Resources for Morphology Learning and Evaluation

Mike Maxwell

Linguistic Data Consortium
3615 Market Street, Suite 200, Philadelphia PA 19104 USA
maxwell@ldc.upenn.edu

Abstract

Recently, there has been a proliferation of research into the acquisition of morphological grammars—that is, grammars and lexicons required for computer-based morphological analysis and synthesis. The approaches to acquiring such grammars range from tools which structure data provided by native speakers and linguists, to unsupervised machine learning. Despite this flurry of research into morphology learning, a means of comparing results among different approaches is largely lacking. This paper describes a test bench for morphology learning, which would assist designers of morphology learning programs by providing both training and evaluation data, and would allow comparison across programs. This paper is simultaneously a description of the projected form of the test bench, and a call for further input.

1. Introduction

Recently, there has been a proliferation of research into the acquisition of morphology by machine, including grammars and lexicons for computer-based morphological analysis and synthesis. Approaches to acquiring such grammars range from tools which structure data provided by native speakers and linguists (such as the Boas system described in Oflazer, 2001 and Zajac, 2001), to unsupervised learning from monolingual texts (Yarowsky, 2000; Goldsmith, 2001; Snover, 2001; and Schone, 2000), from bilingual texts (Yarowsky, 2000; Yarowsky, 2001) or from other resources (Bosch, 1996, 1996; Gaussier, 1999, and Kazakov, 2001).

While research into morphology learning has flourished, what is largely lacking is a means of comparing results—standard data sets, for example, together with a more or less agreed-on set of results that should be derivable from each set. While large quantities of machine-readable linguistic data are available, little if any of it is intended for morphology learning and evaluation. Likewise, although individuals working on morphology learning have sometimes made available data sets usable with their own programs, there is a need for learning and evaluation data that would be usable by a variety of morphology learners, and in particular for comparing different approaches.

The project described in this paper is intended to provide a tool to assist designers of morphology learning programs by providing both training and evaluation data, and which will also facilitate comparison of different approaches to the learning of morphology and phonology. This paper is simultaneously a description of the projected form of the test bench, and a call for further input.

The rest of the paper is organized as follows. Section two discusses the typology of morphology, and how this influences the design of the test bench. Section three describes a set of resources which we believe to be necessary and (perhaps) sufficient input data for systems which purport to 'learn' morphology, and to test the lexicons and grammars which those systems have learned. Section four briefly describes several morphology acquisition systems which already exist or are in the planning stages, and then shows how the resources in section two could be used by those systems in learning and evaluation. Section five lists some remaining questions, while the final section summarizes the paper.

2. Typological Design Criteria

Morphology learning techniques are sensitive to the morphological type of a target language. A morphology test bench should therefore provide data from a typologically varied set of languages. A traditional morphological typology (see e.g. Spencer, 1991) distinguishes the following sorts of languages:

- Isolating
- Fusional (also called "inflectional")
- Agglutinative
- Polysynthetic

Truly isolating languages are uninteresting from a morphology learning perspective, since there is by definition nothing to learn. Fusional and agglutinative languages, on the other hand, should be well represented in a test bench. Polysynthesis is probably rare enough among the world's languages that it can be ignored in the first version of the test bench. Compounding is, however, quite common, so that it will be useful to provide at least one language with makes extensive use of it.

The above terminology is most commonly used with reference to inflectional morphology, but languages differ as well in the degree to which they have derivational morphology. While the emphasis in the test bench will be on inflectional morphology, the languages represented should also exhibit a range of derivational processes.

Languages differ morphologically in a number of other dimensions, including:

- Suffixing languages vs. prefixing languages vs. languages with both suffixing and prefixing
- Degree of phonologically conditioned allomorphy
- Degree of morphosyntactically conditioned allomorphy (primarily stem allomorphy)
- Degree of irregularity (phonologically unpredictable, and therefore lexically listed, allomorphy, generally at the word level)
- Number of inflectional (paradigm and/or declension) classes

The test bench will provide data from languages which differ along these scales as well. However, non-concatenative morphology, including infixation and

reduplication, will probably not be represented, at least in the initial version.

