up MT —The Pin-The-Tail-On-The-Donkey Approach". [Available at http://crl.nmsu.edu/expedition/publications/index.html.]

Bock, J.K. 1987. "An Effect of the Accessibility of Word Forms on Sentence Structures". *Journal of Memory and Language* 26(2):119-137.

Bok-Bennema, R. and A. Groos. 1988. "Adjacency and Incorporation". Everaet, M., A. Evers, R. Huybregts and M. Trommelen, eds., *Morphology and Modularity*. Dordrecht: Foris Publications.

[13] See, for example, the papers from the Cossmodal Attention and Multisensory Integration Conference (http://www.wfubmc.edu/bgsm/nba/IMRF/meeting.html) and Feldman *et al*. 1990.

Booch, G. 1994. *Object-oriented Analysis and Design with Applications*. Redwood City CA: Benjamin Cummings.

Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Dell, G.S. 1986. "A Spreading Activation Theory of Retrieval in Sentence Production". *Psychological Review* 93:283-321.

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

Feldman, J., G. Lakoff, A. Stolke and S. Weber. 1990. "Miniature Language Acquisition: A Touchstone for Cognitive Science". *Proceedings of the 12th Annual Conference of the Cognitive Science Society* 686-693. Cambridge, MA: MIT Press.

Fodor, J.A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

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

Franks, S. and P. Bański. 1999. "Approaches to 'Schizophrenic' Polish Person Agreement". In Dziwirek, K. and C.M. Vakareliyska, eds., *Annual Workshop on Formal Approaches to Slavic Linguistics: the Seattle Meeting, 1998,* 123-43. Ann Arbor: Michigan Slavic Publications.

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

Karmiloff-Smith, A. 1994. "Precis of: Beyond Modularity: a Developmental Perspective on Cognitive Science". *Behavioral and Brain Sciences* 17(4): 693-745.

Levelt, W.J.M. and B. Maassen. 1981. "Lexical Search and Order of Mention in Sentence Production". Klein, W. and W.J.M. Levelt, eds., *Crossing the Boundaries in Linguistics: Studies Presented to Manfred Bierwisch.* Dordrecht: Reidel.

McShane, M., S. Helmreich, S. Nirenburg and V. Raskin. 2000. "Slavic as Testing Grounds for a Linguistic Knowledge Elicitation System". In King, T.H. and I.A. Sekerina, eds., *Annual Workshop on Formal Approaches to Slavic Linguistics: the Philadelphia Meeting, 1999*, 279-295. Ann Arbor: Michigan Slavic Publications.

Maratsos, M. 1992. "Constraints, Modules, and Domain Specificity: An Introduction". In M.R. Gunnar and M. Maratsos, eds., *Modularity and Constraints in Language and Cognition*, 1-23. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Miikkulainen, R. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory.* Cambridge MA: MIT Press.

Minsky, M. 1985. *The Society of Mind.* New York: Simon and Schuster.

Mithun, M. 1984. "The Evolution of Noun Incorporation". *Language* 60:847-95.

Mueller, R. 1996. "Innateness, Autonomy, Universality? Neurobiological Approaches to Language". *Behavioral and Brain Sciences* 19(4):611-675.

Newell, A. and H.A. Simon. 1972. *Human Problem Solving*. Englewood Cliffs NJ: Prentice-Hall.

Nirenburg, S. 1998. "Project Boas: 'A Linguist in the Box' as a Multi-Purpose Language Resource". In *Proceedings of COLING '98.*

Nirenburg, S. and V. Raskin. 1998. "Universal Grammar and Lexis for Quick Ramp-Up of MT Systems". In *Proceedings of 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.

Osherson, D.N. and H. Lasnik. 1990. *An Invitation to Cognitive Science: Language*. Cambridge, MA: MIT Press.

Payne, Thomas E. 1995. “Object Incorporation in Panare”. *International Journal of American Linguistics* 61 (3): 295-311.

Piaget, J. 1955. *The Child’s Construction of Reality*. Routledge.

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

Sheremetyeva, S. and S. Nirenburg Forthcoming. “Towards a Universal Tool For NLP Resource Acquisition”. Proceedings of the Language Resources and Evaluation Conference, Greece, Athens, 31 May - 2 June 2000**.**

Spencer, A. 1995. “Incorporation in Chukchi”. *Language* 71:439-489.

Weggelaar, C. 1986. “Noun Incorporation in Dutch”. *International Journal of American Linguistics* 52(3): 301-305.

Williams, M. 1976. *A Grammar of Tuscarora.* New York: Garland.

Wolfart, H.C. 1981. *Meet Cree: A Guide to the Cree Language*. Lincoln: University of Nebraska Press, 1981.