Kernerman

http://kdictionaries.com/newsletter.html

DICTIONARY News

Number 10 ▮ July 2002

Dictionaries, another Netscape?

Joseph J. Esposito

The article 'The Coming Boom in English Lexicography: Some Thoughts about the World Wide Web (1)', by my friend, Charles M. Levine, is a very nice work. I will add one small item, then expand upon it: it is not exactly true that Microsoft created their own dictionary to avoid paying royalties to Houghton Mifflin. The royalty issue was part of it, but they also wanted more control over the database. The real game for Microsoft is using lexical databases within computer algorithms, as in natural-language processing. No dictionary on the market today is built for that application. In other words, Microsoft now views lexical databases as an aspect of strategic technology, not simply an aspect of marketing. In this respect, Microsoft cares more about their dictionary than about their encyclopedia.

I do not disagree with Levine's comment on how long the "old" dictionary business will be around. Who knows? It's also not important. In the absence of growth, the old business will be strained for capital, which will beget smaller investments, which will in turn hasten the decline. In the short term, this will redound to the benefit of market leaders, such as Merriam-Webster and Oxford University Press, yet people underestimate what bundling with Windows can mean. There used to be – used to be – a company called FTP Software that created a utility that linked a PC to the Internet. Now that utility's clone is built into Windows. Buy any FTP stock lately?

I can add that my grim vision (from a reference publisher's point of view) of Microsoft originated in the 1980s, when I first got involved with dictionaries as the

This contribution is derived from email correspondence following the publication of Charles M. Levine's The Coming Boom in English Lexicography: Some Thoughts about the World Wide Web (Part One), in Kernerman Dictionary News, Number 9, July 2001 (http://kdictionaries.com/newsletter/kdn9-1.html). Levine's Part Two is due next year.

Tenth *Intermezzo* Issue

Editor ▮ **Ilan J. Kernerman**

The views expressed in the articles are those of their authors.

K DICTIONARIES LTD
Nahum 10 Tel Aviv 63503 Israel ▮
tel 972-3-5468102 ▮
fax 972-3-5468103 ▮
kdn@kdictionaries.com ▮
http://kdictionaries.com ▮

Joseph J. Esposito is an independent consultant, specializing in change from one media to another and in developing strategic overview and management discipline. His publishing experience includes, among others, work for Simon & Schuster and Random House, and managing the brands of Merriam-Webster's dictionaries and Encyclopedia Britannica. Mr Esposito initiated the first Internet encyclopedia Britannica Online, became the Britannica CEO, and effected its sale in 1996. He was subsequently CEO of the Internet company Tribal Voice, and has served in various advisory positions, including Board seats at CUseeMe Networks, MIT Press, and Navilinks.
espositoj@att.net

The author's essay, The Processed Book, elaborates on some of the aspects raised here and issues related to the publishing industry. Preview online: http://kdictionaries.com/newsletter/kdn10-probook.html

publisher of Webster's New World, and it consolidated in 1991 while I was running Merriam-Webster and began negotiations with them. (Disclosure: Prior to joining Merriam-Webster, I served as a consultant to Microsoft, though not having anything to do with dictionaries.) In perspective, when I complained about Microsoft bundling a spell checker, with its limited dictionary, into Word ages ago, the techies I knew all laughed at me. Now that most of them have burned through their venture capital after Microsoft "integrated" the gist of their products into Windows, we all cry into our lattes together.

Most discussions of lexicography and dictionaries focus on two items: the future market for printed dictionaries and the opportunities of making dictionaries available in electronic form. The flow of these discussions is predictable. The markets will grow because of:
(a) globalization,
(b) the special place of English (hence English-language lexicography) in the world economy, and
(c) the migration to a knowledge-intensive society – "feed your head", as Jefferson Airplane said.

The next step of this discussion is to try to figure out where the world of print ends and the world of electronics begins. This is an awkward step for publishers because once you go down this path, it is hard to make a case for any mainstream dictionary publisher surviving long-term except for Microsoft, for the simple reason that Microsoft's dictionaries are being integrated with various Microsoft products, giving them instant (and free) ubiquity. This means that even the most distinguished dictionary publishers will soon have to take for granted that every single one of their prospective customers already owns a good- enough Microsoft dictionary. Publishers therefore will (predictably) flock to the niches that Microsoft does not address (sophisticated products, dictionaries of obscure languages and dialects, marketing arrangements with Microsoft's rivals). To my knowledge, no incumbent dictionary publisher has a strategy to deal with this. I would think that the folks at Oxford, Longman, Merriam, etc, would be getting nervous, but there is no evidence that they are. Their outlook seems to be that dictionaries, like diamonds, are forever. I respectfully disagree. All current attempts (except Microsoft's) to put dictionaries into electronic form are nothing more than a limp attempt to extend the life of a failing business model.

The future of the dictionary business is, then, going to look as follows:

First, legacy publishers such as OUP will continue to muddle along, with growth becoming harder to come by except at the expense of their smaller and declining rivals; eventually they will stop publishing for broad markets altogether and the remaining activity will be to focus on the scraps Microsoft leaves on the floor.

Second, Microsoft will create what I will call the Mainstream Dictionary, a good-enough product for most people most of the time. Intellectuals will hate it, but there are not enough of them to matter. Arguably, Microsoft's Encarta Dictionary, a better product than I would have anticipated, is version 1.0 of the Mainstream Dictionary. One of the surprising things that Microsoft did in the creation of Encarta was to go out and hire some exceptional lexicographers, perhaps under the guidance of their hardcopy publishing partner, Bloomsbury Press, in the UK. I say "surprising" because Microsoft's willful ignorance of anything to do with cultural material is astounding. In the USA, Anne H. Soukhanov directed a big part of the operation; in the UK, Faye Carney apparently played a similar role. Both are exceptionally knowledgable dictionary-makers (and both were trained by Merriam, by the way; Carney also worked at Oxford, which has a less rigorous but broader program), but they are not business strategists. The source of Microsoft's dictionary strategy lies elsewhere, perhaps in the company's DNA. Microsoft's competitors should not be distracted by these personnel appointments, as it is not lexicography that can save them but strategy.

Third, an entirely new class of lexical applications will emerge, for which there is no apparent winner at this time, that will be based on machines talking to machines, rather than having dictionaries created for human use. This is important. Nearly all dictionaries nowadays are built with people in mind. And how could it be otherwise, one might ask. But consider what is going on when you want to talk to your car or computer. Voice recognition technology (and its less sophisticated sister, text-to-speech synthesis) requires dictionaries that are built into it, inaccessible to human eyes and ears. Comparably, search technology uses lexical products to find items within huge databases. Who will be the dictionary publishers for such applications? Companies like AT&T, Microsoft, Lucent, and Hewlett Packard. Good-bye, Oxford and Merriam. It was nice to know you, but at some point we all have to move on.

The 21-Language GlobalDix

The first version of GlobalDix multilingual dictionary appeared on CD and online. It includes 20 language versions from Kernerman semi-bilingual English dictionaries, developed by Kielikone in agreement with K Dictionaries and its publishing partners worldwide: Alma Littera, EDDA, H. Aschehoug, Martins Fontes, NTK, Russky Yazyk, SPN, Studentlitteratur, van Dale, WSOY, Zvaigzne ABC.

MOT GlobalDix 1.0 consists of English-English entries that include definitions, examples of use, usage notes, etc, and translation in the different languages. The translations refer specifically to each meaning of the headword, so they are not just a list of equivalents that might vary with context. It is possible to select in which languages to search and receive information, and use the English bridge to translate – indirectly, but accurately – between the languages.

This multi-directional *semi-multilingual* dictionary can serve users who have a working knowledge of English with their communicative needs, multinational corporations, international organizations, universities, schools, language students and others.

GlobalDix is available for Windows NT/98/Me/2000/XP and over the Internet/intranet, as a stand-alone product or as part of Kielikone's MOT Dictionary software that features over 30 more titles from other publishers.

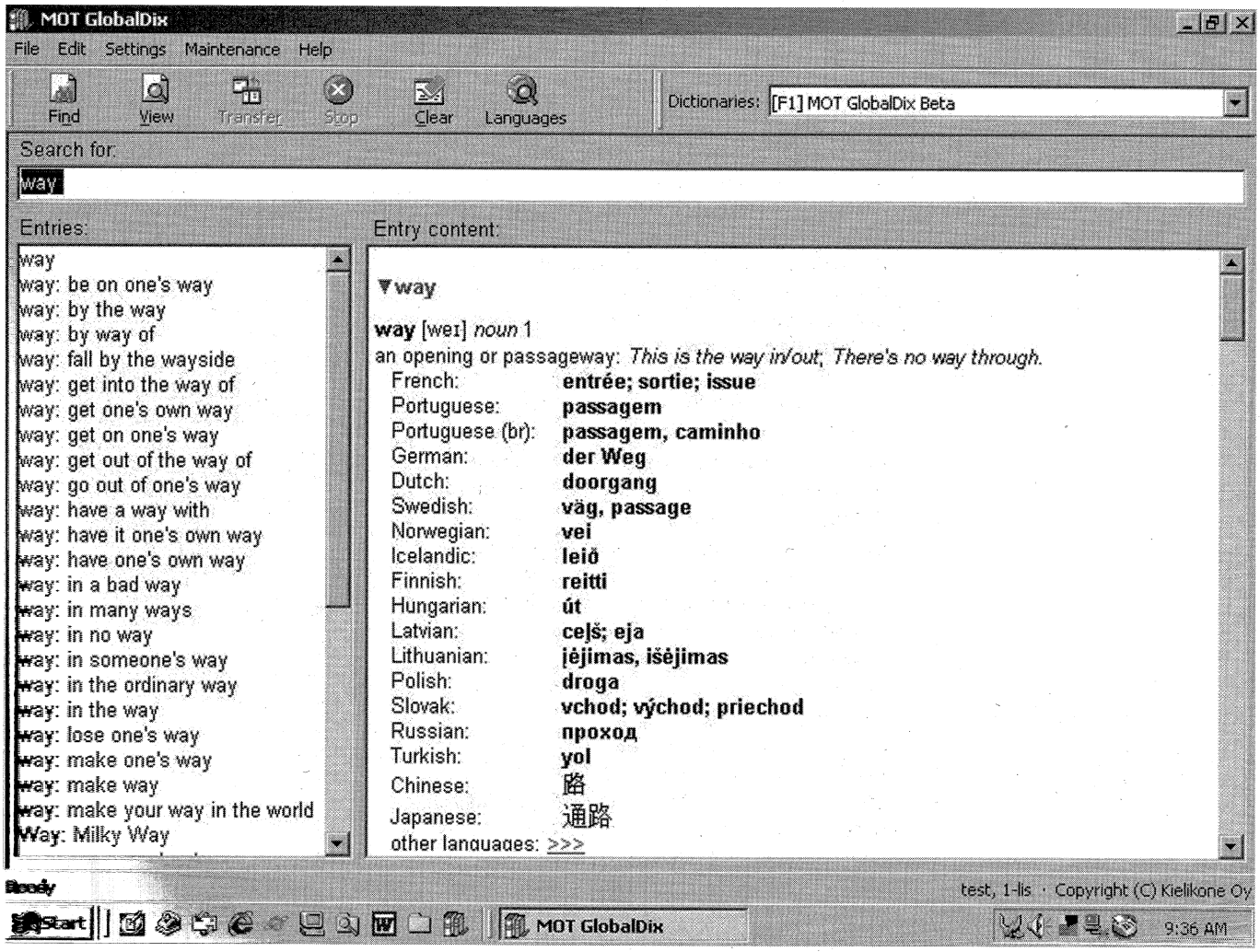
Promotion began in Finland, and first licenses were purchased notably by Nokia for its global organization and Sonera for ISP customers.

New languages and features continue to be added, the translations are being revised and updated, and the software is being upgraded. Version 2.0 will be released also for PDAs, than cellphone.

MOT GlobalDix 1.0

Chinese (Simplified) • Dutch • English • Finnish • French • German • Hungarian • Icelandic • Italian • Japanese • Latvian • Lithuanian • Norwegian • Polish • Portuguese (Brazil • Portugal) • Russian • Slovak • Spanish • Swedish • Turkish

Online free untill 18.08.02:
http://mot.kielikone.fi/mot/koekayttaja/netmot.exe
user: sanakirja
password: 2cool
select (English) GlobalDix



The Benefits of a Product-Independent Lexical Database with Formal Word Features

Janneke Froon and Franciska de Jong

Janneke Froon is a language technology coordinator for Van Dale Data. In 1997 she graduated as a computational linguist at the University of Utrecht, and has since been working for Van Dale using language technology to enhance lexicographic information for dictionaries and language products. She is preparing her PhD thesis, researching the improvement of spell-checkers using lexicographic information, and is working on the development of a large lexicographical database that integrates formal and semantic features.
janneke@vandale.nl

Abstract
Dictionaries can be used as a basis for lexicon development for NLP applications. However, it often takes a lot of pre-processing before they are usable. In the last 5 years a product-independent database of formal word features has been developed on the basis of the Van Dale dictionaries for Dutch. The database has proven to be useful in various NLP applications. This paper describes the history, some advantages, and the constraints in the development, of this database.

1. Introduction

Using traditional dictionaries as a starting point to construct lexicons for NLP applications is obvious. Dictionaries can be deployed in end-user applications such as spelling-correction tools and development tools, for instance phonological lexicons for automatic speech recognition.

Several attempts to apply machine-readable dictionaries have been reported in the literature, for instance Boguraev and Briscoe (1989), Binot and Jensen (1993), Braden-Harder (1993), and Wilks et al (1996). The focus of this work is mostly on the application of semantic information in the dictionaries. Semantic information is only one type of possible information in dictionaries. They can also be used to derive formal word features like hyphenation, pronunciation, word structure and inflection.

This paper describes the development of a lexical database on the basis of the Van Dale dictionaries for Dutch containing formal word features. Reusability of the data has been a major goal while developing this database. Reusability has always been an important concept in the development of lexical databases (cf. Calzolari 1990). For the Van Dale publishing house this concept is important since the information in the database is meant to be used as a source for various dictionaries and for the development of other applications such as text-to-speech systems.

Rather than adapting the dictionary resources to particular applications, a resource is created from which the content needed to realize new products and applications can be extracted. This paper describes the history of the database, the advantages that the database proved to have, and the limitations of the development.

