Structured Data + Automated Selection and Sorting = Dictionary

Rik Schutz
Van Dale Lexicografie bv
Postbus 19232
3501 DE UTRECHT
THE NETHERLANDS
rik.schutz@vandale.nl

Abstract
Van Dale have used and experimented with automated procedures in the dictionary making process. This paper describes and discusses three examples of the reuse of existing lexical material for new (editions of) dictionaries.

1. Distribution of fixed phrases (where will the user find the phrase *national anthem*: under the article **anthem, national** or both?); Sorting word meanings; in a relational database, the order of lexical items like word meanings is not necessarily fixed, whereas a dictionary is always ordered in a particular way;
2. We maintain two data files/dictionaries for each language pair, for example English-Dutch and Dutch-English. It would of course be ideal to have a single data set from which both dictionaries can be generated. We investigated the possibility of reversing and integrating the existing files.

Our preliminary conclusion is that the dictionary producing process could be much faster and better if previous work is recycled. Creating circumstances under which existing material can be reused in a useful and cost-efficient manner is a complicated matter that requires serious investment. The reason we should do it anyway is that in a rapidly changing environment it is impossible to say which properties make dictionaries fit for survival in an uncertain future. The best option therefore is to be flexible.

1. Introduction

Traditionally, conscientious lexicographers compensated for the insufficient time commercial publishers allowed for the job by investing unpaid time. These days, such dedicated professionals are becoming scarce. Moreover, the increased size and complexity of the average dictionary exceeds the skills of even gifted and experienced lexicographers.

Sophisticated editing tools give modern lexicographers smooth, digital access to the sources they need to consult. In theory this would increase the efficiency of the time invested. In practice however, the lexicographer is tempted to consult more sources and check more citations than before.

For a commercial publisher, reducing the human labour factor in the dictionary-making process is a goal in itself. At the same time, we want the quality of our products to be high. And we and to be flexible in making commercial use of the results of our lexicographic labour. Automating manual activities may kill three birds with one stone.

Today, computers put words in alphabetical order, a job previously done by lexicographers (or their children).

But there are other options. The making of new dictionaries or the revision of existing ones is supported more and more by the reuse of previously edited dictionary material.

2. Automated Dictionary-Making

According to Van Dale, the ideal procedure for bringing about a (new edition of a) dictionary is as follows:

- Maintaining a central database – ideally as an ongoing process;
- Determining the profile (content and form) of a specific title;
- Selecting and arranging the data required for that title.

This procedure requires a product-independent database. Product-independent storage means that lexicographical data of all types can be stored in standardised form, independent of title or user profile. Examples of the various types of data are: spelling, pronunciation, definitions, translations, synonyms.

Whether spelling, hyphenation, pronunciation or inflexion can be standardised in a title-independent format, will not be argued. I expect less unanimity about issues like:

- the distinction between homography and polysemy;
- criteria for the distinguishing (sub)meanings;
- the order of meanings within an article;
- the wording of fixed phrases;
- the entry under which fixed phrases should be recorded;
- the order of fixed phrases within an article;
- the question whether X is a translation of Y if Y is a translation of X in the complementary volume.

Dictionaries differ on each of these issues. However, these differences rarely go back to explicit choices accounted for by the lexicographer in the front matter of the dictionary. Dictionaries of approximately the same size and intended for the same user group often differ substantially, even if appearing under the same brand name.

Individual lexicographers seem to follow tradition, their intuition, or personal preferences when deciding about the issues stated above.

In this paper, I will discuss three different examples of lexicographical operations:

- selecting the entry under which a fixed phrase will be found;
- determining the order of meanings within an entry;
- reversing a bilingual dictionary.

In the third edition of the *Van Dale Groot woordenboek hedendaags Nederlands* (comprehensive dictionary of contemporary Dutch), articles were composed as follows.

