Machine Translation, 10:3-4, 1–34 (1996)

Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons

JUDITH KLAVANS klavans@cs.columbia.edu
Center for Research on Information Access, Columbia University, New York, NY 10027

EVELYNE TZOUKERMANN evelyne@research.att.com
A.T.&T. Bell Laboratories, 600 Mountain Avenue, Murray Hill, N.J. 07974

Abstract. This paper describes and discusses some theoretical and practical problems arising from developing a system to combine the structured but incomplete information from machine readable dictionaries (MRDs) with the unstructured but more complete information available in corpora for the creation of a bilingual lexical data base, presenting a methodology to integrate information from both sources into a single lexical data structure. The BICORD system (**BI**lingual **COR**pus-enhanced **D**ictionaries) involves linking entries in Collins English-French and French-English bilingual dictionary with a large English-French and French-English bilingual corpus. We have concentrated on the class of action verbs of movement, building on earlier work on lexical correspondences specific to this verb class between languages (Klavans and Tzoukermann, 1989), (Klavans and Tzoukermann, 1990a), (Klavans and Tzoukermann, 1990b).[1] We first examine the way prototypical verbs of movement are translated in the Collins-Robert (Atkins, Duval, and Milne, 1978) bilingual dictionary, and then analyze the behavior of some of these verbs in a large bilingual corpus. We incorporate the results of linguistic research on the theory of verb types to motivate corpus analysis coupled with data from MRDs for the purpose of establishing lexical correspondences with the full range of associated translations, and with statistical data attached to the relevant nodes.

1. Introduction

This paper addresses the issue of automatic lexicon construction, using a variety of resources including corpora and machine-readable dictionaries. The BICORD system (**BI**lingual **COR**pus-enhanced **D**ictionaries) involves linking entries in Collins English-French and French-English bilingual dictionary with a large English-French and French-English bilingual corpus. Our approach to data mining is to start with linguistic principles to drive the system. The next section presents the issues of bilingual correspondences as they appear in monolingual and bilingual MRD's, and bilingual corpora. Bilingual correspondences are studied from the viewpoint of motion verbs in English and French; examples are given to show not only the typical correspondences from one language to the other, but also some underlying conceptual correspondences of verbs belonging to this category. Issues such as literal and figurative meaning, transitivity, and telicity of motion verbs are addressed. Section 3 presents a theory based on decompositional approaches to motivate the analysis and extraction of motion verbs. Verbs are analyzed into conceptual entities; when this information is identifiable in MRD's, it can be retrieved for processing. Section 4 relates the different approaches used in building lexicons. Statistical

Table 1. Sample Citation for "dance/danser"

English:	The ambassador's contribution was one small party at which a number of us ended up **dancing** on a table.
French:	L'apport de l'ambassadeur s'est résumé à une petite fête ou nous avons fini par **danser** sur une table.

and linguistic methods are discussed with particular attention given to multi-word correspondences. Section 5 describes the algorithm used in the BICORD system. The algorithm makes use of a combination of statistical and linguistic techniques in order to extract the information on motion verbs in both MRD's and bilingual corpora. Once the information is extracted, it undergoes several qualifying tests; eventually, processed information is integrated in a large lexical database, built on the dictionary structure.

Our claim in this paper is that the MRD can be used to help statistical methods by providing a starter list with simple and definite correspondences; the list could be viewed as a clean set of training data. In this way, MRD data can be effectively used to solve some of the one-to-many and many-to-many problems for statistical approaches. At the same time, we observed that, not only was the MRD information incomplete, but also only a partial expression of the typical meaning of the verb was provided. Thus, for lexical analysis and selection, we argue that a non-enhanced MRD will be of limited use. What is necessary is a combination of the text corpus and the MRD data, each of which is inadequate, but which, when combined, creates a rich source of lexical information. Finally, Section 6 addresses larger questions of bilingual correspondences, related to information retrieval and text understanding. Evaluation and applications are suggested to users of such a system.

2. Bilingual Corpus-based Analysis

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. This paper extends earlier work (Klavans and Tzoukermann, 1989), (Klavans and Tzoukermann, 1990a), (Klavans and Tzoukermann, 1990b), in which we reported on a study of a selected sub-set of movement verbs in a bilingual corpus. The corpus consists of 85 million English words (3.5 million sentences) and 97 million French words (3.7 million sentences) from the Canadian Parliamentary Proceedings (the Hansard corpus). Of this, 75 million French and 69 million English words (2,869,041 sentences) have been aligned by sentence (Brown, Lai, and Mercer, 1991). Table 1 gives an example of two aligned sentences.

Among the information given for each file, which represents a separate session of parliament, is the following: speaker name, time, and language comments. The language comments indicate whether the language was French or English in the original, or whether there are sections in a language other than the original one, i.e. if there are French sentences or words inserted in English text, or vice versa.

The goal of this paper is to present the methodology used in the BICORD system; this methodology applies to any lexicon enhanced with corpus information, whether that lexicon is initially derived from a machine-readable dictionary or not. For this study, some representative verbs which have at least one movement sense were selected. For example, Figure 1 shows the verb *commute* in Webster's Seventh.[2]

Motion verbs were extracted from the dictionary based on their hypernyms. Thus, notice that in the intransitive verb *vi* part of speech, sense 3 (indicated in bold in Figure 1), the movement sense is revealed by the hypernym *travel*, itself one of the key indicators of movement. Other verbs with the hypernym *travel* are *barrel*, *bus*, *cannonball*, *coast*, *cruise*, *drift*, *itinerate*, *oscillate*, *peregrinate*, *sail*, *snowshoe*, *tramp*, *trek*, and *zip*. Additional movement verb indicators used for dictionary extraction include *move* and *go* as hypernyms. For example, *zoom*, sense 1, is defined as *to move with a loud low hum or buzz*, and *stagger*, sense 1b, is *to move on unsteadily.* As we explain below in section 5.1, monolingual dictionary data was used to form the initial set of linguistically relevant verbs. Precise details on the method used to collect the linguistic category of movement verbs is described in more detail in (Klavans, 1988).

We then compared the information found in the MRD's with the information found in the bilingual corpus. For example, for verbs like *commute*, discussed in more detail in (Klavans and Tzoukermann, 1989), which do not have a straightfor-

```
+-hdw: commute
|
+-superhom
  +-pos: vt
  +-homograph
  | +-senseid: 1a
  | +-pos: vt
  | +-definition
  | | +-defstring: to give in exchange
  |                 for another
  | +-definition
  |   +-synxref: EXCHANGE
  +-homograph
  | +-senseid: 1b
  | +-pos: vt
  | +-definition
  |   +-synxref: CHANGE
  |   +-synxref: ALTER
  +-homograph
  | +-senseid: 2
  | +-pos: vt
  | +-definition
  |   +-defstring: to convert (as a payment)
  |                 into another form
  +-homograph
  | +-senseid: 3
  | +-pos: vt
  | +-definition
  |   +-defstring: to exchange (a penalty)
  |                 for another less severe
+-homograph
| +-senseid: 4
| +-pos: vt
| +-definition
|   +-synxref: COMMUTATE
  +-pos: vi
+-homograph
| +-senseid: 1
| +-pos: vi
| +-definition
|   +-defstring: to make up for something
+-homograph
| +-senseid: 2
| +-pos: vi
| +-definition
|   +-defstring: to pay in gross
+-homograph
| +-senseid: 3
| +-pos: vi
| +-definition
|   +-defstring: to travel back and forth
|                 regularly
```

Figure 1. MRD entry for *commute* from Webster's Seventh

```
+-hdw: commute

  +-homograph
  | +-homnum: 1
  | +-pos: vt
  | +-sense
  |   +-translat
  |   | +-tran
  |   |   +-word: substituer
  |   |   +-complem
  |   |   | +-srcprep: for
  |   |   | +-srcprep: into
  |   |   | +-trgprep: a
  |   | +-tran
  |   | | +-word: interchanger
  |   | +-tran
  |   |   +-word: échanger
  |   |   +-complem
  |   |   | +-srcprep: for
  |   |   | +-srcprep: into
  |   |   | +-trgprep: pour
  |   |   | +-trgprep: contre
  |   |   | +-trgprep: avec
  |   +-translat
  |   | +-srcnote: Elec
  |   | +-tran
  |   |   +-word: commuer
  |   +-translat
  |   | +-srcnote: Jur
  |   | +-tran
  |   |   +-word: commuer
  |   |   +-complem
  |   |     +-srcprep: into
  |   |     +-trgprep: en
  |   +-collocat
  |     +-colsource
  |     | +-srcnote: Jur
  |     | +-source: commuted sentence
  |     +-coltarget
  |       +-targ
  |         +-target: sentence commuée
  | +-homnum: 2
  | +-pos: vi
  | +-sense
  |   +-translat
  |   | +-tran
  |   | | +-word: faire un /or/ le trajet journalier
  |   | +-tran
  |   |   +-word: faire la navette
  |   |   +-complem
  |   |   | +-srcprep: between
  |   |   | +-trgprep: entre
  |   |   +-complem
  |   |   | +-srcprep: from
  |   |   | +-trgprep: de
```

Figure 2. Partial MRD entry for *commute* from $CR - EF$

ward one-word translation, we found three types of translation: first, cases where most of the main components of the verb concept are present, as in 'se rendre au travail quotidiennement' meaning *to go/get to work on a daily basis*; second, cases where parts of the translation are found, as in 'faire le trajet' *make the trip* with the implied meaning of *back and forth*; and third, cases where a totally different verb from that given in the MRD occurs, such as 'parcourir' *to travel (all over)* or 'voyager' *to travel*. The dictionary definition of *commute* from the English-French side of CR is given in Figure 2.