In order to include languages differing in the ways described above, we will for the most part draw on unrelated languages. However, since some approaches to morphology learning (e.g. Yarowsky, 2001) base learning of a new language on an existing analysis of a related language, some pairs of related languages would be a useful resource. It may be of interest to provide data from languages which can be arranged in a cline of closeness of relationship, such as Spanish, Portuguese, and French.

A set of five to ten languages jointly meeting the above criteria would seem to be a reasonable target for an initial version.

3. Resources and Views

Within each language, a variety of resource types will be provided, as described below; some are for the human user of the test bench, and some are to be used by the learning program itself.

The resources for the learning program's use are not intended to be used in their raw form; rather, a set of 'views' is also described which contain various kinds of information which a learning system or an evaluator might require.

The resources and their views are summarized in the table below. In the section of the table concerning dictionaries, the abbreviation 'SL' refers to 'Source Language'; the glossing language is assumed to always be English. The codes in parentheses after many of the resources and views will be used to refer to the types of information in the sections below.

Projections of views may also be needed, beyond those indicated in the table.[1] For example, a learning program which concerns itself solely with form, not meaning, may require for evaluation a projection of the SL→ English dictionary (DL-2), but without English glosses.

Many of the entries in this table should be self-explanatory; the following sub-sections provide details where the intent may not be so obvious. In addition, the following sections show how the various resources provide the learning and evaluation data required by several morphology learning systems described in the literature. Readers are invited to consider whether the needs of their favorite morphological learner are also met, and to propose changes where this is not the case.

3.1. General Language Information

The general information provided about the language is for the edification of the human user of the test bench, and is not intended to be computer-interpretable. Bibliographic references will include both printed grammars and dictionaries, as well as other linguistic studies.

3.2. Writing Systems

Generic information on alphabetic writing systems will include the typical phonological 'meaning' of alphabetic characters (including multigraphs) where possible. Where that is not possible, a dictionary-based transducer may be provided to map words (both dictionary citation forms and the inflected forms of words) into a phonological representation.

Additional information to be provided about writing systems includes (where applicable) correspondences between upper and lower case (for which a one-way transducer to lower case will be provided), sort orders, and punctuation, as well as a tokenizer.

It may also be desirable to provide a transducer to convert between non-Roman orthographies and a Romanized transliteration of them, for instance for Korean (Hangul) or Tamil. This is especially important for right-to-left writing systems, since it is difficult to work with morpheme level interlinear texts where the source language is written in a different direction from the glossing language.

Some languages may have special-purpose writing systems. For example, speakers of languages whose standard writing systems do not correspond to the standard (lower) Ascii characters often develop Ascii-based (but non-standard) encodings for electronic use (particularly email). (Examples are Arabic, and Spanish without accent marks.) Where feasible, transducers will be provided for converting between alternative and standard encodings. Note that there may be an unavoidable loss in transduction in one direction or the other.

3.3. Grammatical Description

The purpose of the grammatical description is to explain to human users the decisions which have been made elsewhere. In addition, at least Oflazer, Nirenburg and McShane's (2001) approach needs to be explicitly told a certain amount of grammatical information, including the parts of speech, inflectional (morphosyntactic) features[2], and inflectional classes[3].

[1] Some of the views described in the table are already projections of other views, but are distinguished for conceptual reasons.

[2] We intend to use a generic ontology, to avoid theoretical issues as far as possible.

[3] The term 'inflection class' refers to a paradigm or declension class. Deciding how many inflection classes a part of speech in a given language is sometimes controversial, but unavoidable.