2. History

Van Dale Lexicografie is a major lexicographic publisher in the Netherlands. The Van Dale dictionaries are commonly considered to be the most authoritative dictionaries in the Dutch-speaking community. The development of a database using the dictionary files of Van Dale has been a gradual process. In this section, the background of this process is sketched.

2.1 *Twenty years ago*

Until the early 1980s, paper dictionaries published by Van Dale each had their own author. The authors were responsible for the contents of the dictionary. The role of the editors was to check the data for textual correctness and ensure that the books got printed and sold. Since each author was responsible only for one dictionary, the contents of the dictionaries were not related, causing unintended differences. At best, the same implicit lexical standards were adhered to by all authors.

2.2 *Fifteen years ago*

The situation changed during the mid-80s. A new series of bilingual dictionaries from Dutch to three languages (English, French and German) was derived from the same dictionary file of common Dutch words. The potential advantages were immediately recognized by the NLP-community and attempts were made to use this high quality source of data for the development of NLP-systems, e.g. in the machine translation project Rosetta (Rosetta 1994).

However, the in-principle ideal situation didn't last long. The different dictionaries got different authors who were again responsible for their own dictionary only. Although editors checked the material that the authors produced, the dictionaries started to drift apart and therefore many defects were reintroduced, such as inconsistencies of products and the fact that products were not taking advantage of each other's corrections. As a result of the independent editing of the dictionaries the same work was sometimes done more than once and some work could never be done, because it was too expensive for a single dictionary and it was impossible to combine the various efforts.

2.3 *Seven years ago*

In 1995 the Dutch government changed the rules of spelling. For Van Dale this

implied a spelling adaptation of Dutch words in about 50 books, containing 6 million words. This had to be done in a very short period, since users would have stopped buying if new editions were expected soon. Besides the time pressure, the job to be done caused a problem too. The editors of dictionaries may be experts in the lexicographic area, but not in the new spelling regulation that had to be adhered to and internalized.

These new circumstances forced the publisher to adopt a new working method: creating a product-independent spelling database. For this aim, it was still necessary to look at the spelling of about half a million different words. However, at least, this had to be done only once, and not over and over again for all books.

The use of such a product-independent database of spellings proved to have many advantages and was soon followed by other product-independent databases, such as one with hyphenation information and another with pronunciation information. These databases were integrated into a single database with formal word features.

The database had to overcome problems like those described in Quazza and Van den Heuvel (2000), as phonemic information in dictionaries has a limited usability because it is available only for exceptional words and base words, not for all related words.

2.4 *Current applications*

The Van Dale database with formal word features has been used in various applications. The Dutch text-to-speech system *Fluency*[1] uses the phonemic transcriptions in the database. The speech synthesis of *Fluency* has been used in several of Van Dale's electronic dictionaries and in the *Fluency e-mail-reader*, a tool which automatically announces and reads aloud e-mail messages[2]. Furthermore, the database is the basis of the *Van Dale Spellingcorrector* (VDS 2000), a spell-checker for Dutch. These two applications are examples of end-user applications.

Another type of use that illustrates the importance of product-independent databases for the NLP research community is the application in development projects. For instance, the Druid project[3] and the ECHO project[4] use the pronunciation information of the database to build an acoustic model for a system for Dutch speech recognition. This speech recognition module is meant to play a role in the development of technology for spoken document retrieval, particularly in video retrieval (see Ordelman et al 1999).

A similar product-independent approach has been used in the development of the VLIS database[5] from an earlier database that contains semantic word features. This semantic database has been used in the Dutch version of EuroWordNet (see Vossen et al 1999), and is now available under license for commercial use. In cross-language retrieval tools the semantic database has proven to be valuable, in particular in the development of the disambiguation method applied in the Twenty-One search engine, which has been evaluated at several TREC-conferences[6] (cf. Hiemstra and Kraaij 1999, Hiemstra and De Jong 1999).

3. Advantages

The database with formal word features proved to have many advantages, three of which are illustrated here. Firstly, the consistency of products is easily attainable. Secondly, the information in the database is richly encoded. Finally, the information is flexible. As the illustration below will underline, these aspects are beneficial for the production of printed dictionaries, for the development of NLP-products targeting the end-user market, and for the level of support for NLP research teams.

3.1 *Consistency of products*

For Van Dale it is important that the products are consistent. As explained above, Van Dale is an authority in the field of lexicographic information. Therefore, its credibility and authority can be damaged if different products manifest different information. If, for instance, the 'Groot Woordenboek der Nederlandse Taal' (Large Dictionary of the Dutch Language, Geerts and Den Boon, 1999) contradicts the dictionary 'Hedendaags Nederlands' (Contemporary Dutch, Van Sterkenburg, 1996) in the pronunciation of a word, the user of these dictionaries cannot rely on the information anymore.

Word information like hyphenation and pronunciation is sometimes difficult to describe and causes differences. Even spelling, which seems to be strictly regulated and therefore unambiguous, has many uncertainties. For instance, the use of hyphens in words like *on-line-verbinding* (*on line connection*) is not unambiguously prescribed.

The consistency of dictionaries is guaranteed if the information is drawn from the same source every time a new dictionary file is assembled. The dictionary doesn't contain the information itself, but only a dynamic link to the information that is in the central database. When a new edition is prepared, information from the central database is imported into the dictionary. The imported information

Franciska M.G. de Jong teaches language technology at the Computer Science Department of the University of Twente, Enschede, and works for TNO-TPD in Delft as a consultant in the area of multimedia technology. Her background is in theoretical and computational linguistics, and she worked as an assistant researcher at the Faculty of Arts of the University of Utrecht (1980-1985) and as a senior researcher at Philips Research on the Rosetta machine translation project (1985-1992). She is frequently involved in international program committees, expert groups and review panels, and has initiated a number of EU projects. Professor de Jong is currently coordinating several projects aimed at multimedia indexing and retrieval, and chairs the Advisory Board of Van Dale Lexicografie.
fdejong@cs.utwente.nl

About van Dale Data

For over 100 years Van Dale Lexicography has been recognized as the foremost dictionary publishing source in the Netherlands. Since 1989 it has been publishing electronic dictionary applications.
Van Dale Data BV has been an independent enterprise of van Dale Lexicografie BV since 1999, focusing on the management and commercial operation of linguistic databases and their applications within language and speech technology.
Van Dale is part of the Veen Bosch en Keuning publishing group.

www.vandaledata.nl

Van Dale Lexicographical Information System (VLIS)

- semantic hierarchical network
- multilanguage information
- phraseology, idioms
- classification
- word attributes

Contents

- 170,000 Dutch word definitions, 1,070,000 translations
- 145,000 semantic relationships
- 225,000 examples, 525,000 translations
- 250 thematic labels

Applications

- lexicographical products for different media
- multilingual dictionaries
- multiple text retrieval and analysis techniques
- automatic classification and summarizing of texts
- development of databases
- building of indexing tools

cannot be edited in the product file itself. If changes are needed, for instance because errors are found, they have to be stored in the central database. All dictionaries will profit from the corrections.

Not only do book dictionaries thereby become consistent, but so do all applications derived from the database. New insights are shared in every product. The overall quality of the products can reach a higher level.

3.2 *Richness of data*

The second advantage is the richness of the database. A product-independent database will tend to represent data on a more abstract level than when the data are assembled for a special product, thereby resulting in a richer resource. The most important reason is that while working on the database, it is often not clear at first which information will be needed in which product. It is not desirable to leave information out just because it is not needed at the moment when the database is constructed.

The next two examples illustrate the benefits of rich codes in phonemic representations and in hyphenation marks.

The first example is the representation of underlying phonemes while representing pronunciation. In the pronunciation of *bezettoon* (*busy signal*) only one *t* is heard. The *t* of *bezet* (*busy*) disappears because of degemination with the *t* in *toon* (*signal*). If the representation is needed for the phonemic transcription in a dictionary, one *t* will do. If the representation is used to synthesize the pronunciation, a single *t* will sound unnatural, and the presence of a second *t* has to be indicated. A code indicating such a special *t* can cause the dictionary generator to delete it, while causing the speech synthesis tool to pronounce the *t*'s in a special way. The same code can be used in different products.

A second example originates from the hyphenation of words. An investigation of hyphenation for Dutch showed that it was better to indicate syllable boundaries instead of hyphenation, although syllable boundaries coincide often but not always with hyphenation positions. The reason is that there is a Dutch hyphenation rule that prohibits hyphenating on a position that would cause a syllable of one letter to be separate from the rest of the word on a new line. So *radi-o* (id.) and *a-demen* (*breathe*) are not allowed. The rule also applies when a single-letter-syllable is separated from the rest of a compounding part or derivational part. Therefore *radi-otoestel (radio+toestel: radio set)* and *bea-demen (be+ademen: insufflate)* are not allowed. Lexicographers who enter the hyphenation marks serve two aims: firstly, they have to indicate the syllable boundaries; secondly, they have to check whether a single-letter-syllable will be created. In the encoding, this distinction has to be kept. For *ademen* this causes the encoding *a:de-men*, where a colon indicates a syllable boundary that doesn't coincide with a hyphenation position.

In conclusion, every stage in the production of the word information should be represented. If every step of the thinking process is explicitly encoded, it is possible to correct the result without having to recall what was going on. Besides, a rich representation has an advantage in itself. The maintainability of the database greatly improves if rich representations are being exploited. Using rich codes, it is possible to infer which processes are responsible for the formation of, for instance, pronunciation or hyphenation. By checking the soundness of these processes, the quality of the data can be improved.

3.3 *Flexibility of data*

The third advantage of a database is the flexibility of the data. When needed for new products, the information in the databases is readily available, and because of the richness of the data there will be no obstacles in adapting it to a new product. Therefore the database may aspire to do things with the available data that are otherwise unattainable.

Gibbon (2000) points out that phonemic transcripts from machine-readable dictionaries require "extensive pre-processing" before they can be used in system lexicons. However, in the Van Dale database the phonemic information is simply there, readily applicable in a variety of products.

An example is the use of phonemic information in the *Van Dale Spelling-corrector*, a spell-checker for Dutch which benefits from this information in two different ways. The first is the use of phonemic information in the assembling of a list with predicted errors in the spell-checker. This list is used to detect quickly and properly correct the spelling errors in the list. A lot of predictions about errors can be made on the basis of problematic spelling patterns. For instance the *c* in Dutch is often confused with *k*, resulting in well-known errors like *kontakt* for *contact* and *aktie* for *actie*. A large group of errors is caused by writers staying too close to modern pronunciation, disregarding historical aspects of the spelling of certain words. For instance the *b* in *ambtenaar* is often incorrectly

omitted, because it cannot be heard. Another example is the word *quitte* that has a spelling which is very different from its pronunciation /kit/, resulting in the erroneous spelling error *kiet*. These spelling errors can be predicted if phonemic representations are used. A whole class of plausible errors can be incorporated in the spell-checker, that are beyond reach if phonological information isn't available.

The second way pronunciation information is used in the *Van Dale Spellingcorrector* is in finding homophones. Homophones are words that are pronounced similarly, but have different spellings. Examples in Dutch are *biljart* (*game of billiards*) and *biljard* (*number, thousand billion*), *boxer* (*type of dog*) and *bokser* (*someone who boxes*), and in English *discrete* (*separate*) and *discreet* (*tactful*). These words cause problems, because writers tend to mix them up, writing for instance *biljart* when the number is meant. A spell-checker can be improved if attention is paid to the difficulties with homophones, by using the pronunciation information in the databases.

Due to, among others, the use of pronunciation information in various ways, the *Van Dale Spellingcorrector* can compete with spell-checkers for Dutch that are provided with word processors. Without the information in the database, pronunciation information would have been out of reach because of the high costs. The database provides an affordable opportunity to incorporate into a spell-check this valuable information source.

4. Constraints

Although the development of a multi-purpose database has many advantages, it has a price in both time and money. If information has only one purpose, and can be used in one product only, the cost effectiveness is not optimal. However, every time the information is reused, the return on investment potentially increases, and for some applications the use of a product-independent database may be the only source of data that is affordable.

For a company, making profit is crucial, and it is tempting to choose making money in the short term. The development of a multi-purpose database undoubtedly has advantages, but especially in the long run. There is thus always the risk that developments are stopped for economic reasons, just before the end-goal is reached, because the remaining work isn't profitable enough. Collaboration with non-profit institutions, such as NLP-research groups with research capacity and/or knowledge, can then be an incentive for sustained resource development.

5. Conclusion

Building a multi-purpose database for formal word features in Dutch, or any other language, is a difficult and expensive task. However, the more the information is used in applications, the cheaper the information gets. The product-independence of the database pays itself back in the long run.

The advantages of such a product-independent database are indisputable. The information in the dictionaries is easily available, more consistent and rich, which benefits any application using the database. However, endurance is demanded of the developing companies to make these advantages commercially viable.

Notes

1 http://www.fluency.nl
2 http://www.emaillezer.nl
3 http://dis.tpd.tno.nl/druid
4 http://pc-erato2.iei.pi.cnr.it/echo
5 Vlis is the Van Dale lexicographic information system and semantic network, in existence since 1992.
6 The Twenty-One search engine is distributed in the Netherlands by Irion Technologies. http://www.irion.nl

References

Binot J. and K. Jensen. 1993. 'A Semantic Expert Using an Online Standard Dictionary.' In *Natural Language Processing: The PLNLP Approach*, K. Jensen et al. (ed.). Dordrecht: Kluwer Academic Publishers, pp. 135-149.

Boguraev B. and E. Briscoe (ed.). 1989. *Computational Lexicography for Natural Language Processing.* Harlow: Longman Group UK.

Braden-Harder L. 1993. 'Sense Disambiguation Using Online Dictionaries.' In *Natural Language Processing: The PLNLP Approach*, K. Jensen et al. (ed.). Dordrecht: Kluwer Academic Publishers, pp. 247-263.

Calzolari N. 1990. 'Lexical Databases and Tetextual Corpora: Perspectives of Integration for a Lexical Knowledge-Base.' In *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, U. Zernik (ed.). Hillsdale, NJ: Laurence Erlbaum, pp. 191-208.

Geerts G. and C. Den Boon. 1999. *Van Dale Groot Woordenboek der Nederlandse Taal.* Utrecht: Van Dale Lexicografie.

Gibbon D. 2000. 'Computational Lexicography.' In *Lexicon Development for Speech and Language Processing*, F. Van Eynde and D. Gibbon (ed.). Dordrecht: Kluwer Academic Publishers, pp. 1-42.

