LEXICOGRAPHICAL INFRASTRUCTURE AND GOVERNMENT: THE ACTION PLAN FOR THE PRODUCTION OF DUTCH BILINGUAL DICTIONARIES

WILLY MARTIN
Free University of Amsterdam

INTRODUCTION

This paper consists of four parts:

- In the first section, lexicography and its main actors will be introduced.
- Secondly, the CLVV and its ideas, plans and projects will be presented.
- Third, special attention is directed to the role infrastructure (tools, models and strategies) can play in bilingual lexicography.
- Finally, the paper will conclude by trying to generalize from the concrete case dealt with, and point towards some future developments.

1. LEXICOGRAPHY

In order to have a common basis to start from, we will use the definition of lexicography as found in Heestermans, viz., "[lexicography is] de tak die zich bezighoudt met de manier waarop de woordenschat wordt verwerkt in een woordenboek" ('that branch [of activities] that deals with the way words are processed in a dictionary') (Heestermans 1976, 39-57).

Although others such as Lutzeier (1995) broaden this definition by explicitly combining theory and praxis (see e.g. his definition: "Unter Lexikographie verstehen wir die Theorie und Praxis des Schreibens von Wörterbüchern", 'By the term Lexicography we define the theory and practice of compiling dictionaries'), it is obvious that both in the broad sense (taking metalexicography as part of lexicography), and in a more narrow one (taking the practice only into account), the central object of interest in lexicography remains the same, namely *dictionaries*.

In an attempt then both to synthesize and to supersede the above-mentioned definitions one could define lexicography as "the process or profession of writing or compiling dictionaries in a sound and scientifically justified way, by making use of adequate techniques and technologies, which results in the representation or description of (aspects of) a vocabulary of a language (or parts thereof) in function of

(groups of) users". This way lexicography is basically seen as a multifaceted activity, a "scene" in which several participants are involved. The definition of this activity could then be represented in the form of the following schema or frame (for the relation between frames and definitions see Martin 1994, 237-256; Martin 1998):

LEXICOGRAPHY	
IS AN	ACTIVITY
SUBTYPE	MAKE/PRODUCE
AGENT	LEXICOGRAPHERS
CO-AGENT	METALEXICOGRAPHERS
AFFECTED OBJECT	VOCABULARY (PARTS/ASPECTS)
RESULT	SCIENTIFIC DESCRIPTION
FORM	BOOK, CD ROM, DATA BASE
MEANS	INFORMATION PROCESSING TOOLS
AGENT	IT-DEVELOPERS
BENEFICIARY	USERS
OTHER PARTICIPANTS	PUBLISHERS, SPONSORS

Table 1: "Frame-like" definition of Lexicography

Of course, the above frame is a somewhat simplified and schematized representation of "reality". It does not always make the relationship between the several slots explicit. The relationship between lexicographers and metalexicographers, for instance, is a complex one: sometimes lexicographers take up the role of metalexicographers fully, other times only partially, sometimes the two roles are disjunctive, etc.

However, what becomes explicitly clear from this "frame-like" definition is that lexicography (nowadays) involves different actors, viz.:

- users
- lexicographers
- metalexicographers
- IT-developers
- publishers
- sponsors

In the next section we will take up the role and function of the "other participants" as they have remained unspecified up till now.

2. GOVERNMENT AND BILINGUAL LEXICOGRAPHY: THE CLVV AS A CASE-IN-POINT

2.1 CLVV: BACKGROUND

Lexicography was, and still is, for the most part, a commercial undertaking: some dictionaries belong to the best selling books in the world and so it is obvious that publishers play an active and important role in their production and distribution. When government intervenes, its role is mostly much less active and prominent. National governments, for example, often do play a role as subsidizers of large, scientific, monolingual, so-called "national" dictionaries such as the *WNT* (*Woordenboek voor de Nederlandse Taal*/'*Dictionary of the Dutch Language*') in the Netherlands and Flanders, or the *Dictionnaire de la Langue Française* in France. That governments can play a role in the production of bilingual dictionaries and that their role need not be restricted to that of a passive sponsor or subsidizer is, however, much less known and obvious.