Figure 2 shows the headword *commute* from the English-French portion of Collins-Robert dictionary. Homograph 1, the transitive verb sense, refers to the *substitute* sense, as shown in Table 2 from the Hansards.[3]

commute ≃ 'commuer'	
English:	The motion is silent on the royal prerogative of pardon, by which Cabinet can **commute the death penalty** to life imprisonment.
French:	La motion passe sous silence la prérogative royale de grâce en vertu de laquelle le cabinet peut **commuer la peine de mort** en emprisonnement à vie.
commute ≃ 'commuer'	
English:	Will the present Government or future Governments who are opposed to the death penalty be able **to commute such a sentence**, and if so, is not the present debate absolutely meaningless?
French:	Le gouvernement actuel ou des gouvernements futurs opposés à la peine de mort pourront-ils **commuer cette condamnation**? Dans l'affirmative, le présent débat n'est-il pas dénué de sens?
commute ≃ 'faire la navette'	
English:	Whether they **are commuting** to and from places of employment ...
French:	Qu'ils **fassent la navette** entre leur domicile et leur lieu de travail ...
commute ≃ 'banlieusards'	
English:	If we ... impose this tax on people who have no other way to get from where they live to where they work, or **people who commute**...
French:	Si l'on entend imposer cette taxe aux gens qui n'ont pas d'autre moyen de se rendre à leur travail, **les banlieusards**,...

Table 2. Sample citations of *commute* from the Hansard corpus

The algorithm which picks up movement senses did not select homograph 1 of *commute* since there are no indicators of movement, such as the hypernym *go, travel*, or prepositions such as *between, from* etc. Notice that the two translations of the movement sense of *commute* in Figure 2 accurately represent two translations of the commute concept, that is, *to travel back and forth (usually) on a daily basis*. The first translation expresses the daily concept via 'journalier' *daily* and the travel concept via 'trajet' *trip*. The second translation expresses the concept of going back and forth via 'la navette' meaning *shuttle*.

Some sample citations from the corpus, some of which reflect these dictionary translations, are given in Table 2.

In general, French correspondences to English movement verbs, such as *mosey, limp, drift, zoom* typically consist of a general motion verb 'entrer/sortir/aller/avancer' with an adverbial or prepositional modifier showing manner, e.g. 'nonchalamment', 'sans se presser', 'à toute vitesse', 'à la dérive', etc. As seen in these examples, English tends to incorporate the motion concept and manner into one verbal lexical item. The term incorporation is used to capture a common linguistic phenomenon when several underlying conceptual elements of meaning may occur "bundled" into one or more surface lexical items, i.e. incorporation captures the fact that there is often a complex mapping between underlying conceptual structure and the surface expression of these underlying elements. Similar to the case of incorporation of motion and manner in English is the incorporation of the cause concept in movement verbs in English, e.g the Webster's Seventh verb [*march* 4 vt (a) (*Mil*)] as in *to march troops*, is translated in *CR* as 'faire marcher (au pas) les troupes' or [*gallop* 3 vt *horse*] is translated as 'faire galoper', in the sense *cause the horse to gallop* or *make the horse gallop*. The linguistic basis for the approach we have taken in this project is grounded in the very notion of incorporation and in the well-documented linguistic fact that languages exhibit regular correspondences reflecting the surface expression of underlying conceptual elements. In the case of English and French, the fact that the concepts MOTION and MANNER tend to be expressed in French as two surface units, a general verb for MOTION and an expression or adverbial for MANNER, rather than as one unit, as in English (either a single verb or a phrasal verb), and that the general and basic motion verbs 'entrer/sortir/aller/avancer' tend to carry the MOTION concept to the surface, enables our approach to succeed. It is possible to identify the set of motion verbs both in English and French by drawing on this fact. In addition to the semi-automatic identification of the set of verbs in this class, it enables the discrimination of the motion sense from other non-motion senses, thus permitting further semi-automatic distinctions to be made in the lexical data base.

A comparison with an example from the use of surface morphological structure of English words for the identification of a category of linguistic elements might be useful in clarifying the theory behind our approach. In English, the suffix -*tion* (and its variants -*ation*, -*ion*, and so on) can be used to find the set of potential nominalizations. In this case, a surface spelling is used to identify a set of linguistic elements, which can then be used to find associations between related verbs and nouns.

Table 3. Sample correspondences of movement verbs in English and French from Collins-Robert

amble	aller (or) marcher d'un pas tranquille
canter	aller au petit galop
crawl in/out	entrer/sortir en rampant (or) à quatre pattes
dart in/out	arriver/partir comme une flèche
glide (person)	circuler à pas feutrés (or) comme en flottant
glide (vehicle)	s'avancer en douceur (or) silencieusement
hike	aller (or) marcher à pied
jog	faire du jogging
lope	courir en bondissant
lope in/out	entrer/sortir en bondissant
mosey along	aller (or) marcher sans se presser
ramble	se promener au hasard
shamble	marcher en traînant les pieds
stroll	se promener nonchâlamment
trudge in/out	entrer/sortir péniblement (or) en traînant les pieds
whiz	aller à toute vitesse en sifflant

Clearly, using the clue of spelling alone will overgeneralize, giving incorrect candidates such as *nation* or *ambition*. However, the principle is what is important to focus on. In the case of BICORD, surface verbs such as 'entrer/sortir/aller/avancer' are used to identify the set of potential movement verbs in French, and then that set is used to further identify and confirm corresponding English candidates in the MRDs. Certainly, the procedure has flaws, due to polysemy as discussed below, but it has proven quite successful. For example, Table 3 shows some examples of typical correspondences, all taken from CR, using a seed list generated automatically from the Webster's Seventh taxonym dictionary (Klavans, Chodorow, and Wacholder, 1990). As the table shows, the list of English verbs identified by this technique is strikingly accurate. Each verb is clearly a manner of movement verb, one which, as the theory would predict, incorporates the manner and motion concepts into one verb or phrasal verb, whereas the French elements are clearly distinct.

The phenomenon of incorporation in verbs of motion is well-studied in the linguistic literature, such as in (Talmy, 1985), (Jackendoff, 1987), (Levin and Rappaport, 1988), (Talmy, 1975), (Gruber, 1965). For example, manner of motion verbs such as *float* are in fact ambiguous. Consider example (1) from Carter (1988):

(1) The bottle will float under the bridge.

In one sense there is no goal, that is, the bottle is afloat for a while at a location under the bridge; in another interpretation, there is a goal interpretation, that is the bottle will float along some trajectory towards the bridge until it arrives at some location under the bridge. Tenny (1994) extends arguments of Jackendoff (1990) showing that only verbs describing a manner of motion can undergo this ambiguity, and that new coinages can be coerced towards this interpretation, e.g.

(2) Sue will skate-board under the bridge.

Table 4. Hansard citation for multi-word correspondences

English:	Our young nation now **marches forward** into the 21st century, a century that will belong to Canada.
French:	Notre jeune pays se **dirige** maintenant **d'un pas assuré vers** le XXIème siècle, un siècle qui appartiendra au Canada.

Sentence (2) exhibits the same ambiguity as sentence (1). In earlier work, Tenny (1992) argued that manner of motion verbs are distinct from two other major verb classes, change-of-state and incremental theme verbs, in the way they are formed; motion verbs typically add arguments whereas the other classes typically add predicates. Tenny is concerned with the interaction of manner of motion verbs with aspectual structure, but her point about the properties of the motion verb class are relevant to our choice of motion verbs as a cohesive and significant category, both from a lexical conceptual and practical point of view. Tenny (1994) argues that motion is a fundamental linguistic notion, a semantic primitive necessary for the correct representation of a set of verbs and necessary for predicting verb interpretation and alternations.

These kinds of linguistically complex multi-word correspondences often cause problems in the lexical transfer component of machine translation systems because the correspondence is not simply one to one. In the case of movement verbs, there are a range of possible translations using adverbials, prepositional phrases, and adjuncts that can occur at random distance from the head verb. Table 4 shows an example from the corpus where the English *march forward* corresponds to the French 'se dirige ... d'un pas assuré', meaning literally *to direct oneself ... with a confident or sure step.* Notice that the English verb *march* has no direct translation, nor does *forward*, the adverbial particle associated with the movement verb. Similarly, the French verb 'se diriger' has no literal correspondence in the English *march*. The concept 'd'un pas assuré' is spelled out explicitly in French in a manner prepositional phrase, whereas this concept is incorporated in the English *march forward*.