Word Attributes Database and Language Technology

- spelling and hyphenation
- expansions
- word class
- frequency
- context relationships
- pronunciation
- transcriptions
- morphology

Quantity

- 250,000 Dutch keywords
- 1,250,000 expansions

Quality

- checks using language rules
- relationships between words
- inheritance of attributes

Enrichments

- editorial expertise
- parameterization
- corpus
- frequency
- reverse engineering
- hyphenation, expansions

Applications

- electronic dictionaries
- language tools
- speech applications
- rhyme engine
- games

Speech Technology

- pronunciation indication
- lexicon of more than 180,000 word forms
- rules for unknown words
- rules for interpreting numbers, punctuation, etc.
- prosody generation
- rules for length of sounds in context
- rules for sentence melody

Diphone Synthesis

- diphone
- diphone database
- MBROLA synthesizer

Applications

- talking dictionaries
- tools for the handicapped
- games
- fluency e-mail reader
- telephony and Internet

Hiemstra D. and F. De Jong. 1999. 'Disambiguation Strategies for Cross-language Information Retrieval.' In *Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries: ECDL' 99*. Heidelberg: Springer-Verlag, pp. 274-293.

Hiemstra D. and W. Kraaij. 1999. 'Twenty-One at TREC-7: Ad-hoc and Cross-Language Track.' In *Proceedings of the seventh Text Retrieval Conference TREC-7*. NIST Special Publication 500-242, pp. 227-238.

Ordelman R., A. Van Hessen and D. Van Leeuwen. 1999. 'Dealing with Phrase Level Coarticulation (PLC) in Speech Recognition: A First Approach.' In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*. Cambridge: Cambridge University Press, pp. 64-68.

Quazza S. and H. Van den Heuvel. 2000. 'The Use of Lexica in Text-to-Speech Systems.' In *Lexicon Development for Speech and Language Processing*, F. Van Eynde and D. Gibbon (ed.). Dordrecht: Kluwer Academic Publishers, pp 207-233.

Rosetta M.T. 1994. *Compositional Translation.* Dordrecht: Kluwer Academic Publishers.

Van Sterkenburg P. 1996. *Van Dale Groot Woordenboek van Hedendaags Nederlands.* Utrecht: Van Dale Lexicografie.

VDS. 2000. *Van Dale Spellingcorrector voor MS-Word.* Utrecht: Van Dale Lexicografie.

Vossen P., L. Bloksma and P. Boersma. 1999. *The Dutch Wordnet.* Amsterdam: The University of Amsterdam.

Wilks Y., B. Slator and L. Guthrie. 1996. *Electric Words – Dictionaries, Computers, and Meanings.* Cambridge, MA: MIT Press.

SOCRAT Dictionary

The Russian software developer, Arsenal Company, is releasing PDA versions of various language editions of the Kernerman Semi-Bilingual English Dictionaries series as part of its SOCRAT machine translation system.

Arsenal has developed a special compression algorithm to deal with the large volume of data in the dictionaries. While the initial lexical structure of the dictionaries has been preserved, special attention was granted to assure good readability on small screens. Font formatting and styling, coloring, and other tips and tricks were used to accomplish this task.

The working area is divided into three logical parts: Word-list, Word Article, and Settings. The navigation is implemented using Windows classic tabs. Words can be entered either via the keyboard or by handwriting recognition. It is possible also to connect directly from another application by copying a word to the clipboard and clicking the SOCRAT icon.

To begin with, the following language versions are available: French, German, Hungarian, Italian, Polish, Portuguese, Russian, Slovak, Spanish.

The SOCRAT dictionaries can be used with a wide range of mobile computers using PalmOS and Windows CE operating systems. Here is a brief list of compatible devices and system requirements:

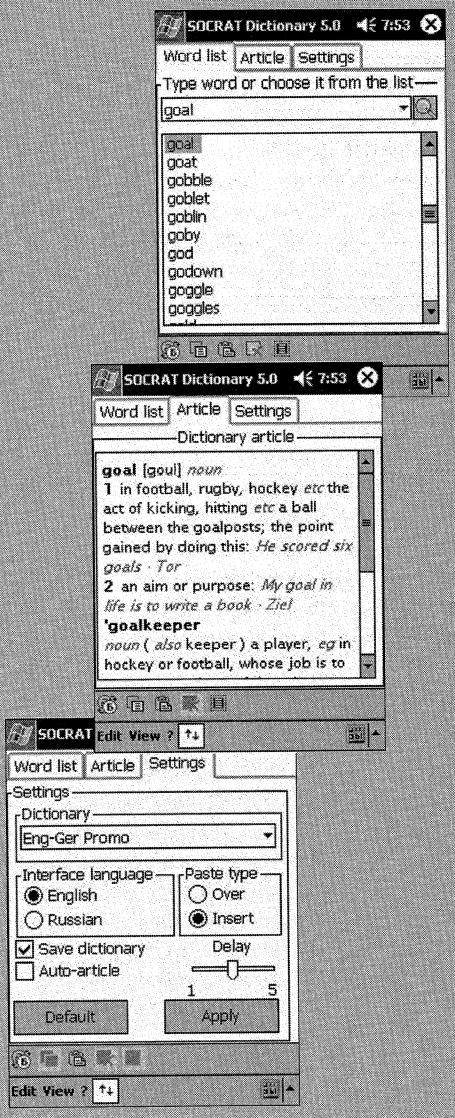
- Palm: PalmOS 3.5, Windows 98/2000/NT/Me, 2.5 MB of Available RAM

Compatible Devices: Handspring Visor Deluxe, Handspring Visor Edge, Handspring Visor Pro, IBM Workpad c500, IBM Workpad c505, Palm M105, Palm M125, Palm M500, Palm M515, Palm VIIx, Palm Vx, Samsung Smartphone SPH-I300, Symbol SPT 1700, TRGpro

- Pocket PC: Windows CE 3.0, Active Sync 3.0, 4 MB of Available RAM

Compatible devices: @migo 600-C, Audiovox Maestro PDA1032, Audiovox PDA1032C, Cassiopeia E-115, Cassiopeia E-125, Cassiopeia E-200, Cassiopeia E-2000 (Japan), Cassiopeia E-700, Cassiopeia E-750, Cassiopeia EM500, Compaq iPAQ 31 36 3700 Series, Compaq iPAQ 3800 Series, HP Jornada 560 Series, NEC MobilePro P300, O2 xda, Toshiba E570, Toshiba Genio-e

www.arssoft.com



Benedict: an EU Project for an Intelligent Dictionary

Mika Herpiö

Benedict is a large development project for an intelligent dictionary, which started as part of the second last call of the EU's 5th Framework IST (Information Society Technologies) research program. The driving force behind the project is the language technology firm Kielikone, also known for the new MOT GlobalDix multilingual dictionary that is based on the Kernerman Semi-Bilingual Dictionaries series. Kielikone is responsible for the software development and coordination of Benedict. Also involved are the University of Tampere, Gummerus Publishers and Nokia from Finland, and the University of Lancaster and HarperCollins Publishers from the UK.

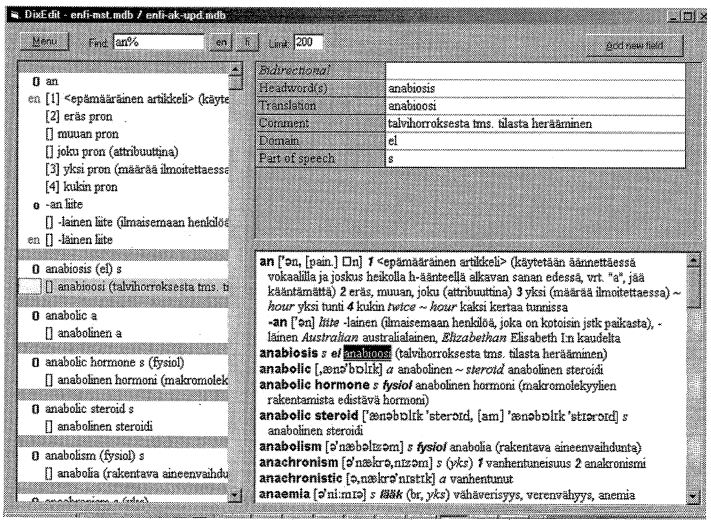
Benedict attempts to break the mental barriers caused by the tradition of the dictionary as a printed product, which still restrict its development in the electronic era. One such obstacle has been the notion that dictionary entries can appear only in one form: the one in which they are printed. The new Benedict dictionary will no longer look the same for each user – it will adapt to different users, even to different texts the users work with. How can such an adaptation be obtained so as to truly benefit the user? This is the challenge about to be solved during the three years of the project.

The Benedict product will provide an interactive user-specified access interface that tailors the dictionary content to user specifications, multi-layered entry structure, links to corpus data, and syntactically- and semantically-based corpus search tools in the dictionary database. Benedict is particularly aimed to cater for the demands of the multilingual corporate world. It will also transform the approach to economic space consumption, which has restricted the amount of data that can be accessed through a dictionary.

The new Benedict dictionary will not be created from scratch. The basis for its software development is MOT 3.0 – Kielikone's current dictionary engine that has been on the market (Windows, intranet/Internet, mobile) for more than 10 years. The new features will be integrated gradually in future editions of MOT. Since Kielikone and the University of Lancaster have considerable experience in language technology applications, the Benedict project will be able to employ state-of-the-art HLT (Human Language Technology) components in creating the final product. One of these components is the semantic tagger developed in Lancaster, which might have a vital role in directing the user to the relevant information in the dictionary entry. Kielikone has already developed also parsers, lemmatizers and other HLT solutions for the Finnish language, which will be applied in this project as well.

Significant sub-projects of Benedict include dictionary projects by Harper-Collins and Gummerus, in which content will be developed especially, but not entirely, for electronic use. The University of Lancaster has another big project to further develop their English semantic tagger and to develop a Finnish semantic tagger in collaboration with the University of Tampere – as an experiment for developing semantic taggers for more languages.

Kielikone plans to develop by-products in the project, like a system for updating the dictionary content, which applies user logs of the web dictionary. Perhaps more interesting for the dictionary community will be DixEdit, a structured content editor especially developed for lexicographic content. This will enable XML output of the dictionary, while the user doesn't have to input a single tag herself/himself. DixEdit shows the content in WYSIWYG view during the editing process, and the validity of the data is automatically surveyd all the time. The first version of DixEdit is available for Windows 98/2000/XP (a sample screen appears below).



Mika Herpiö graduated from the Helsinki University of Technology. He is the director of business development and a partner in Kielikone Oy.
mika@kielikone.fi

Kielikone is the leading Finnish language engineering company. It develops generic linguistic modules as a basis for language technology products such as dictionary software, machine translation software, terminology management software, and spell and grammar checkers for Finnish, on Windows, Internet/intranet, and mobile platforms.

http://mot.kielikone.fi/benedict

English-Japanese Lexicography and the *Unabridged Genius*

Kosei Minamide

Kosei Minamide teaches English linguistics at Osaka Women's University. He studied discourse analysis and educational linguistics at the University of London from 1986 to 1987 and attended the first International Lexicography Course held at the University of Exeter in 1987. Professor Minamide has been engaged in editing English-Japanese dictionaries, is the chief editor of *Taishukan's Unabridged Genius English-Japanese Dictionary* (2001) and *Genius English-Japanese Dictionary* (3/e 2001). His current interest is in the incorporation of findings and insights of linguistics, specifically cognitive linguistics, pragmatics and discourse analysis, into lexicographical description.
mkosei@kcn.ne.jp

1. Introduction

In April 2001, *Taishukan's Unabridged Genius English-Japanese Dictionary* was introduced in Japan and has been promoted by its publisher as a revolutionary addition to the competitive English-Japanese dictionary market. The *Unabridged Genius* contains 255,000 entries and bears the hallmarks that have distinguished the *Genius* brand for decades: accessible and readable description style; helpful usage notes; comprehensive coverage of new computer and Internet terms along with newly coined words in science, politics, business, etc; updated definitions, especially in the fast-moving areas of technology; and the addition of entries on people and places of worldwide note. These features will make the dictionary a useful companion for Japanese who wish to explore the English language in all its dimensions.

From my viewpoint as an editor of the *Unabridged Genius*, I will review the history of English-Japanese lexicography and discuss its particular features and problems. This task is undoubtedly a daunting one, and in the limited space available I can offer only a brief and limited overview, commenting on a few English-Japanese dictionaries which stand as landmarks in the history of English-Japanese lexicography, and including discussion of the *Unabridged Genius.*

2. The English-Japanese Dictionary roots

A vast number of English-Japanese dictionaries, almost exclusively designed for Japanese learners of English, have been published since the year 1862, when *Eiwa-Taiyaku-Shuchin-Jisho* (*A Pocket Dictionary of the English and Japanese Languages* [sic]), which can bear the honor of being the first printed English-Japanese dictionary, was edited by several Dutch-Japanese interpreters, with T. Hori as the chief editor. They used, as their primary sources, the Dutch-English part of Picard's *A New Pocket Dictionary of the English and Dutch Languages* (2nd edition, 1857), adopting from it about 35,000 English entry words, and a few Dutch-Japanese dictionaries, most significantly Katuragawa's *Oranda-Jii* (*A Dutch-Japanese Dictionary,* 1855-58), which was relied on for translation c the Dutch definitions into Japanese. Wh was Dutch involved in editing a English-Japanese dictionary? Until Japa abandoned its national isolation policy i 1855 and began to trade with the Wes the only window initially opened to th outside world was to the Netherland Though extremely limited in number, Dutc books on medicine, surgery, pharmacolog astronomy and some other related area were brought in by Dutch merchant through this interaction which continue for some time on the Island of Dejim in Nagasaki. Dutch was at that tim practically the only foreign language wit which Japanese people, more specifically very limited number of Japanese who wer allowed to study Dutch by the Tokugaw Shogunate, were in contact. Several kinc of Dutch-Japanese dictionaries, such a *Oranda-Jii*, were produced solely fc decoding purposes prior to the advent c English-Japanese dictionaries.

