Willy Martin

A Dutch recipe for the production of bilingual dictionaries

1. Introduction

In the announcement letter to this conference the following is stated:

> Nowadays people do not necessarily turn straight to a dictionary for information about words. They key their enquiry into a search engine, text a friend, or take any one of a number of other (chiefly online) routes to discover the information they want.
>
> With this in mind, the tenth EFNIL Conference will investigate the different ways in which people access lexical information – both in their own language and in other languages – and how governments, language institutions, publishers, and others go about the business of compiling and disseminating this lexical information in the first place.
>
> This theme cuts right to the heart of linguistic diversity of Europe. Each of our states has different ways of addressing the issues of language: some are more advanced along this road than others. Different peoples and different languages may need different solutions, but are there things we can learn from what our colleagues elsewhere in Europe are doing?

As you will have noticed, this passage ends with a question mark. So, in all honesty, I am not sure that you will but I really do hope that you can learn or unlearn from a project that was started in Flanders and the Netherlands in 1993, and that is finished by now, 20 years later, although, like a gardener's, a lexicographer's work never comes to an end. Therefore in this talk, I will not only deal with the past, but at the end also turn 'back to the future' as will become clear from the survey below:

1. Introduction
2. Government Policy and Dictionaries: the CLVV as a case in point
   - 2.1 Background
     - What, Why and How?
     - Tasks & Goals
   - 2.2 Specific Issues
     - Selection and Prioritization Criteria
     - Results
     - Infrastructure
3. Lessons and Remarks for the Future

2. Government policy and dictionaries: the CLVV as a case in point

2.1 Background

National governments often play a role in lexicographical matters as subsidizers of large, scientific, monolingual, so-called 'national' dictionaries, such as the *WNT* (Woordenboek der Nederlandsche Taal (Dictionary of the Dutch Language)) in the Netherlands and Flanders, the *Trésor de la Langue Française* in France or the *Deutsches Wörterbuch* in Germany. However, when it comes to bilingual dictionaries their role is much less obvious and much less known.

In 1993 the governments of the Netherlands and Flanders took a clear stand on that issue as the ministers of education of both countries decided to install an intergovernmental body of lexical experts in order to improve and stimulate the production of *bilingual dictionaries* and lexical databases with Dutch as a source or target language.

At its installation in March 1993 the CLVV (**C**ommissie voor **L**exicografische **V**ertaal **V**oorzieningen = Committee for Interlingual Lexicographical Resources) was given the following tasks:

1. establish Action Plans for the realization of a program of Dutch bilingual dictionaries,
2. define priorities based upon these APs,
3. evaluate project proposals, both on the level of contents and on that of management,
4. have technological projects carried out so to further lexicographical interlingual resources,
5. look for co-sponsoring and co-financing (partner countries, EU, government, trade, industry, ...),
6. supervise both quality and progress of the approved projects,
7. give advice on how to proceed,
8. give advice to those projects which were not prioritized or not subsidized.

Actually, these tasks were based on a report written previously in 1991 (see Martin/Theeuwes 1991), which drew up the state of the art with regard to bilingual dictionaries of Dutch and in which the following principles upon which to found a governmental subsidizing and intervention policy were recommended:

(1) Government should only intervene where the private market fails.
[In other words, one should be sure that the lacking dictionaries or lexical databases will not be published in a reasonable time by a private publisher unless there is a form of government support].

(2) Government support should only be given when the social merits are larger or at least equal to the social costs.
[Therefore, the intended dictionaries or lexical databases, should represent a real need, a.o. being expressed by the number of potential users. Next to that, other social merits dictionaries can bring along are positive external effects such as cultural identification, social integration, R&D activities, etc. These also should be taken into account].