Examples like this illustrate some of the subtleties and difficulties in constructing lexical correspondences for movement verbs.

3. Theoretical Grounds for Motion Verb Extraction: a Decompositional Approach

A diagnostic test for motion verbs can be defined as follows: a motion verb can be used in the frame "VERBing is a way of moving", enhanced with directional, manner, or other adverbials, e.g. "VERBing is a way of moving from X to Y" or "VERBing is a way of moving at *n* speed", and so on (Cruse, 1986). For example, verbs which qualify according to these criteria include *walk*, *limp*, and *zoom*, as in:

(3) Walking is a way of moving from x to y [in a certain manner]....

(4) Limping is a way of moving from x to y [in a certain manner]. . . .
(5) Zooming is a way of moving from x to y [at a certain speed]. . .

On the other hand, verbs such as the following do not, and should not, qualify in their literal senses:

(6) * Belonging is a way of moving from x to y.
(7) * Arguing is a way of moving from x to y.
(8) * Baking is a way of moving from x to y.

We take a decompositional approach to the underlying structure of movement verbs (Dowty, 1979), (Talmy, 1985), (Cruse, 1986), (Jackendoff, 1987), (Levin and Rappaport, 1988). Theoretically, the position is based on the claim that word meanings, i.e. the concepts that words label, are constructed from semantic components that recur in the meanings of different words. Words are clearly semantically related to one another in systematic ways, as revealed by tests of entailment and inference. Thus, words are not considered to be unanalyzed atoms but can be decomposed into a set of recurrent conceptual features.

Two decompositional analyses of lexical items place particular focus on the motion concepts, (Dowty, 1979) and (Jackendoff, 1987). Without reproducing the full analyses here, Dowty (1979) proposed three aspectual operators DO, BECOME, and CAUSE and suggests that verbs are defined from basic stative predicates in terms of these operators. The operator DO is analyzed by Dowty as a binary relation between individuals and properties, as for movement in:

(9) DO(*j*,MOTION)

This is to be interpreted that something that *j* does causes *j* to be in motion. Thus, MOTION is an activity created from states via the DO relation. For the purposes of this paper and regardless of later revisions to initial claims, the important point is that MOTION is a basic relation, one which it is expected will be relatively pervasive in the dictionary of a language, in several surface lexical items.

Jackendoff (1987) proposed that the set of conceptual primitives includes such entities as THING (or OBJECT), EVENT, STATE, ACTION, PLACE, PATH, PROPERTY, and AMOUNT. Such primitives can be expanded into more complex expressions. Of the basic primitives for movement verbs are EVENT and PATH:

$$\text{EVENT} \rightarrow \left\{ \begin{array}{l} [_{\text{Event}}\,\text{GO(THING, PATH)}] \\ [_{\text{Event}}\,\text{STAY(THING, PLACE)}] \end{array} \right\}$$

$$\text{PATH} \rightarrow \left[{}_{\text{Path}} \left\{ \begin{array}{l} \text{TO} \\ \text{FROM} \\ \text{TOWARD} \\ \text{AWAY-FROM} \\ \text{VIA} \end{array} \right\} \left(\left[\left\{ \begin{array}{l} \text{THING} \\ \text{PLACE} \end{array} \right\} \right] \right) \right]$$

Figure 3. Basic primitives for movement verbs (Jackendoff 1987)

Thus the EVENT function can be expanded into either of the two Event-functions GO and STAY. The arguments of GO reflects the THING in motion and the PATH it traverses. Rules such as the PATH expansion in Table 3 define the conceptual structure that underlies meaning of sentences, such as:

(10) John ran into the room.

$$\text{EVENT} \rightarrow [_{\text{Event}}\text{GO}\,([_{\text{Thing}}\text{JOHN}]\,,\;\; [_{\text{Path}}\text{TO}\;\;([_{\text{Place}}\text{IN}\;\;([_{\text{Thing}}\text{ROOM}])])])]$$

Jackendoff (1990) uses a MOVE function to analyze sentences which describe an object's motion without implying a measure or terminus, for verbs like *wiggle*, *dance*, *spin*, *wave*:

(11) Willy wiggled.
(12) Debbi danced.
(13) The top spun.

For verbs implying a PATH, there is both a MOVE and GO function. Jackendoff shows how certain verbs incorporate into their meaning the PATH and PLACE functions, e.g. *enter* denotes *going into somewhere*; he provides some proposals on linking rules between conceptual structure, argument structure, and thematic roles, demonstrating that a fine-grained compositional approach to meaning provides answers to theoretical issues in understanding thematic relations in linguistic theory. The relationship between the theory and our application was discussed on section 2, namely that the ability to mine the MRDs for the set of motion verbs and the ability to distinguish movement from non-movement senses of polysemous verbs rests on the theoretical underpinnings that determines the underlying structure of primitives. An understanding of this theory has driven this research, and has enabled the structuring of lexical entries in the enhanced lexical data base.

Several approaches to machine translation have built on Jackendoff's primitives, as well as on (Hale and Keyser, 1986), (Levin and Rappaport, 1988), and others. For example, Dorr (1992) demonstrates that the decompositional approach in machine translation permits the definition of a recursive translation mapping which treats verbs and their arguments as compositional units. Thus, the underlying decompositional properties of a verb, along with the underlying decompositional features of arguments can be considered during each step of translation. Dorr analyzes translation *divergences*, i.e. cases where the natural translation of one language into another results in a very different form than that of the original. For example, a case of *conflational* divergence (studied by Talmy (1975), Talmy (1985)) is found in the English-Spanish translation:

(14) I stabbed John $\Leftrightarrow$: Yo le di puñaladas a Juan.
'I gave knife-wounds to John'

In this example, Dorr shows how necessary components of meaning of the action are conflated in one language, in this case English, but occur as separate components in another language, in this example Spanish. Thus English uses a single word *stab* for the two Spanish words 'dar' *give* and 'puñaladas' *knife-wounds*. The *knife-wounds* component of meaning is not overtly realized in English but is incorporated into the verb *stab*.

This example is analogous to the movement verb cases where one verb in English incorporates the motion and aspects of motion, such as speed, manner, or direction, whereas in French these meaning components occur as separate words and phrases. Dorr shows that such examples are translated naturally by a compositional approach, which readily lends itself to the analysis of components which may or may not be explicitly realized in the surface lexical item. Whereas Dorr argues for an interlingual approach to machine translation based on the compositional analysis of meaning, the task we have addressed has a somewhat different focus, that of constructing a lexicon based on the compositional approach to meaning. We utilize componential analysis as the basis for our exploitation of the MRD data, and we then build on this data using the compositional approach.

The question of how fine-grained a decompositional meaning analysis should go is one that continues to puzzle linguists and computational linguists alike; a full discussion of the formal representation of lexical meaning is beyond the scope of this paper. However, from an empirical point of view, we were able to extract observations from decompositional approaches to movement verbs, and to use those generalizations to mine the machine-readable dictionaries for related verbs. We build on theoretical arguments about the compositional nature of movement verbs, and extract some performance criteria for creating and expanding our initial verb set.

We suspected that the decompositional nature of movement verbs coupled with the linguistic observations of Talmy (1985) and others on the cross-linguistic variations between verbs of motion in languages like English and French or Spanish would have an important empirical consequence, namely that such concepts would necessarily be part of the dictionary definitions of these verbs. Although we would not go so far as to claim that the lexicographer's goal is to represent word meaning in terms of the underlying conceptual features, our hypothesis was that such features might naturally appear as part of the individuation of lexical items. This hypothesis turned out to be correct, and enabled us to initiate our exploration of the set of movement verbs in several MRD's.

4. Statistical and symbolic approaches in building lexicons

Combining linguistic and statistical methods is becoming increasingly common in computational linguistics, especially as more corpora become available (Klavans and Resnik, 1996). In the monolingual lexicon construction literature, e.g. (Church and Hanks, 1990), (Calzolari and Bindi, 1990), (Brent, 1993), (Klavans and Resnik, 1996), purely statistical procedures have been shown to achieve fairly solid results. However, there is now a movement in the direction of incorporating prior linguistic knowledge for the monolingual lexicon, as discussed in (Pustejovsky, Bergler, and Anick, 1993) demonstrating ways to combine MRD and corpus data for the acquisition of lexical knowledge.

In the arena of automatic bilingual lexicon construction, early statistical approaches (Brown et al., 1988), (Brown et al., 1990), argued for exclusively statistical non-linguistic methods to induce translations, although later developments reveal an increasing need for lexicographic and linguistic knowledge (Brown et al., 1993). For example, Catizone, Russell, and Warwick (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. Church (1993) uses a different technique to align texts ultimately for extracting correspondences, again a purely statistical, although effective, approach for the text. In contrast, Sadler (1989) proposes parsing aligned corpora into dependency trees, which form the structures upon which lexical correspondences are suggested to the user. Pustejovsky, Bergler, and Anick (1993) discusses ways to extract semantic information form MRD's and combine it with statistical data to acquire lexical knowledge needed for applications such as sublanguage lexicons.

