The BICORD System

Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries[1]

Judith Klavans
IBM T.J. Watson Research
Yorktown Heights, N.Y. 10532

Evelyne Tzoukermann
A.T. & T., Bell Laboratories,
Murray Hill, New Jersey 07974

ABSTRACT

Our goal is to explore methods for combining structured but incomplete information from dictionaries with the unstructured but more complete information available in corpora for the creation of a bilingual lexical data base. This paper concentrates on the class of action verbs of movement, and builds on earlier work on lexical correspondences between languages and specific to this verb class. The languages we explore here are English and French. We first examine the way prototypical verbs of movement are translated in the Collins-Robert (Collins 1978, henceforth CR) bilingual dictionary. We then analyze the behavior of some of these verbs in a large bilingual corpus. We take advantage of the results of linguistic research on verb types (e.g. Levin, to appear) coupled with data from machine readable dictionaries to motivate corpus-based text analysis for the purpose of establishing lexical correspondences with the full range of associated translations and then attach frequencies to translations.

1. Background. As NLP systems become more robust, large lexicons are required, providing a wide range of information including syntactic, semantic, pragmatic, morphological and phonological. There are difficulties in constructing these large lexicons, first in their design, and then in providing them with the necessary and sufficient data. These problems have recently been the topic of intense research (Klavans 1988, Boguraev and Briscoe 1989, Boguraev et al. 1989, Zernick 1990). Moreover, an important sub-area of computational lexicon building that has barely been approached is that of bilingual lexicon construction (Calzolari and Picchi 1986, Rizk 1989).

2. Motion Verbs. In this paper, we report on data for movement verbs (or motion verbs). The class of English motion verbs and their translations into Romance languages has been widely discussed from various points of view including theoretical, structural (Talmy 1985), and applied (Atkins et al. 1990, in preparation). English generally incorporates movement and cause or manner into a single lexical item whereas languages like French do not. For example, in CR *stroll* is translated as 'se promener nonchalamment', 'flâner' and *stroll in/out etc.* as 'entrer/sortir/s'éloigner sans se presser' or 'nonchalamment'. Notice that in French, the translation typically consists of a general motion verb 'entrer/sortir/aller/avancer' with an adverbial or prepositional modifier showing manner, e.g. 'nonchalamment' or 'sans se presser'. Similarly, in English, causation in movement is often incorporated, e.g the English verb *march* as in *to march the troops* is translated in CR as 'faire marcher (au pas) les troupes'. These multi-word correspondences often cause problems in the lexical transfer component of machine translation systems.

3. Bilingual Corpus-based Analysis. In earlier work (Klavans and Tzoukermann 1989), we reported on a study of a selected sub-set of movement verbs in a bilingual corpus. The corpus consists of 85 million English and 95 million French words from the Canadian Parliamentary Proceedings (the Hansard corpus). Of this, 75 million French and 70 million English words are aligned by sentence (Brown et al. 1988). For example:

```
      SENTENCE #:          357748
The ambassador's contribution was one small
party at which a number of us ended up
dancing on a table.

L'apport de l'ambassadeur s'est résumé
à une petite fête ou nous avons fini
par danser sur une table.
```

Figure One : Sample Citation

Some representative verbs which have at least one movement sense were selected. We compared the extent of the information found in the bilingual corpus with the information found in the CR machine-readable dictionary (MRD). For verbs like *commute* which do not have a straightforward translation, we found either (1) all the components of the verb concept, as in 'se rendre au travail quotidiennement'; (2) parts of the translation, as in 'faire le trajet'; or (3) a totally different verb from that given in the MRD, such as 'parcourir' or 'voyager'.

We observed that, not only was the MRD information incomplete, but also only a partial ex-

[1] This work was completed at IBM, T.J. Watson Research, although the second author is currently at A.T. & T., Bell Laboratories.

pression of the typical meaning of the verb was provided. In the past, since printed dictionaries have been subject to the constraints of time and space, they have not always been able to offer full information about entries. However, with electronic dictionaries and lexical data bases, this should no longer be a restriction. In fact, given more and richer information, we envision a move away from the flat hierarchical structure of dictionaries to a more network-like representation of lexical knowledge.