The governments of the Netherlands and Flanders rather take up the role of model country here, as their language policy is directed towards an active role as meta-actor with regard to the production of bilingual dictionaries with Dutch as one of the languages, at least in these cases where need arises. This is the case when one is sure that the products aimed at will not be achieved within a reasonable time lapse and with a sufficient quality level by private enterprise, and when the social merits of these products at least equal the social costs.

Actually, the active policy of the Netherlands and Flanders in lexicographical matters is inspired by the fact that both countries regard linguistic infrastructure as being as important as other basic pieces of infrastructure.

Just as it is important for governments to care for a good road infrastructure, so it is equally important to care for a good dictionary infrastructure. Dictionaries, like roads, are connecting means: the monolingual ones connect people within one country or one linguistic community, whereas the bilingual ones create the possibility to come into contact with people from abroad. In other words, good monolingual dictionaries are important from the point of view of care for one's own language, whereas good bilingual dictionaries strenghten the position and enhance the possibilities of two linguistic communities, thus creating direct communicative links between these communities. This is particularly important for the so-called lesser-used languages, which otherwise have to use an *interlingua*, such as English or another major language to come into contact with each other. If one wants to create equal possibilities for all citizens in a community (e.g. in the EU), which will allow them to take part in the information society, the use of language should not be a hindrance, but a help. These considerations have brought the governments of the Netherlands and Flanders to take up the role of an active meta-actor (one that has an impact on the primary actors in the field) in the area

of bilingual lexicograghy. This has led them to establish in 1993 a Committee called CLVV (Commissie voor Lexicografische Vertaal Voorzieningen/'Committee for Lexicographical Interlingual Resources') in order to advise on, define, streamline and co-ordinate their policy with regard to bilingual dictionaries. In what follows, we will briefly summarize some facts and figures about the CLVV.

2.2 CLVV: SOME FACTS AND FIGURES

For a somewhat more elaborate description/characterization of the CLVV and its policy we refer to Martin 1995. We here restrict ourselves to some basic facts and figures, taking them up point by point:

- CLVV = Commissie voor Lexicografische Vertaal Voorzieningen ('Committee for Lexicographical Interlingual Resources')
- Established by the Ministers of Education of both the Netherlands and Flanders on March 1, 1993, for the duration of 3 (first term) and 5 (second term) years
- Task: define an action Plan (and carry it out) for a period of about 10 years (1993-2002) for the development of bilingual dictionaries with Dutch as source or as target language, implying:
  – inventory of needs
  – prioritization and selection of languages
  – launching of projects
  – evaluation of projects
  – support for projects (financial and infrastructural)
  – search for (co-)financing
  – guidance and control of projects
  – advice to projects
- General policy lines:
  – public resources will only be used for the financing of those projects
    - for which the market of private enterprise fails, and
    - of which the social merits are larger or at least equal to the social costs
  – projects should lead to
    - multifunctional lexical Data Bases
      - made in a cost-effective way
      - based on modern technology
- Selection of languages should be based on real needs as indicated by demographic, economic, educational, cultural, scientific and political parameters; should take into account the geographical context (perceived as four concentric circles and moving from the inner- to the outermost circle: in the case of Dutch this is: the Dutch-speaking area ---> the EU ---> Europe ---> World)

- Budget
  – Dutch/Flemish financing = 12,000,000 Dutch guilders spread over the period 1993-2001
  – Co-financing is needed (mostly from partner-countries)
  – Distribution of budget:
    - 80% for lexicographical labour
    - 20% for infrastructure (mainly tool development)
    - project costs [considering the fact that desk dictionaries (Data Bases) are aimed at macro (: 40,000+) items; rich micro (: 2 volumes: A-B. B-A)] varying from 300,000 to 1,000,000 Dutch guilders.
- Secretariat: 2501 HN The Hague (Dutch Language Union), Lange Voorhout 19.