Kay and Röscheisen (1993) discuss the problem of multi-word correspondences in terms of developing statistical algorithms for text-translation alignment. As with other statistical approaches, Kay and Röscheisen (1993) note about their work that "the present algorithm rests on being able to identify one-to-one associations between certain words, notably technical terms and proper names.... The most interesting further developments would be in the direction of loosening up this dependence on one-to-one associations ...because this would present a very significant challenge...." Although such limitations are inherent in a word-to-word based system, this discussion clarifies precisely the type of problem to arise. To continue, Kay and Röscheisen (1993) consider the possibility of extending their methodology to handle one-to-many and many-to-many associations. They claim that their methods can at least capture latent information in one-to-many associations in text, but they note the serious combinatorial problems which would result. What is required is serious modification to the indexing method used so that the approach would be adequately efficient.

Smadja, McKeown, and Hatzivassiloglou (in press) discuss the many-to-many mapping problem in terms of different types of collocations across languages. Their system, Champollion, endeavors to overcome the multi-word translation problem by searching for significant collocations in one language of an aligned bilingual corpus, and then producing candidate translations. Such a system would be a candidate to enhance BICORD, with the exception that the initial set of collocations would be derived from a machine-readable dictionary, rather than from the noisy data resulting from the statistical processing of text. After using the clean, but incomplete, data from dictionaries, the remaining corpus could be processed to pick up whatever collocational information remained.

5. The BICORD System - Description of the Algorithm

Our approach involves a combination of standard linguistic methodology using MRDs, enhanced with some statistical techniques. MRDs have served as useful resources in many natural language applications. Although they are limited in size and in internal cohesion (see (Atkins, Kegl, and Levin, 1988)), nonetheless the structure imposed by lexicographers has proven to be useful for a variety of purposes. For example, Byrd et al. (1987) show how dictionaries can be parsed and used to extract semantic and syntactic knowledge; Boguraev (1991) shows how the semantic and subject codes of the Longman Dictionary of Contemporary English can be used to create syntactic frames for a GPSG lexicon; Boguraev et al. (1989) show how both bilingual and monolingual dictionaries provide extensive implicit knowledge; Klavans, Chodorow, and Wacholder (1990) illustrate the building of a knowledge base, with different semantic links, from Webster's Seventh and from Longman's.

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. As for the latter application, i.e. enhancing a statistical system with data from the MRD, new approaches to self-organizing systems are beginning to take into account MRD data to set values or to alter values in computation (Brown et al., 1993). The former 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 can be ordered by frequency, orderings which can easily be updated depending on the sublanguage corpus. The enhanced MRD would be more complete in retaining the original structure and content from the lexicographer's expertise, by containing correspondences not found in the original dictionary, and in suggesting statistically probable translations in context.

The following figure gives a schematic overview of the BICORD system.[4]

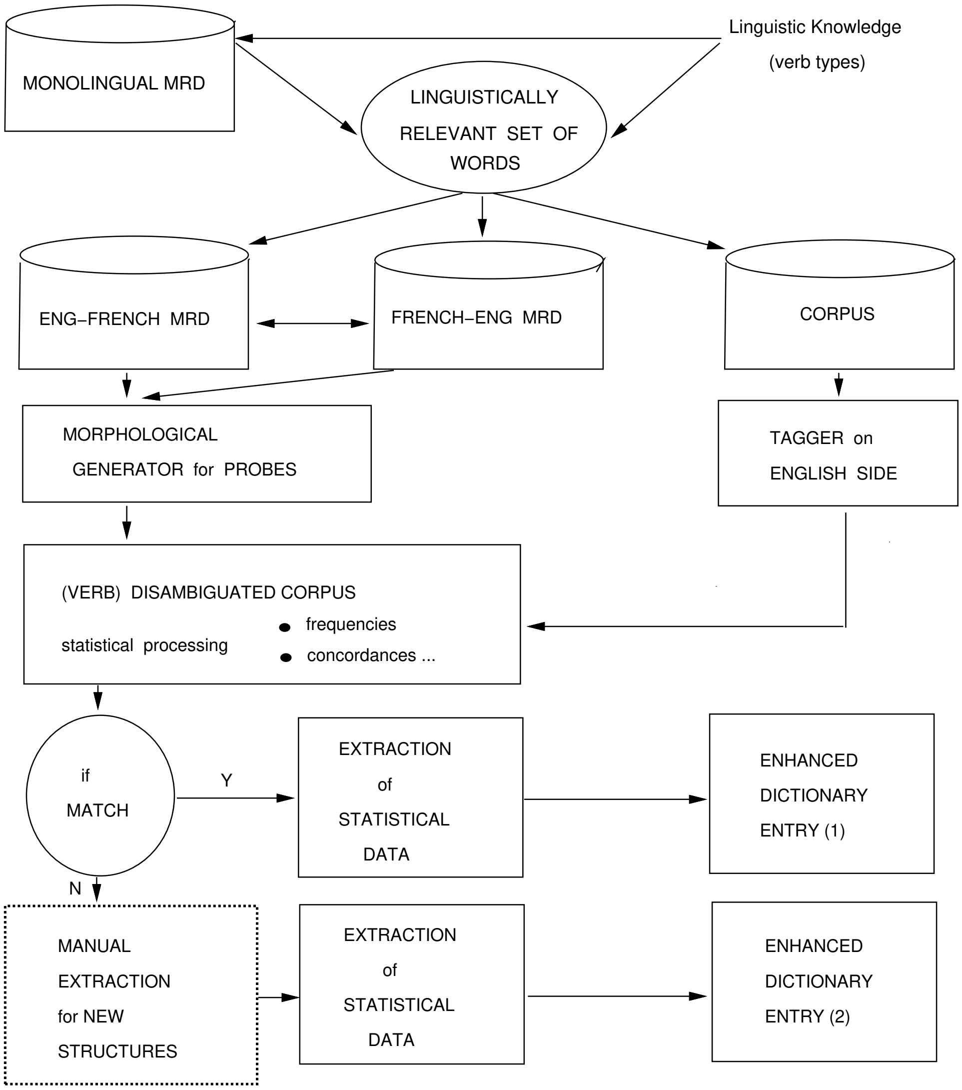

Figure 4. The BICORD system

This section gives a step by step discussion of the modules of the system; each module functions independently and thus could be used similarly independently. The input and output of the system can be viewed as follows:

Input: let $\mathcal{D}$ be the set of all dictionaries and $\mathcal{C}$ the set of all corpora. Let $\mathcal{D}_{\mathcal{EF}}$ be an English-French bilingual dictionary, and $\mathcal{D}_{\mathcal{FE}}$ be the corresponding French-English side of a dictionary, in this case CR (Atkins, Duval, and Milne, 1978). Let $\mathcal{C}_{\mathcal{H}}$ be the Hansard corpus, described above. Let W be the set of words of a language L. A subset of W, in this case $W_{V(motion)}$ for movement verbs was chosen to search over $\mathcal{C}_{\mathcal{H}}$.

Output/Goal: a bilingual corpus-enhanced dictionary.

The algorithm works as described in the next section.

5.1. Step 1: Define the Test Set

5.1.1. Select Initial Test Set: A decompositional approach

Since the goal of this research was to test one-to-many and many-to-many mappings between French and English, and since the decision was made to focus on the set of movement verbs, the first task consisted of identifying the set of verbs themselves. The criteria used to ensure that a verb is a member of this semantic class were set out in Section 3, based on Cruse (1986).

An initial list was derived from the taxonym dictionary (Chodorow, Byrd, and Heidorn, 1985), (Klavans, Chodorow, and Wacholder, 1990), created from Webster's Seventh (Gove, 1963), using techniques based on the theory of decompositionality of movement verbs described in Section 3.[5] Verb senses with "move" as the hypernym or superordinate were extracted from Webster's Seventh Dictionary, as were verbs with the extracted hyponyms as hypernyms. For example, in the following, the hypernym of the definition is *move.*

(15) **drift** vi 1b: to **move** or float smoothly and effortlessly
(16) **walk** vi 2a: to **move** along on foot: advance by steps
(17) **zoom** 1: to **move** with a loud low hum or buzz

We then considered verbs with *drift*, *walk*, etc. as the hypernym and expanded the potential set, as in:

(18) **float** vi 2a: to **drift** on or through or as if on or through a liquid
(19) **limp** vi 1a: to **walk** lamely
(20) **mince** vi 3b: to **walk** with short steps in a prim affected manner
(21) **stroll** vi 1: to **walk** in a leisurely or idle manner

This gave a list of about five hundred (500) verbs, most of which were clearly movement verbs in all senses. Further details on the initial set of lexical primitives, and the way these primitives were derived is more fully explained in (Klavans, 1988) and in Section 3.