4. Related Research. Combining linguistic and statistical methods is becoming increasingly popular in computational linguistics especially as more corpora become available.[2] Work in this vein ranges from the syntactic and semantic to the lexical. For example, Atkins 1987 demonstrates convincingly that with corpus data, the lexicographer can attack the difficult problem of word senses in a systematic way. Church and Hanks 1989 and Church et al. 1990 develop a battery of statistical methods to induce linguistic regularities. They identify coocurrence relations by computing statistics (e.g. by use of mutual information, t-score) over millions of words of text. Their approach is focussed on monolingual rather than bilingual corpus analysis, and constitutes a significant contribution to lexical research. On more syntactic note, Dagan and Itai 1990 use statistical methods over linguistically parsed text (Jensen 1986) to resolve anaphoric reference.

In the arena of automatic bilingual lexicon construction, Catizone et al. 1989 take two corresponding texts (English and German) and develop algorithms to determine lexical alignments by using statistical methods over texts combined with the optional support of an MRD. In contrast, Sadler 1989 proposes parsing aligned corpora into dependency trees, which form the structures upon which lexical correspondences are suggested to the user. The early stages of the construction of the Bilingual Knowledge Base (BKB) rely heavily on human input but gradually becomes more automatic as data is collected. Using purely statistical techniques, Brown et al. 1988 make use of the Hansard bilingual corpus for the purpose of building a machine translation system. Such a system is a good example of using exclusively statistical non-linguistic methods to induce translations.

5. The BICORD System - Bilingual Corpus-based Dictionary. Our approach involves a combination of standard linguistic methodology using MRD's, enhanced with some statistical techniques. Dictionaries are often discounted because they are built on basis of introspective intuition rather than purely on objective observation of data. However, our underlying assumption is that the insights that a dictionary encodes and represents should not be disregarded (although there are some limitations resulting from the structural organisation). This is a controversial assumption. Even though, in the past, dictionaries have been built solely on the basis of intuition, current trends are to use corpus-driven criteria, as, for example, in the Collins COBUILD dictionary (1987). Without question this is a step in the right direction towards completness and accuracy of coverage of the language as it actually occurs. However, the limitation of corpus analysis is that subtle linguistic intuitions about word behavior (such as "negative evidence") cannot be obtained from corpora; in other words, what is disallowed in the language may never be discovered. Thus we disagree with the claim of Garside, Leech, and Sampson 1987 that the survival of both descriptive and theoretical computational linguistics lies primarily in statistical analysis. We take the more moderate view that both approaches (linguistic and statistical) are essential if the language is to be characterized accurately and in its entirety.

We extracted occurrences of several movement verbs (called "probe" strings) from the English side of the Hansard corpus. The criteria used to ensure that the verb was a member of this semantic class is described in Atkins, Boguraev and Klavans 1990 (in preparation). The test set of verbs was *drift*, *dance*, *commute*, *emigrate*, *immigrate*, *ascend*, *descend*, *circle*, *sail* and *glide*. The probe string was used to search in CR, both for translations and collocations under the entry itself, and also for French headwords in the French side of the dictionary with the probe as a translation. The extracted corpora, consisting of the set of English citations containing the probe string (in any morphological shape) and the corresponding French sentence, is called a "probe corpus". A statistical tagger (Tzoukermann and Merialdo 1989) was used to assign a part of speech to the English side of the corpora. Translations and collocations were abstracted automatically from the parsed version of CR (see Neff and Boguraev 1989) using LQL (Neff et al. 1988). For illustration, a partial entry for *dance* is:

```
+-hdw: dance
+-superhom
  |...
  +-homograph
    +-homnum: 2
    +-pos: vt
    +-sense
      +-translat
      | +-argument: waltz etc
      | +-word: danser
  | ...
```