This first English-Japanese dictionar was produced in response to the urge need to learn about Western culture in th wake of the US navy's visit in 1853, ar as mentioned just above, it was compile by several Japanese interpreters of Dutc who had little or no experience of speakir or hearing English actually used by nativ speakers. In Europe, the origin of mo bilingual dictionaries can be traced back the practice in the early Middle Ages c writing interlinear glosses. These glosse mostly Latin-English and French-Englis put together, rearranged and enlarge developed into glossaries, finally in bilingual dictionaries. In Japan, howeve just as we have seen, the fir English-Japanese dictionary was a liter fusion of Dutch-English and Dutc Japanese dictionaries, a fusion dictionaries compiled on the basis of total different principles and assumptions. Th uniquely edited dictionary was abridge and enlarged in subsequent editions ar many imitated or pirated versions we published. Crude and inaccurate by today standards, these dictionaries had their ov personality as a result of policy decisio taken by the dictionary editors, and th made their own distinctive contribution early studies of English in Japan.

The collapse of the Tokugawa Shogunat was followed by the Meiji Restorati

(1868), established in the desire for 'an enriched domain and strengthened military power'. The government therefore eagerly encouraged scholars to 'translate' the West. The number of Japanese who had opportunities to hear or read real English gradually increased and English began to be taught at a considerable number of schools. Copies of dictionaries compiled by William Lobscheid, John Ogilvie, P. A. Nuttall, Noah Webster, and a series of revisions and abridgments of these dictionaries, were brought back by people dispatched abroad to study Western culture or imported by foreign-book traders. By this time English-Dutch dictionaries had been totally discarded as dictionary resources, replaced by these English-Chinese and English-English dictionaries. By taking full advantage of the newly introduced repositories, English-Japanese dictionaries have made remarkable progress in format, content (pronunciation, definition, sense division, illustrations, usage labels, etc), typography, quality of paper, printing and binding.

3. English-Japanese Lexicography in the 20th Century

In the Meiji Period (1868-1912), and the subsequent Taisho (1912-1926) and the early Showa Periods (1926-1940), the government was not only eager to import Western culture through books but also invited experienced foreign scholars and scientists as teachers and engineers, among whom were Harold Palmer and A. S. Hornby, the great pioneers of ELT in the 1930s and 1940s who made Japan a test ground for ELT innovations, just as Michael West did in India (Cowie 1999). Fully aware that the existing dictionaries, whether monolingual or bilingual, failed to meet the needs of his Japanese students, Hornby published the first monolingual EFL dictionary, *Idiomatic and Syntactic English Dictionary (ISED*, 1942), in which he refined and elaborated lexicographic devices which were first introduced by Palmer in his *Grammar of English Words* (1938), such as construction patterns, the difference between countable and uncountable nouns, and syntactic patterns of 24 anomalous finites. This dictionary was republished by OUP under the name of *A Learner's Dictionary of Current English* (1948), which underwent many revisions to finally become *Oxford Advanced Learner's Dictionary (OALD)*. Today there proliferate on the global market various monolingual learner's dictionaries which incorporate Hornby's ideas of verb patterns and countable/uncountable noun distinction, and West and Endicott's ideas of a limited defining vocabulary, creating a unique identity and carving out a niche in the EFL market. They are exported to Japan and all over the world. It is worthy of note, however, that the root of these modern sophisticated learners' dictionaries trace back to Japan and India.

Influenced by Palmer and Hornby, English-Japanese dictionaries began to change. They were more or less made on the model of *ISED* and some other monolingual general dictionaries such as *The Concise Oxford Dictionary of Current English* (1911), *The Pocket Oxford English Dictionary of Current English* (1924), shortened and updated versions of Webster's *An American Dictionary of the English Language* (1859) and Daniel Jones' *An English Pronouncing Dictionary* (1917). Digesting lexicographic information from these more advanced and sophisticated resources, the Japanese lexicographers began to improve their dictionaries by supplementing them with illustrative examples, illustrations, encyclopedic information, analyses of learner errors, brief etymologies, and grammar and usage notes at many entries, in addition to performing their basic work of defining English words in Japanese and presenting pronunciation in IPA, modified IPA or respelling systems. Worthy of special note among the general dictionaries published prior to *ISED* are *Sanseido's Concise English-Japanese Dictionary* (1922) and *Kenkyusha's New English-Japanese Dictionary* (1927). The Sanseido dictionary, mostly following the British tradition of dictionaries for words and encyclopedias for facts, was so popular that a part of the title *'konsaisu'* (= *concise*) was metonymically used for a long time to refer to a small-sized English-Japanese dictionary in general. Revised and updated, this dictionary has developed into the current 13th edition (2001), still sought after by people who prefer a handy dictionary they can grab to look up a word when reading to a learner's dictionary with complicated grammar codes and lengthy usage notes. The Kenkyusha dictionary, on the other hand, was a large-scale volume of about 100,000 entries, with encyclopedic features. This dictionary has been updated and further enlarged several times and in March of this year a 6th edition appeared, expanded to 260,000 entries. Like the earlier editions, it enjoys a unique position of authority in the matter of accuracy and sophisticated presentation of pronunciations, etymologies and definitions of words, particularly technical terms for which a group of expert consultants were employed.

Taishukan's Unabridged Genius English-Japanese Dictionary
Tomoshichi Konishi and Kosei Minamide (eds.)
Taishukan-Shoten
Tokyo, 2001
ISBN 4-469-04132-7
2528pp.
265 x 190 x 70 mm.
www.taishukan.co.jp

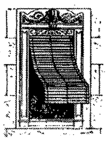
a column from the *Unabridged Genius*

Jackie

Jack·ie, ~y /dʒǽki/ 名 **1** (豪・NZ・俗)(侮蔑) ジャッキー《先住民の男；(PC) native Australian, native New Zealander》. **2** [Jackie]『ジャッキー』《英国の少女向けの雑誌》.
Jáckie Róbinson Róokie of the Yéar Awárd 〔野球〕→ Robinson.
sít úp like Jáckie ずうずうしくえらそうに座る.
j~ hànger (南ア) 〔鳥〕 =fiscal shrike.
j~ hàngman (南ア) 〔鳥〕 =fiscal shrike.
jáck-in-a-bòx 名 =jack-in-the-box.
jáck·ing 名 (英俗) (うわさ)話をすること.
jáck-in-òffice 名 (複 jacks-) [時に J~] 尊大な小役人.
jáck-in-the-bòx 名 (複 ~·es, jacks-) [時に J~] **1** (ふたを開けると人形などが飛び出す)びっくり箱；(しばしば)上下に躍動する物[人]. **2** (英俗) 梅毒《◆ pox の押韻俗語》.
jáck-in-the-grèen, Jáck-in-the-Grèen 名 (複 ~, jacks-) (英) 青葉のジャック《昔 May Day の祭りで, 青葉で覆った木の枠に入って, またはそれを支えて踊った人》.
jáck-in-the-púlpit 名 (複 ~s, jacks-) 〔植〕 **1** *Arisaema triphyllum* 《北米産のサトイモ科テンナンショウ属の一種》. **2** (英) = cuckoo-pint.

jáck·knife 名 (複 -knives) **1** ジャックナイフ (cf. clasp knife). **2** 〔水泳〕 =~ dive. **3** 〔統計〕 ジャックナイフ法《全データから1変数のみを除いた計算を何度か繰り返し, データの変動性を推定する方法》. ── 動 自 **1** 中央部で折れ曲る；〈トレーラーが〉(連結部で)エビ型に折れ曲る；直角になる. **2** 〔水泳〕 エビ型飛込みをする. ── 他 …をジャックナイフで切る；…を(2つに)折り曲げる.

ja·lop·y, ~py /dʒəlɑ́pi/ 名 (略式) ぼろ自動車[飛行機] 《◆ jallopy, jollopy, jaloopy, jalupie などともいう》
ja·louse /dʒəlúːz/ 動 他 (スコット) …を怪しいと思う (suspect)；…と推測する.
jal·ou·sie /dʒǽləsi, -siː | ʒǽluziː, ニニ, dʒəlúːsi/ 『フランス』 名 **1** 板すだれ. **2** ガラス製よろい戸.
*****jam**[1] /dʒǽm/ 『初 18c: jam[2] (傷つける, つぶす)から?』 名 U (『種類』 C) **1** ジャム((米) jelly) ‖ strawberry ~ イチゴジャム / a spoonful [jar] of ~ スプーン1杯[1びん]のジャム. **2** (主に英略式) 楽な[手間のかからない]こと.



jalousie 1

jám todáy → JAM[1] tomorrow.
jám tomórrow (英略式) (苦労の後の)明日の幸せ；(実現されそうにもない)バラ色の未来《◆ Lewis Carroll の *Through the Looking Glass* 中の句；jam today は「すぐにほしい物」を表す》 ‖ There is little point in promising people ~ *tomorrow* when they are losing their jobs because of the recession. 不況のため失業者が続出している時に明日の幸せを約束してもほとんど意味がない《◆ 英国のThatcher 元首相の言葉》.
wánt jám òn it (英略式) 利点は多いがまだ不満に思う ‖ D'you *want* ~ *on it*? さらにまだ何がほしいのか.
── 動 (-mm-) 他 **1** 〈果物〉をジャムにする. **2** 〈パンなど〉にジャムを塗る.

jam[3] 動 自 (米俗) **1** […と]セックスをする[with]. **2** すばやく動く, 急いで立ち去る《◆ jam it ともいう》. **3** コカインをかぐ. ── 名 (俗) **1** セックス；セックスの相手；ワギナ(vagina). **2** (米俗) コカイン；アンフェタミン(amphetamine). **3** (豪) きざな態度, うぬぼれ ‖ lay [put] on ~ きざな態度をとる[うぬぼれる]. **4** (米) 異性愛の人.
~ ràg (英俗) 衛生的なタオル；生理帯.
~ ràid (英俗) 生理期間.
~ tàrt (豪俗) (1) 女性, 恋人. (2) おなら.
JAM, jam (略) (Eメール) just *a* minute.
Jam. (略) *Jamaica*; 〔聖書〕 *James*.
JAMA (略) Journal of the American Medical Association 米国医師会会報.
jam·a·dar /dʒʌ́mədɑ̀ːr/ 名 **1** (インド) 下級官吏；〔史〕 (英領インドの)英国陸軍インド兵の将校. **2** (家や事務所の)掃除夫, 清掃人.

(reduced from the original)

During World War II, English was denounced as an enemy's language and English-Japanese lexicography declined. After the war, over-simplified and carelessly edited dictionaries or wordlists proliferated, and it took about two decades for the lexicography to rise from the ashes, recover its strength, and produce modern, comprehensive and practical dictionaries which surpass those published before the war.

Worthy of special note is *Kenkyusha's New Collegiate English-Japanese Dictionary* (1967), equipped with elaborate codes of sentence patterns and countable/ uncountable noun distinction labels. Since it appeared, "a war of learner's dictionaries" has been unceasingly carried out among publishers who wish to pre-empt the market of high school and college student users. Dictionaries aimed at this market have become increasingly oriented toward the facilitation of encoding under the influence of notional, functional and communicative teaching approaches.

4. The *Genius*' breakthrough

The fact that linguistic theories have been rarely mentioned in English-Japanese dictionary prefaces demonstrates that most lexicographers have tended to confine themselves to collecting examples or, more often than not, borrowing them directly or indirectly from other dictionaries and classifying meanings on the basis of conventional methods with little or no background linguistic knowledge. To those lexicographers linguistics is a remote, abstract and even frivolous discipline, which makes little or no practical contribution to dictionary making. To linguists, on the other hand, lexicography is often lamentably unscientific, uneasily poised between academic discipline and the commercial world. However, as demonstrated by recently-published dictionaries, for example, *Macmillan English Dictionary for Advanced Learners* (2002), which describes metaphoric expressions on the basis of Lakoff and Johnson (1980), an increasing number of lexicographers have come to realize that lexicography should be reconciled with developments in some fields of linguistics, particularly in such fields as computer linguistics, pragmatics, discourse analysis and cognitive semantics. Fully aware of the importance of a closer acquaintance with these areas of linguistics, the *Genius* editorial staff tried to incorporate insights and findings of these fields of linguistics into the *Unabridged Genius*.

Over the past two decades, computerized corpora have played a more and more important role in editing dictionaries, to the point where lexicography is arguably incomplete without a significant component devoted to corpus linguistics. Popular myth is that the larger the corpora, the better they are suited to lexicographic task. This might lead to a conclusion that any corpus used as a database for dictionary editing should be as large as such major corpora as the British National Corpus (BNC). We are doubtful about this conclusion. Such corpora is liable to drown us in data, presenting an overwhelming number of examples usually shown in the form of KWICK-concordance where a key word is centered in a fixed-length field (e.g. 80 characters). Semantic interpretation of the examples thus presented is very difficult for non-native speakers, because meaning is usually negotiated, depending on the context and structure of the text where it occurs. Longer contexts are needed for accurate interpretation of examples. In preparation for compiling the *Unabridged Genius*, therefore, we constructed an informally produced corpus containing over 20 million words of contemporary American English, consisting of a spoken corpus of 10 million words (from interviews, newscast, TV discussions, etc) and written corpus of 10 million words (from newspapers, magazines, etc).

Compared with established corpus resources such as the BNC, which are designed to be representative, our corpus is insignificantly small and not well balanced in terms of text (or genre) types, selection of entries and decision of the order of definitions within entries. As suggested by Tribble (1997), however, the computer-driven research works best when its use is integrated. There are now available for integration various kinds of corpora, freely accessible by individuals on the Internet. Also obtainable is a vast amount of information on collocation and usage from such search engines as Google. Full text search of CD-ROM encyclopedias will serve as a coherently structured and usable resource. This increasing availability of linguistic data stored on the web and on CD-ROMs, coupled with a simple but very powerful search tool, will compensate for non-native lexicographers' limited exposure to language in use and make it possible to look at natural English in quantities large enough to see recurring patterns in texts of all kinds and to offer users up-to-date coverage of the language. These digital resources can replace the luxury of multiple exposures to English over time and in a variety of meaningful contexts, which are usually denied to non-native lexicographers. They will help to reduce the

long-term unilateral dependence on English monolingual dictionaries for lexicographic information and change our traditional way of bilingual dictionary editing. In this light, the *Unabridged Genius* may not represent a modest change, but, arguably, a revolutionary step in the history of English-Japanese lexicography in particular and bilingual lexicography in general.