Resources		Views
General Language Information	Name(s) of language	Generic Views
	Geopolitical information	
	Bibliographic References	
	Pointers to computational resources	
Writing systems	Description	Generic Views
	Transducers	See text
Grammatical Description (G)	Morphology (GM)	POSs (1)
		Inflectional features by POS (2)
		Inflection classes by POS (3)
	Morphophonology (GP)	Generic View
	Named entity mapping, Abbreviations	Generic View
	Syntax (GS)	Generic View
Dictionary (D)	Lexeme Dictionary (DL)	English → SL dictionary (1)
		SL → English dictionary (2)
		SL Lexemes belonging to each inflection class (3)
	Affix Dictionary (DA)	English → SL dictionary (1)
		SL → English dictionary (2)
		Paradigm of Affixes (3)
Texts (T)	Monolingual Texts (TM)	Native orthography (1a)
		With word breaks (1b)
		With morpheme breaks (2)
	Bilingual Texts (TB)	Unaligned (1)
		Aligned at 'segment' level (2)
		Aligned at word level (3)
		Divided/ glossed/ aligned at morpheme level (4)
	Multilingual Texts(TX)	(same as for bilingual text)
Morphological Transducer (X)		Parse of word (1)
		Generate word from lexeme + morphosyntactic features (2)
		Paradigm of a stem (3)
		Stems derived from a stem (4)
		Random surface words (5)
Tagger (Tg)		Tagged text (1)

Table 1: Types of Data in the Test Bench

For expository purposes, I divide the grammatical description into three parts: morphology, morpho-phonology, and syntax.

The morphology description should be written at the level of detail of a grammar sketch in a typical bilingual dictionary, emphasizing inflectional morphology and productive derivational morphology, including inflectional features, inflection classes, slots for inflectional affixes[4], and allomorphy.

The morpho-phonology and syntax descriptions are included for the user's edification. It is not a requirement that a morphology learning program discover the same set of phonological rules that are given in the morpho-phonology description, provided the correct surface forms are generated: weak equivalence is the goal, not strong equivalence. Likewise, the syntax sketch can be quite minimal. Relevant distinctions between literary, informal written, and spoken language should also be mentioned, along with dialectal differences.

Named entity mapping includes information about how names are rendered in the language, including transliteration of foreign names. Abbreviatory conventions also merit mention (although this may only be available for English, and perhaps a few other languages).

3.4. Dictionaries

The dictionaries will give the sort of information provided by a typical bilingual dictionary, save that information on semantics is minimal. I distinguish between a lexeme dictionary (for morphemes belonging to major parts of speech) and an affix dictionary.

In the case of the SL → English view of the lexeme dictionary (DL-2), the SL side should include for each entry at a minimum the following: citation form (or

[4] The theoretical status of slots is uncertain, but for the practical purposes envisioned here, this should not be an issue. The slots need to be labeled (if only with a number) so they can be referenced by the affix dictionary (DA).

forms), part of speech[5], inflection class, irregular stems, irregular inflected forms (together with their morphosyntactic features), and glosses.[6] Since the information is provided in electronic form, the dictionary can also provide for each entry a link to that lexeme's paradigm, generated on the fly by the transducer (see section 3.6, "Morphological Transducer"). The English → SL lexeme dictionary (DL-1) is simpler, and is intended only as an index to the SL lexemes; in particular, no information about English inflection classes or irregular forms is given.

The affix dictionary is similar in concept to the lexeme dictionary, but contains somewhat different information. Minimal information for inflectional affixes includes the form (including allomorphs and their conditioning properties), whether the affix is a prefix or suffix, the part of speech and inflection class(es) to which the affix attaches, the slot in which it attaches, and a gloss (corresponding to the affix's morphosyntactic features).

Minimal information for derivational affixes includes form (allomorphs), prefix or suffix status, mapping between input and output parts of speech and inflection classes, and a gloss.[7] Derivational affixes which are not perfectly productive will need to list the subset of stems (lexemes in the lexeme dictionary) to which they attach.

In addition, the affix dictionary will provide what is referred to in the table as a 'paradigm of inflectional affixes.' What is meant here is that for each inflection class, the dictionary will give (or generate) a skeleton paradigm of inflectional affixes with a placeholder for the stem.[8] Phonologically conditioned allomorphy in affixes is a problem for this view, since the placeholder cannot condition the allomorphy. One solution would be to present default allomorphs (chosen arbitrarily, if necessary), and to allow the user to choose different phonological properties of the stem and see the effects on the affix allomorphs.