At the level of the concept of the dictionary, the publisher and editor in chief:

- set criteria for distinguishing homographs;
- set criteria for arranging the meanings within an article;
- decided on the way fixed phrases will be distributed among articles (see for details the paper by Jaap Parqui in this same volume);

- set criteria for a lemma to deserve pronunciation;
- selected the various types of semantically related lexical entities (synonyms, hypernyms, antonyms) to be included in the articles.

At data level, they determined that:

- any new items to be included in the new edition would be added to the database containing all the lexical entities (word meanings plus fixed phrases) from the Van Dale contemporary dictionary series with Dutch as L1;
- the editorial staff would select a volume of lexical entities that fit a volume of 1,600 pages.

The above criteria were used, together with a number of other, similar criteria, to extract a file from the database containing the content of the new dictionary. The file was then typeset and printed.

3. The Placement of Fixed Phrases

As a consequence of the approach outlined above, the new edition differs substantially from the one previous, even in cases where no information was added or deleted.

Many articles now list a great number of fixed phrases that were scattered over a number of other entries in the last edition. For example: the first meaning of **nationaal** now has 22 fixed phrases, where before it had only 6. All 22 were available in the previous edition, but then they were spread among several entries.

Because each word in each fixed phrase has been coded in the database (with the exception of grammatical words), distributing fixed phrases is rather easy, as I'll show by means of the example *nationale kampioenschappen.*

In the database, a fixed phrase links to each of the words used in that phrase. The set of rules that we created for the printed version of *Groot woordenboek hedendaags Nederlands* created a hierarchy in word category: noun, adjective, verb. The first noun (in our example **kampioenschappen**, the only noun present) determines under which article the phrase will enter the dictionary. The phrase will also be placed under the second word in the hierarchy, in this example the adjective **nationaal**.

Within the article **kampioenschap**, the second word in the hierarchy serves as guideword, which is used to place the expressions in alphabetical order. In our example ***nationale*** *kampioenschappen* comes before ***open*** *kampioenschappen.*

Within the **nationaal** article, the noun **kampioenschappen** serves as guideword. Here also, phrases are placed in alphabetical order of guideword. *Nationale* ***kampioenschappen*** therefore comes after *nationaal* ***inkomen*** but before *nationale* ***reserve.***

4. Sorting Meanings within an Article

For the *Groot woordenboek hedendaags Nederlands* we investigated whether automated sorting of word meanings was feasible and desirable. We set the following general rules for the sequence of word meanings:

- General comes before specific
- Frequent comes before rare
- Current usage comes before obsolete

These rules were also used as instructions for the editors of the previous edition. What made them difficult to follow, is that characteristics such as the ones mentioned here are not so easy to measure. Measuring word frequency is quite easy; there are some sophisticated corpus tools around. Measuring the frequency of word meanings is a different matter altogether; unfortunately the corpus tools are not yet advanced enough to help us out here.

For the new edition, we carried out an experiment. We wanted to see if we could use the features of the lexical entities available in the database to determine the order of word meanings within an article. We calculated 'generalness points' and assigned them to each word meaning and then ranked the word meanings according to their score. We used two different points systems and considered a third.

In our database, each lexical entity has its place in the semantic hierarchy of the database: each lexical entity is linked to one or more hyponyms and in each cluster of synonyms, one has been given the status of central term. We can measure the 'generalness' by assigning points for each hyponym and synonym linked to a word meaning. The higher the number of points, the more general the word meaning.

The second system we used to measure the generalness of a lexical entity, was to count the number of fixed phrases related to it in the database. Word meanings occurring in a large number of phrases score higher on the generalness scale than those occurring only in one or two or even none.

A third system to determine whether a word meaning is general or specific, would be to take into account whether it is supported in the database by a domain-specific subject label. If a word meaning is labelled as medical or technical, it is likely to be less general than a meaning without such an indication. A label such as 'obsolete' could be given a negative value.