2.3 CLVV-PROJECTS (STATE-OF-THE-ART AS FROM 1996 ONWARDS)

2.3.1 *Bilingual Dictionary Projects*

Swedish – Dutch v.v.:	published 1996
Italian – Dutch v.v.:	to be finished end 1998
Arabic (MSA) – Dutch v.v.:	to be finished mid 2000
Arabic (MSA) – Dutch v.v. (learner's dict.):	to be finished end 1998
Turkish – Dutch v.v.:	to be finished mid 1998
Turkish – Dutch v.v. (learner's dict.):	to be finished mid 1998
Dutch – Polish:	to be finished end 1999
Polish – Dutch:	to be finished mid 2000
Dutch – Hungarian:	to be finished end 1999
Hungarian – Dutch:	finished beginning 1998, in production
Dutch – Czech:	published 1997
Czech – Dutch:	to be finished mid 2000
Danish – Dutch v.v.:	to be finished 1999
Finnish – Dutch v.v.:	planning stage
Greek – Dutch v.v.:	to be finished mid 2001
Portuguese – Dutch v.v. :	to be finished mid 2001
Norwegian – Dutch v.v.:	to be finished end 2000
Rumanian – Dutch v.v.:	to be finished mid 2000
Estonian – Dutch v.v. :	to be finished mid 2000
Sranan – Dutch v.v. :	to be finished end 2000
Korean – Dutch:	to be finished end 1999
Dutch – Indonesian:	to be finished end 2000

2.3.2 *Infrastructural Projects*

OMBI:	version 4.6 to be finished mid 1998 (Dictionary Editor)
RBN:	second phase to be finished end 1998 (Reference Lexicon of Dutch to be used as Dutch Source Input and (Partial) Target Output)

3. GOVERNMENTS AND LEXICOGRAPHICAL INFRASTRUCTURE

In the preceding section we have stated that the Dutch and Flemish governments have opted for an active and coherent policy with regard to bilingual dictionary production. This active policy becomes apparent not only from the fact that the governments themselves use an Action Plan in which priorities are defined, based on indications of real needs, but also on the fact that financial resources should not only be used for dictionary projects as such, but also for the development of generic tools and models with which to construct bilingual dictionaries in a cost-effective and yet high-quality way. In doing so, one can tackle not only concrete, *hic-et-nunc* needs, but anticipate future ones as well. In order to reach that goal, the CLVV has created development tools such as the OMBI and frameworks of linking such as the "hub-and-spoke" model. In the following sections, we will turn our attention to these tools.

3.1 OMBI

OMBI, the Dutch acronym for Omkeerbare Bilinguale woordenboeken ('Reversible Bilingual Dictionaries'), is an editor for bilingual dictionaries which, like most dictionary editors, is a device to guide, structure and correct input data according to a pre-defined grammar. As such its advantages are: consistency, correctness and efficiency.

Next to the fact that OMBI shows all characteristics which dictionary editors in general have, it also has the following features which makes it rather exceptional and, in some cases, even unique; namely:

- with OMBI, language pairs can be reversed with a very high degree of precision;
- with OMBI, different kinds of dictionaries can be derived from one Data Base;
- with OMBI, one can easily add a third language and thus create multilingual resources;
- OMBI has different kinds of import facilities: among others, it can re-use existing dictionary input (which will be converted to OMBI-SGML; of course, the better structured the input data, the easier for OMBI to convert the data);
- OMBI is generic and not language-specific in that one can change/add/delete

fields, field names, values and rules for calculi from the DDD (Default Database Design);

- OMBI offers various search possibilities;
- OMBI yields SGML-output.

Of all features of OMBI, the reversal function is certainly the most innovative one. However, in order for OMBI to function optimally, it has to meet all of the requirements for so-called linkable reference lexicons (see Martin e.a. 1998). This means that it expects the lexicographer to specify "words" not only at form level (as FUs = Form Units), but at meaning level also (as LUs = Lexical Units) (e.g. by giving a short meaning description or rèsumè). In other words, OMBI distinguishes two different types of units: the Form Unit (FU = the word or string as physical entity) and the meaning(s) associated with that word (LU). Reversal can never be successful unless translation is done at meaning level as opposed to string level. Furthermore, OMBI not only links two semantic units (LUs) to each other, it also forces the lexicographer to specify the nature of the links. The translation relation or link is analysed according to four relevant parameters which influence reversibility. The latter only holds if certain conditions are met. The four parameters are the following:

- conceptual equivalence (e.g. E. *river* is a hyperonym of F. *rivière*)
- pragmatic contrast (e.g. E. *bike* differs in style [it is informal] from F. *bicyclette* [which is neutral])
- variant status (e.g. E. *bike* will have as its main translation equivalent in French *vélo*, while *bicyclette* will be a variant)
- lexicalization status(e.g. German *Sprachverwirrung* is fully lexicalized whereas E. *confusion of tongues* is semi-lexicalized).