While a dictionary is primarily useful for evaluation of morphology learning, some learning strategies require a subset dictionary to serve as a 'seed' for learning. It is neither feasible nor necessary for the system to explicitly provide such a subset, since how much of a subset would be appropriate is application-dependent. Rather, creation of a subset (here and elsewhere) is left up to the end user. (It might, however, be useful to provide frequency data for lexemes, as a basis for choosing a subset.)

Like texts (see below), dictionaries may be provided in multilingual form, i.e. with entries including not only English and the target language, but also in some language related to the target language. Such multilingual dictionaries are probably not useful for evaluation purposes, since the test bench will normally be used to evaluate morphology learning of a target language with reference to English. But a plausible bootstrapping technique would be to use a small multilingual dictionary as a seed lexicon.

[5] I assume here that there will be a separate dictionary entry for each part of speech to which a lexeme belongs.

[6] Since the focus is on machine learning, glosses (as opposed to full definitions) are sufficient. The glosses in the dictionary should be consistent with those used in bilingual text (TB-4).

[7] Where the boundary between inflection and derivation is unclear, a slot for derivational affixes may be appropriate.

[8] This is similar to Goldsmith's (2001) notion of 'signatures'. However, Goldsmith's program does not distinguish between inflectional and derivational affixation.

3.5. Texts

The test bench will also include text resources. Texts can be classified as monolingual (TM), bilingual (English and SL, abbreviated TB), and multilingual (TX). The latter are texts which, in addition to the SL and English, have a translation into some other language. As discussed above, the reason for providing multilingual texts is to provide a learning mechanism for situations where a grammar and/or dictionary is available in a related language, and the morphological learning program is expected to create the SL analysis by modifying an existing analysis for the related language.

Bilingual texts will be divided into morphemes, with separate 'lines' for aligned morpheme, word, and segment (sentence or verse, with free translation) glosses. From these aligned and glossed texts, the user can project bilingual texts aligned only at coarser levels (or unaligned), as well as monolingual texts of various sorts, as required for various learning strategies. Monolingual texts are therefore not treated here as a distinct resource, but rather as a view of bilingual texts.

The description in the table above also distinguishes monolingual texts in 'native orthography' (TM-1a), and texts with word breaks indicated (TM-1b). This distinction is only relevant for the situation where word breaks are not indicated in the conventional orthography. The same distinction can be made for unaligned bilingual and multilingual texts (TB-1 and TX-1) and those aligned at the segment level (TB-2 and TX-2), but is not shown in the table above. Again, monolingual texts in these two forms can be derived from bilingual texts by projection.

An assumption is that text annotation is unambiguous. For example, each morpheme in bilingual text glossed at the morpheme level has a single gloss (unlike the lexicon). That is not always the case, but it is not clear how true ambiguities in text glossing should be indicated.

3.6. Morphological Transducer

The test bench will also provide a morphological transducer for each language, to allow both parsing and generation (including generation of the paradigm of a stem).

Note that applying the transducer to SL text may not give the same result as the pre-parsed texts (TM-2, TB-4 and TX-4). In particular, the transducer will frequently find ambiguous parses where no such ambiguity is indicated in the parsed texts (presumably because the latter has been disambiguated using the context).

Two of the 'views' produced by the transducer deserve mention. The set of stems derived from a stem (view X-4) refers to a list of all uninflected stems which can be derived by the addition of a single derivational affix to the given stem. (Note that this view should be applicable recursively.)

The paradigm of a stem (X-3) refers to a structure in which all inflected forms of the stem are given for each cell of the paradigm, together with the inflectional features that generate each cell. This view can be accessed from the source language dictionary for lexemes (as discussed above), but it can also be applied to the output of view X-4, i.e. to derived stems not listed in the lexicon.

A mapping will be provided between the user's view and the input (for generation) or output (for parsing) of the transducer. That is, paradigm cells are defined by their inflectional features, e.g.:

[Ergative [Person 1
Number Singular]
Absolutive [Person 2
Number Plural]
Aspect Incomplete
]

—which might not correspond to the order or number of affixes, e.g.:

ya-h-koltay-at-ik
INC-ERG:1-help-ABS:2-PL

(the example is from Tzeltal). In summary, the user's view of the paradigm should abstract away from issues such as the linear sequence of morphemes, zero affixes, and extended exponence.