5.1.2. Refine the Test Set

One of the most serious ongoing problems in any corpus analysis is polysemy. The power of the dictionary lies in the fact that senses are tagged, but the sense-tagging is diluted once the headword is extracted and used without regard to sense, a problem which has been addressed in Klavans, Chodorow, and Wacholder (1990). For example, although *run* is certainly a movement verb in some senses from Webster's Seventh, as shown below, it is not a movement verb in all senses. The verb *run* as in (22), (23), and (24) passes the test for movement verbs as discussed in Section 3:

(22) **run** vi 1$1a: to go faster than a walk
(23) **run** vi 1$1b: (of a horse) to move at a fast gallop
(24) **run** vi 1$3a: to go rapidly or hurredly

However, in the following usages, the verb *run* does not pass the test:

(25) **run** vi 1$9a: to continue in force or operation
(26) **run** vi 1$13a: to lie in or take a certain direction
(27) **run** vi 1$14a: to occur persistently

The sense used in (25) is the one used in sentences like *the contract has two more years to run*, the sense in (26) is *the boundary line runs East*; in (27) the sense is the one appropriate to usages such as *musical talent seems to run in his family.* None of these senses indicates movement from place to place, as found in (22), (23), or (24).

It could be argued that the nature of the hypernyms in (22)-(24), *go*, *move*, and *go* respectively, is inherently different from the hypernym type in (25)-(27), *continue*, *lie (in)*, and *occur*. The fact that each of the hypernyms in the first set is essentially a movement-type verb could be suggestive that the defined word has a movement sense. Indeed, in our own filtering of dictionary senses, we have used less ambiguous cases of movement verb hypernyms to identify movement senses. However, hypernyms such as *continue*, *lie (in)*, and *occur* could just as easily be used for movement senses. The following examples are from Webster's Seventh:

(28) **carry off** vi 1$2: to continue one's course or activity in spite of hindrance or discouragement
(29) **follow through** vi 1$1: to continue a stroke or motion to the end of its arc
(30) **scatter** vb 1$4: to occur or fall irregularly or at random

As these examples show, a movement verb sense, such as for *carry off*, *follow through*, or *scatter* in (28)-(30), can have the same hypernym as for the non-movement senses of *run*, given in (25)-(27).

As a consequence, to assume that a verb with a movement sense will, as a hypernym, necessarily reveal other movement verb hyponyms could result in dangerous overgeneralizations. At the same time, to use very high frequency verbs to drive the system would increase the chances of failure considerably, despite the superficial advantage of coverage, since it thereby opens the door for extensive polysemy which would degrade results. The criteria for degree of polysemy was based on dictionary structure, since it is an obvious fact that higher frequency content words tend to be the most polysemous as reflected in number of dictionary senses (Atkins, 1987). For this reason, the set was reduced to extract a small test set of representative verbs of medium frequency since they are assumed to be of limited polysemy. For example, compare *run*, *walk*, *commute*, and *zoom*. In Webster's Seventh, *run* has 3 homographs: 15 senses for the verb, with an additional 10 senses for verb-particle constructions; 11 senses for the noun; and 3 for the adjective usage, giving a total of 33 senses across homographs. In comparison, a verb like *zoom* has 2 homographs and a total of 4 senses. In the tagged version of the Wall Street Journal representing 61.8 million tokens, *run* as a verb occurs 7365 times, whereas *zoom* as a verb occurs 19 times. Thus, assuming that the number of dictionary senses naturally reflects polysemy and frequency, verbs with between two and twenty-five dictionary senses were chosen, as a rough measure of polysemy. Examples from the initial list are:

`ascend`	`drift`
`circle`	`emigrate`
`commute`	`glide`
`dance`	`immigrate`
`descend`	`sail`

5.2. Step 2: Search through $\mathcal{D}_{\mathcal{EF}}$ and $\mathcal{D}_{\mathcal{FE}}$

The selected lexical items were used to search in $\mathcal{D}_{\mathcal{EF}}$, both for translations and collocations under the entry itself. To expand the list of translations, we then searched for the French headwords from in $\mathcal{D}_{\mathcal{FE}}$ with the English words in the translation field. For example, given the English movement verb *dance*, extracted initially from the monolingual English dictionary as shown in Figure 5, two sets of translations were extracted. One is the obvious set of translations found for *dance* in $\mathcal{D}_{\mathcal{EF}}$. In addition, we used the query capabilities of the Lexical data base Query Language (LQL) (Neff, Byrd, and Rizk, 1988) to search for all headwords in $\mathcal{D}_{\mathcal{FE}}$ that had *dance* in the translation field. Translations and collocations were abstracted automatically from the parsed version of $\mathcal{D}_{\mathcal{EF}}$ and $\mathcal{D}_{\mathcal{FE}}$ (Neff and Boguraev, 1989) using LQL (Neff, Byrd, and Rizk, 1988). To illustrate, Figure 5 shows a partial entry for *dance* from $\mathcal{D}_{\mathcal{EF}}$, where the dots represent elided dictionary material. Translations found in $\mathcal{D}_{\mathcal{EF}}$ were 'danser', 'entrer/sortir etc. joyeusement', 'gambader', 'sautiller'. These translations were then used for building the search list.

To expand the search list, Figure 6 shows a partial entry from $\mathcal{D}_{\mathcal{FE}}$ with *dance* as a translation. The query here searched for a French headword with the value of the part of speech (pos) attribute as vi or vt (intransitive or transitive verb) which had *dance* in the translation field.

As shown in Figure 6, the French verb 'gambiller' was found to have *dance* as a translation. However, 'gambiller' was not given as a translation of *dance* under the entry for *dance*. Bilingual dictionaries are known to be assymetric as shown by Rizk (1989), and we took advantage of this assymetry to create a more comprehensive search list.

```
+-hdw: dance
+-superhom
  | . . .
  +-homograph
    +-homnum: 2
    +-pos: vt
    +-sense
      +-translat
      | +-argument: waltz etc.
      | +-word: danser
  | . . .
  +-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 5. Partial MRD entry for *dance* from $\mathcal{D}_{\mathcal{EF}}$