The active policy of the Netherlands and Flanders in bilingual lexicographical matters has also been inspired by the fact that both countries considered bilingual dictionaries as important pieces of basic infrastructure, comparable to road infrastructure, offering people the possibility to come into contact with people from abroad, creating direct communicative connections between two linguistic communities. This fact is particularly important for so-called less-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, for instance in the EU, allowing them to take part in the information society, the use of a language should not be a hindrance, but a help.

Next to the fact that the CLVV had to formulate a program of *concrete lexicographical projects*, it also had to define the general policy lines for a *coherent, anticipatory* and *economically justified* policy. Coherence in this context implied abandoning the ad hoc*, first come, first served* subsidizing policy and replacing it by a systematic approach based an a plan, so, for the subsidizing bodies to know better what to do first and what to do later or not at all.

Furthermore, in order to anticipate needs and tackle them in an economic way, public funds, according to the CLVV, should be used to finance the development of *multifunctional and re-usable electronic lexical databases*. This point-of-view implied, among others, the following:

It should be possible to derive from the same database:

1. both graphical and electronic products,
2. several types of dictionaries,
3. the reverse part of it,
4. bidirectional dictionaries.

It should also be possible to link the languages involved in the original bilingual databases with other languages outside of it, so to yield new databases going beyond the languages of the original language pair (see below: the Hub-and-Spoke Model).

It goes without saying that this was a very ambitious programme. Given the limited budget (6.5 to 7 million euro) prioritization criteria were needed to make a selection of projects.

2.2 Specific issues

2.2.1 Selection and prioritization principles

Such as it is obvious that government should not intervene where the private market was successful, as is, for instance, the case for bilingual dictionaries such as Dutch-English, Dutch-French, Dutch-German and Dutch-Spanish, so too, in other cases the market was much less attractive and successful. Consequently, the CLVV has defined criteria with which to prioritize languages in order to come to an argued selection of languages.

In order to do so the *metaphor of dictionaries as connecting roads* has been further developed and made concrete by taking the *geographical context* into account. The area in which Dutch was spoken (the Netherlands and Flanders) was taken as the innermost of a series of ever growing *concentric circles*: the EU, Europe, the World. As the core or innermost circle itself showed already an internal, non-homogeneous structure (meaning that within that area already different languages were spoken), there should not only be roads from the inner towards the outer circles and vice versa, but also within the inner circle itself (see figure 1, also see Martin 2007).

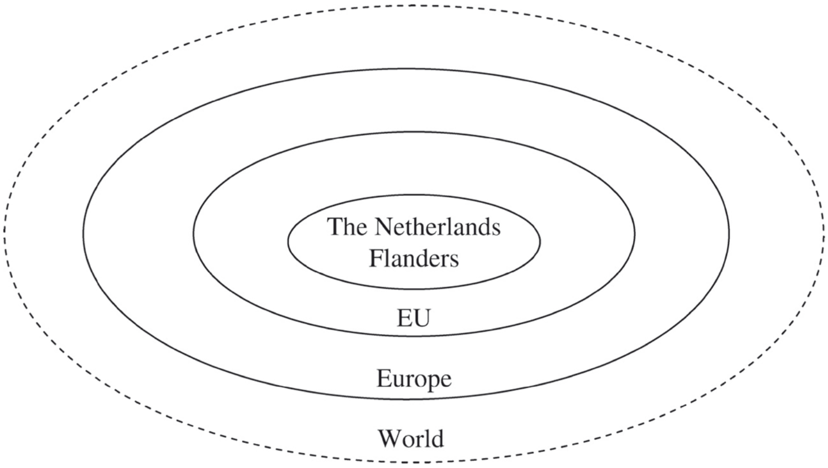

Figure 1: Language areas as concentric circles

As dictionaries were considered to be *connecting roads,* so languages were considered to represent *communities living in a certain area, wanting to come into contact with other communities.* The stronger the need for the contacts between the two communities, the greater the need for *roads* to make these contacts possible and easy. In order to measure the strength of this need, use was made of the following types of *indicators*:

Indicators for the need of bilingual dictionaries

1. Sociodemographic indicators
2. Economical indicators
3. Educational indicators
4. Cultural and Scientific indicators
5. Political indicators

To understand better what is meant with the respective *indicators* some examples are given for each of them in what follows.