3.7. Tagger

A tagger would prove useful to disambiguate morphological parses in text. However, it is not clear that we will always be able to provide such a tagger, although doing so may be nearly trivial for languages with complex morphologies.

4. Morphology Learning Systems and the Views they Require

There are two ways the test bench can be used with a morphology learning system: as a provider of data for learning, and as an evaluator of what the system has learned.

The following table summarizes the views which a diverse set of learning approaches described in the literature (or with which I am otherwise familiar) require, both for training and for evaluation. These approaches were chosen for their variety; no attempt was made to cover every program or project dealing with morphological acquisition. The subsections following provide commentary on the entries in this table. Not mentioned in this table are 'General Language Information' or information on writing systems, since the former is mostly for human use, while the latter will be needed for most programs (if for no other reason than to make sense of the output).

Learning Program	Discovery	Evaluation
Linguistica (Goldsmith, 2001)	TM-1b or X-5	DL-2; DA-3; X-3,4
Expedition/ Boas (Oflazer, 2001)	GM-1,2,3 DL-2,3; X-1,3	DL-2, X-3
Phonological Learner (Albright, 1999)	X-3	X-3
Stealth-to-Wealth (SIL)	TB-2,3	DL-2, X-3,4 DL-1 or 2; DA-1 or 2
Learning from Bitexts	TB-2	Same as for Stealth-to-Wealth

Table 2: Resources Required by various Programs

Two other factors should be mentioned. At training time, it may be desirable to intentionally introduce noise, in the form of probabilistically incorrect data. This would be for the purpose of imitating a human consultant, who could be expected to make occasional mistakes, or to mimic real text data, which may contain typos, spelling mistakes, dialectal variants, etc. However, it is not clear just how this spurious data should be created, short of introducing random spelling errors. For example, suppose it was desirable to make an error in one of the forms of the paradigm of a certain lexeme. Assuming the paradigm is provided by a transducer (X-3), without having the transducer's grammar, it is not simple to change the membership of a lexeme from one inflection class to another, nor to 'forget' an irregular cell of that paradigm.

It might also be desirable at training time to output information in 'dribbles.' The Expedition/ Boas system, for example, will probably best be served by working with one inflectional class at a time. It is not clear what sort of interface is called for here; the learning system could ask for another information 'chunk', but it is not obvious what a 'chunk' is, whether there should be an order to the presentation of chunks, or what the API would be.

I turn now to the individual learning systems listed in the table above.

4.1. Linguistica

Goldsmith's Linguistica program performs unsupervised learning of morphology from monolingual input. As described in Goldsmith (2001), this input is in the form of a monolingual text with wordbreaks (TM-1b). However, it is not apparent that the input need be actual texts; a random stream of wordforms, X-5, would probably suffice as well.

As output, Linguistica produces a list of stems together with their 'signatures' (the set of affixes with which they appear). No distinction is made between inflectional and derivational affixation, hence a signature is not quite the same as a paradigm. Evaluation will thus require a list of lexemes belonging to each inflection class (DL-3), to be used with a transducer to generate the inflectional paradigm of those stems (X-3) and the list of stems derived from the listed stems (X-4). It would also be useful to compare the paradigms of inflectional affixes (DA-3) with the signatures returned by Linguistica, bearing in mind that the latter includes both inflectional and derivational affixes.

Evaluation of Linguistica's recall requires the system to keep track of which lexemes were present in the texts given at training time. A similar requirement exists for keeping track of information given to the other systems discussed below, but I will not mention it for each individual case.

4.2. Expedition/ Boas

Boas is a knowledge elicitation system used in the Expedition project at New Mexico State University's Computing Research Laboratory, described in Oflazer, Nirenburg and McShane (2001). It is intended to be used by a team consisting of a linguist and a native speaker.

