Dissecting a Dictionary

Krista Varantola
University of Tampere
33014 University of Tampere, Finland
krista.varantola@uta.fi

Abstract
This paper has two main aims. The first is to investigate how a dictionary database for an English-Finnish (E-F) dictionary can be recycled and reused for a Finnish -English (F-E) dictionary by manipulating the original data; the second aim is to study the possibilities which dissecting the dictionary data in various ways will offer to the editors of dictionaries. In other words, the issue is how dictionary content and consistency could be controlled and monitored by means of analyzing the accumulating data and how different dissections of the data could help editors and lexicographers during the compilation process.

1 Background

The dictionary database under investigation in this study consists of data that was originally used to produce a set of electronic English-Finnish (E-F) and Finnish-English (F-E) dictionaries. This electronic dictionary collection is widely used in Finland and popular both as a single user CD-ROM version and as an intranet-based, on-line version in businesses and academic environments. One reason for the popularity of this set of electronic dictionaries is naturally that English is today a lingua franca in business and academic contexts and that most Finns working in those environments must regularly read and produce texts in English. It is, however, somewhat unclear how this dictionary database came about, i.e. there are scant records of how the data was compiled and what went into it, what type of editorial principles and compilation guidelines were used, whether they were followed consistently and how experienced the lexicographers employed were. Furthermore, it is also unclear what kind of other databases have, in the course of time, been merged into the master database and incorporated in the final electronic dictionary products.

Until recently, the language technology company in charge of the state-of-the-art technology of the electronic product also managed the compilation of the dictionary content. The situation has now changed and an experienced print dictionary publisher is involved as content provider. All further development thus takes place in cooperation between the language technology company and the print dictionary publisher. The aim of the publisher is, in the near future, to produce a set of user-friendly E-F and F-E print dictionaries which are based on the existing electronic data.

The genesis of the electronic and print versions is thus unusual, because the common direction is from print to the electronic format. We all know that this tradition has often resulted in the complaint that electronic dictionaries are merely electronic versions of the print dictionaries with few innovations that utilize the potential of the electronic format. This is no longer entirely true and at the moment, the two formats are often developed as separate concepts.

When publishing bilingual dictionaries for a small language community such as Finnish, publishers have to be very cost-conscious. It is, for instance, not feasible to think in terms of producing two separate sets of E-F and F-E dictionaries for the target audiences, one set for the Finnish speakers with passive (comprehension) needs and active (production) needs in English and the other for the non-Finns with comprehension and production needs in Finnish.

One way of overcoming this problem of having to keep in mind both active and passive users is to incorporate a large number of usage examples and their translations in the entries of both the E-F and the F-E dictionary. And this is indeed what has been done in the E-F print dictionary. It is now at its completion stage and the original data has been edited and systematized and is now ready for the next macro-editing stage. Corpus examples have been added to a large number of entries to illustrate the use of the English headword in its typical contexts and with its typical collocates. The examples are to a large extent based on sentences from the BNC which have in turn been translated into Finnish. The principle has been to provide translations that are both accurate and idiomatic Finnish. The translations will thus provide information on the range of use of the English headword and also give an idea of what a corresponding Finnish sentence would be.

It could be claimed that the addition of examples and their translations is this dictionary's answer to the contextual turn which dictionary production in general has taken over the past decade or so. The reason for this turn is obvious. Contextualization became possible thanks to large corpora and tools which enable a multi-faceted analysis of corpus data. The contextual turn is also something that dictionary users, particularly professional users such as translators, warmly welcome. On the other hand, print dictionaries have much more limited possibilities in exploiting contextual information than electronic dictionaries. The decisions that print dictionary editors take about the inclusion of contextual material thus have to carefully balanced by taking into account both the value of the information provided and space restrictions.

2 The Research Question

The question now is how the present E-F database could be used in the process of compiling the material for the F-E dictionary that is next on the production line. If the E-F material could be automatically or semi-automatically manipulated to provide reliable data for the F-E dictionary, major savings could be made during compilation. The main task is thus to explore to what extent the dictionary database could be reversed. Another task is to find out how the database could be manipulated for macro-editing purposes and future development. In other words, we need to examine what types of displays of the database content would help the editors

- to improve the consistency of the content of the final product,
- to control the length of the dictionary,
- to systematize and balance entry information
- to detect unwanted bias in content selection, etc.