As important as a computer corpus is a native speaker's intuition. In recent years some linguists express their doubts about intuition and introspection as linguistic data on the ground that the sample of the language which native speakers, however fluent and competent they are, have met is only a fraction of the ever-changing entity called English and that corpus data therefore often provide overwhelming evidence which contradicts their intuitive judgments. It should be noted, however, that empirical data, however large they may be, do not provide us with 'negative' information (evidence of non-use), just 'positive' information (evidence of actual use). In order to obtain negative information which is often very useful to non-native learners, we devised techniques to elicit such intuitive data along the lines explored by Greenbaum and Quirk (1970). Fortunately we had a competent informant in Prof. L. H. Schourup, an American colleague of mine and a prominent linguist in the fields of pragmatics and poetics. He succeeded in shedding new light on ill-formed patterns of form and use which would have remained unnoticed if we had looked only at the surface data revealed by the computational analysis.

Because of the complicated problems concerning copyright and the extreme difficulty of finding entirely suitable examples in the corpus, we had most of the illustrative examples invented by native speakers. It might be true that one does not study all of botany by making artificial flowers, but the metaphor is not necessarily valid. Whereas artificial flowers are dead, invented examples can be of great usefulness provided they are contrived in such a way as to make apparent the restrictions and constraints on the use of a word. In this respect such examples better serve our purposes than corpus-driven examples which often do not make sense out of context.

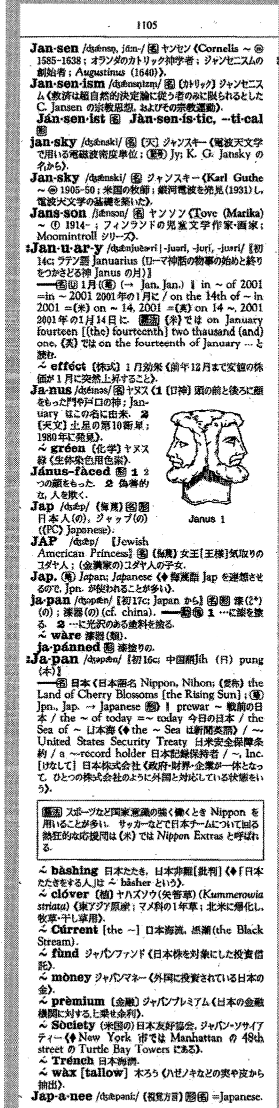
Since it is basically an expanded version of the smaller *Genius English-Japanese Dictionary*, which is mainly designed for senior high school and college students, the *Unabridged Genius* is from the start far more approachable and user-friendly both in format and content than any other preceding large-scale dictionary which is primarily intended for adult users, and it gives much more space to the detailed description of core vocabulary. In the light of recent findings of pragmatics and discourse analysis, it also gives much space to such 'small' lexical items as *oh, ah, uh, um, well, yes, yeah, okay, you know, I mean, in fact,* and *indeed.* Information is included about the meaning (emotional or cognitive) of such items, their function in turn-taking, and the expression of emotion, hesitation, emphasis, and (dis)agreement. These items share interesting pragmatic properties to which preceding dictionaries have not paid much attention because these lexical items tend to be regarded as superfluous or even undesirable by many speakers.

It has been generally believed that in order to be fluent speakers of English, non-native learners have to master grammar and memorize as many English words as possible: by so doing they will become able to create an indefinite number of sentences by putting memorized words in a grammatical framework. In reality, however, this procedure often results in sentences which are grammatically correct but pragmatically inappropriate. In recent years, we have come to realize that native speakers sort out memorized phrases which are appropriate to context rather than create new sentences by grammatical rules. These phrases stored in a mental lexicon as prefabricated chunks are often referred to as lexical phrases (Nattinger and DeCarrico 1992). Native speakers store these lexical phrases naturally and unconsciously in their mental lexicon as they grow older and select phrases most appropriate to relevant context whenever necessary to express complex ideas very simply and yet precisely. But this does not apply to non-native learners, whose exposure to naturally-occurring English is limited. What is worse, these lexical phrases, unlike the so-called idioms, tend to slip by unnoticed and not to be stored for re-use. We have therefore attempted to record as many of these prefabricated phrases as possible, and bring them to the users' attention and help them to notice which phrases are more communicatively or pragmatically important.

5. Conclusion

Many arguments are raised against the use of bilingual rather than monolingual dictionaries: the former are claimed to be nothing but a list of possible translations for English words often with little information about which meaning applies in which context and with no guidance about the grammatical patterns they operate in. These

a column from the *Unabridged Genius*



1105

Jan·sen /dʒǽnsn, jáːn-/ 名 ヤンセン《Cornelis ～ 1585-1638；オランダのカトリック神学者；ジャンセニスムの創始者；Augustinus (1640)》.

Jan·sen·ism /dʒǽnsnìzm/ 名〔カトリック〕ジャンセニスム《救済は超自然的決定論に従う者のみに限られるとした C. Jansen の宗教思想，およびその宗教運動》.
Ján·sen·ist 名 **Jàn·sen·ís·tic, ~ti·cal** 形

jan·sky /dʒǽnski/ 名〔天〕ジャンスキー《電波天文学で用いる電磁波密度単位；(略) Jy; K. G. Jansky の名から》.

Jan·sky /dʒǽnski/ 名 ジャンスキー《Karl Guthe ～ 1905-50；米国の技師；銀河電波を発見(1931)し，電波天文学の基礎を築いた》.

Jans·son /jǽnsən/ 名 ヤンソン《Tove (Marika) ～ 1914- ；フィンランドの児童文学作家・画家；Moomintroll シリーズ》.

Jan·u·ar·y /dʒǽnjuèəri | -juəri, -juri, -juəri/〖初14c; ラテン語 Januarius (ローマ神話の物事の始めと終りをつかさどる神 Janus の月)〗
—名 U 1月 (略) (→ Jan., Jan.) ‖ in ～ of 2001 = in ～ 2001 2001年の1月に / on the 14th of ～ in 2001 = (米) on ～ 14, 2001 = (英) on 14 ～, 2001 2001年の1月14日に. 語法 (米)では on January fourteen [(the) fourteenth] two thousand (and) one, (英)では on the fourteenth of January … と読む.
～ efféct (株式) 1月効果《前年12月まで安値の株価が1月に突然上昇すること》.

Ja·nus /dʒéinəs/ 名 ヤヌス《**1**〔口神〕頭の前と後ろに顔をもった門や戸口の神；January はこの名に由来. **2**〔天文〕土星の第10衛星；1980年に発見》.
～ gréen (化学) ヤヌス緑《生体染色用色素》.

Jánus-fàced 形 **1** 2つの顔をもった. **2** 偽善的な，人を欺く.

Janus 1

Jap /dʒǽp/ (侮蔑) 名 形 日本人(の)，ジャップ(の) ((PC) Japanese).

JAP /dʒǽp/ 〖Jewish American Princess〗 名 (侮蔑) 女王[王様]気取りのユダヤ人；(金満家の)ユダヤ人の子女.

Jap. (略) *Japan; Japanese* 《◆侮蔑語 Jap を連想させるので，Jpn. が使われることが多い》.

ja·pan /dʒəpǽn/ 〖初17c; Japan から〗 名 形 漆(ˢ)(の)；漆器(の) (cf. china). —動 他 **1** …に漆を塗る. **2** …に光沢のある塗料を塗る.
～ wàre 漆器(類).
ja·pánned 形 漆塗りの.

Ja·pan /dʒəpǽn/ 〖初16c; 中国語 Jih (日) pung (本)〗
—名 日本《日本語名 Nippon, Nihon; (愛称) the Land of Cherry Blossoms [the Rising Sun]；(略) Jpn., Jap. → Japanese 形》‖ prewar ～ 戦前の日本 / the ～ of today = ～ today 今日の日本 / the Sea of ～ 日本海《◆the ～ Sea は新聞英語》/ ～-United States Security Treaty 日米安全保障条約 / a ～-record holder 日本記録保持者 / ～, Inc. [けなして] 日本株式会社《政府・財界・企業が一体となって，ひとつの株式会社のように外国と対応している状態をいう》.

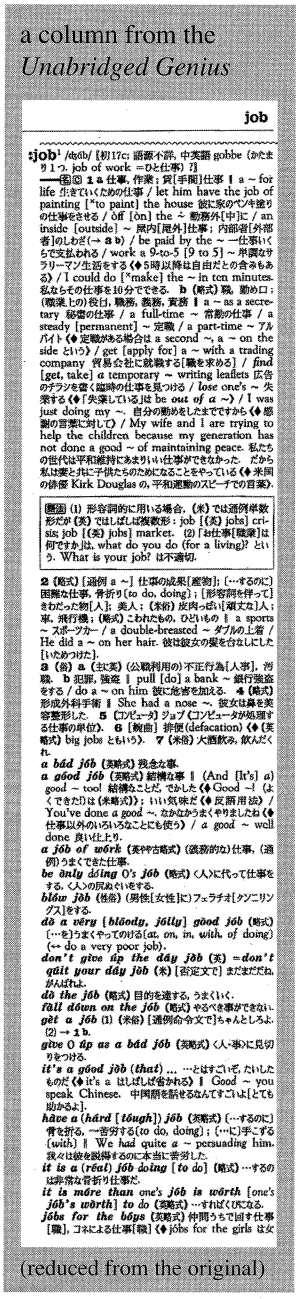
語法 スポーツなど国家意識の強く働くとき Nippon を用いることが多い. サッカーなどで日本チームについて回る熱狂的な応援団は(米)では *Nippon Extras* と呼ばれる.

～ bàshing 日本たたき，日本非難[批判]《◆「日本たたきをする人」は ～ bàsher という》.
～ clóver (植) ヤハズソウ(矢筈草) (*Kummerowia striata*)《東アジア原産；マメ科の1年草；北米に帰化し，牧草・干し草用》.
～ Cúrrent [the ～] 日本海流，黒潮 (the Black Stream).
～ fùnd ジャパンファンド《日本株を対象にした投資信託》.
～ mòney ジャパンマネー《外国に投資されている日本の金》.
～ prèmium (金融) ジャパンプレミアム《日本の金融機関に対する上乗せ金利》.
～ Sòciety (米国の)日本友好協会，ジャパン・ソサイアティー《◆New York 市では Manhattan の 48th street の Turtle Bay Towers にある》.
～ Trénch 日本海溝.
～ wàx [tallow] 木ろう《ハゼノキなどの実や皮から抽出》.

Jap·a·nee /dʒæ̀pəníː/ (視覚方言) 形 名 = Japanese.

(reduced from the original)

a column from the *Unabridged Genius*



job

:job[1] /dʒɑb/ [初17c; 語源不詳. 中英語 gobbe (かたまり1つ, job of work =ひと仕事) ?]

——名 **1 a 仕事**, 作業; 賃[手間]仕事 ‖ a ~ for life 生きていくための仕事 / let him have the job of painting [*to paint*] the house 彼に家のペンキ塗りの仕事をさせる / **off** [**on**] the ~ 勤務外[中]に / an inside [outside] ~ 院内[屋外]仕事; 内部[外部]者のしわざ(→ **a b**) / be paid by the ~ 一仕事いくらで支払われる / work a 9-to-5 [9 to 5] ~ 単調なサラリーマン生活をする(◆5時以降は自由だとの含みもある) / I could do [*make*] the ~ in ten minutes. 私ならその仕事を10分でできる. **b** (格式) 職, 勤めロ; (職業上の)役目, 職務, 義務, 責務 ‖ a ~ as a secretary 秘書の仕事 / a full-time ~ 常勤の仕事 / a steady [permanent] ~ 定職 / a part-time ~ アルバイト(◆定職がある場合は a second ~, a ~ on the side という) / get [apply for] a ~ with a trading company 貿易会社に就職する[職を求める] / *find* [get, take] *a* temporary ~ writing leaflets 広告のチラシを書く臨時の仕事を見つける / *lose* one's ~ 失業する(◆失業している]は be *out of a* ~) / I was just doing my ~. 自分の務めをしたまでですから(◆感謝の言葉に対して) / My wife and I are trying to help the children because my generation has not done a good ~ of maintaining peace. 私たちの世代は平和維持にあまりいい仕事ができなかった. だから私は妻と共に子供たちのためになることをやっている(◆米国の俳優 Kirk Douglas の, 平和運動のスピーチでの言葉).

[**語法**] (1) 形容詞的に用いる場合, (米) では通例単数形だが (英) ではしばしば複数形: job [(英) jobs] crisis job [(英) jobs] market. (2) [お仕事[職業]は何ですか]は, what do you do (for a living)? という. What is your job? は不適切.

2 (格式) [通例 a ~] 仕事の成果[産物]; (…するのに)困難な仕事, 骨折り (to do, doing); [形容詞を伴って]されたった物[人]; 美人; (米俗) 皮肉っぽい[頭文字]人; 車, 飛行機; (略式) これもの, ひどいもの ‖ a sports ~ スポーツカー / a double-breasted ~ ダブルの上着 / He did a ~ on her hair. 彼は彼女の髪を台なしにした[いためつけた].