From the standpoint of using a morphology test bench for learning with Boas, the test bench must provide the

same information a native speaker would. The initial information elicited by Boas from the informant includes

> …the parameters for which a given part of speech inflects (e.g., Case, Number), the relevant values for those parameters (e.g., Nominative, Accusative; Singular, Plural), and the licit combinations of parameter values (e.g., Nominative Singular, Nominative Plural). The informant then posits any number of paradigms…[Oflazer, 2001, pg. 65]

In addition, it appears the informant is expected to know what cell of the paradigm holds the citation form. Since many native speakers will not know this, Boas advises them to look at a printed grammar of their language. The focus of Boas as a learning program, then, is not on acquiring the grammar per se, but rather on acquiring the lexicon, in the broad sense: a list of lexemes including their assignment to inflection classes, together with any irregular forms, and ultimately the general phonological rules which explain most allomorphy.

The grammatical information listed above is provided in the test bench as the morphological portion of the Grammar Description (GM). However, as discussed earlier, this information is mostly intended for human consumption; no API into this grammatical description is envisioned. Even if some of the information (parts of speech, inflectional features, and inflectional classes) were provided in computer-readable form (e.g. as an XML file), the form-based interface to Boas described in Oflazer (ms.) does not lend itself to receiving input through such an API (although presumably another interface to Boas could be built).

Thus, the morphology test bench is best suited to evaluating Boas' acquisition of the lexicon. With regard to this lexical data, Boas requires the user to provide all the forms of at least one member of each inflection class (referred to as the 'primary example'; this is the paradigm of a stem, item X-3). Additionally, Boas expects the informant to provide citation forms (or roots) and inflection class affiliation for other stems (given in DL-3),[9] as well as any irregular forms (DL-2). Finally, Boas generates morphophonological rules to predict new forms, thereby avoiding exhaustive elicitation. Inevitably these rules will under- or over-apply, so the user must decide whether wordforms generated by the system are correct. The test bench can mimic the user in by using its morphological transducer (X-1) to parse the wordforms presented by Boas, verifying that each wordform is indeed the desired inflected form of the stem in question.

Boas employs a finite state transducer representing lexical items, affixes, and constraints on their co-occurrence and allomorphy. Testing this transducer involves two sorts of tests: parsing known wordforms (analogous to measuring the recall of the grammar + lexicon), and generating wordforms from lexemes plus morphosyntactic features (analogous to measuring precision).

To evaluate Boas's ability to parse, the test bench can pass to Boas wordforms taken from the paradigms of lexemes in the SL dictionary (DL-2 and X-3; recall that these paradigms are produced by the test bench's own morphological transducer). Verification then consists of checking agreement between the parse returned by Boas, and the stem + inflectional features of the cell from which the wordform was generated.

To evaluate generation, Boas would produce the paradigms of lexemes it has learned, and these paradigms would be compared with the actual paradigm given by the test bench for each lexeme (X-3).

Since Boas generates a set of ordered morphophonological rules, it might be interesting to compare this set with the rules provided in the test bench (GP). However, as described above, there is no requirement that these should be the same (or even similar); it is enough that the correct surface forms be generated for each combination of a lexeme and a set of inflectional features. Likewise, while it might be interesting to compare the set of inflection classes discovered by Boas with the set provided in the test bench (GM-3), this is not a requirement for evaluation. (Since the test bench is providing the inflection classes to Boas, these are in any class unlikely to differ greatly.)

4.3. Stealth-to-Wealth

'Stealth-to-Wealth' ('S2W') is a term for a general approach being developed by the Summer Institute of Linguistics (SIL), of which no real implementation exists as yet. Given the lack of anything testable, the purpose in mentioning it at all is to see what resources such a program would require from a morphological test bench.

S2W is intended to assist a field linguist (a linguist with perhaps less training than the average linguist in academia) to analyze and describe the morphology and phonology of a language, with assistance needed on the part of more trained linguists only for the more difficult analytical problems. The approach envisions an object-oriented database for storage of information, together with a parser/ generator to test the grammar. The power of the S2W system for helping a field linguist do analysis comes in part from its object-oriented nature, such that the system 'knows' what data means, and is therefore able to reason about it.