1. The fact that, in countries such as the Netherlands and Flanders, not only the national language (Dutch) is spoken, but also languages of large groups of immigrants, such as Turks and Moroccans, is a strong argument in favour of a *connecting road* between these language communities.
2. Import and export trade serve, among others, as indicators for economical relationships between communities.
3. One could argue that the fact that one wants to learn the language of another community is a *derived* or *indirect* indicator, yet *educational* indicators are of importance to get an idea of the size of the need.
4. Some communities have a high cultural or scientific prestige or influence so that need arises to come into contact with them.
5. The fact that a political organization such as the EU not only grows in breadth (more countries), but also in depth (more domains that fall under its administration), makes it of extreme importance for countries and languages to find their way and define their position vis-à-vis the other.

Although the mentioned indicators only indirectly point at relationships and their strengths, yet they are of great help to select and prioritize between the otherwise quasi-unlimited mass of languages. Against this background, the concentric circles shown before have been *filled* as in figure 2:

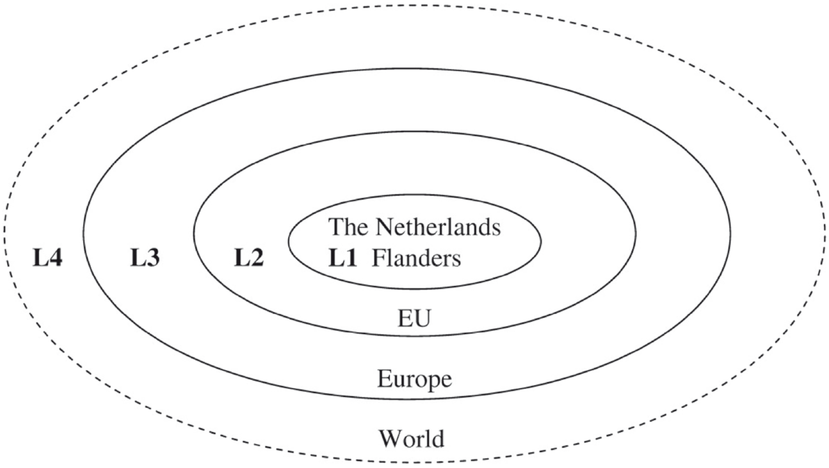

Fig.2: 'Prioritized' languages from a Dutch point-of-view
L1 = Arabic, Sranantongo,Turkish
L2 = (a) Danish, Finnish, Greek, Italian, Portuguese, Swedish (EU languages before 2004)
(b) Estonian, Hungarian, Polish, Czech (EU languages from 2004 onwards)
L3 = Norwegian, Russian, Rumanian
L4 = Indonesian, Korean

Notice that no circle has absolute priority, although certain indicators play a more prominent role in one circle compared to another (sociodemographic in the innermost, economical and political in the second and third, cultural, political and economical in the fourth).

2.2.2 Results

In the end 20 dictionary projects have been realized under the auspices or with the support of the CLVV, which itself resorts under the Dutch Language Union (Nederlandse Taalunie), the official institution for the Dutch Language. CLVV-paper dictionaries vary as to size and contents, the average one however contains 45,000 entries with a fairly rich microstructure and has about 800 to 1000 pages per volume. As a rule there is both a p- and an e-version. Table 1 gives a survey of the results with beginning and ending of editing phase and date of publication (+ publisher).

