Machine Translation, 10:3-4, 1–34 (1996)
© 1996 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

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

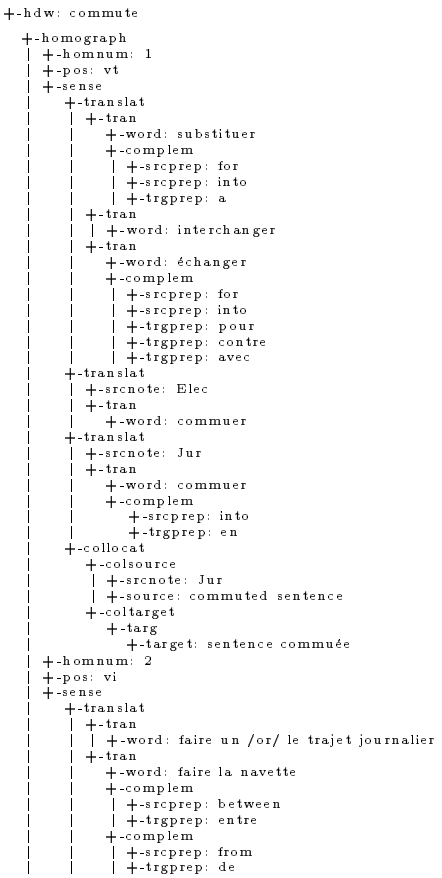

Figure 2. Partial MRD entry for *commute* from $CR-EF$

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. 'nonchâlamment', '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*