The S2W discovery approach is meant to be driven off the process of doing interlinear glossing of texts. In a field situation, the linguist has the help of a native speaker to gloss the meaning at the sentence level, and perhaps at the word level. To simulate this process, the test bench should provide glossing at the sentence level (TB-2) and (perhaps on demand) at the word level (TB-3).[10]

From the process of glossing, the S2W method is intended to build a bilingual morpheme dictionary (and word dictionary, for irregular forms), and a humanly interpretable morphological grammar and phonology in computationally implemented form. The phonology can exist at various levels of sophistication, ranging from simple statements about allomorphy, to ordered phonological rules. Similarly, the morphosyntax and morphotactics can range from ad hoc to sophisticated. From the standpoint of a morphology test bench, then, the emphasis at evaluation time is on weak equivalency, not strong equivalency of grammars. In particular, it is not necessarily the case that morpheme breaks will be

[9] Future work may include learning the assignment of citation forms to inflection classes.

[10] Seldom if ever would text divided, glossed and aligned at the morpheme level be available in a realistic learning situation.

identical between the grammar produced by the S2W process and that described in the test bench.

Evaluating an S2W grammar at a weak level will therefore be similar to evaluating a Boas grammar. That is, an S2W grammar can be evaluated by having the S2W parser analyze wordforms taken from the paradigms of lexemes in the SL dictionary (DL-2 and X-3), and ensuring that the resulting parse agrees with the stem + inflectional features of the cell from which the wordform was taken. Likewise, to evaluate generation, the S2W transducer would produce the paradigms of lexemes it has learned, and these paradigms would be compared with the actual paradigm given by the test bench for each lexeme (X-3). In addition, since the S2W process is intended to discover derivational morphology (in addition to inflectional morphology), it will be necessary to compare the stems which can be productively derived from a given stem in the S2W analysis with the derivations given by the test bench (X-4).

At a stronger level of comparison, it may be of interest to compare the morpheme dictionary produced by the S2W process with the test bench's bilingual dictionary (English → SL lexeme dictionary and/or SL → English lexeme dictionary, DL-1or 2, together with the corresponding affix dictionary, DA-1 or 2).

Another product of the S2W process is bilingual text divided, glossed and aligned at the morpheme level. This differs from what would result if an accurate morphological parser were used to do the same task without supervision, in that the human is assumed to have disambiguated the result. This output can be evaluated against the same parsed and disambiguated text in the test bench (TB-4).

4.4. Learning from Bitexts

A method which might prove successful for morphology acquisition, but which to my knowledge has not been tried, is learning from bitexts (bilingual texts). This method begins with an aligned bitext segmented at no finer a level than the sentence level, with corresponding segments in the KL and the UL linked. The KL text must exist in stemmed form, i.e. without any inflectional affixes. (Even better would be a KL text consisting of sense-tagged stems, i.e. stems which represent a single sense.)

Given a stemmed KL text, the next step is to find alignments between stems in this text and possible stems in the UL text. Clearly there will be a large degree of ambiguity in such alignment, so proposed alignments will be probabilistic.

Once some number of stem-level alignments has been done (perhaps several hundreds), it should be possible to begin looking at the remainders of each word in the UL, which are potential affixes or sequences of affixes. Several automated or semi-automated methodologies might be used to search for candidate affixes, but determining the meaning of such affixes would probably be human-directed, assisted by appropriate views (such as concordance views of individual affixes).

Finding stem-level alignments and finding inflectional affixes in the UL will be a mutually reinforcing process: stripping potential affixes off words in the UL will leave the remainders of those words as potential stems.

The data to allow the simulation of this learning from bitexts would be bilingual text aligned at a 'segment' level (TB-2). Evaluation of this method would use the same resources as would evaluation of the Stealth-to-Wealth methodology.

4.5. Other Kinds of Learning

I have not proposed a use for bilingual text divided, glossed and aligned at the morpheme level (resource TB-4), apart from its possible use for evaluating the same sort of text produced as a side effect of the Stealth-to-Wealth methodology. Nor have I proposed a use for a tagger (resource Tg). However, in most real uses of a morphology learner (e.g. as part of a machine translation system), real texts must not only be parsed but also disambiguated. It may be that part of that disambiguation could be done by a tagger. While none of the systems I have discussed as potential users of a morphology test bench actually does this, it seems a logical extension. The resulting disambiguated morphologically parsed texts could then be evaluated against the test bench's parsed and disambiguated bilingual texts (TB-4).