3 (俗) **a** (主に英) (公職利用の)不正行為[人事], 汚職. **b** 犯罪, 強盗 ‖ pull [do] a bank ~ 銀行強盗をする / do a ~ on him 彼に危害を加える. **4** (略式) 形成外科手術 ‖ She had a nose ~. 彼女は鼻を美容整形した. **5** (コンピュータ) ジョブ(コンピュータが処理する仕事の単位). **6** [婉曲] 排便 (defacation) (◆(英略式) big jobs ともいう). **7** (米俗) 大酒飲み, 飲んだくれ.

a bád jób (英略式) 残念な事.

a góod jób (英略式) 結構な事 ‖ (And [It's] a) good ~ too! 結構なことだ, でかした(◆Good ~! はくできた!]は(米略式)); いい気味だ(◆反語用法) / You've done *a good* ~. なかなかうまくやりましたね(◆仕事以外のいろいろなことにも使う) / *a good* ~ well done 良い仕上り.

a jób of wórk (英やや古略式) (退屈的な)仕事, (通例)うまくできた仕事.

be únly dóing *O's jób* (略式) (人)に代って仕事をする, (人)の尻ぬぐいをする.

blów jób (俗俗) (男性[女性]に)フェラチオ[クンニリングス]をする.

dò a véry {*blóody, jólly*} *gòod jób* (略式) (…を)うまくやってのける[at, on, in, *with*, of doing) (↔ do a very poor job).

don't gíve úp the dáy jòb (英) =***don't quít your dáy jòb*** (米) [否定文で] まだまだだね, がんばれよ.

dò the jób (略式) 目的を達する, うまくいく.

fáll dówn on the jób (略式) やるべき事ができない.

gèt a jób (1) (米俗) [通例命令文で]ちゃんとしろよ. (2) → **1 b**.

give O úp as a bád jób (英略式) (人・事)に見切りをつける.

it's a góod jób (*that*) … …とはすごい, たいしたものだ(◆it's a はしばしば省かれる) ‖ Good ~ you speak Chinese. 中国語を話せるなんてすごいよ[とても助かるよ].

have a (*hárd* [*tóugh*]) ***jób*** (英略式) (…するのに)骨を折る, 一苦労する(*to* do, doing); (…に)手こずる(*with*) ‖ We had quite *a* ~ persuading him. 我々は彼を説得するのに本当に苦労した.

it is a (*réal*) ***jób doing*** [*to do*] (略式) …するのは非常な骨折り仕事だ.

it is móre than *one's jób is wórth* [*one's jób's wórth*] *to do* (英略式) …すればくびになる.

jóbs for the bóys (英略式) 仲間うちで回す仕事[職], コネによる仕事[職] (◆jobs for the girls は女

(reduced from the original)

arguments are, however, often made by those who have in mind a small bilingual dictionary where demands of space result in drastic and misleading simplification. It must be stressed therefore that this criticism does not apply to most of the modern English-Japanese dictionaries. We have been fully aware that word meanings are not simply equations between the two languages, but that they grow out of and depend on specific uses and contexts. We have attempted to reinforce supportive decoding/encoding information with example sentences and phrases and indicators of context and grammar, adding entries of learned words and technical, biographical and geographical words which are typically missing from monolingual learners' dictionaries, elaborating usage labeling – temporal, geographical, cultural and functional. Thus our dictionaries, whether pedagogical or general, are not just in the business of juxtaposing English words and their Japanese equivalents. They have reached a stage in which they serve as learning tools which develop the lexical and linguistic competence of the Japanese users of English. This does not deny, however, that monolingual dictionaries have several decisive advantages over bilingual ones. We just argue that bilingual dictionaries are not without their advantages over monolingual ones: monolingual dictionaries complement rather than replace bilingual dictionaries.

From this discussion it seems reasonable to conclude that the *Unabridged Genius*, with its data collected from various types of corpora, represents an important forward step beyond the traditional dictionary editing which has depended too heavily on British and American dictionaries, to the point where English-Japanese lexicography can be considered to be, in its own right, mature and autonomous in theory and practice.

Early this year a "new" fusion-type dictionary – *Wordpower Ei-Ei-Wa Jiten* (*Wordpower Fully-bilingual Dictionary*) was published, a bilingualized version of *The Oxford Wordpower Dictionary* (2nd edition, 2000). Though there is a difference between *fully-bilingual* and *semi-bilingual*, the basic idea and method apparently derive from the Kernerman semi-bilingual dictionaries (Kernerman 1994). This E-E-J-J dictionary would be the best for learners at a certain level, for the simple reason that it presents an English definition as well as Japanese translation. We will have to see how autonomy and fusion, contradictory on the surface but complementary at the base, will contribute to the development of English-Japanese lexicography in Japan.

References

Cowie, A. P. 1999. *English Dictionaries for Foreign Learners: A History.* Oxford: Oxford University Press.

Humblé, P. 2001. *Dictionaries and Language Learners.* Frankfurt am Main: Haag & Herchen.

Greenbaum, S. and R. Quirk. 1970. *Elicitation Experiments in English: Linguistic Studies in Use and Attitude.* London: Longman.

Kernerman, L. 1994. 'The advent of the semi-bilingual dictionary.' In *Password News.* (http://kdictionaries.com/newsletter/kdn1-l.html)

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live By.* Chicago: The University of Chicago Press.

Minamide, K. 1998. *Eigo-no-Jisho-to-Jishogaku* (*English Dictionaries and Lexicography*). Tokyo: Taishukan-Shoten.

Nattinger, J. R. and J. S. DeCarrico. 1992. *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Rundell, M. 2001. 'Teaching lexicography or training lexicographers?' In *Kernerman Dictionary News*, 9. (http://kdictionaries.com/newsletter/kdn9-3.html).

Sinclair, J. M. (ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Sinclair, J. M. 1991. *Corpus Concordance, Collocation*. Oxford: Oxford University Press.

Tribble, C. 1997. 'Improving corpora for ELT: quick-and-dirty ways of developing corpora for language teaching.' Paper presented at the first international conference: Practical Applications in Language Corpora. Poland: University of Lodz. (http://ourworld.compuserve.com/homepages/christopher_tribble/palc.htm)

Dictionaries

Fowler, H. W. and F. G. Fowler. 1911. *The Concise Oxford Dictionary of Current English.* London: Oxford University Press.

Fowler, F. G. and H. W. Fowler. 1924. *The Pocket Oxford Dictionary of Current English.* London: Oxford University Press.

Hori, T. et al. 1862. *Eiwa-Taiyaku-Shuchin-Jisho* (*A Pocket Dictionary of the English and Japanese Languages* [sic]). Edo (Tokyo): Yosho-Shirabedokoro.

Hornby, A. S. et al. 1942. *Idiomatic and Syntactic English Dictionary* (*ISED*). (Photographically reprinted and published as *A Learner's Dictionary of Current English* by Oxford University Press, 1948; retitled *The Advanced Learner's Dictionary of Current English* (*OALD*), 1958; 6th edition, 2000). Tokyo: Kaitakusha.

Iwasaki, T. and Y. Koine. 1967. *Kenkyusha Eiwa-Chujiten* (*Kenkyusha's New Collegiate English-Japanese Dictionary*). Tokyo: Kenkyusha.
Jones, D. 1917. *An English Pronouncing Dictionary.* London: J. M. Dent & Sons.
Kanda, N. and H. Kanazawa. 1922. *Shuchin-Konsaisu-Eiwa-Jiten* (*Sanseido's Concise English-Japanese Dictionary*). Tokyo: Sanseido.
Katuragawa, H. 1858. *Oranda-Jii* (*A Dutch-Japanese Dictionary*). 5 vols. Facsimile edition. Tokyo: Waseda University.
Kihara, K. 2001. *Konsaisu-Eiwa-Jiten* (*Sanseido's Concise English-Japanese Dictionary*). 13th edition. Tokyo: Sanseido.
Konishi, T. and K. Minamide. 2001. *Genius-Eiwa-Jiten* (*Genius English-Japanese Dictionary*). 3rd edition. Tokyo: Taishukan-Shoten.
Lobschied, W. / Tuda, S. et al. 1879. *Eika-Wayaku-Jiten* (*A Dictionary of the English, Chinese and Japanese Languages, with Japanese Pronunciation*). Tokyo: Yamanouchi-Fuku Shuppan.
Nuttall, P. A. 1882. *Routledge's Desk Dictionary of the English Language.* London: George Routledge & Sons.
Ogilvie, J. 1864. *A Comprehensive English Dictionary.* London: Blackie & Son.
Okakura, Y. 1927. *Shin-Eiwa-Daijiten* (*Kenkyusha's New English-Japanese Dictionary*). Tokyo: Kenkyusha.
Palmer, H. 1938. *A Grammar of English Words.* London: Longmans, Green & Co.
Picard, H. 1857. *A New Pocket Dictionary of the English and Dutch Languages* (2nd edition. Revised and augmented by A. B. Maatjes). Zalt-Bommel: John Noman & Son.
Rundell, M. 2002. *Macmillan English Dictionary for Advanced Learners* (MEDAL). Oxford: Macmillan Education.
Shimaoka, T. 2002. *Wordpower Ei-Ei-Wa Jiten* (*Wordpower Fully-bilingual Dictionary*). Tokyo: Zoshinkai-Shuppan.
Takebayashi, S. 2002. *Shin-Eiwa-Daijiten* (*Kenkyusha's New English-Japanese Dictionary*). 6th edition. Tokyo: Kenkyusha.
Webster, N. 1859. *An American Dictionary of the English Language* (Revised and enlarged by C. A. Goodrich). Springfield: George and Charles Merriam.
West, M. P. and J. G. Endicott. 1935. *The New Method English Dictionary.* London: Longmans, Green & Co.

Sexy Dictionary

Ilan J. Kernerman

Dictionaries and sex are worlds apart, but the world is changing and words change. As it becomes more outspoken, sex and its words are increasingly attributed to denote anything but sex, and *sexy* now means *interesting*, *exciting*, *fashionable* (cf. LDOCE3, MEDAL, OALD6, and appendix).

Sex is physical and natural, sensual and essential; it embodies attraction, temptation, passion, satisfaction; it is instinctive, almost irresistible, and grabs instant attention. Now wonder, then, that in these times of consumerism, it is (as has maybe always been) the great teaser in marketing, promotion and advertising, applied for the sale of *sexy* products, ideas, etc.

Dictionaries stand in stark contrast; fruits of the mind, of studious thought and scientific labor, rationalizations, verbalizations, lists and order all put in frames and formations. Often this makes them staid and estranged, cumbersome and complex, desireless and repelling – lacking some basic intuition that seems inherent in normal lively communication. What waste of wonderful human knowledge!

Does it have to be so? Why do many fear dictionaries, use them only if they must or not at all? Can't a dictionary be handy and friendly? Tell us about what lexicography and poetry share in common, relate dictionaries to reality and dream; give just the right information simply, clearly, fully; not patronize as a sacred know-it-all scripture, but literally speak the language of the user and be sensitive to the media that is used.

Let it appeal, seduce, be pleasant to use and easy to understand, stimulate and gratify, help to wisen us up, to broaden horizons and lead on to new ground – be a *sexy dictionary*, not just a provocative title.

Ilan J. Kernerman manages K Dictionaries and edits Kernerman Dictionary News.

Thanks to Ramesh Krishnamurthy for researching *sexy* in the Bank of English. An appendix containing his findings and analysis is online: http://kdictionaries.com/newsletter/kdn10-sexy.html

This contribution is a revised extract of an article in *Studs*, 2, 2001. Lund: Studentlitteratur.

A Tale of Two Tongues: Language and Lexicography in Norway

Language is a subject which nobody can feel indifferent to, because of its subjective and personal character as a part of human nature. In Norway, which has two different and officially equal written standards of Norwegian – bokmål and nynorsk – every person has to make their choice between the two, and at the same time get to know the other standard as well. This linguistic division reflects different approaches to history, nationalism, democracy, etc. Ruth Vatvedt Fjeld and Olaf Almenningen, both from the University of Oslo, try to analyze here their own language situation from different points of view: one as a user of the majority standard, bokmål, and the other as a user of the minority standard, nynorsk.

The Norwegian language background

Olaf Almenningen

Olaf Almenningen is from Voss in Western Norway, and has been the administrative leader of the 'nynorsk' organization, Noregs Mållag, from 1976 to 1987. He is a researcher at the Department of Scandinavian Studies and Comparative Literature, Section for Norwegian Lexicography and Dialectology, University of Oslo, and is working on the monolingual dictionary project, Norsk Ordbok.
olaf.almenningen@inl.uio.no

The current language situation in Norway is the result of our national history as a colony or province of Denmark (1450-1814), and then dependent on the Swedish king as a part of the common union with Sweden (1814-1905). However, in 1814 Norway set up her own constitution, and therefore in many ways we can say that modern Norwegian history starts here. This concerns the language issue as well, as Danish came to be the one and only officially written language in the Norwegian regions during the long Danish rule from 1450 to 1814, while the former Middle Ages Norwegian language (Old Norse) faded away more or less from 1350 onwards. These language roots survived in the following centuries first and foremost in the form of rural dialects throughout the countryside. In the almost 200 years since 1814, we have tried to solve the difficult language problems (at least we like to consider them difficult ourselves!) caused by our history; and in this effort mainly two strategies were followed.

From 1814 to 1890 the dominant Danish language, gradually with more Norwegian elements in it as time went on, was the only official alternative in schools and in society as a whole. It was a more than obvious task, then, to try to "Norwegianize" this tongue by reforming its main spelling, regulate the morphology in a homebound direction, and absorb a typical Norwegian vocabulary and leave out the Danish words and expressions. The leader of this reform-strategy was Knud Knudsen (1812-1895), a headmaster and teacher of Norwegian. He leaned very much on the ortophonic theories in his Norwegian language teaching, and strongly asserted that the difference between spoken and written forms should be as small as possible. Knudsen claimed that his reformed future Norwegian language should be based mainly on the spoken language of the educated classes. This reformed language standard later got its official name 'bokmål' (which, among other things, indicates that it is the most common written standard in Norway).

The other strategy was one of a more revolutionary character; namely, to form or restore a completely new Norwegian written language based on modern dialects, whose roots could be traced back to the Old Norse. In fact, the "original" Norwegian language had survived as rural (and, to a certain extent, urban) dialects throughout the colonial period from 1450 onwards.

This more revolutionary solution became the task of Ivar Aasen (1813-1896), a self-taught language genius from Sunnmøre in northwestern Norway. His strategy required a broad political and cultural movement in order to succeed, as its final goal was to replace the then-dominant Danish language completely. Aasen himself was humble and cautious in launching the "new" Norwegian language (called 'nynorsk' in Norwegian), and asserted that the most important thing was for his language to exist as an alternative to the main Dano-Norwegian written standard.

In 1885 the Norwegian national assembly (Stortinget) decided that both standards were to be considered equal, as regards official rights, and that both were to be taught in schools and used in education. As a result of this parliamentary decision, from 1892 local school authorities could choose whether to use the nynorsk or the bokmål standard in primary schools. From 1902 all teachers had to learn to write and teach both standards, and some of this system was implemented also in college from 1907 onwards. In the years after 1905, when Norway finally broke out of the Swedish union, the nynorsk standard gradually extended its role as the written language in the local government. The history of nynorsk as an officially used written language is therefore not older than 100-150 years. In 1930 the Stortinget had

to pass an act which regulated the use of both standards in official documents, such as stamps, passport, money, laws and information from the state and its institutions.

From 1900 onwards the central school authorities tried to diminish the language gap between nynorsk and bokmål to make the difference between them as small as possible. These efforts included both spelling and morphology in words that were common in both standards. The Stortinget thus passed language planning acts for both standards in 1917, 1938 and 1959, also trying to include Norwegian dialects into both, hoping that the two standards would gradually grow into one common written Norwegian language. This ambition was not always approved by the majority of the population, especially not among the dominant conservative groups of the bokmål users. Nowadays the main view among politicians, teachers and writers as regards the existence of two standards is one of mutual acceptance.