The values of these parameters are used in a calculus which leads to certain results. OMBI makes, among others, use of the following conceptual equivalence reversal rules:

A translation link from L1 to L2 may be reversed with respect to conceptual equivalence according to the following table (where Ø indicates an irreversible translation):

L1 ⇨ L2		L1 ⇨ L2
complete equivalence	⇨	complete equivalence
interlingual hyperonymy	⇨	interlingual hyponymy
interlingual hyponymy	⇨	interlingual hyperonymy
related equivalence	⇨	related equivalence
sybstitution by explanation	⇨	Ø
substitution by borrowing	⇨	Ø
substitution by near equivalence	⇨	substitution by near equivalence

Summarizing, one can state that OMBI stores data about form units and lexical units of two languages and that also the inter- and intralingual links between these data are considered to be data. As a result, very useful dictionary articles for the B-A part can be automatically generated by OMBI by properly linking A-items to B-items. In the following, some sample data are presented in order to give an idea of what one may expect when using an editor such as OMBI.

The input is chosen to exemplify the following cases:

1. *one-to-one-cases:* meaning that both the source and the target language have one FU (+LU), see *kaneel - cinnamon.* As a rule, the output, which is yielded completely automatically, is nearly perfect: one could keep it without further manual intervention but for, in the case of *cinnamon*, the replacement of *cinnamon stick,* which one could put into the macro-structure or otherwise delete. One of the nice things about this entry is that its examples make clear how one can "individuate" or singularize an uncountable item such as *cinnamon*: cf. cinnamon *stick*, a *pinch* of cinnamon. In cases like the above, the automatic reversal yields a (quasi-) "perfect" and "complete" B-A entry, which in many cases will be "better" than what is usually to be found in comparable English-Dutch dictionaries.

2. *one-to-many-cases:* meaning that one FU (having several meanings or LUs) in the source language yields different FUs (+LUs) in the target language. From this point of view there is a divergence between source and target language. See Dutch *kaart* and the reversal of it yielding English *card, chart, map1, map2, ticket,* and *menu*.

D kaart E	card chart map1 map2 ticket menu

Table 2: A Case of Divergence

Actually because of the fact that, in contrast to the *kaneel - cinnamon* case, where the one-to-one relation was between monosemes, the relation is now between one polysemous FU and many polysemous FUs, the results from the reversal cannot be "final" in one go. So e.g. the Dutch word *bon* also will yield *ticket* (in the meaning "penalty"), and so English *ticket* will "grow" into a fuller and more complete entry, once Dutch words such as *bon, bekeuring, lot, ontslagbriefje*, etc. next to *kaart* will have been dealt with.

D	E
kaart	card
bon	… …
lot	ticket

Table 3: Cases of both Divergence and Convergence

The result of the reversal of a FU like Dutch *kaart* yields for such English FUs as *card, chart, map* etc. interesting, though only partial entries. Material for the completion of such entries as e.g. English *ticket* should come from other Dutch FUs such as *bon, lot,* etc. This material is then *merged* into one article/entry as we will demonstrate with the next case.

3. *many-to-one-cases:* meaning that the source language has many FUs (+LUs) which have to be combined into one FU from the target language, see *paar* 1 and *paar* 2 versus *couple*.

