The Use of Machine Readable Dictionaries in the Pangloss Project*

Stephen Helmreich, Louise Guthrie, and Yorick Wilks

Computing Research Laboratory
Box 30001
New Mexico State University
Las Cruces, NM 88003-0001

Machine Readable Dictionaries (MRDs) contain much useful information about language. Researchers have worked for the last decade on ways to extract this information for language processing systems. But processing dictionaries for use in natural language computation is itself a difficult problem. Transforming information from a version designed for human readers to one usable by computers is a slow process that, for example, may require one to define a "semantics" of font changes on a dictionary printer tape (e.g., that small bold capital letters mean that the word so capitalized is a synonym of the headword). Some fairly straight-forward information, such as word lists, part of speech lists, and orthography indications have been used by various language processing systems, yet the rather obvious use of MRDs -- to allow the automatic construction of lexicons for natural language processing systems, and in particular for MT systems -- has not been within our reach until recently.

[Farwell, Wilks and Guthrie 1992] described work which used the *Longman Dictionary of Contemporary English* (LDOCE) to construct lexical entries semi-automatically for one of the lexicons in the ULTRA machine translation system [Farwell and Wilks 1990]. Currently, the CRL is working together with the Center for Machine Translation at Carnegie-Mellon and the Information Sciences Institute at the University of Southern California on *Pangloss*, a Department of Defense-funded machine translation project. Some of the goals of that project are to use MRDs in the construction of its lexicons, to improve the techniques used in [Farwell et. al 92] and to develop techniques for using bilingual dictionaries to construct the individual language lexicons automatically.

The *Pangloss* project is founded on an interlingual, knowledge-based approach combined with statistical methods. Automatic acquisition of lexical information is used at all three levels of the initial system: analysis (Spanish), interlingual/ontological, and generation (English). For this extraction/acquisition task we have used both LDOCE and Collins Spanish-English and English-Spanish bilingual dictionaries.

In this paper we address issues raised by the call for papers from the perspective of this project and also describe in some detail the actual work so far carried out.

Lexical levels - an overview of the lexicons

The interlingual nature of the *Pangloss* system lends itself to a stratified approach to lexicon building: lexical information, for both English and Spanish, contains primarily syntactic and morphological information, together with a pointer to an interlingual conceptual token containing semantic information. In our initial system there are three levels for which automated acquisition is useful: the Intermediate Representation Lexicon based on ULTRA, the Spanish lexicon for analysis (also originally based on ULTRA), and the English lexicon used by the generator, Penman [Penman Natural Language Group, 1989]. During the second stage of the project, we use LDOCE extensively for generating entries in the joint ontology, and in the initial processing of Spanish input. We discuss each level separately.

I. Spanish Lexical Entries

Nouns:

```
se_form(san,josé,ts,f,_A,san_jose_x).
se_form(abarrote,ts,m,_A,groceries0_0).
se_form(abarrotes,tp,m,_A,groceries0_0).
se_form(abrigo,ts,m,_A,coat1_1).
se_form(abrigo,tu,m,_A,shelter1_2).
se_form(abrigos,tp,m,_A,coat1_1).
```

Spanish entries for nouns and pronouns have the representation above. "se_form" identifies the entry as a noun or pronoun. The components of the five-tuple encode the lexical item, person and number information, gender information, case information and the corresponding interlingual token. Noun senses make up more than half the senses in a standard dictionary and will be a large part of the final *Panglosss* Spanish lexicon. The entries were created in a semi-automatic way, using the Collins Spanish/English dictionary as an aid. We are developing procedures to create these entries entirely automatically from the machine-readable version of the Collins dictionary. The procedures for supplying the first four components (lexical item, person and number, gender, and case)

* This research was supported at the New Mexico State University Computing Research Laboratory by NSF Grant IRI-9101232 and through DoD Contract No. MDA904-91-C-9334.

from the dictionary have been developed and we are investigating techniques for associating the correct interlingual token.

Verbs as well as adjectives are represented in the Spanish lexicon of *Pangloss* as shown in Figure 1. "sr_form" identifies the entry as a verb or adjective. The ten-tuple representation encodes the information below whenever it is appropriate. Use of a prolog-style variable ("_A" for example) indicates components which are not germane to the entry. The ten fields indicate the lexical token, whether the verb is stative or dynamic, agreement information (person and number), agreement information (gender), other morphological information, information on tense, aspect, mood, and voice as well as the corresponding interlingual token. At present these entries were created interactively, with the use of the Collins Bilingual Dictionary. Procedures are being developed to create entries automatically from the dictionary: We have completed procedures to derive automatically the agreement information (person, number and gender), tense, aspect, mood, voice and morphological information. We are investigating methods for deriving the remaining features as well as techniques for associating the interlingual token.