Project	Start-End	Publisher + Date
Dutch-Arabic v.v. Learners	1996-2000	Bulaaq, 2001
Dutch-Arabic v.v. Translation	1997-2002	Bulaaq, 2003
Dutch-Czech	1997	Leda, 1997
Dutch-Danish v.v.	1997-2001	Gyldendal, 2004

Project	Start-End	Publisher + Date
Dutch-Estonian	1997-2011	Pegasus, ?
Dutch-Finnish v.v.	2002-2007	Het Spectrum, 2012
Hungarian-Dutch	1995-2000	Akademiai, 2000
Dutch-Hungarian	1999-2002	Grimm, 2002
Dutch-Indonesian	1997-2002	KITLV, 2004
Dutch-Italian v.v.	1996-2001	Van Dale, 2001
Korean-Dutch	1999-2006	Hankuk University of Foreign Studies Press, 2007
Dutch-New Greek v.v.	1998-2007	Het Spectrum, 2008
Dutch-Norwegian	1998-2002	Boekwerk, 2007
Dutch-Polish v.v.	1996-2002	Pegasus, 2008
Dutch-Portuguese v.v.	1998-2002	Het Spectrum, 2004
Dutch-Rumanian	1997-2005	Pegasus, 2007
Russian-Dutch	1997-2002	Pegasus, 2002
Dutch-Sranantongo	2004	Het Spectrum, 2004
Dutch-Swedish v.v.	1995	Van Dale, 1996
Dutch-Turkish	1994-2007	Leiden University Press, 2012

Table 1: CLVV's bilingual dictionary projects

2.2.3 CLVV's approach to material and immaterial infrastructure

Financial resources were not fully used for the concrete dictionary projects as such, 20% of them being used 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 could not only tackle *hic et nunc* needs but anticipate future ones as well. Moreover, dictionary projects not elected for financial support could apply for the use of the generic tools developed, such as OMBI and the RBN. In what follows the main characteristics of both these tools are mentioned. For more details on OMBI we refer the reader to the literature (Martin/Tamm 1996; Maks 2007).

OMBI: main characteristics

OMBI = editor for ***OM*** keerbare ***BI*** linguale Woordenboeken (editor or Dictionary Writing System for Reversible Bilingual Dictionaries).

OMBI is a device to guide, structure and correct input data according to a pre-defined grammar (= comparable to other editors).

OMBI was (in 1996), and still is, innovative in that it has a REVERSAL FUNCTION which reverses language pairs with GREAT PRECISION.

OMBI links at SEMANTIC LEVEL and specifies the LEXICAL RELATION (for instance: hyponym, hyperonym, synonym etc. plus their constraints) between semantic units.

E.g.: KAART = card (for games), card (piece of paper), ticket (for entrance), map (topographical) when reversed becomes: CARD (for games) = kaart, (piece of paper) = kaart; TICKET (for entrance) = kaart; MAP (topographical) = kaart [+ examples under the correct lexical unit].

Another piece of material infrastructure, namely the RBN, is briefly presented in what follows (for more details see van der Vliet 2007).

Material infrastructure: RBN

RBN = ***R***eferentie ***B***estand ***N***ederlands (= Reference Database of Dutch).

Meant as an exemplary lexical database for the production and understanding of Dutch (monolingual lexical database).

Multifunctional (not only being of use for monolingual lexicographical purposes, but for bilingual and NLP-purposes as well).

Corpus-based.

Size: 45,000 entries, very rich and explicit microstructure.

The CLVV considered the existence of a high-quality editor and of a good monolingual lexical database of Dutch as a *condition sine qua non* for the *accessibility* of Dutch in a multilingual context. The fact, for instance, that an ‘own’ database has been constructed offered various advantages as shown below:

1. the possibility to focus on relevant aspects not present in other resources (such as the systematic treatment of collocations, complementation, semantic typology of items etc.),
2. the possibility to compare and link the different projects,
3. the possibility to offer data to non-prioritized projects,
4. independence vis-à-vis other providers such as publishing houses (no monopoly position).

Examples of *immaterial* CLVV infrastructure are the *linking method* (not translating items from one language into another, but linking two monolingual resources at semantic level, for more details see Martin 2003), and the *hub-and-spoke model* which is a model to go *beyond bilinguals* and move from bilingual dictionary making to multilingual dictionary construction. Figures 3, 4 and 5 illustrate in a simplified way what is meant by this (also see Martin 2004).