When looking at *couple* one will observe that the data contained in the Target Language entry have their origin in more than one Source Language entry (the fact that here the form *paar* seems the same does not matter: there are 2 *paar* entries, *paar* 1 and *paar* 2.) Meanings 1 and 2 stem from *paar* 1 (and so do the accompanying examples), whereas meaning 3 (+ examples) comes from *paar* 2. The result has been obtained without any manual post-editing as table 4 makes clear.

paar 1, meaning 1	pair, meaning 1
	couple, meaning 1
meaning 2	couple, meaning 2
	couple, meaning 3
paar 2, meaning 1	

Table 4: A Case of Convergence (paar 1 / 2 couple)

3.2 THE "HUB-AND-SPOKE" MODEL

As one will have observed, in using OMBI one does not translate an item from language A to another one from language B, but, instead of that, one explicitly links a LU from A to one from B so to be able to calculate the reversibility of the link. Actually in doing so one can distinguish three possible kinds of input, viz.:

- no data for A, nor for B are given, one has to create entries in A and link them to entries in B, which one has to create also;
- data for both A and B are given, one has to link only;
- data for either A or B are given, the data for B or A have to be entered (links included).

An optimal situation is that where both data for A and B are given and one only has to link. For that purpose, the CLVV has developed a Reference Lexicon for Dutch, which can be used as one of the monolingual sources (Dutch) that can be linked to another monolingual source (e.g. Portuguese, Greek, Danish etc.). Ideally such a monolingual source is what we have called elsewhere (see Martin e.a. 1998) a Linkable Reference Lexicon (LRL). For the CLVV, up till now the objective has been to link two LRLs in as economical and efficient a way as possible, and thus to derive bilingual dictionaries as front-ends. If, however, one widens the scope from bi- to multilingual resources, one is no longer dealing with two languages but three or more.

Consider the following situations:

In table 5 there are LRLs connected with each other in a reversible way. In table 6 a third language has been added, i.e. one of the LRLs of table 5 (B) has now been linked to another LRL (C), thus generating a new bilingual set (BC, CB).

A	B

Table 5: Reversible Links between two languages A and B

A	C
B	

Table 6: Two reversible bilinguals, linked to one common hub

Language B, which is linked with languages A and C, we will call the *hub*, in analogy with air traffic organization, in which certain airports act as centres (*hubs*) from which flights from other airports (*spokes*) operate.

The key claim of the "hub-and-spoke" model as the CLVV uses it, is that, by carefully linking the lexical items of each of two different spoke-languages to one and the same hub-language, the items of the spoke-languages themselves can be linked together in a cost-effective way (see A-C in table 7).

A	C
B	

Table 7: Exploiting the "hub-and-spoke" model: links between spokes (A ↔ C) inferred

The hub (B in table 7) functions then as *tertium comparationis*, and, as long as the links show values with which one can calculate in a meaningful way, one can derive the third bilingual dictionary in an almost automatic way.

For the rest of this section, we will restrict ourselves to a simplified exemplification of the "hub-and-spoke" model taking just 5 lexical items (one from French, two from English and two from Dutch) into account and looking at the result of linking them in a "hub-and-spoke" configuration.

Input: 3 LRLs, a French, Dutch and English one, containing 1 (French) and 2 (Dutch, English) lexical items, respectively.

Desired Output: 3 Bilingual "Dictionaries" containing these items. The procedure will be the following: first French will be linked to Dutch (with reversal), then Dutch will be linked to English (with reversal). Finally French will be automatically linked to English and *vice versa*. This is schematized in table 8.

French		English
pleurer	*cry*	
		weep
Dutch		
huilen		
wenen		

Table 8: Reversing and Linking via one Hub

Step 1: *Linking French to Dutch*

– French *pleurer* 1 *verser des larmes*

– Dutch *huilen* 1 *tranen storten*
links *pleurer* 1-*huilen* 1 : complete conceptual equivalence, lexicalized, no pragmatic difference, main translation equivalent

– Dutch *wenen* 1 *tranen storten*
links *pleurer* 1-*wenen* 1 : complete conceptual equivalence, lexicalized, pragmatic diff.: formal/archaic, variant translation equivalent

– Result (when processed by (an) OMBI (-like editor)):

pleurer = *huilen* ⇨ ↑ *wenen*
(reversal:)
huilen = *pleurer*
wenen = ↓*pleurer*

(↑ and ↓ are signs used to indicate more, resp. less formal usage; the double arrow (⇨) separates main from variant translations).