```
+-hdw: gambiller
  |
  +-superhom
  | . . .
  +-homograph
  | +-pos: vi
  | +-sense
  |    +-translat
  |       +-word: to dance
  |       +-word: jig
```

Figure 6. Partial MRD entry with *dance* as a translation from $\mathcal{D}_{\mathcal{FE}}$

Table 5. Sample Citation for *drift*

English:	We have been **drifting** along and acting as though everything is going okay with the environment while the environmental crisis is mounting in Canada and around the world.
French:	On **laisse les choses aller à vau-l'eau** et on agit comme si tout allait bien dans ce domaine alors que la crise de l'environnement empire au Canada et partout dans le monde.
English:	You cannot continue to allow this to **drift**, Mr. speaker.
French:	On ne peut continuer à **laisser les choses aller à la dérive**, monsieur le président.
English:	Should we ever lose the Province of Quebec then Canada will be that much poorer and we will have a **tendency to drift** in the direction the U.S. has taken.
French:	Si le Canada perdait le Québec, il en serait gravement appauvri et il **risquerait de s'engager** dans la même voie que les Etats-Unis.

5.3. Step 3: Processing Citations from $\mathcal{C}_{\mathcal{H}}$

5.3.1. *Extract Relevant Citations from $\mathcal{C}_{\mathcal{H}}$*

Relevant citations were extracted from the corpus taking the basic list described in 5.1 and expanded by dictionary probing as described in 5.2. A simple morphological system was used to expand baseforms into inflected forms to query the corpus. To illustrate, consider again the verb *dance*. The character strings in the translation and collocation fields were extracted from $\mathcal{D}_{\mathcal{FE}}$ and $\mathcal{D}_{\mathcal{EF}}$. These strings were filtered to remove function words and some common words, such as 'faire' (*to make* or *do*), and morphological variants were generated. Some examples for 'danser' and 'gambader' *dance* are 'danser/dansa/dansera ...', 'gambader/gambadons ....'.

A text was extracted, consisting of the set of English citations from $\mathcal{C}_{\mathcal{H}}$ containing the probe string in any morphological form and the corresponding French sentence. There was a maximum of 1146 citations for any medium-frequency movement verb. For example, Table 5 gives several examples of aligned citations for *drift*. Notice that in the first citation, *drift* corresponds to the phrase 'laisse les choses aller', which more literally corresponds to *let things go along* or *to let things slide*. This phrase is modified by the manner propositional phrase 'à vau-l'eau', a somewhat obsolete phrase meaning to go *with the current*. In the second citation for *drift*, the correspondence includes the literal translation for *drift*, namely 'dériver', but in a different syntactic configuration. That is, instead of the direct translation, *to drift* is translated as 'laisser les choses aller à la dérive', *to let things go adrift* (literally *"go to the drift"*). In the third example, the phrase *have a tendency to drift in the direction* is translated with the expression 'il risquerait de s'engager dans la même voie', literally *there is a risk of embarking on the same path.*

Table 6. Tagged English Text

Citation #1:	we	are	dancing	upon	eggshells...	
Tagger output:	PP*S	VBR*	VVG1*	I*	N*2	
Citation #2:	the	politician	who	liked	to	dance...
Tagger output:	AT*	N*1	P*Q	VVPAST*	TO*	VVI*
Citation #3:	...Russian	people	dancing	rather	than	fighting.
Tagger output:	J*	N*1	VVG1*	R*R	I*	VVG1*

5.3.2. Tagging Text with a Statistical Tagger

A statistical tagger (Tzoukermann and Merialdo, 1989) and (Merialdo, 1994), was used to assign part of speech to the English side of the relevant corpus. In this way, words with a verbal part-of-speech were disambiguated from other parts-of-speech. The tagger used to preprocess the corpus was trained on one million words and based on a trigram model. The model estimates the parameters of lexical probabilities $P(w_i|t_i)$ (probability of a word given its tag) and contextual ones $P(t_i|t_{i-2}, t_{i-1})$ (probability of a tag given the two previous ones). The model was trained using Maximum Likelihood tagging which chooses the most likely tag for each word in the sentence; this was performed by using the Forward Backward algorithm which, when applied iteratively, improves the tagged output. The tagger utilizes a set of tags, which is a reduction of the Lancaster Treebank (Leech, Garside, and Atwell, 1983) tagset. From 159 original tags from the Treebank set, 76 tags were used for the present tagger. By random sampling, the error rate for part of speech tagging was determined to be about 3%. This constitutes the first step in disambiguation, enabling lexical correspondences. Some illustrative fragments for *dance* are given in Table 6.

Notice that part-of-speech tagging was applied only to the English side of the corpus to cull out verbal usages. This was done for two reasons: first, at that time, we had no tagger for French, but we did have one for English. More importantly, for this particular set of verbs, using a tagger to eliminate non-verb usages from the French side would be inapplicable. This is due to the fact that motion verbs in English carry the meaning of the concept expressed in non-verbal adjuncts in French. Therefore, the query on French motion verbs would result in a set of very general verbs, such as *entrer, sortir*, etc., rather than specific movement verbs such as for English *gallop*, *limp*, or *zoom*.

Efficient tagging has become a topic of intense interest recently in the computational linguistics community (Church, 1989), (DeRose, 1988), (Kupiec, 1989), (Brill, 1992). Tagging is both a preliminary step to effective parsing, and, at the same time, is an end in itself since correct part of speech assignment is a particularly challenging problem in a morphologically limited language like English. The use of part of speech assignment as a stand alone module permits a variety of filters on text. This project demonstrates one possible usage of part of speech disambiguation. Of course, a powerful tool for further disambiguation and structural analysis might be to parse the text in its entirety. This would entail some changes in our approach given the increase in information due to parsing, but might possibly increase the amount of data which could be accurately extracted from the corpus and incorporated into the lexical data base.

5.4. Step 4: Create Lexical Data

Data from $\mathcal{D}$ are utilized to drive our first pass at linking and filtering pairs common to both data resources. Citations that have lexical correspondences already provided by the MRD are extracted from the probe corpus.

5.4.1. First Pass Output

An extended lexicon was then built, using the structure already provided by the bilingual dictionary (CR), where the frequencies are computed over these matches. An example of a partially enhanced entry for *dance* is found in Figure 7. The figure shows the structured dictionary entry for *dance* with the addition of corpus information. Notice that dictionary nodes are now identified with a prefix "**d_**", and corpus derived nodes with "**c_**". Thus, the two information sources and the scope of the information is clearly distinguished by the specified source prefix. The figure also shows that the inflectional forms of "danser" as a verb occur with the following frequencies: "danser" in the infinitive at .44; in the past form at .17; and in the future tense, at .05. Notice that the sum total of the frequency does not equal 1; this is due to the fact that only 70% of the corpus citations have been placed into the dictionary lexical structure. Of course, with iteration and refinements as shown below, this number increases. New information is placed at the relevant node, low in the tree if there is no ambiguity of attachment or scope, and, if necessary, higher in the tree until evidence is found to permit the information to be moved down in the structure. In Figure 7, the information under the new homograph number "c_2,c_3" applies to the definitions "d_2" and "d_3". The "c_2,c_3" value can be viewed as an expression of the fact that this information applies to both verb homographs "d_2" and "d_3". Indeed, if the corpus data were unambiguously relevant to "d_2" or "d_3", it would be found under that node itself. That is, an additional node was added to the MRD structure to insert information about 'danser' at the highest level. Since the transitivity of a verb cannot currently be reliably determined automatically from tagged text, there is no evidence to motivate placement elsewhere. Thus, the data is inserted high in the tree, at the homograph level. However, if the text were parsed or if heuristics were applied, then it might be known whether the usage were transitive or intransitive, and thus whether these usages could be placed correctly under homograph "d_2", the transitive sense or homograph "d_3", the intransitive sense. Additionally, in homograph "d_3", *dance around* is found in the corpus, translated by "gambader" with a frequency of .02, and translated by "sautiller" with a frequency of .02. This is indicated by the new node c_collocat, shown in bold in Figure 7 under "gambader", and in the new c_collocat node under "sautiller" in this enhanced lexical data structure.

Each correspondence that matched one of the MRD probes was counted, stored, and removed from the probe corpus. For example, of the 109 citations of *dance* as a verb, 52 sentences matched the MRD correspondences. The remaining data from the corpus can then be iteratively gathered and used to create an even fuller entry, such as shown in Figure 8. As shown in the figure, the amount of information that has been inserted at the different nodes now sums up to 89%, which is substantially more than the 70% of Figure 7.

```
+-hdw: dance
+-superhom
   . . .
   +-homograph
     +-homnum: c_2, c_3
     +-pos: v
     +-sense
       +-c_ translat
        +-word: danser
          +-inflect: inf
          +-freq: .44
        +-word: danser
          +-inflect: past
          +-freq: .17
        +-word: danser
          +-inflect: fut
          +-freq: .05
          . . .
   | . . .
   +-homograph
   | +-homnum: d_2
   | +-pos: vt
   | +-sense
       +-d_ translat
       | +-argument: waltz etc.
       | +-word: danser
   | . . .
   +-homograph
   | +-homnum: d_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  : .02
           |
       | +-target: sautiller
           | +-c_ collocat
           |   +-source: to dance round
           |   +-inflect: past
           |   +-freq  : .02
       |
       +-d_ collocat
       | +-source: the child danced away /or/ off
       | +-target: l'enfant s'est éloigné
       |            en gambadant /or/ en sautillant
       | . . .
```

Figure 7. Partial Enhanced Entry

```
+-hdw: dance
|
+-superhom
  | ...
  +-homograph
  | +-homnum: c_2, c_3
  | +-pos: v
  | +-sense
  |    +-c_translat
  |    | +-word: danser
  |    | | +-inflect: inf
  |    | | +-freq: .44
  |    | +-word: danser
  |    | | +-inflect: past
  |    | | +-freq: .17
  |    | +-word: danser
  |    | | +-inflect: fut
  |    | | +-freq: .05
  +-homograph
  | +-homnum: d_2
  | +-pos: vt
  | |
  | +-sense
  |    |
  |    +-d_ translat
  |    | +-argument: waltz etc.
  |    | +-word: danser
  | ...
  +-homograph
  | +-homnum: d_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   : .02
       |   |
       | +-target: sautiller
       |   | +-c_ collocat
       |   |   +-source: to dance round
       |   |   +-inflect: past
       |   |   +-freq   : .02
       |
       +-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   : .08
         +-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 8. Fuller Enhanced Entry

In the case of the English and French verbs *dance*/'danser', vt or vi is a pure syntactic difference. In fact, *dance*/'danser' permits indefinite object deletion, as in:

(31) The boy *danced* the waltz. 'Le garçon dansait la valse'.
(32) The boy *danced.* 'Le garçon dansait'.

although *dance*/'danser' does not participate in transitivity alternations:

(33) * The waltz *danced.* * 'La valse dansait'.

In other cases, the difference between the transitive and intransitive surface structure can mask deep semantic difference. For example, a verb like *gallop* does participate in transitivity alternations; in the following example, *gallop* in (35) has a different interpretation than in (34). In (34), the girl is causing the horse to gallop across the field. Sentence (35) is ambiguous; either the object is deleted from (34) and the girl is still on the horse, or the girl is herself galloping across the field. In (36), the object of the causative in (34), namely *horse*, is in movement, i.e. *the horse galloped.*

(34) The girl galloped the horse across the fields.
(35) The girl galloped across the fields.
(36) The horse galloped.

Furthermore, in the transitive example (34), the subject is the AGENT of the action, and the object is the THEME. Examples (35) and (36) are both intransitive, but they differ in terms of the linking between grammatical relations and grammatical function. In (35), the subject is the AGENT, whereas as in (36), the subject is the THEME. The relationship between the transitive (34) and the intransitive (35) is that the surface object is deleted, namely *horse*, but the subject remains the AGENT. The relationship between the transitive (34) and the intransitive (36) is that of a transitivity alternation; the THEME moves from surface object to surface subject position. Indications of transitivity alternations are not always clear in MRDs, as discussed in (Atkins, Kegl, and Levin, 1988) and (Neff and Boguraev, 1989). In contrast, all senses of verbs such as 'gambader' *to leap* and 'sautiller' *to hop* are intransitive (as determined by a look-up in $\mathcal{D}$), so frequencies can be automatically placed under the associated homograph three, under the assumption that the dictionary has represented the senses accurately. Notice also that corpus derived information is placed under the relevant d_collocat node for 'gambader' and 'sautiller' since these are cases where matches occurred on the target term, but the source is different.