Not all parses can be disambiguated by the part of speech of the whole word. In particular, syncretism in a paradigm cannot be disambiguated in this way. But it should be possible to train a program which would disambiguate paradigm syncretism (and presumably other types of ambiguity) using text glossed at the morpheme level. Again, this might be a natural extension from word-level tagging.

Morphologically analyzed text (as opposed to analyzed words out of context) might also be used as a training method for a general morphology learner method that would learn the meaning of wordforms without explicitly parsing them into morphemes. Such a method might be an extension of work that has been done on morphology learning in the connectionist paradigm.

5. Remaining Issues

Some issues which remain to be resolved include:

- Choice of languages
- Quantity of data
- The API to the data
- Representation of morphosyntactic properties (features), given that there is no universally agreed-on ontology for annotation or glossing
- Representation of complex morphosyntactic features
- Representation of morpheme breaks (which may be controversial) and 'zero morphemes'
- Ensuring commensurability between the dictionary and grammar on the one hand, and the input/output of the transducer on the other, given that the transducer may have been developed independently of the dictionary and grammar

6. Summary

The objective of a morphology test bench would be to assist designers of programs for learning computationally implemented morphological grammars by providing learning data, and to make it possible to evaluate and compare such programs.

This paper has described a set of resources, and several views of those resources, which would seem to be useful components of such a test bench. Several outstanding issues were described as well.

Finally, input is solicited concerning the type and variety of data which should be included.

7. References

Albright, Adam; and Bruce Hayes. 1999. "An Automated Learner for Phonology and Morphology". Ms. http://www.humnet.ucla.edu/humnet/linguistics/people/hayes/learning/learner.pdf.

Bosch, Antal van den; Walter Daelemans; and Ton Weijters. 1996a. "An Inductive-Learning Approach to Morphological Analysis". Paper presented at *Sixth CLIN Meeting*, University of Antwerp, Belgium.

Bosch, Antal van den; Walter Daelemans; and Ton Weijters. 1996b. "Morphological analysis as classification: An inductive-learning approach". Paper presented at *NeMLaP-2*, Bilkent University, Turkey.

Gaussier, Eric. 1999. "Unsupervised learning of derivational morphology from inflectional lexicons". Paper presented at *Workshop in Unsupervised Learning in Natural Language Processing*, University of Maryland.

Goldsmith, John. 2001. "Unsupervised Learning of the Morphology of a Natural Language". *Computational Linguistics* 27:153-198.

Kazakov, Dimitar; and Suresh Manandhar. 2001. "Unsupervised Learning of Word Segmentation Rules with Genetic Algorithms and Inductive Logic Programming". *Machine Learning* 43:121-162.

Oflazer, Kemal; and Sergei Nirenburg. "Practical Bootstrapping of Morphological Analyzers". Ms. http://crl.nmsu.edu/expedition/publications/oflazer.ps.

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

Schone, Patrick; and Daniel Jurafsky. 2000. "Knowledge-Free Induction of Morphology Using Latent Semantic Analysis". Paper presented at *CoNLL-2000*, Lisbon, Portugal.

Snover, Matthew G.; and Michael R. Brent. 2001. "A Bayesian Model for Morpheme and Paradigm Identification". Ms. http://lsrg.cs.wustl.edu/acl01bs.pdf.

Spencer, Andrew. 1991. *Morphological theory: an introduction to word structure in generative grammar*. Oxford, UK ; Cambridge, Mass.: Basil Blackwell.

Yarowsky, David; and Richard Wicentowski. 2000. "Minimally Supervised Morphological Analysis by Multimodal Alignment". Paper presented at *ACL-2000*, Hong Kong.

Yarowsky, David; Grace Ngai; and Richard Wicentowski. 2001. "Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora". Paper presented at *First International Conference on Human Language Technology Research*.

Zajac, Rémi. 2001. "Morpholog: Constrained and Supervised Learning of Morphology". Paper presented at *CoNLL-2001*, Toulouse, France.