The current language situation in Norway is roughly as follows: 15% of the pupils in primary schools have nynorsk as their first choice, and the rest have bokmål. In addition, pupils from the ages of 15 to 19 have to learn to write the standard different to their own. Thus, all pupils eventualy learn to read and write both language standards. As a whole, the total "nynorsk population" today amounts to about 600,000, whereas the rest – about 3.5 million – stick to bokmål as their first choice. In local government, 117 municipalities have in 2002 chosen nynorsk, and about 180 bokmål, whereas 150 haven't made any definite choice (these are mostly bokmål users, though). The nynorsk regions are mostly scarcely populated rural areas in the western parts of the country and in the eastern valleys, whereas the cities and urban areas mostly stick to bokmål. The nynorsk written standard has to be considered a minority language in Norway today – although it is the majority as far as spoken language is concerned – representing the rural and most of the urban dialects. It has its strongholds in literature, theatre, broadcasting and art, whereas bokmål dominates the press, main publishing houses and magazines, and other important fields of society such as economy, industry and modern communication.

Current situation and lexicographic overview

Ruth Vatvedt Fjeld

Bokmål and nynorsk are two varieties of the same language. Since Norway is a geographically large country, it has many oral dialects too. In the new Norwegian state established in 1814 (or 1905, when it became independent from Sweden), it was a strong wish for the new nation to have its own national language. By and by it was launched as a principle that every person should have the right to choose her or his own written variant, as close to the oral dialect as possible, and this is still an important credo in Norwegian language planning. As a result, the two written standards favor dialects from either Eastern or Western Norway. The Eastern dialects are closest to bokmål, which geographically as well as demographically covers the main parts of the country, while nynorsk is a Western standard, above all reflecting Aasen's own dialect from Sunnmøre, in northwestern Norway.

The wish behind inventing a modern Norwegian language, and keeping the written standards as close to the oral language as possible, was to satisfy political and democratic goals. The standards are designed to give most Norwegians the options needed to express their real mother tongue in an acceptable written form. This has made the language situation complicated, and in both standards of modern Norwegian a large variety of inflectional paradigms is allowed, with minute differences.

Although the goals are well motivated, the system offers several pedagogical problems. As all inhabitants are obliged to learn both official standards, this language policy has been in the background of a long quarrel between the language users. Since the government in 1885 stated as an undisputable principle that the two standards were to be considered as equal in every respect, and therefore should be allowed in all kinds of private and official writings, the cause of most disagreements has been about when the minority standard nynorsk ought to be used. Many complaints were made about poor availability of textbooks and public information on both standards. The Nynorsk Movement (målrørsla) is an active and dedicated organisation claiming that they were neglected and oppressed. In 1981 a new Legal Act was issued, restating the principle

Ruth Vatvedt Fjeld is from Aremark in Eastern Norway, and has been teaching Nordic linguistics and lexicography in the University of Oslo since 1991. She is the leader of the Association for City Language (Bymålslaget) and a language councillor of the Norwegian State Broadcasting Corporation. Professor Fjeld has edited dictionaries and written extensively on lexicography and linguistics, and organized the first Nordic Conference on Lexicography in Oslo 1990, which led to the foundation of the Nordic Association of Lexicography.
r.e.v.fjeld@inl.uio.no

a column from Gundersen's *Norsk Ordbok*

* nynorsk
\+ bokmål
A both
[ ] riksmål

fjølment

***fjølment A -**
fjør|a el +fjær|a (om fuglefjør helst fjør)
***fjøral A**
***fjør|e -a, +[fjør|e -a], fjær|e -a**
fjør|e -a/+-et, +fjær|e -a/-et (av fjør, +fjær)
***fjør|e V -a, +[fjør|e -a], +fjær|e -a/-et (om sjøen)**
***fjør|en -e/-i/[-ent] fl -ne**
fjøret(e) A -, *fjørut jfr +fjæret(e)
fjørfe|et
fjøring|a, +fjæring|a/-en
***fjørut A - jfr fjøret(e)**
fjørvekt|a/+-en, +fjær-
fjøs|et fl -, bf -a/+-ene
fjøsstell|et
fjåg A
f. Kr. = før Kristus, før Kristi fødsel
fl. = *fleire/+flere, *fleirtal/+flertall
flabb|en
flabbet(e) A -, *flabbut
flag|e -a (vindkast m.m.)
flag|e V -a (→ flage ovafor)
flagellat|en
flageolett|en
flagg|et fl -, bf -a/+-ene
flagg|e -a/+-et
flaggermus|a jfr *skinnvengje
+flaggstang|a, *-stong|a fl -stenger
flagr|e -a/+-et
flak|et fl -, bf -a/+-ene
flak|e -a/+-te (danne flak)
flak|e -a/+-et (gape; rive)
flakk|e -a/+-et
flakn|e -a/+-et
flakong|en
flakr|e -a/+-et
flaks|en (hell)
flaks|et (det å flakse m.m.)
flamber|e -te
flamingo|en
+flamlender|en fl -e[r], bf -ne el flamlending|en
flamm|e *-a/+-en
flamm|e V -a/+-et
flammebjørk|a
flammet(e) A -, *flammut
flamsk A -
Flandern
flan|e -a/+-te
flanell|en
flaner|e -te
flanet(e) A -, *flanut
flanke|n
flanker|e -te

from 1885 with additional provisions regulating in minute detail when information should be given in both standards, when only one standard was required, and which standard had to be chosen.

This legislation was all based on the wish to safeguard the right of the weaker part, the minority, in the name of democratic language planning. The unalterable principle of closeness between spoken and written languages as decisive for good language usage made it necessary to accept the overwhelming diversity. It is therefore noteworthy what one of the most eager present-day actors for nynorsk in language planning writes: "Through its literary tradition, containing folk literature as well as belles lettres, it has a register of expressions – and indeed a particular expressiveness – which is not matched by bokmål; many bokmål users support nynorsk for this reason." (Vikør 2001).

The quotation shows that some users of nynorsk do not see their written standard just as a technical variety, but closely attach to it other values of political and personal kind, and that democratic thoughts are overruled by dedication to their own standard.

But the disagreement between nynorsk and bokmål is not the whole picture of the Norwegian language situation. As a consequence of our political history and the Norwegian legislation, at least 25% of the programmes have to be in nynorsk in the public broadcasting, and all public information has to be printed in both standards. Each public office has to respond in the standard that was used to approach it, which means that all officials must be able to write both standards. This is expensive for a small linguistic society, and official language planning has therefore tried to implement one common Norwegian standard, the so-called samnorsk (common Norwegian) standard. Up till 1972 this was the explicit aim among leading linguists and language planners in Norway.

In 1938, a big step was taken to turn this plan into reality, with a new spelling standard for both bokmål and nynorsk. The main goal was to create a common standard built on the vernaculars in both Western and Eastern Norway, quite a radical political undertaking for that time.

Unfortunately, World War II came short after, and language politics were for 5 long years given little attention. Besides, the Quisling-regime launched its own language planning of a national-romantic kind, which gave no support to the samnorsk solution. After 1945, the 1938-standard was implemented in schoolbooks and all public information. But the linguistic climate has considerably altered since 1938, and the bourgeoisie in particular did not accept these radical forms, which reflected the language of the working class rather than the heritage from the historical Danish-speaking upper class.

A culturally conservative movement, Riksmålsforbundet and Det Norske Akademi for Sprog og Litteratur (The Norwegian Academy of Language and Literature) are two private organisations that established an elitist private norm, closer to the Danish written standard, disallowing many forms normally spoken by modern Norwegians. Consequently, now there are at least three or four standards of written Norwegian – nynorsk, bokmål, samnorsk and riksmål – each representing different ideologies and goals in language planning.

The most important difference between the two official standards (bokmål and nynorsk) is of lexical, and even more of morphological nature. Hence the lexicographical situation in Norway is odd. Because of the instability of the norms, the Norwegian Language Council is allowed to change the orthography quite often. In earlier years, this occured twice a year, but since 1991 such change may take place every four years. This has made the Norwegians extremely tolerant to orthographic changes. In addition, each norm has several optional variant forms, and school children are allowed to choose freely among the variants, which is a great challenge to the teachers.

It is no wonder, then, that orthographic word lists are a big issue in Norwegian lexicography, presenting the variety of the standards, with new editions appearing very often. Each standard has several lists on the market, and publishers have a good income from such word lists. However, this word list industry has led to a neglect of other kinds of lexicographical works, with very few dictionaries being published. The most comprehensive is *Norsk Riksmålsordbok* (1937-1957, with supplements 1995), documenting the unofficial riksmålsstandard. A desk dictionary of the same kind, *Riksmålsordboken*, appeared in 1977. As late as in 1986, the first two standard dictionaries for the official varieties of bokmål and nynorsk were presented, *Bokmålsordboka* and *Nynorskordboka.*

As the riksmål variety was going to be documented in *Norsk Riksmålsordbok*, a great effort was taken in 1930 to launch the project *Norsk Ordbok*, which was to document the nynorsk standard and the rural dialects in one dictionary. The first

volume was published in 1950, and the edition has now reached the letter H. Recently the project received fresh and generous funding so it can be completed by 2014, the bicentennial of the new Norwegian state founded in 1814.

There has been one attempt to include the two standards bokmål and nynorsk in one word list, *Norsk ordbok* (1966) by Dag Gundersen, inspired by Einar Haugen's bilingual dictionary *Norwegian-English Dictionary* (1965). Lemmas acceptable only in bokmål were marked with a special tag, and nynorsk lemmas with another tag. Lemmas without tags were accepted in both standards. This dictionary showed well how much the two standards have in common, and maybe made it too clear that the quarrel between the two parties could have been solved given some efforts and cooperation. Gundersen's dictionary was an initiative approved by all Norwegians who wanted to unify the two standards into one common but flexible standard. And it was a pedagogical masterstroke, as pupils could easily find out how close to their own dialect they could come without breaking the rules for the standard chosen. Still, it had no financial success because it was not approved by the authorities for use in schools, and so it has long been unavailable. Probably the attitudes by the leading language planners already in the 1960s had been changed from working towards one common Norwegian standard to strengthening two clearly different official standards. In June 2002 the government finally stated that the unification of bokmål and nynorsk in one standard is no longer an official goal in language planning.

Today it is impossible to know what this language policy will lead to. It is beyond question that bokmål is the most used standard, but it is difficult to give exact figures, and estimates differ from 85 to 92% of the population. As the standards are numerous and variable, this is no simple issue, and it is also a question of how to count. Many nynorsk users change to bokmål after school, especially well educated young people moving to the urban parts of the country. There is on the other hand a continuum of standards ranging from more or less private ones, from conservative nynorsk close to Aasen's original, via radical standards close to modern oral variants, all the way to conservative standards close to the Danish written in the 19th century – the Golden Age in Norwegian literature.

The Norwegian language is fairly young, and therefore has no long tradition in bilingual lexicography. Danish dictionaries have fulfilled the needs for translation and L2-learning, though some of the best among these dictionaries were actually written by Norwegian lexicographers. In this way also Norwegianisms in Danish were included, and give us valuable information about how Danish was practiced by the Norwegians in the last two centuries.

In modern lexicography there are several bilingual dictionaries between Norwegian and English, German and French. In later days also Spanish, Portuguese, Russian, Turkish, Vietnamese, Serbo-Croatian and Persian have been available. Recently the linguistic situation in Norway has changed radically with new immigrants from East Asia and elsewhere, which raises the need of several new dictionaries.

There are still a lot of tasks to be solved in Norwegian lexicography, especially in LSP dictionaries and among the Nordic languages. And all dictionaries need to have two editions, one for bokmål and one for nynorsk! Some of the bilingual dictionaries are made after Haugen's principle of including both standards with special tags, but this tradition is more seldom followed today.

References

Norwegian Language Council. 2002. Statement on no longer aiming to unify the two standards: http://www.odin.dep.no/kkd/norsk/publ/otprp/043001-050003/index-dok000 and http://www.odin.dep.no/kkd/norsk/publ/otprp/043001-050003/index-dok000-b-n-a.html

Vikør, L. 2001. 'Northern Europe: Language as Prime Markers of Ethnic and National Identity.' In *Language and Nationalism in Europe,* Stephen Barbour and Cathie Carmichael, (eds.). Oxford: Oxford University Press.

Dictionaries

Gundersen, D. 1966. *Norsk ordbok.* Oslo: Universitetsforlaget.

Guttu, T. 1977. *Riksmålsordboken.* Oslo: Kunnskapsforlaget.

Haugen, E. 1965. *Norwegian-English Dictionary.* Oslo: Universitetsforlaget.

Hellevik, A. m.fl. 1966- . *Norsk Ordbok.* Oslo: Det Norske Samlaget.

Hovdenak, M. m.fl. 1986. *Nynorskordboka.* Oslo: Det Norske Samlaget.

Knudsen, T. og A. Sommerfelt m.fl. 1937-1957. *Norsk Riksmålsordbok.* Oslo: H. Aschehoug & Co.

Landrø, M. I. og B. Wangensteen. 1986. *Bokmålsordboka.* Oslo: Universitetsforlaget.