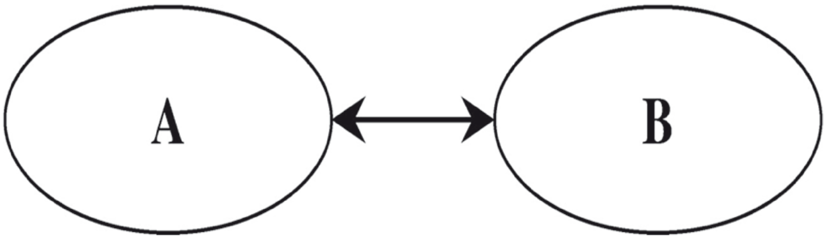

Fig. 3: The Hub-and-Spoke Model: linking two languages

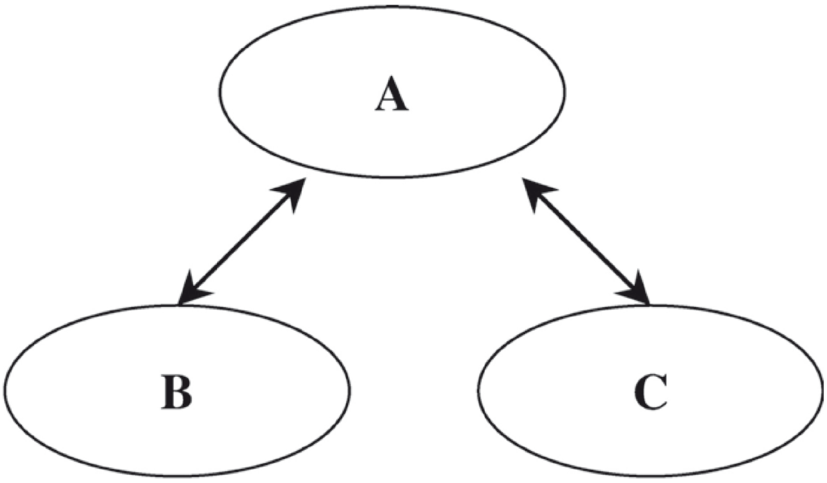

Fig. 4: The Hub-and-Spoke Model: linking two languages to one ‘hub’ language

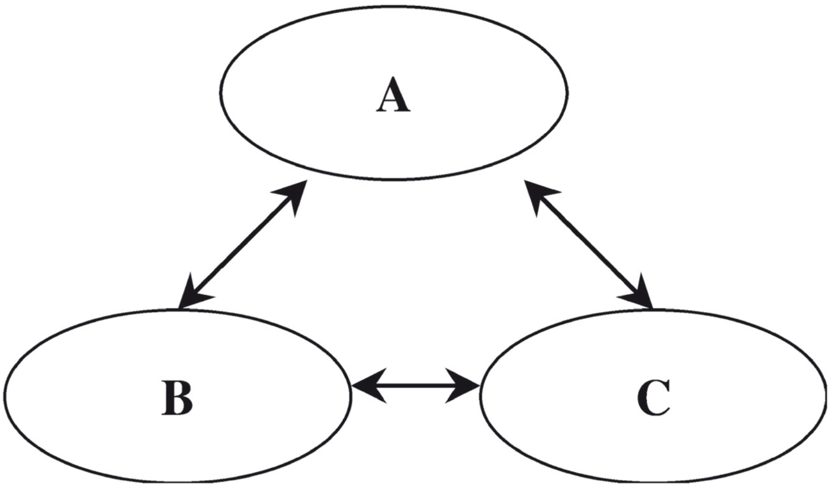

Fig. 5: The Hub-and-Spoke Model: Inferring links between two ‘spoke’ languages

3. Lessons and remarks for the future

In this last section I want to draw attention both to some strong points of the CLVV-project and to some of its weaknesses, and end with a couple of remarks on the changed and changing lexicographical scene.