For example, we have obtained as a resource a bilingual (Spanish/English) list of some 3000 financial terms compiled by *Barron's*. This list of Spanish terms with corresponding English terms allows us to identify the relevant possible sense tokens from LDOCE. Where a collocation or term is not listed in Longman's, an alphabetic rather than numeric tag is assigned to the word sense.

In the second stage of the project, we are also using Collins bilingual dictionaries to provide information needed for pre-processing Spanish text. For example, we have developed a part-of-speech tagger for Spanish, based on word lookup in Collins. To do this, it was necessary to obtain the citation form for inflected words. For this purpose, we used verb classification information to develop a look-up table for all the inflectional forms of verbs listed in Collins.

II. Intermediate Representations:

The entries corresponding to nouns and verbs in the intermediate representation are based on the intermediate representation of ULTRA which have been described in [Farwell et al. 1992]. The entity entries in Figure 2 have been derived entirely from the machine readable version of LDOCE, except for the third component which is a manually assigned semantic category. "ir_spec_ent" entries correspond roughly to nouns. The fields encode a semantic category, whether the noun is proper or common, whether it is mass or count, the LDOCE semantic category (provided from the machine readable version of the dictionary), and the LDOCE domain category (provided for some senses in the machine readable version of LDOCE). For nouns which do not appear in LDOCE (such as "San Jose" above), the entries were created manually. The first component represents the headword, homograph and sense in LDOCE. If there is only one homograph or sense, it is labelled "0". Thus, "grocery0_0" indicates that it is the first (and only) sense of the first (and only) homograph of the headword "grocery". Headwords not appearing in LDOCE are given an arbitrary homograph/sense tag as can be seen in "san_jose_x".

The entries in the intermediate representation corresponding roughly to verbs and adjectives were also generated semi-automatically. See [Farwell 1992] for details. In the example entry above, "ir_spec_rel" indicates a relational entry (verb or adjective). The fields mark the sense token, whether the sense is dynamic or stative, a semantic classification for the verb, the semantic roles of its arguments (subject, direct object, indirect object, if any), and the semantic classification of the entities generally filling those roles.

During the second stage of the *Pangloss* project, this Intermediate Representation Lexicon will be incorporated within an Ontology created by combining the already substantive ontologies of Penman's Upper Model, CMU's Ontos, and the semantic coding hierarchy of LDOCE semantic classification codes for nouns. Into this upper-level network will be placed the LDOCE senses, using the network of disambiguated genus terms described in [Bruce and Guthrie 1992].

During the second stage of the project, the Spanish analysis system will access these entries in the Ontology rather than the Intermediate Representation Lexicon to obtain semantic information. We also plan that the LDOCE semantic codes will be used as a basis for sense disambiguation, in addition to the hand-coded ULTRA-based semantic categories.

III. English Lexical entries for generation

As the entries for Penman (Figure 3) are more readable, we will not describe their content in detail. Given a list of proposed Penman English lexical entries, we have developed procedures for deriving from LDOCE the inflected forms of nouns, verbs, adjectives and adverbs.

Interdependence and Uniformity of Representations

For theoretical reasons, these different lexical representations--Spanish lexicon, English lexicon, and Intermediate Representation lexicon (Ontology)--are kept separate. The representations differ according to the type of information contained (syntactic vs. semantic) and according to use (analysis vs. generation). The second-stage Ontology, however, will be accessed by both the analysis and generation modules, and will also serve as

the basis for inferencing within the Interlingual representation, which we call the TMR (Text Meaning Representation).

Use of Automated procedures

In addition to the automated procedures described above, we have extracted various phrasal lexicons for English/Spanish, and Spanish/English, both from the Collins bilingual dictionaries and from corpora. We have developed a lexicon of phrases and idioms from LDOCE and, by parsing the definitions of LDOCE [Slator 1988] and developing procedures for disambiguating the hypernyms of noun senses in the dictionary, we have created a network of over 20,000 noun sense representations [Bruce & Guthrie 1992], [Guthrie et al. 1990] which is being used in the ontology of *Pangloss*. We are also developing a procedure for identifying and sense disambiguation *typical* arguments of verbs, where this information is provided by LDOCE. In short, our efforts have been directed toward providing scalability in encoding explicit linguistic and semantic information that techniques relying on implicit information have been able to approach using purely statistical methods [Brown et al., 1990].

Sharing lexicons

Part of our effort has been directed to extracting information from MRDs on-line for the specific purposes of this project. Undoubtedly that will continue, particularly as new MRDs become available. Another part of our effort, however, has been directed to the general problem of making on-line dictionary information more accessible: processing each dictionary and its entries into a format that makes extraction of *any* information much simpler and faster. For each dictionary, of course, this machine-tractable (as opposed to merely machine-readable) format differs, if only because different types of language and different types of dictionaries contain different kinds of information. Therefore, at a general level, we are in the process of constructing a lexical data base, which can contain the information from several dictionaries.