In the following sections, I shall discuss methods and data presentation modes that could speed up dictionary compilation and highlight the content of the database in ways that would

be helpful to the dictionary editor. The examples are based on an interim version of the database that is being used for the E-F dictionary.

3 Reversing the Dictionary

In a different context, I have claimed that bilingual dictionaries are actually a contradiction in terms, but because they exist and are well-liked, this claim cannot be taken too seriously (Varantola 2002:36). The claim is naturally based on the general observation that there are few if any real synonyms in a language, words that could replace each other in all thinkable contexts without preference for one or the other alternative. If this is true about one language, how could there be fully equivalent words between two languages? In that sense, dictionary equivalents or translations are actually types of keys to the meaning of the headword rather than fully exchangeable building blocks in another language.

However, as Oppentocht and Schutz (2003:225) point out,

> «Very often a translation Y for word X in section L1-L2 is not even an entry in the complementary part L2-L1 of the dictionary, and if Y is an entry, it does not always have X as a translation. It must be said that in many cases this is due to careful judgement on the part of the bilingual lexicographer, *but in many more it is simply the heritage of a period in which dictionaries were compiled with inadequate tools and too little time for checking and comparing.*» (My italics)

And indeed, there are a number of entries in bilingual dictionaries for which a single equivalent is considered sufficient. These would presumably also be good candidates for reversal. Oppentocht and Schutz further state that

> «In the year 2003 it is still common practice in many publishing houses to work with autonomous editors for each title. We predict that soon a central and co-ordinated storage of lexical data will replace this procedure.»

3.1 Finnish Equivalents in the E-F Dictionary

The analysis shows that the raw data contains over 80, 000 Finnish equivalents and that the majority of them occur only once in the equivalent field. This Finnish equivalent list can be matched against the tentative headword list of the F-E dictionary or against other bilingual or monolingual dictionaries for comparison and for decisions about the intended coverage of the F-E dictionary.

It is not surprising that the number of equivalents per letter varies a great deal, in the same way as the number of headwords per letter varies in any dictionary. What is interesting, however, is that the different letters show internal consistencies as well. For example, K is the most common initial letter and in the raw data the number of Finnish equivalents (types) starting with K is over 12, 000. Out of them, well above 60% are equivalents used only once in the E-F data. The number of equivalents starting with H is over 5, 000 and again about 60% occur only once. The tendency seems to be the same throughout. The share of one-off equivalents seems to account for around 60% of the equivalent types.

This list of single occurrences per letter would obviously be a good candidate list for reversal. However, a thorough clean-up would need to be done first to get rid of all non-headwords in Finnish. On perusal, the first impression is that these are mainly paraphrases of English headwords that have no clear equivalent in Finnish. Many are multiword

explanations and ad hoc compounds describing culturally-bound English headwords or phrasal headwords. Thus it would seem sensible to start the cleaning up operation from the Finnish equivalent list first, before going on to the full entry in the E-F dictionary.

There is obviously no reason why this hypothesis of reversability should not also be extended to equivalents that occur more than once. Even they may provide useful data for the reversion process and also provide ideas of potential cross-reference needs for related expressions, as well as suggest senses and shades of meaning that may have been neglected in existing dictionaries which are based on traditional and more piecemeal compilation principles.

Furthermore, the equivalents which occur a great number of times need to be studied separately to see why they have found their way into the equivalent field of so many E-F entries. The first results seem to indicate that these words are very general words and often parts of a paraphrased entry. Such words are, for instance, the Finnish «equivalents» for *break, drive, give, go, good, hard, hand (n), put, reach, take, time, use, etc.*

A spot check of these common, semantically relatively empty words, revealed that they are typically explanatory additions to more specific equivalents. They are also used to imply that a more general translation or paraphrase of the headword may also be adequate. However, they are rarely given as first translations of the headword. Indeed, one gets the impression that in many cases the lexicographer has just tried to cover his or her back by using them as translation equivalents. Their usability for reversal purposes seems to be minimal. As a matter of fact, they are often far too general even as translation equivalents and could in many cases be deleted from the E-F database, at least for the print version where space is an issue. So in a sense they may turn out to beuseful for an unexpected purpose, a further editing or cleaning up of the original (E-F) database.