5.4.2. Second Pass Output

Further analysis of the remaining probe corpus is pursued by observing cooccurrences both over tags and lexical items. For example, with *dance*, looking at imme-

Table 7. Right context over tags

VERB	CATEGORY	%
dance	prep	77.829
dance	other	22.171

diate right context over tags reveals verb-preposition patterns, as shown in Table 7:

Considering the lexical items that correspond to those tags, the majority of these cases are for the preposition *to*. Taking into account cooccurrences over a larger window of five words, idioms are revealed like *dance to ... tune*, which is not found in either $\mathcal{D}_{\mathcal{FE}}$ or $\mathcal{D}_{\mathcal{EF}}$, 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 as shown in (Church et al., 1991). Assisted by such evidence, a new set of collocations can be inserted in the lexicon; a fuller entry for *dance* is shown as follows in Figure 8.

6. Research Questions

The transfer of the information from the corpus to a structured lexicon raises two kinds of problems, some of which are common to other areas of text processing, such as information retrieval and text understanding:

- the automatic matching problem,
- the structure of the lexical entry problem.

The first problem is the same as that which beleaguers information retrieval systems, namely, determining exactly what counts as a valid match. The goal is to include all and only correct and relevant matches, and to exclude all incorrect and irrelevant matches. The ideal would be to accomplish this in a more fully automatic way, with as little manual processing as possible. However, for some of our probes, we obtained as little as 5% match from corpus using dictionary probes, leaving 95% of the corpus citations to be processed. Other probes gave us 75% successful hits. An incremental method to expand the probe set would help solve this problem. In addition, adding syntactic as well as lexical semantic clues might give better extraction results, and allow easier insertion of the corpus data into the lexicon, therefore reducing the task of manually inserting and integrating the remaining data. The most radical move would be to eliminate the manual step altogether, but experience shows that even with the most statistically powerful systems, multi-word entries which are not fixed idioms and which display a range of translations fail to be correctly associated (see Brown et al. (1988) for a discussion of failures.) Therefore, we chose the route of incorporating lexicographical and linguistic knowledge directly into the process of lexicon building.

Related to the problem of matching is that of the variability in lexical correspondences. One of the more curious types of example we came across seems to incorporate translator deixis into the translation. Consider an example with *emigrate*:

Table 8. Sample Citation for *emigrate*

English:	It is reminiscent of the rules in place before World War I which said anyone who wished to **emigrate** to Canada from the Indian subcontinent could do so provided they came here directly.
French:	Cela nous rappelle les règles en place avant la première guerre mondiale qui prévoyaient que toute personne venant du sous-continent indien et désirant **immigrer** au Canada pouvait le faire, pourvu de venir ici directement .

This example for *emigrate* shows an interesting phenomenon, namely the frequent confusion between *emigrate* and *immigrate* in the Hansard corpus which we suspect may arise from the point of view of the speaker and/or the translator. The French 'immigrer' appears as *emigrate*, not *immigrate*; conversely, the English *emigrate* does not appear as French 'émigrer'. This example illustrates some of the complexities in building a correspondence based lexicon, but also demonstrate the richness that the corpus can bring to the lexical task.

The second problem, that of the structure of the lexical entry, troubles linguists and lexicographers, both of computational and non-computational persuasions. First, it is not straightforward to know with which field to associate a correspondence. For example, in *dance*, does *dance around* go under a separate translation field or is it related to the collocation field with *dance about*? Furthermore, as shown above, automatic insertion of matched information often has to be attached at the higher node. This is due to the fact that there is no syntactic analysis of the corpus. For example, in Figure 7 and 8, there is no syntactic clue that indicates whether *dance* is transitive or intransitive, so a general higher node needed to be created.

In addition, there is the important issue of distinguishing figurative and literal usages, particularly where figurative uses might be frequent. For example, in Table 9, the verb *dance* and its translation show the figurative sense of the verb, although the correspondence in this case is correct for both the literal and figurative usage.

6.1. Sublanguage Corpora

Since the Hansard corpus consists of the Canadian Parliamentary proceedings, it contains a number of juridical and parliamentary terms, usages, and structures. This is typical both of sublanguages holding for a limited domain (Grishman and Kittredge, 1986), and of genres specific to contextual situations. The Hansards could be considered a sublanguage since it has the definite sublanguage characteristics of containing a specialized lexicon and of exhibiting particular syntactic

Table 9. Figurative Uses of "dance/danser"

English:	with visions of summer holidays **dancing** in their heads
French:	avec, **dansant** dans leurs têtes, des visions de vacances d'été
English:	I think every Conservative Member of Parliament must have a huge closet full of shoes because they are all up on their tippytoes **dancing** around the issues...
French:	Je pense que chaque membre conservateur du Parlement doit avoir une énorme armoire pleine de chaussures car ils sont tous sur la pointe des pieds **en train de danser** autour de problèmes...

structures. For example, consider the humorous example of the most frequent translation of the verb *hear* as "bravo" as computed by the IBM Machine Translation project [6], rather than the expected "écouter" or "entendre". Sublanguages are different from the standard language in that their lexicons often contain specialized usages. At the same time, syntactic differences abound as well. For example, a typical structure in the Hansards is *Mr. Speaker, order please, sixty seconds.*, whereas this construction is not at all common in the standard language. The social context of the the Hansards creates an unusual speech situation, one which results in a specific genre of the sequential parliamentary lecture. Thus, the Hansards reflect both a sublanguage and a genre.

The flexibility inherent in the BICORD system allows 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. Sublanguage lexicons generally have narrow meanings for general terms (such as *boot* in computerese being specialized and distinct from *boot* in military-ese). This flexible feature of BICORD allows a representation of the lexical correspondences of general and specific texts in one data structure. It also permits comparison between sublanguages. The result will be an enriched lexicon built over a variety of corpora to reflect the actual usages of the words or phrases in context, both in general and in sublanguage usage, and ranging over different genres.

6.2. Granularity of lexical entry

Furthermore, there is the question of how much data should be included. If the decision is made to include only that which falls above a given threshold, then what is the statistical (or manual) cut-off for significance in lexicographic and linguistic correspondences? If the threshold is too low, an extensive expansion of the lexical entry structure would be required to include the wide range of unique correspondences, thus weakening the general structure of the entry and the generalizable nature of the content. In other words, the enrichment of the lexicon could cause a related explosion of superfine detail in the dictionary structure. Additionally, some new context fields should be added to the collocation nodes, but determining the criteria for selecting them automatically is difficult. The issue of complexity and exhaustiveness in lexical structures created by BICORD is not resolved, but we view it as an important challenge in our current work.

7. Applications

This work can be applied in at least three different ways. The first one involves use of the output of BICORD system. Bilingual on-line MRD's are now frequently used by the general population. If a user needs to access a bilingual dictionary, it is becoming increasingly common to access an on-line dictionary, particularly with the upsurge of multimedia computers supporting large lexical databases. Due to well-structured information, on-line dictionaries are used extensively, but on-line corpora, although becoming more and more accessible and abundant through the network resources, are often not utilized for their lexical richness. Therefore, a system such as BICORD, which incorporates the information of both sources and is not limited to the MRD content alone, can be of a great use.

Secondly, another application of the BICORD system is within the context of professional translation. As more and more bilingual MRD's are available to human translators, it is extremely helpful for the translator to access a lexical database enhanced with information from corpora. Among several needs, the most useful information is a lexical database that captures frequently occuring expressions. Being able to access such information allows the translator to be consistent throughout texts, and allows groups of translators to achieve inter-textual consistency. If the BICORD system uses a particular corpus, this will benefit that particular set of translators; for example, Hansard translators could largely benefit from the lexical resource built from the Hansard corpus. In a similar way, multi-word correspondences giving several translations with frequency counts associated with them allows the translator to make a judgment or a choice for the translated item.

Thirdly, techniques developed in the BICORD system can be very useful in lexicographic research. In the context of a lexicographer developing a new dictionary, one could think of using an already existing dictionary structure in order to first enhance the lexical database with other MRD information, and second, to enhance it with multiple corpora information. Although BICORD exemplifies the use of bilingual resources, the same approach can easily be applied to the construction of monolingual lexical databases. The computational linguist can make use of the algorithms and apply them to any comparable database. Finally, the output of BICORD can be used directly by systems for processing multilingual data, such as transfer-based machine translation systems. Limits presented in the BICORD system can be overcome by the use of parsers operating at a syntactic level and not only at a word part-of-speech level. Challenges lie in developing methods that permit the quick updating of machine translation dictionaries.