3.1 Strong points

3.1.1 High productivity based on a well-founded plan

All in all in a period of less than twenty years, thirty bilingual volumes each containing 800 to 1000 pages have been finished. Without false modesty one can state that this is a result that even large language communities can envy.

This achievement certainly has contributed to strengthen the position of Dutch and its ‘partner’ languages not only at a linguistic, but also at a social, cultural, economical and political level. Moreover, the fact that most of the dictionaries published up till now are already sold out, requiring a second edition, proves that the dictionaries answered real users' needs, filling existing gaps for specific target groups.

3.1.2 Re-usable infrastructure

The CLVV has been active in the field of dictionary tool development and lexicographical infrastructure. Both OMBI and RBN can be regarded as advanced tools in this respect. Furthermore, the CLVV has considered dictionaries to be 'derived' products, underlying which were databases that should be reusable not only 'within' but also 'beyond' the Dutch context. This implies that, as a rule, the databases for Danish, Finnish, Portuguese etc. are reusable either alone or in combination with languages different from Dutch, thus opening perspectives for co-operation within a framework such as EFNIL. Another positive side effect of the choice for re-usable infrastructure is that for projects that are not selected for subsidizing, support still can be given in the form of infrastructure.

3.1.3 Metalexicographical impact

Next to the fact that the CLVV has been an active player on the lexicographical scene, it also has played a prominent metalexicographical role. In its projects the CLVV has tried to bring along its own, innovative view on the lexicon. Models such as the Hub-and-Spoke Model, functions such as reversing at semantic level (in tools such as OMBI) and the systematic organization of collocational data such as in the RBN, therefore are all grounded in the same concept of a relational, frame-based lexicon. Closely connected to this is the benchmarking function of the CLVV.

Because of the fact that the CLVV was quite successful on the lexicographical market – in fact no publisher in the Netherlands and Flanders was more active during the last decade – it served as a kind of benchmark in the field with a strong impact on quality assessment, standards and evaluation procedures in the Netherlands and Flanders.

3.2 Weak points

3.2.1 Time-management

Although the CLVV was quite successful in having its ambitious program carried out, it should have taken measures to come to a better time-management. As a rule projects needed more time than originally planned for. Bearing in mind that lexicographical projects are notorious in this respect and also taking into account the complexity of the task, the delay in most cases (two to three years) was still reasonable. Yet the CLVV could have done better if

- tools and infrastructure had shown less teething troubles (see further under 3.2.2);
- it had preferred, for certain projects, a 'slow' rather than a 'quick' start (see further under 3.2.3);
- publishers had been involved earlier in the project.

In most cases publishers were involved when the editing work was already finished with the result that often time got lost in coming to a final, both printed and electronic, version which fitted style, format and other desiderata of the publisher.

3.2.2 Infrastructural teething troubles

The fact that the CLVV had to develop its own infrastructure (RBN and OMBI) made its projects, though perhaps more innovative and challenging, also more vulnerable. The development from prototype to product (as, for instance, in the case of OMBI) took more time than expected and so, in the beginning, the teams did suffer from teething troubles of an infrastructure 'under construction'. Of course, things became better as time went by and the tools themselves profited from an intensive use but, with hindsight, it would have been better if the first projects could have disposed of less 'experimental' versions of both RBN and OMBI.

3.2.3 The know-how gap

The development of tools and data is one thing but perhaps as important as providing for infrastructure is providing for the necessary know-how to handle it. With hindsight, it would have been rewarding both for the teams and for the CLVV if the latter had provided for regular workshops in which people could have learned in a more systematic way how to make use of the infrastructure. The CLVV now provided for a help desk to solve infrastructural problems. This support certainly proved its value. However, next to that, a broader, more general kind of support in the form of workshops and/or master classes would have reached more people and enhanced expertise in a more systematic way.