In the middle category, we find words which are semantically more «meaningful» but still relatively general. This means that the interpretation of the exact sense also highly depends on the context they are used in. A good and illustrative example is the verb *hajottaa* (core sense ='*break*') in Finnish. An existing large F-E dictionary gives five basic sense divisions for *hajottaa*:

1 disperse, scatter, break up, dissolve, dismiss, disrupt, dismiss, disband, separate
2, break down, dismantle, take to pieces, undo, untie
3 decompose
4 break, knock down, tear down, demolish
5 dissipate, dispel, dissolve, disintegrate

The analysis of the database shows that *hajottaa* occurs over 30 times in the 'equivalent' field of the database, usually on its own and as a potential translation equivalent to the following 24 verbs:

bang, break, decompose, degrade, demolish, diffuse, disband, disperse, dissociate, dissolve, divide, knock, pull, pull to pieces, ***rip****, scatter,* ***smash****,* ***split****, spread, take, tear,* ***trash****, unbuild, undo*

Had the editors of the above print dictionary had this type of information at their disposal, they might have considered including the ***rip, smash, split, trash,*** type of meaning potential in the entry of the Finnish headword *hajottaa.*

In the present dictionary project, the lexicographers have access to this type of cross-referenced information and can thus go on to the English corpus material to look for illuminating usage examples. They can do this by using the English headwords of a particular translation equivalent in the E-F dictionary as search words for concordancing. Gathering multiple and multi-faceted information in this way also gives the lexicographers a chance to work out potential sense divisions more systematically and according to the editorial principles agreed on. This type of information will also help the editors in deciding which equivalents to include in various types and sizes of dictionary. Furthermore, it can be assumed that the editors will gain other novel insights about equivalence when they compare the results of the reversed list with headword lists compiled according to more traditional methods.

3.2 Reversability of Examples

The purpose of the examples in the E-F dictionary is to make it more user-friendly and to cater, in addition to comprehension needs, also for production needs. Dictionary use studies have shown (Atkins & Varantola 1997) that experienced users often look for reassurance in a dictionary entry. They may have an equivalent in mind when writing in their L2, but they are unsure of its collocational behaviour or range of use. If a bilingual L2-L1 dictionary gives them adequate and insightful usage examples, users may find the answer they are looking for already in the bilingual dictionary without having to go on to a monolingual L2 dictionary for reassurance. Furthermore, if the user of a bilingual dictionary between a world language and a small language is a non-native speaker of the small language, the translated usage examples will give this user an idea of how things are expressed in the small language which has no tailored production dictionaries for non-native speakers.

In the E-F dictionary, the bulk of the usage examples come from the BNC. The Finnish examples are translations of them (see above). It seems quite feasible to also use the same example corpus in the F-E dictionary, which is basically a production dictionary. The examples can be organized in the form of a parallel corpus so that it is easy to move in either direction when reorganizing them for the F-E dictionary.

4 Macroediting

4.1 The Example Data

The way the example corpus is organized also makes it possible to reorganize the example data alphabetically or any other systematic way on the basis of the English headword and the entry in which the example occurs, for example:

Myrsky **hajotti** aidan ~. /**asunder**
Tämä saattaa johtaa parlamentin ***hajottamiseen.*** */dissolution*
Parlamentti on ***hajotettu.*** */dissolve*
Yhtye ***hajotti*** *hotellihuoneen palasiksi.* */smash*
Punkrockin soittajat ***hajottivat*** *konserttisalin.* */trash*
Poliisi yritti ***hajottaa*** *väkijoukkoa.* */disperse*

In addition, it is possible to view all the other examples in which the search word also appears, although it is not the focus word. In this way the lexicographer can check even the nuances of those examples for reuse in other contexts.

An editor can exploit the different types of statistical analyses and dissections of the example databases and gain novel insights and new profiles of the dictionary content. The dissections will also enable macro-editing at different stages of the production process. An existing database can be edited in advance to form the basis of a new dictionary and consistency checks can be done after the first draft version of the whole dictionary becomes available.

4.2 Consistency

All dictionary production teams aim at writing systematic and consistent entries. Detailed editorial guidelines are used to reach this aim but nevertheless individualistic, non-systematic solutions tend to creep in during the compilation process. The individual lexicographers' use of usage notes and labels can, however, be followed «on-line» while the work is in progress by means of cross-tabulations and selective listings. For example, the use of special field labels can be monitored. In this way it is possible to detect whether individual lexicographers have developed idiosyncratic labelling habits or whether inconsistencies have crept in during the compilation process. Gaps and bias in the coverage of different fields will also come to light. It is thus much easier for the editor to determine, if individual lexicographers have adopted idiosyncratic labelling habits, or to decide that a certain field is poorly presented or over-represented.