[2] For example, the ACL Data Collection Initiative (ACL/DCI) coordinated by Dr. Mark Liberman at A.T.& T. Bell Laboratories was established to make corpora of all shapes and sizes more widely available to the research community.

```
+-homograph
| +-homnum: 3
| +-pos: vi
| +-sense
    |...
    +-collocat
    | +-srcnote: fig
    | +-source: to dance in/out etc
    | +-target: entrer/sortir etc joyeusement
    |
    +-collocat
    | +-source: to dance about
    | +-source: to dance up and down
    | +-target: gambader
    | +-target: sautiller
    |
    +-collocat
    | +-source: the child danced away /or/ off
    | +-target: l'enfant s'est éloigné
    |           en gambadant /or/ en sautillant
    |...
```

Figure Two: Partial MRD entry for dance

Also, the French words 'gambiller' and 'guincher' have *dance* as a translation. Probes had a maximum of 1146 citations, with a maximum of 25 senses and collocations in CR (a rough measure of polysemy).

The tagger used to preprocess the corpus was trained on 1 million words (about 42,000 sentences) tagged manually and provided by the tree bank of Lancaster University (Garside, Leech, and Sampson 1987). Our version has 81 tags, a subset of the tree bank tags. Of these tags, 52 are categorial (such as VV*1 for infinitival form of a non-auxiliary verb) and 29 are lexically bound, some of the latter being bound to a class of one (e.g. I0* is for the preposition *of*), and some are bound to a small sub-class of category (such as PP*S for "personal pronoun subject"). Some tags (such as N*1 "singular noun") provide morphological information, as well as categorial. The program, based on a trigram model, computes the probability of a word in relation to its tag and assigns the tag that corresponds to the highest likelihood. In its simplest form:

$$p(T|W) = p(W|T) * p(T)$$

that is, the probability of a tag given its word corresponds to the product of the probability of observing the word given its tag by the probability of observing the tag. By random sampling, we determined the error rate for part of speech tagging to be about 3%.

In this way, examples of sample strings as a verb were separated from the nominal uses. This is the first step in disambiguation, enabling lexical correspondences. To give an idea of size, there were 293 citations (about 12,000 words) with the string *dance* in its four morphological forms in English. The distribution by part of speech for these citations is:

Category	Citations	%
VERB	109	37
NOUN	174	59
ADJ	10	3

The distribution varies by probe; for example, of the 34024 citations for the string *"move"* (and its variants), 26218 usages were labelled as verbs (77%), 7412 as nouns (22%), and 394 (11%) as adjectival. Some illustrative fragments for *dance* are:

```
we      are dancing upon eggshells...
PP*S    VBR* VVG1*    I*    N*2

the politician who  liked  to  dance...
AT*  N*1        P*Q VVPAST* TO*  VVI*

 ...Russian people dancing rather than fighting.
      J*    N*1    VVG1*   R*R     I*   VVG1*
```