Bilingual dictionaries

Although our work with the Collins dictionaries has not yet been as extensive as that with LDOCE, it appears that at least as much useful information for MT purposes is available there. We have already made use of some monolingual information, such as morphological information and syntactic information. Monolingual semantic information in the form of synonyms, subject areas, or preferred collocations is perhaps more accessible and more extensive than in LDOCE. Phrasal lexicons have been extracted from Collins for both example-based translation and for coding multi-word lexical items for the *Pangloss* parser/analyzer.

MT mismatches and divergences

Mismatches between source and target languages are numerous and exist at every relevant semantic level of language (lexical, syntactic, pragmatic, cultural). Dealing appropriately with mismatches at a lexical level is a vital task for any MT system. Mismatches at the lexical level (sense splits, sense overlaps, complete mapping failures) are well-known problems for MT. Within the interlingual model of the current *Pangloss* system, these sense mismatches are not resolved at the lexical level. There is no direct mapping from source language lexical items to a set of target language lexical items. Every language-specific lexical entry maps uniquely into an interlingual concept token. If disambiguation of the source text is not or cannot be accomplished sufficiently well to provide the distinctions necessary for the generation module during analysis, the generator itself must make the distinctions based on the intermediate representation. This procedure (as employed in the ULTRA system) is described in [Helmreich, Jin, Wilks, and Guillen 1992].

Conclusion

The (semi-) automatic use of MRDs for applications like MT remains a desirable goal, but cannot as yet be claimed as proved, though we have certainly given what we believe to be the first operational proof of the concept. We do not suppose that very general MRDs (like LDOCE and Collins) will suffice in the absence of domain-specific lexicons described from bilingual corpora (parallel and, more intriguingly, non-parallel) by empirical methods. What we have described here is no more than a core lexicon. In a separate project (Tipster) we are investigating not only how to join MRD-derived to corpora-derived lexicons but how to obtain the latter by "tuning" the former against the domain corpus itself [Anick & Pustejovsky 1990; Cowie et al. 1992, 1993].

As is well known, there is at the moment no way of evaluating the performance of a lexicon separately from the whole project of which it is a part, other than by comparing final output performance with different versions of a lexicon at different times (while the rest of the system is kept fixed). We believe fuller evaluation methods for lexical systems will be derived, but at the moment this project (of automatic lexical provision for *Pangloss*) is subject to the evaluation of *Pangloss* itself, which is conducted periodically against other DoD MT systems funded under the same overall scheme.

```
sr_form(identifica,dyn,ts,_A,fin,prs,simp,indic,actv,identify0_1).
sr_form(identifica,dyn,tu,_A,fin,prs,simp,indic,actv,identify0_1).
sr_form(identificada,dyn,ts,f,psp,_A,_B,_C,pasv,identify0_1).
sr_form(identificadas,dyn,tp,f,psp,_A,_B,_C,pasv,identify0_1).
sr_form(identificado,dyn,_A,_B,psp,_C,perf,_D,actv,identify0_1).
sr_form(identificado,dyn,ts,m,psp,_A,_B,_C,pasv,identify0_1).
sr_form(identificados,dyn,tp,m,psp,_A,_B,_C,npasv,identify0_1).
sr_form(identifican,dyn,tp,_A,fin,prs,simp,indic,actv,identify0_1).
sr_form(identificando,dyn,_A,_B,prp,_C,prog,_D,actv,identify0_1).
sr_form(identificar,dyn,_A,_B,inf,_C,simp,_D,actv,identify0_1).
sr_form(identificaron,dyn,tp,_A,fin,pst,simp,indic,actv,identify0_1).
sr_form(identificará,dyn,ts,_A,fin,fut,simp,put,actv,identify0_1).
sr_form(identificará,dyn,tu,_A,fin,fut,simp,put,actv,identify0_1).
sr_form(identificarán,dyn,tp,_A,fin,fut,simp,put,actv,identify0_1).
sr_form(identificó,dyn,ts,_A,fin,pst,simp,indic,actv,identify0_1).
sr_form(identificó,dyn,tu,_A,fin,pst,simp,indic,actv,identify0_1).
```

Figure 1. Pangloss Spanish Lexicon: verbs

Entities:

```
ir_spec_ent(san_jose_x,prop,loc,c,_Lc,_Ld).
ir_spec_ent(groceries0_0,nrm,p_obj,c,sol_or_liq,open).
ir_spec_ent(grocery0_0,nrm,place,m,abstract,open).
```