8. Conclusion

To sum, this paper presents techniques for merging information from machine-readable dictionaries and corpora into a useful lexical data base. We present an algorithm based on linguistic and statistical principles for building enhanced dictionary entries. We discuss the issues and problems in achieving this goal. Future research will address questions of more sophisticated corpus analysis, and the incorporation of results into the lexical data base structure.

Notes

1. We have greatly benefited from the detailed and careful comments of several anonymous reviewers, whom we thank for their time and insights.
2. The definition entries are taken from the parsed version of on-line dictionaries at the IBM TJ Watson Research Center. The dictionary entry parser was written by Mary Neff with Bran Boguraev (Neff and Boguraev, 1989). The query language was written by Roy Byrd. We are grateful to them for the use of the system and the data. We have sometimes simplified the presentation of the dictionary entries for the purpose of clarity.
3. For access to the Hansard corpus and its alignment, we acknowledge both the Speech Recognition Group of the IBM TJ Watson Research Center, and Ken Church and Bill Gale of AT&T Bell Laboratories.
4. The system as presented here was designed, developed, and built at the IBM TJ Watson Research Center, where the first author was then a member of the Lexical Systems Group and the second author was then in the Continuous Speech Group.
5. We are grateful to B.T.S. (Sue) Atkins and Beth Levin who first encouraged us to explore movement verbs.
6. P. Brown, personal communication.

References

Atkins, Beryl T. 1987. Semantic ID tags: Corpus evidence for dictionary senses. In *Proceedings of the Third Conference of the University of Waterloo.* Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research.

Atkins, Beryl T., Alain Duval, and Rosemary C. Milne. 1978. *Collins Robert French Dictionary: French-English. English-French.* Collins Publishers, London.

Atkins, Beryl T., Judith Kegl, and Beth Levin. 1988. Anatomy of a verb entry: from linguistics theory to lexicographic practice. *International Journal of Lexicography*, 1:84–126.

Boguraev, Branimir. 1991. Building a lexicon: The contribution of computers. *International Journal of Lexicography*, 4(3).

Boguraev, Branimir, Roy Byrd, Judith Klavans, and Mary Neff. 1989. From structural analysis of lexical resources to semantics in a lexical knowledge base. In *First International Lexical Acquisition Workshop*, Detroit, Michigan. International Joint Conference on Artificial Intelligence.

Brent, Michael R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Third Conference on Applied Computational Linguistics*, Trento, Italy.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A statistical approach to language translation. In *Proceedings of the Twelfth International Conference on Computational Linguistics*, Budapest, Hungary.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P., J. Lai, and R. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the Twenty-ninth Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California.

Brown, P., S. Della Pietra, V. Della Pietra, M. Goldsmith, J. Hajic, R. Mercer, and S. Mohanty. 1993. But dictionaries are data too. In *Proceedings of DARPA*, Princeton, New Jersey.

Byrd, Roy, Nicoletta Calzolari, Martin Chodorow, Judith Klavans, Mary Neff, and Omneya Rizk. 1987. Tools and methods for computational lexicology. *Computational Linguistics*, 13(3):219–240.

Calzolari, Nicoletta and Remo Bindi. 1990. Acquisition of lexical information from a large textual Italian corpus. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finland.

Carter, Richard. 1988. On movement (written in 1984). In Beth Levin and Carol Tenny, editors, *On Linking: Papers by Richard Carter*, volume 25 of *Lexicon Project Working Papers*. MIT Press, pages 231–252.

Catizone, Robert, Graham Russell, and Susan Warwick, 1989. *Deriving Translation Data from Bilingual Text*. ISSCO, Geneva, Switzerland, unpublished manuscript.

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the Twenty-third Annual Meeting of the Association for Computational Linguistics*, pages 299–304.

Church, Kenneth W. 1989. A stochastic parts program noun phrase parser for unrestricted text. In *IEEE Proceedings of the ICASSP*, pages 695–698, Glasgow.

Church, Kenneth W. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the Thirty-first Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Church, Kenneth W., Patrick Hanks, D. Hindle, and W. Gale. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Using on-line Resources to Build a Lexicon.* Lawrence Erlbaum.

Cruse, D. Alan. 1986. *Lexical Semantics.* Cambridge University Press, Cambridge, England.

DeRose, Stephen. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Dorr, Bonnie J. 1992. The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7(3):135–193.

Dowty, David. 1979. *Word Meaning and Montague Grammar.* Reidel, Dordrecht.

Gove, Philip B., editor. 1963. *Webster's Seventh New Collegiate Dictionary.* G.& C. Merriam Company, Springfield, Mass.

Grishman, Ralph and Richard Kittredge, editors. 1986. *Analyzing language in restricted domains: Sublanguage description and processing.* Lawrence Erlbaum.

Gruber, J. S. 1965. *Studies in Lexical Relations.* Ph.D. thesis, The Massachusetts Institute of Technology, Department of Linguistics, Cambridge, Massachusetts. published later 1976 as Lexical Structures in Syntax and Semantics, North-Holland, Amsterdam.

Hale, Kenneth and Jay Keyser. 1986. *Some Transitivity Alternations in English.* Center for Cognitive Science, The Massachusetts Institute of Technology.

Jackendoff, Ray S. 1987. The status of thematic relations in linguistic theory. *Linguistic Inquiry*, 18(3):369–411.

Jackendoff, Ray S. 1990. *Semantic Structures.* MIT Press, Cambridge, MA.

Kay, Martin and Martin Röscheisen. 1993. Text translation alignment. *Computational Linguistics*, 19(1):75–102.

Klavans, Judith L. 1988. Complex: A computational lexicon for natural language systems. In *Proceedings of the Twelfth International Conference on Computational Linguistics*, Budapest, Hungary.

Klavans, Judith L., Martin Chodorow, and Nina Wacholder. 1990. From dictionary to knowledge base via taxonomy. In *Proceedings of the Sixth Conference of the University of Waterloo.* Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research.

Klavans, Judith L. and Philip Resnik, editors. 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* MIT Press, Cambridge, Mass.

Klavans, Judith L. and Evelyne Tzoukermann. 1989. Corpus-based lexical acquisition for translation systems. In *Proceedings of the Sixth Israeli Conference of Artificial Intelligence and Computer Vision*, Tel Aviv, Israel.

Klavans, Judith L. and Evelyne Tzoukermann. 1990a. Linking bilingual corpora and machine readable dictionaries with the BICORD system. In *Proceedings of the Sixth Conference of the University of Waterloo.* Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research.

Klavans, Judith L. and Evelyne Tzoukermann. 1990b. The BICORD system: Combining lexical information from bilingual corpora and machine readable dictionaries. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finland.

Kupiec, Julian. 1989. Augmenting a hidden markov model for phrase-dependent word tagging. In *Proceedings of the 1989 DARPA Speech and Natural Language Workshop*, pages 92–98, San Mateo, California. Morgan Kaufmann.

Leech, Geoffrey, Roger Garside, and Erik Atwell. 1983. Automatic grammatical tagging of the LOB corpus. *ICAME News*, 7:13–33.

Levin, Beth and Malka Rappaport. 1988. On the nature of unaccusativity. In *Proceedings of New England Linguistic Society.*

Merialdo, Bernard. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Neff, Mary and Bran Boguraev. 1989. Dictionaries, dictionary grammars and dictionary entry parsing. In *Proceedings of the Twenty-seventh Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.

Neff, Mary S., Roy J. Byrd, and Omneya A. Rizk. 1988. Creating and querying hierarchical lexical data bases. In *Proceedings of the Second Applied Association for Computational Linguistics Conference*, pages 84–92, Austin, Texas.

Pustejovsky, James, Sabine Bergler, and Peter Anick. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358.

Rizk, Omneya, 1989. *Sense Disambiguation of Word Translation in Bilingual Dictionaries: Trying to Solve the Mapping Problem Automatically.* Unpublished M.A. thesis, Courant Institute of Mathematical Sciences, New York University, New York.

Sadler, Victor, 1989. *The Bilingual Knowledge Bank: A New conceptual basis for MT.* BSO/Research, unpublished manuscript, Utrecht.

Smadja, Frank, Kathleen McKeown, and Vasileios Hatzivassiloglou. in press. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics.*

Talmy, Leonard. 1975. Semantics and syntax of motion. In J.P. Kimball, editor, *Syntax and Semantics*, volume 4. Academic Press, New York, NY, pages 181–238.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen, editor, *Language Typology and Syntactic Description: Grammatical categories and the Lexicon.* Cambridge University Press, Cambridge UK.

Tenny, Carol L., 1992. *How Motion Verbs are Special.* University of Pittsburgh, Department of Linguistics, unpublished manuscript.

Tenny, Carol L. 1994. *Aspectual Roles and the Syntax-Semantics Interface.* Kluwer Academic Publishers, Dordrecht.

Tzoukermann, Evelyne and Bernard Merialdo, 1989. *Some Statistical Approaches for Tagging Unrestricted Text.* IBM, T. J. Watson Research Center, Yorktown Heights, New York, unpublished manuscript.