Data from CR are utilized to drive our first pass at filtering out pre-linked pairs common to both data resources. Citations that have lexical correspondences already provided by the machine-readable dictionary are extracted from the probe corpus. For example, consider again the verb *dance*. The character strings in the translation and collocation fields are extracted from CR; these strings are filtered to remove function words and some common words (such as 'faire' (*to make* or *do*), morphological variants are generated. Some examples for *dance* are 'danser/dansa/dansera ..., gambader/gambadont ....' Probe translations and collocations from CR are then ready to be used to automatically match strings in the French side of the corpus. Each correspondence that matches one of the MRD probes is removed from the probe corpus, stored, and counted, leaving a reduced probe corpus. For example, for 109 citations of *dance* as a verb, 52 sentences matched the MRD correspondences, as shown in Figure One. An extended lexicon can then be built, using the structure already provided by CR where the frequencies are computed over these matches. For example, an initial partial enhanced entry for *dance* is:

```
+-hdw: dance
+-superhom
  |...
  +-homograph
  | +-homnum: 1
  | +-pos: v
  | +-sense
      +-c_translat
        +-word: danser
        | +-inflect: inf
        | +-freq: 44%
        +-word: danser
        | +-inflect: past
        | +-freq: 17%
        +-word: danser
        | +-inflect: fut
        | +-freq: 5%
        |...
  |...
  +-homograph
  | +-homnum: 2
  | +-pos: vt
  | +-sense
      +-d_translat
```

```
    | +-argument: waltz etc
    | +-word: danser
| ...
+-homograph
| +-homnum: 3
| +-pos: vi
| +-sense
    |...
    +-d_translat
    | +-context: person
    | +-context: leaves in wind
    | +-context: boat on waves
    | +-context: eyes
    | +-word: danser
    |
    +-d_collocat
    | +-srcnote: fig
    | +-source: to dance in/out etc
    | +-target: entrer/sortir etc joyeusement
    |
    +-d_collocat
    | +-source: to dance about
    | +-source: to dance up and down
    |
    | +-target: gambader
      | +-c_collocat
      |  +-source: to dance around
      |  +-inflect: present
      |  +-freq   : 2%
      |
    | +-target: sautiller
      | +-c_collocat
      |  +-source: to dance round
      |  +-inflect: past
      |  +-freq   : 2%
    |
    +-d_collocat
    | +-source: the child danced away /or/ off
    | +-target: l'enfant s'est éloigné
    |           en gambadant /or/ en sautillant
    |...
```

Figure Three: Partial Enhanced Entry

Notice that dictionary nodes are now identified with a prefix "d_", and corpus motivated nodes with "c_". New information is placed at the relevant node, low in the tree if there is no ambiguity of attachment or scope, and higher in the tree if necessary until evidence is found to permit the information to be moved down in the structure. For example, an additional node is added to the MRD structure to insert *danser* since *danser* is a translation both in homograph 2 and in homograph 3. Since transitivity of a verb cannot be determined automatically, there is no evidence to motivate placement so the data is inserted high in the tree, at the homograph level. In contrast, 'gambader' and 'sautiller' are always intransitive (as determined by a look-up in CR), so they can be automatically placed under homograph three. Notice also that corpus derived information is placed under the relevant d_collocat for 'gambader' and 'sautiller' since these are cases where matches occurred on the target term, but the source is different.

The Hansard, being the Canadian Parliamentary proceedings, contains a number of juridical and parliamentary terms, usages, and structures, a typical feature of any sublanguage. However the flexibility inherent in the BICORD system would allow a repetition of the same process over different sublanguages. As other texts are used, frequencies can be updated in two ways, by counting all frequencies into a general score, and also by keeping separate frequencies linked to the source text. This feature allows a representation of the lexical correspondences of general and specific texts in one data structure. It also permits comparison between sublanguages. The result would be a balanced lexicon built over a balanced variety of corpora to reflect the actual uses of the words or phrases in context.

Further analysis of the remaining probe corpus is pursued by observing cooccurences both over tags and lexical items. For example, with *dance*, looking at immediate right context over tags reveals verb-prep patterns:

VERB	CATEGORY	%
dance	prep	77
dance	other	22

Moving from tag cooccurences to lexical items, the majority of these cases are for the preposition *to*. Including coocurrences over a larger window of five words, idioms are revealed like *dance to ... tune*, which is not found in CR, either under *tune* or *dance*. These and other patterns can be discovered by statistical analysis over tags and lexical items in the reduced probe corpora. Therefore, a new set of collocations can be inserted in the lexicon; an entry for *"dance"* enhanced further is shown as follows:

```
+-hdw: dance
|
+-superhom
  |...
  +-homograph
  | +-homnum: 1
  | +-pos: v
  | +-sense
  |   +-c_translat
  |   | +-word: danser
  |   | | +-inflect: inf
  |   | | +-freq: 44%
  |   | +-word: danser
  |   | | +-inflect: past
  |   | | +-freq: 17%
  |   | +-word: danser
  |   | | +-inflect: fut
  |   | | +-freq: 5%
  +-homograph
  | +-homnum: 2
  | +-pos: vt
  | |
  | +-sense
  |   |
  |   +-d_translat
  |   | +-argument: waltz etc
  |   | +-word: danser
  | ...
  +-homograph
  | +-homnum: 3
  | +-pos: vi
  | +-sense
      |
```

```
+-d_translat
| +-context: person
| +-context: leaves in wind
| +-context: boat on waves
| +-context: eyes
| +-word: danser
|
+-d_collocat
| +-srcnote: fig
| +-source: to dance in/out etc
| +-target: entrer/sortir etc joyeusement
|
+-d_collocat
| +-source: to dance about
| +-source: to dance up and down
|
| +-target: gambader
  | +-c_collocat
  |  +-source: to dance around
  |  +-inflect: present
  |  +-freq   : 2%
  |
| +-target: sautiller
  | +-c_collocat
  |  +-source: to dance round
  |  +-inflect: past
  |  +-freq   : 2%
|
+-c_collocat
  +-source: to dance  to
  +-argument: (the) tune (of)
  +-freq    : 11%
  +-target : se mettre au diapason
  +-target : compléter le quatuor
  |...
+-c_collocat
| +-source: to dance  around
  +-freq    :  8%
  +-target : tourner autour du pot
  +-target : aller et venir
  |...
|
|
+-d_collocat
| +-source: the child danced away /or/ off
| +-target: l'enfant s'est éloigné
|           en gambadant /or/ en sautillant
|...
```

Figure Four: Fuller Enhanced Entry

It is not always the case that the remaining corpus data can be easily inserted in the lexicon and in fact, we encountered a few problems during this process. First, it is not straightforward to know with which field to associate the resulting correspondences. For example, in *dance*, does *dance around* go under a separate translation field or is it related to the collocation field with *dance about*? Second, some new context fields should be added to the collocation nodes, but determining the criteria for selecting them automatically is not always evident. Further, there is a question of locating and integrating robust new data from the corpus into the already existing structure.

6. Applications and Future Plans. A system such as BICORD can be used in two complementary ways: to enhance an MRD with statistical data and, conversely, to enhance a statistical system with data from an MRD. The first application can be viewed in the light of a lexicographer's workstation; it can also be viewed as a contribution to the choice of lexical item made by the component responsible for lexical transfer in a machine translation system. Translations and collocations in the original MRD are ordered by frequency, orderings which can easily be updated depending on the sub-language corpus. The enhanced MRD is more complete in containing correspondences not found in the original dictionary, and in suggesting new statistically significant translations. As for the second type of application, systems such as described in Brown et al. 1988 which use purely statisfical approaches to infer translations from a bilingual corpus can benefit directly from the information already given in the MRD. This information can be used to preset values in the computation of correspondences, rather than letting the system learn values already discovered.

Future work depends on testing these two applications, namely that MRD-based lexical transfer will proceed more accurately given statistical information and that statistical implementations, given enhanced MRD data, will demonstrate improved performance in determining lexical correspondences.

Acknowledgements: We thank members of the Speech Recognition Group at IBM for cleaning and maintaining the Hansard corpus. In particular, we acknowledge help from Bernard Merialdo.

References

1. Atkins B. T., (1987) "Semantic ID Tags: corpus evidence for dictionary senses", In *Proceedings of the Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary The Uses of Large Text Databases* , Waterloo, Canada, pp. 17-36.
2. Atkins, B.T.S., B. Boguraev and J. L. Klavans (1990, in preparation) "From Machine-Readable Dictionaries to a Lexical Knowledge Base: a Discussion of Some Issues with Particular Reference to Verbs of Motion", in J. Pustejovsky (ed.), *Semantics in the Lexicon*, Kluwer, Dordrecht.
3. Boguraev, Bran and Ted Briscoe (1989) *Computational Lexicography for Natural Language Processing*, Longman : London.
4. Boguraev, Branimir, Byrd, Roy, Klavans, Judith, and Neff, Mary (1989, to appear) "From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base", paper presented at IJCAI, to appear as a chapter in *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, MIT Press, Uri Zernik, editor.