Relations:

```
ir_spec_rel(identify0_1,dyn,act,agnt,pat,none,human,p_obj,none).
```

Figure 2. Intermediate Representation Lexicon: entities and relations

Penman noun:

```
(lexical-item
 :name LOS-ANGELES
 :spelling "Los Angeles"
 :sample-sentence ""
 :features (NOUN NOINFLECTIONS PROPERNOUN COUNTABLE NOT-PERIOD
  NOT-DETERMINERREQUIRED NOT-PROVENANCE)
 :comments "city"
 :date "10/24/88 10:21:02"
 :editor "HOVY")
```

Penman Verb:

```
(lexical-item
 :name IDENTIFY
 :spelling "identify"
 :sample-sentence "The navy identified the new Russian ship as the Gorky."
 :features (VERB INFLECTABLE LEXICAL NOT-CASEPREPOSITIONS
  OBJECTPERMITTED NOT-TOCOMP QUESTIONCOMP PARTICIPLECOMP NOT-MAKECOMP
  BAREINFINITIVECOMP NOT-COPULA PASSIVE THATCOMP NOT-THATREQUIRED
  NOT-SUBJUNCTIVEREQUIRED NOT-ADJECTIVECOMP
  NONE-OF-BITRANSITIVE-INDIRECTOBJECT EXPERIENCEVERB PERCEPTION MIDDLE
  OBJECTNOTREQUIRED NOT-SUBJECTCOMP UNITARYSPELLING VISUAL IRR
  PLURALPASTFORM SECONDSINGULARPASTFORM PLURALFORM FIRSTSINGULARFORM
  EDPARTICIPLEFORM THIRDSINGULARFORM INGPARTICIPLEFORM PASTFORM
  SECONDSINGULARFORM FIRSTSINGULARPASTFORM THIRDSINGULARPASTFORM)
 :properties ((INGPARTICIPLEFORM "identifying")(EDPARTICIPLEFORM
  "identified")(PLURALPASTFORM "identified")(THIRDSINGULARPASTFORM
  "identified")(SECONDSINGULARPASTFORM "identified")
  (FIRSTSINGULARPASTFORM "identified")(PASTFORM "identified")
  (PLURALFORM "identify")(SECONDSINGULARFORM "identify")
  (FIRSTSINGULARFORM "identify")(THIRDSINGULARFORM "identifies"))
 :comments "process"
 :date "10/26/88 12:08:51"
 :editor "HOVY")
```

Figure 3. Penman entries

References

Anick, P., and J. Pustejovsky (1990). Knowledge Acquisition from Corpora. COLING-90(2):7-12.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin (1990). A Statistical Approach to Machine Translation. *Computational Linguistics,* 16:79-85.

Bruce, R., and L. Guthrie (1992). Genus Disambiguation: A Study in Weighted Preference. COLING-92(4):1187-1191.

Cowie, J., T. Wakao, L. Guthrie, W. Jin, and J. Pustejovsky (1992). CRL/NMSU and Brandeis: Report on the Diderot System as Used for the Tipster 12 Month Evaluation. TIPSTER 12-month meeting, Plenary Session Notebook, San Diego, California.

Cowie, J., L. Guthrie, W. Jin, and J. Pustejovsky (1993). The Diderot Information Extraction System. To be presented at PAC-LING 93, Vancouver, B.C.

Farwell, D. et Y. Wilks (1990). ULTRA: a Multi-lingual Machine Translator. *Memoranda in Computing and Cognitive Science,* MCCS-90-202, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Farwell, D., L. Guthrie, et Y. Wilks (1992). The Automatic Creation of Lexical Entries for a Multilingual MT System. COLING-92(2):532-538.

Guthrie, Louise, Brian Slator, Yorick Wilks, and Rebecca Bruce (1990). Is there content in Empty Heads? COLING-90(3):138-143.

Helmreich, S., W. Jin, Y. Wilks, R. Guillen (1992). Research Issues in Machine Translation at the Computing Research Laboratory. *Memoranda in Computing and Cognitive Science,* MCCS-92-242, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Penman Natural Language Group (1989). The Penman Reference Manual. The Penman User Guide. The Penman Primer. Information Sciences Institute at the University of Southern California, Marina del Rey, CA.

Procter, P., L. Ilson, J. Ayto, et al. (1978). *Longman Dictionary of Contemporary English.* Harlow, UK: Longman Group Limited.

Slator, Brian M. (1988). Constructing Contextually Organized Lexical Semantic Knowledge-bases. *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-88),* Denver, CO, pp.142-148.

Smith, C., with M. Bermejo Marcos and E. Chang-Rodriguez (1990). *Collins Spanish-English, English-Spanish Dictionary.* Glasgow, UK: Harper-Collins Publishers