Step 2: *Linking English to Dutch*

– Dutch	*huilen*	1	*tranen storten*	
wenen		1	*tranen storten*	
– English	*cry*	1	*shed tears*	
links	*cry*1-*huilen*1//*wenen*1		: complete concept. equiv.// comp. concept.equiv. lexicalized//lexicalized no pragm.diff.//formal/archaic main transl. eq.//variant transl. eq.	
	weep	1	*shed tears*	
links	*weep*1-*huilen*1//*wenen1*		: complete concept. equiv.// comp. concept. eq. lexicalized// lexicalized pragm.diff: less formal// no pragm.diff. var.trans.eq. // main transl.eq.	
– Result				
huilen		= *cry*	⇨	↑ *weep*
wenen		= *weep*	⇨	↓ *cry*
(reversal)				
cry		= *huilen*	⇨	↑ *wenen*
weep		= *wenen*	⇨	↓ *huilen*

Step 3: Generating French – English links and v.v.

pleurer	= *cry*	⇨	↑ *weep*
cry	= *pleurer*		
weep	= ↓ *pleurer*		

In the above example, we have pair-wise tried to make clear that, by carefully linking three LRLs automatically, we can get not only the two pairs and their reversal, but also the third pair and its reversal. What one needs is an editor such as OMBI, which can take links as data and calculate with them (links are typed, and the type of link has an influence on the behaviour of the pairs under reversal and under multiple linking). The results, of course, are dependent on the nature and correctness of the links: not always can the resulting equivalence candidates be taken without human post-editing. However, the results obtained up till now look very promising, both from a qualitative and from a quantitative (economic) point of view.

4. GENERAL CONCLUSIONS

In this paper we have tried to give an insight into the concrete policy of the Dutch and Flemish governments concerning biligual dictionary construction. We hope that from this concrete case the following general issues have become clear:

A coherent action plan is to be preferred to an *ad hoc*, incidental subsidization policy; the latter works on a first come, first served basis, the former tries to prioritize on a sound basis and to anticipate possible needs. In order to achieve this, not only concrete lexicographical products, but also the development of infrastructural tools play an important role. A co-ordinating lexicographical platform such as the CLVV, as a rule, does not itself carry out any projects, but sees to it that the ones which are needed most are carried out. It could be interesting for governments of other countries to follow the example of both Holland and Flanders and to create comparable bodies.

Such co-ordinating platforms in other countries would be welcome because co-operation and cost-sharing between countries and languages can be facilitated that way and lead to more and better results: think of the application of the "Hub-and-Spoke" Model in this respect, superseding bilateral interests.

If governments cease to play a passive role in lexicographical matters, and instead become more active meta-actors, they can link short-term to long-term actions/realizations. In this respect, the creation of a generally available and generic infrastructure such as dictionary editors with reversal functionalities, is a piece of infrastructure that not all private enterprises can afford to develop themselves or place at everyone's disposal. Such tools therefore can be developed with government funds as freeware, for them to be used not in just one *hic-et-nunc* situation/project, but in a broader and more wide-ranging context.

References

HEESTERMANS, J. 1976. Een kritische kanttekening en een schets voor een lexicografische theorie. In *De Nederlandse lexicografie tussen handwerk en machine*, ed. P. van Sterkenburg. Groningen: H.D. Tjeenk Willink.

LUTZEIER, P.R. 1995. *Lexikologie. Ein Arbeitsbuch*. Tübingen: Stauffenburg Verl.

MARTIN, W. 1994. Knowledge-Representation Schemata and Dictionary Definitions. In *Perspectives on English. Studies in honour of professor Emma Vorlat*, eds K. Carlon e.a. Leuven; Paris: Peeters.

——. 1998. Frames as definitions models for terms (paper read at the ProCom Conference, Vienna August 1998).

MARTIN W. & A. TAMM. 1996. OMBI: An editor for constructing reversible lexical databases. In *Euralex '96* (Conference Proceedings, Göteborg 1996), eds M. Gellerstam, J. Järborg, S.G. Malmgren, K. Noren, L. Rogström & C. Röjder Papmehl, 675-685. Göteborg: Göteborg University, Department of Swedish.