5. Brown, P.,J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin (1988) "A Statistical Approach to Language Translation". *4th Conference on Computational Linguistics, Coling* , Budapest, Hungary.
6. Calzolari, N and E Picchi (1986) "A Project for a Bilingual Lexical Database System", *Advances in Lexicology*, Second Annual Conference of the UW Centre for the New Oxford English Dictionary, 79 – 92.
7. Catizone, Robert, Graham Russell, and Susan Warwick (1989) "Deriving Translation Data from Bilingual Text", unpublished ms. , ISSCO, Geneva, Switzerland.
8. Church K. and P. Hanks (1989) "Word Association Norms, Mutual Information and Lexicography", *Proceedings of the Association for Computational Linguistics* , Vancouver, Canada.
9. Church, K., W. Gale, P. Hanks, D. Hindle (1990, to appear) "Parsing, Word Associations, and Typical Predicate-Argument Relations", in Zernik ed..
10. Collins Cobuild English Language Dictionary (1987), John Sinclair, ed. Collins Publishers: London.
11. Collins. 1978. *Collins Robert French Dictionary: French-English. English-French.* Collins Publishers: London.
12. Dagan, Ido and Alon Itai (1990) "Automatic Acquisition of Constraints for the Resolution of Anaphora Reference and Syntactic Ambiguities" unpublished ms., Computer Science Department, Technion, Haifa, Israel.
13. Garside, R., G. Leech, and G. Sampson, eds. (1987) *Computational Analysis of English: a corpus-based approach* Longman : London and New York.
14. Jensen, Karen (1986) "PEG 1986: A Broad-coverage Computational Syntax of English," Unpublished paper. IBM Research: Yorktown Heights, New York.
15. Klavans, J. L. (1988) "COMPLEX: A Computational Lexicon for Natural Language Systems", *Proceedings of the 12th International Conference on Computational Linguistics.* Budapest, Hungary.
16. Klavans, Judith and Evelyne Tzoukermann (1989) "Corpus-based Lexical Acquisition for Translation Systems" *Proceedings of the Sixth Israeli Conference of Artificial Intelligence and Computer Vision, Tel Aviv, Israel..*
17. Levin, Beth. (to appear) "The Representation of Semantic Information in the Lexicon," in D. Walker, A. Zampolli, N. Calzolari, eds., *Automating the Lexicon -- Research and Practice in a Multilingual Environment.* Cambridge, England: Cambridge University Press.
18. Neff, M. S., R. J. Byrd, and O. A. Rizk (1988) "Creating and Querying Hierarchical Lexical Data Bases," *Proceedings of the Second ACL Conference on Applied NLP*, Austin, Texas, 84-92.
19. Neff, M. and B. Boguraev (1989) "Dictionaries, Dictionary Grammars and Dictionary Entry Parsing", *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 91 – 101.
20. Rizk, O. (1989) "Sense Disambiguation of Word Translations in Bilingual Dictionaries: Trying to Solve the Mapping Problem Automatically". Master's Thesis, Courant Institute of Mathematical Sciences, New York University, N.Y.
21. Sadler, Victor (1989) "The Bilingual Knowledge Bank: A New conceptual basis for MT" unpublished paper, BSO/Research, Utrecht.
22. Talmy, Leonard (1985) "Lexicalization Patterns: Semantic Structure in Lexical Forms", in T. Shopen, ed., *Language Typology and Syntactic Description: Grammatical categories and the Lexicon.* Cambridge University Press: Cambridge, England.
23. Tzoukermann, Evelyne and Bernard Merialdo. 1989. "Some Statistical Approaches for Tagging Unrestricted Text", unpublished ms., IBM, T. J. Watson Research Center, Yorktown Heights, New York, 10532.
24. Zernik, Uri (1990, to appear) *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, Lawrence Erlbaum Associates Incorporated: Hillsdale, New Jersey.