3.3 The future

In what follows some remarks are given which should be taken into account when considering CLVV's heritage, its consolidation and follow-up.

1. The situation with regard to bilingual dictionaries for the Dutch speaking community anno 2012 is quite different from that anno 1993. The first priority now no longer is the *production* of bilingual dictionaries but their *consolidation* and *updating/outdating*. In other words, government, in co-operation with publishing houses, should see to it that the products delivered remain usable and up-to-date. A *wiki-environment* with online dictionaries as *freeware* could offer a solution here.
2. The follow-up and actualization just mentioned does not only regard the data but the *tools* and *infrastructure* as well. Tools not only need to be constructed, they also need to be *maintained*.
3. Government intervention does not stop once the aimed at products are delivered. Next to an action plan for the production, an *action plan for the consolidation, updating and follow-up of the 'deliverables' is now needed*.

   A new business plan should foresee and aim at the *integration/linking of the several resources developed* up till now in one all-encompassing database so that corrections/updating can be done centrally instead of for each database separately.
4. The fact that within the community of linguists the conviction has grown that, contrary to what was believed by generative grammarians, language cannot be fully captured by rules, has strengthened the *position of corpora as instruments of lin-*

guistic description, next to grammars and dictionaries. Consequently, the question of the specific role and contents of each of these three components becomes more and more stringent.

5. There is not only the challenge for the lexicographer to find out what should be put in the dictionary and what should be left in corpus and grammar, he should also find out how to explore the huge mass of data he is confronted with, nowadays. Next to quantitative devices, *qualitative exploration devices* such as *frames* or any other *semantic corpus query system,* can offer a solution to this problem.
6. One of the most striking differences between lexicography now and lexicography some twenty years ago is the switch from p(aper)- to e(lectronic)- or internet dictionaries. Modern dictionaries should not be replicas of p-dictionaries but constructed in their own right, from which p-dictionaries can be derived, if needed, in this order and not the other way round (also see Fuentes-Olivera/Bergenholtz 2011).

4. References

Fuertes-Olivera, P./Bergenholtz, H. (eds.) (2011): *e-Lexicography. The Internet, digital initiatives and lexicography*. London/New York: Continuum.

Granger, S./Meunier, F. (eds.) (2008): *Phraseology. An interdisciplinary perspective.* Amsterdam/Philadelphia: John Benjamins.

Maks, I. (2007): OMBI: The Practice of Reversing Dictionaries. In: *International Journal of Lexicography* 20*,* 259-274.

Martin, W. (2003): Lexicography, lexicology, linking and the Hub-and-Spoke Model. In: Botha, W. (ed.): *Festschrift vir Dirk van Schalkwyk.* Stellenbosch: WAT, 268-285.

Martin, W. (2004): Simullda, the Hub-and-Spoke Model and Frames or How to make the best of three worlds? In: *International Journal of Lexicography* 17, 175-187.

Martin, W. (2007): Government policy and the planning and production of bilingual dictionaries. In: *International Journal of Lexicography* 20*,* 221-237.

Martin, W. (2008): A unified approach to semantic frames and collocational patterns. In: Granger, S./Meunier, F. (eds.): *Phraseology. An interdisciplinary perspective.* Amsterdam/Philadelphia: John Benjamins, 51-65.

Martin, W./Tamm, A. (1996): OMBI: an editor for constructing reversible lexical databases. In: M. Gellerstamm et al. (eds.): *Euralex 1996 Proceedings*. Göteborg: Göteborg University, 675-685.

Martin, W./Theeuwes, J. (1991): *Lexicografische Vertaalvoorzieningen in het Nederlandse Taalgebied.* Den Haag: Ministerie van OC&W.

van der Vliet, H. (2007): The Referentiebestand Nederlands as a multi-purpose lexical database. In: *International Journal of Lexicography* 20, 239-257.