4.3 Potential Bias

Likewise, it is possible to study the scope and extent of the vocabulary used in the examples, and notice whether the examples display sexist or otherwise offensive tendencies. Another possibility is to investigate whether the vocabulary used in the examples is too narrow, too simple or too complicated. In the present example data, for example it is intriguing to learn that personal pronouns and words expressing marital, family or blood relations are so common that a WordSmith Tools analysis marks them as keywords, when the English example data is compared with a reference corpus consisting of journalism. Thus

He, I, His, She, Her, You, My, They, We, Me, Him, Your, Our, I'm, He's, Us, She's

Husband, Wife, Grandmother, Mother's, Motherhood, Mothers, Mother-in-law, Stepmother, Godmother,Grand-mother's, Grandmother's, Granmother, Mother, Step-mother, Father, Father's, Grandfather, Godfather

are all keywords, because they appear so frequently in the example data. Whether this is sensible is an entirely different matter, but the keyword analysis highlights the issue and allows the editor and lexicographers to adopt a different policy if so desired.

A superficial collocational analysis displays another interesting tendency. *He* is used over 3000 times in the examples, whereas *she* is used in just under 2000 cases. The collocational patterns of these pronouns vary a fair amount. *Hair, children child, beauty* collocate with *she* in the examples and not at all with *he. Drink, drinking and drinks, murder, cigarette, death* are entirely associated with *he*-behaviour. *Smile, face* and *eyes* co-occur much more frequently with *she*, and *car, career* and *game* clearly collocate with *he. Family, school, love, pain* and *money* are, however, gender-neutral collocates in this database. It is too early to say whether this is a bias that reflects the collocational patterns of the BNC, or whether the choices reflect subliminal choices, «real-life» examples or taboos in society which the lexicographers, both female and male, have not been aware of.

The ways to study dictionary anatomy are numerous, as long as the database allows the use of filters and cross-tabulations and thus enables free variation in the way in which the data is displayed. In other words, it should be possible to list, for instance, all the headword nouns that end in *-ogy* and have the usage label *'medicine'* and have been prepared by a certain lexicographer. In this way, it would be possible to monitor the consistency of any data type information and make spot-checks where necessary. Moreover, the different displays of data and macro-level checks will naturally help to improve the guidelines while the actual work is in progress. The fact that the editor can peruse different data type lists separately will also contribute to overall systematization and help to counteract and combatindividualistic and idiosyncratic analysis habits.

The length of the entries can also be monitored at macrolevel. Examples that are too long can be found and edited together with their translations. Frequency information can be applied to modify the headword list. Furthermore, the reversed headword list can be compared with other available headword lists and special field glossaries. Concordancing can be used to study collocations and potential bias in the selection of examples. Alphabetical wordlisting and concordancing also reveal the alternative spelling variants that have been used in the dictionary database.

5. Conclusion

Overall, the application of various corpus analysis tools to dictionary content data will help in the maintenance of the database. Together with the dictionary editing tool, they provide a powerful toolset both for micro-level and macro-level dictionary work. For this study, most of the manipulations had to be made in a fairly complicated fashion, because the analysis

tools are not integrated with the editing tool and the data thus needs to be pre-edited and imported to different tools to produce the desired profiles. This problem should, however, be fairly easy to overcome. Furthermore, if BNC-style comprehensive corpora and WordSketches- style software become available for other languages than English, dictionary compilation and editing could be automated to an increasing degree. This in turn should make dictionary production a much more affordable business even in smaller language communities.

References

Atkins, B.T.S and K. Varantola. 1997. «Monitoring Dictionary Use». IJL 10/1(International Journal of Lexicography). 1-45.

Kilgarriff, A. and D. Tugwell. 2002. 'Sketching Words' in Marie-Hélène Corréard (ed.) Lexicography and Natural Language Processing A Festchrift in Honour of B.T.S. Atkins. Euralex 2002. 125-137.

Oppentocht, L. and R. Schutz. 2003. 'Developments in Electronic Dictionary Design' in Piet van Sterkenburg (ed.) A practical Guide to Lexicography. Amsterdam. John Benjamins. 228-239.

Varantola, K. 2002. Use and Usability of Dictionaries: Common Sense and Context Sensibility? in Marie-Hélène Corréard (ed.) Lexicography and Natural Language Processing A Festchrift in Honour of B.T.S. Atkins. Euralex 2002. 30-44.