We did not put this third points system into practice, but limited ourselves to the first and second systems (points for *semantically related items* and points for *the number of related fixed phrases*, respectively). We then used the points given to each word meaning to automatically determine the order of word meanings. In the article **mol**, for example, the two meanings were placed in the following order (the number at the beginning of the line represents the number of points):

1 mole (animal) – 1 fixed phrase (*blind als een mol*)
0 mole (spy) – no phrases, no preferred synonym status

In the article **boom**, the scores and consequently the order of word meanings turned out as follows (again, the number at the beginning of the line represents the number of points):

41 tree – 23 fixed phrases; 18 hyponyms
4 pole – 4 hyponyms
0 diagram – no phrases, no preferred synonym status

In approximately 10% of all polysemous articles, the order of the word meanings changed in comparison to the previous edition. In the majority of cases, the new sequence was closer to the chosen principles.

However, many examples resulted in changes in the sequence of word meanings that weren't improvements. For example, the article **roos** now looked like this:

1. rose (flower); 2. bull's-eye; 3. rose (plant)

Most lexicographers and dictionary users would prefer the two related meanings 'plant' and 'flower' to follow one another. (This could of course be solved by re-adjusting the rules.)

A second drawback of automated sorting was that we were confronted with the legacy of a now-rejected defining style that was still present in some of the older definitions. Some definitions are worded by building on the previous definition. If we then change the order of the word meanings, and thereby the order of the definitions, we're faced with a serious problem. If we take the example **publicatie** (publication) the new sequence would be as follows:

1. any paper or book that contains such printed information; 2. the act of making information or stories available to people in a printed form

This is obviously not acceptable.

Thirdly, we were faced with a tight time schedule. Programming for the automated ordering of meanings would have required a serious workload for our programmers.

We therefore decided to reject this innovative use of our database for determining the sequence of word meanings, and to fall back instead on the order available from the previous edition of the dictionary. However, we learned from our experiment: automated sorting of meanings is not out of the question, but it does require an even more sophisticated database to obtain satisfactory results.

5. Reversing Bilingual Dictionaries

5.1 The Rough Way

When Van Dale started its series of bilingual dictionaries from scratch in the late seventies, the English-Dutch volume was written first, before the complementary Dutch-English volume. The same was true for the German and French bilingual volumes. The Dutch-Other Language (L2) volumes were all based on the same inventory of current Dutch.

After the L2-Dutch volumes were finished, the editors of the Dutch-L2 volumes were given a skeleton of Dutch dictionary articles: entry, part of speech, short indication of meanings and fixed phrases. They were also provided with an automatically generated reversal of the complementary volume. At that time, around 1980, automated processes had

not reached the level that we have gotten used to in the last two decades. Each Dutch translation was simply multiplied by the number of words it contained and then put in strictly alphabetical order.

For example, the English phrase words to that effect in the article **effect** resulted in the following reversed list:

```
woorden    woorden van die strekking     words to that effect
van        woorden van die strekking     words to that effect
die        woorden van die strekking     words to that effect
strekking woorden van die strekking     words to that effect
```

Then, the entire reversals list was put in alphabetical order. If we were to look up the word 'strekking', we would find a large number of reversed translations, including the one reversed from *words to that effect*:

```
strekking                                   extension
strekking de strekking van een zin          the sense of a sentence
strekking hij begrijpt de strekking nooit   he always misses the
point
strekking wat was de strekking van de brief?what did the letter
purport?
strekking woorden van die strekking     words to that effect
...
```

Because the Dutch skeleton lacked the phrase *woorden van die strekking*, the reversal was added tot the Dutch-English volume, with the original English phrase as its translation.

Although ploughing through the huge quantities of often useless 'suggestions' was a laborious job, the Dutch skeleton for each of the three bilingual dictionaries (English, German and French) was substantially enriched. However, due to the huge quantity of output, the tight time schedule and the lack of experience with this kind of material, a few things went wrong too.

Unfortunately, a lot of valuable material was neglected, buried under the enormous quantities of waste. Also, enthusiasm about a beautiful translation sometimes resulted in rather useless Dutch being added to the dictionary. In Dutch-French the phrase *een miljoen oude franken* ('a million ancient francs') was added to the entry **frank** with the translation *une brique*. Some twenty years after the French currency was re-valued at a factor of one hundred, the addition of a French slang translation for a highly unlikely Dutch phrase wasn't a particular improvement.

One of the important things we learned from this exercise, is that it makes sense to distinguish between a translation equivalent and a descriptive translation. The latter should per definition not to be included in the complementary volume.

5.2 The Do-It-Yourself Way

Ever since the bilingual dictionaries have been available on CD-ROM (in both directions), users can decide for themselves whether to consult the dictionary in the obvious, traditional direction, or the other way round. When I need an English translation for a Dutch word, I often go directly into the English-Dutch volume. By using the full text search mode, I am given all the English entries that contain the Dutch string I wish to translate. Not only do I encounter all the useless material that was rightly neglected by the lexicographers who worked with the reversed lists when the dictionary was created, but I often find very useful translations.

I'll give just one example of this advantageous look-up method. If I want to translate the Dutch word woordenboek, the Dutch-English volume provides me with *dictionary*, *lexicon* and *wordbook*. However, the English-Dutch volume also gives me *thesaurus* among the list of entries containing *woordenboek*. In those contexts where woordenboek is used in the sense of thesaurus (Dutch has no specific term) thesaurus is obviously the best choice.

My personal experience with the CD-ROM encourages me to think in a new direction: we might increase the value of our dictionaries by organising easy access to such useful translation equivalents in the 'wrong' volume. The traditional division into two volumes keeps users away from useless translations, but it also stops them stumbling onto treasures.

5.3 The Indecent Way

In a very recent experiment, we produced a non-branded series of cheap bilingual dictionaries. We created three sets of bilingual dictionaries to and from Dutch using one multilingual but unidirectional database. We used essentially the same reversal technique as outlined above, but it was applied with a better understanding of what could go wrong, using more powerful computers, better programming skills and superior content. The content was better in the sense that 'examples' were divided into categories such as proverb, fixed phrase, collocation and illustrative example sentences. That way, illustrative examples could be excluded from the reversal. It was also better in the sense that descriptive translations were now coded as such, so that the automated reversal could be limited to true translation equivalents.

Language technology enabled us to convert a list of rough reversal output into structured dictionary articles. In some cases, the resulting articles were very impressive. Despite the fact that we reversed a concise dictionary, the article dictionary that resulted from the reversal contained the following items:

biographic ~, explaining ~, bilingual ~, crossword ~, desk ~, foreign-language ~, multilingual ~, polyglot ~, rhyming ~, technical ~, specialist ~.

By way of comparison, in our comprehensive English-Dutch dictionary the entry *dictionary* does not have a single collocation, idiom or example and only three of the compounds listed above occur in the entry word list.

The reversal is obviously incomplete. For example, *pocket dictionary*, *pronouncing dictionary* and *translation dictionary* are missing from the range of dictionaries. Also,

reversals lack all English that is not the result of earlier translations of Dutch items into English.

Even so, if we compare our 'indecent' English-Dutch dictionary, containing roughly 30,000 entry words, to a dictionary of the same size, we might say the results are quite acceptable.

6. Conclusion

Like animals and plants, dictionaries need to be fit to survive. In a rapidly changing environment it is impossible to say which properties make dictionaries fit for survival in an uncertain future. The best option therefore is to be flexible.

The dictionary producing process can be much faster and better if work that has been done previously is recycled. Creating the circumstances in which existing material can be reused in a useful and cost-efficient manner is complicated and requires serious investment. The reason we should do it anyway is not that we want cheap success, or because we are of the illusion that the work of expensive lexicographers could ever be eliminated. Instead, the expertise and creativity of lexicographers should be used to bring the quality of the results of automated recycling up to an acceptable level, and beyond.