Noreng, H. 1995. *Norsk riksmålsordbok* V-VI. Oslo: Kunnskapsforlaget.

a column from Gundersen's *Norsk Ordbok*

* nynorsk
\+ bokmål
A both
[ ] riksmål

***flanut A -, flanet(e)**
flanør|en
flar|e -a/+-te
***flas|et, +flass|et**
***flas|e V -a, +flass|e -a/-et**
***flaset(e) A -, *flasut**
flask|en
flask|e -a
flask|e -a/+-et (f- opp; f- seg)
flaskebrott|et, *-brot|et
flaskefôr|et
flaskehals|en
flaskepost|en
+flass|et, *flas|et
+flass|e -a/-et, *flas|e -a
+flasset(e), *flaset(e), *flasut
flat|a/-en
flat A
flatbrød|et
flatbygd|a
flat|e -a/+-en
flat|e -a/+-et (f- ut)
+flateinnhold|et
flatemål|et
flatlendt A -
flatn|e -a/+-et
flatseng|a
flatter|e -te
flau A +-t[t]/*-tt
***flaum|en, *[flom|men], +flom|men**
***flaumljos|et, *-lys|et, *[flom-], +flomlys**
flause|n
flegma|et
***flegmatikar|en, +flegmatiker|en**
+fl -e[r], +bf -ne
flegmatisk A -
***flein|e -a**
fleinskalla A -, +fleinskalle|t - fl-de/-te
fleinskalle|n
fleip|en
fleip|e -a/-te/+-et
fleipet(e) A -, *fleiput
***fleirdobbel(t) A som *dobbel, dobbelt**
***fleire flest, +flere**
***fleirfaldig A - jfr +flerfoldig**
***fleirtal|et, +flertall**
***fleirtalsval|et**
***fleirtydig A -, +fler-**
fleis|en
flekk|en *fl -er/[-ar]
flekk|e -a/+-et (av flekk)
+flek|ke -te, *[flekke] jfr *flekkje
Flekkefjord
flekket(e) A -, *flekkut
***flekkje flekte/flakte el *[flekke]**
flekktyfus|en
***flekkut A -, flekket(e)**

Developing the Personal Dictionary

Ian Kemble

Ian Kemble is Deputy Head of the School of Languages and Area Studies at the University of Portsmouth. His main teaching subject is German, but increasingly his teaching reflects his scholarly interests in computer applications to language teaching and translation, on which he has published.
ian.kemble@port.ac.uk

Abstract
Apart from the dictionary, the computer has become the other essential tool of the translator. The article below is an account of a one-semester study unit (Computer Assisted Translation), available to undergraduate students in their final year, in which the computer is used to generate a glossary of technical terms, thereby replicating to some extent the experience of the modern lexicographer.

Background

When a group of foreign language teachers of the School of Languages and Area Studies first looked at the possibility of using computers in teaching in the mid-Eighties, two approaches presented themselves: either one went 'behind the screen' and learned a programming language (in those days it was a form of BASIC) in order to develop one's own teaching materials, or one stayed 'in front of the screen' and used commercially available, dedicated or authoring language learning software packages. I joined the group which opted for the latter, the end-user approach.

In many respects this perspective paved the way for the next important phase of our computer literacy development, the applications phase. This took the form of employing computer applications packages such as word processing in our teaching. For me the obvious application was to translation[1]. A resources book was developed for Year 2 students which took the form of 30 texts of progressive difficulty for translation from English into German[2]. Students would prepare for the discussion in class of the translated text by producing a draft version with the word processor in a small open access computer laboratory. The discussion of the text would then take place. The follow-on took the form of producing an improved version of the text, again with the word processor and, at the same time, a draft of the next text was prepared. The improved version was checked by the teacher. It was a simple idea, but it proved effective.

At the same time as this innovative class was launched I became increasingly aware of developments in lexicography involving corpora[3], and the idea of a course unit which combined translation with lexicography was born: the Computer Assisted Translation Unit, known as the CAT unit.

The CAT unit involves the students in two projects: a glossary and a machine translation project. In the following the focus will be on the former.

In outline, the first project involves students in developing a short **glossary** of 20 entries. Students identify a printed technical text of 5,000 words which becomes their electronically-stored **corpus**. The corpus is first submitted to a **concordancer** and then to a **dictionary generator**. The process of moving from a raw text or corpus to a glossary is analyzed in an **evaluative report** of approximately 2,500 words. The four main components of the project are examined in detail below.

The corpus

A corpus of 5,000 words is, of course, tiny. I can still remember digesting with some degree of incredulity the announcement that the Birmingham COBUILD project employed a corpus of some 20 million words. Nowadays, of course, large corpora consist of several hundred million words and there is talk of the first one billion word corpus (Landau 2001). However, our experience has shown that 5,000 words constitute an adequate corpus. Of course, small corpora present problems, just as large ones do. Words with a frequency of just one are quickly reached and, even with the limited objective of identifying just 20 glossary entries, some terms have to be selected from those with the lowest frequency. Students have to justify the selection, and inclusion in the glossary, of such low frequency terms.

As to the corpus itself, since these are foreign language students the corpus is in the foreign language, the glossary direction being foreign language > native language (L2-L1). In the first few years (the unit has been offered since 1990), there was an insistence that the corpus should be technical in the pure or applied scientific sense of the word. With time, a different perception has been reached, in which socio-scientific corpora of a more general nature are also accepted. Most texts are hybrids of common and technical language. It is pointed out to students that different text types generate different problems from the perspective of dictionary compilation. Providing students are aware of the issues and have some idea of how the problems can be tackled, they are usually in a position to resolve the remaining issues which may occur, and to justify their decisions.

The School of Languages and Area Studies at the University of Portsmouth is one of the largest departments in the UK teaching five languages (English, French, German, Italian and Spanish) and associated studies to more than 900 students on over 20 courses.
www.hum.port.ac.uk/slas/main.htm

Concordancer

Initially, Micro-OCP (Oxford Concordance Package) was used. With the advent of Windows a package, compatible with the new environment, has been identified, namely MonoConc[4]. The basic features of the concordancer programme are utilized: frequency lists, the standard KWIC (keyword in context) concordance, more sophisticated concordances, such as left- and right-sorted concordances. The majority of students approach the task in a reductionist way, and are, perhaps, not always willing to experiment with the facilities of the package in a dynamic way, but there are always the minority who are more adventurous and realize that concordances can be quickly produced and analyzed.

Hitherto, no statistical measures have been included in the project, such as the lexical density measure or the type:token measure, but that is set to change (see section on The Future below).

The procedure is the normal one, initially to produce two frequency lists: a first based on frequency of occurrence and a second based on alphabetical listing. With such small corpora the alphabetical list comes into its own, grouping together words with the same stem and thereby creating a potential for glossary entries to be identified. The project guidelines require students to comment in the evaluative report on the way the two lists complement each other in terms of the information they provide for the lexicographer.

Secondly, a variety of concordances is produced which allows the identification of suitable multi-word terms for an entry into the glossary. The patterning in language we know as collocation has been described in terms of a cline (Carter 1987, Chapter 3, p. 63 in particular) with relatively loose patterning at one end of the spectrum (unrestricted collocation) and much more fixed patterning at the other (restricted collocations). Small corpora of 5,000 words are likely to produce much looser patterns, possibly ones which are unique to the text in question. However, their inclusion in the glossary can be justified in terms of the frequency of occurrence in the corpus in question. Students, on the other hand, veer away from relatively unresrticted collocations, despite their legitimate inclusion in the glossary, and prefer to be influenced – understandably, though unjustifiably – by the orthodoxy of the general dictionary. The concordancer produces not only single word terms and multi-word terms, but sample contexts in which the terms in question are used. This is particularly useful for the next phase, namely the compilation of the glossary. All information (e.g. frequency lists and concordances) is stored as separate files on a disc, supported by hard copy notes.

Dictionary generator

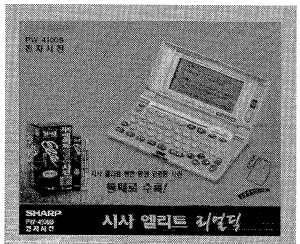
Again, the original software has been replaced by Windows software. Currently, TRADOS 95 Multiterm software is used which, among other things, has the advantage of being an industry standard.

TRADOS is a flexible package allowing the user to define the precise nature of the entry. In brief, there are three types of attribute, from which the user selects the features which are appropriate to the 'shape' of her or his glossary entry. They are index fields, which allow the languages to be used in the glossary to be specified; index fields which contain, for example, definitions, source information or notes; and, attribute fields, which contain information which can be classified, e.g. gender, subject field, etc.

The package is designed primarily for single workstation use and does not appear to operate completely free of technical intervention in a network environment. But the problems, small in number, are easily resolved by the technical support staff. Students have the task of defining the user of the glossary and designing their glossary accordingly. At the same time, they are required to include a number of definitions and uses of the term in context. Examples of the use in context are provided by the concordancer, the glossary definition has to be researched by the student, as does the equivalent of the term in the target language. In this way, a balance has to be achieved between flexibility (the student decides on the glossary entry) and rigour (the student justifies the glossary entry).

The glossary

The glossary originally contained 40 terms, but that has been progressively reduced to 20. It has to be borne in mind that the task is completed within six weeks, one half of the taught semester. In feedback sessions in the past, students have commented on the disproportionate amount of time devoted to the project, hence it has been made manageable and has, thereby, created the potential for projects of a higher quality. The glossary contains single terms, which are usually selected on the basis of a number of criteria which typically include: frequency of occurrence, level of technicality, keyword status, and multi-word terms which are provided by an analysis of the concordances which are generated.



SHARP PW-4100S
Sharp Korea
Seoul, 2001
Electronic Hand-Held Device containing the SI-SA ELITE English-English-Korean, English-Korean and Korean-English Dictionaries, with TOEIC material (by YBM Si-sa)
http://kdictionaries.com/products/medium-elect/med-el-korean.html

Evaluative report

The project evaluation provides the key to the project. It takes the form of a report in which the student reflects on the process of developing a glossary from a raw corpus. A justification for the decision taken at each stage is provided: selection of corpus, analysis of frequency lists, analysis of the concordances, determination of the shape of the glossary, construction of the glossary entries, etc. Its analytical content is a significant criterion in the marking of the project.

The unit has two other characteristics which are worthy of mention: **delivery** and **skills development**.

Delivery

The unit is delivered in a very different way to the standard practice of lecture and seminar, since technology is involved. All sessions take place in the computer laboratory. Information sessions precede workshop sessions. In the former, lexicography is introduced with particular reference to modern developments involving the computer, in the latter – which are more frequent – students work on their project with support from the teacher. Learning is experiential.

Skills development

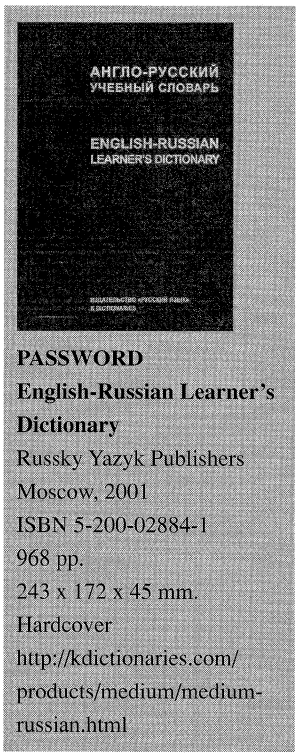
The unit has the advantage of assisting students with the development of skills, particularly those of analysis, problem-solving and IT. In the UK, evidence of key skills development is an important item on the agenda of the quality assurance agencies.

The Future

The unit has proved popular with students who are attracted by its vocational orientation and its potential to develop their IT skills. The School has recently embarked on a research project, the aim of which is to identify what students actually do in terms of vocabulary acquisition and development. The major finding is that only a third of the total student cohort sees itself as effective when it comes to vocabulary learning. As a consequence, a series of measures is to be taken to remedy the situation. These include the production by the School of a 'Guide to Vocabulary Learning' and the introduction into the Year 1 Study Skills programme of a session on 'Making Effective Use of the Dictionary'. The approach which is being adopted is characterized by a desire to achieve a balance between flexibility (allowing learners to develop their own vocabulary learning strategies) and rigour (insistence on the development of a personal dictionary). In the future, students will enter the final year of their programmes with an understanding of the principles of vocabulary organisation and have evidence to show for it in the form of a personal dictionary in the traditional printed format. The CAT option will then enable them to develop their own personal dictionary in electronic format.

Conclusion

We are not in the business of training lexicographers, but we are, and will be increasingly, able to provide students with both an understanding and experience of vocabulary across a range of contexts, including the organisation of vocabulary in a personal dictionary.



PASSWORD
English-Russian Learner's Dictionary
Russky Yazyk Publishers
Moscow, 2001
ISBN 5-200-02884-1
968 pp.
243 x 172 x 45 mm.
Hardcover
http://kdictionaries.com/products/medium/medium-russian.html

Notes

1. Interestingly, translation both as a vehicle for developing language skills and as an end in itself has survived the 'communicative revolution' which has characterized language teaching methodology in the UK for the past two decades. Indeed, translation has not only survived but has thrived.
2. *Translating for Pleasure* was first produced in 1987 and was followed by a second edition ten years later. Despite the aging of a number of the texts it continues to sell in the market place. This is evidence that there is a definite niche for such a resource.
3. The Collins COBUILD dictionary, first published in 1987 (now in its third edition, 2001, HarperCollins, Glasgow), was the pioneering dictionary of the new generation of dictionaries in the UK.
4. MonoConc for Windows, developed by Michael Barlow, is published by Athelstan (http://athel.com).

Short Indicative Bibliography

Brierley, W. and I. Kemble. 1991. *Computers as a Tool in Language Teaching.* Chichester: Ellis Horwood.
Carter, R. 1987. *Vocabulary.* London: Unwin Hyman.
Kemble, I. 1996. *Translating for Pleasure.* Portsmouth: Hampshire Open Learning Unit.
Landau, S. I. 2001. *Dictionaries, The Art and Craft of Lexicography.* (Second Edition.) Cambridge: Cambridge University Press.

The Corpus Revolution in EFL Dictionaries

Ramesh Krishnamurthy

1. Introduction

The early history of monolingual EFL dictionaries was described in detail by A.P. Cowie in Kernerman Dictionary News (2000). In his article, Cowie said: "And the authenticity of the grammatical claims made about English, and of the examples selected, has been improved beyond recognition by the use, since the early 1980s, of large-scale computer-stored corpora of English, the best known of which are the British National Corpus and the Bank of English."

This paper will describe the revolutionary impact that the use of large computer-held corpora has had on EFL dictionaries since the 1980s, with special reference to the Cobuild project (home of the Bank of English corpus) at Birmingham University, which in many ways pioneered the developments. As Michael Lewis has said (2001): "The first Cobuild dictionary changed the face of dictionary making, and the way some of us thought about vocabulary, for ever."

2. Traditional sources of lexicographic evidence

For centuries, lexicographers had to rely on their own and their colleagues' intuitions and language experience as the basis for their descriptions of language. They also frequently made use of descriptions in previously published works, thus perpetuating any errors and inaccuracies.

However, individual intuition and experience are subject to limitations. As John Sinclair has said: "Users of a language are not necessarily accurate reporters of usage, even their own" (1987); "Using a language is a skill that most people are not conscious of; they cannot examine it in detail, but simply use it to communicate" (1995); and "There are many facts about language that cannot be discovered by just thinking about it, or even reading and listening very intently" (1995). Even highly-skilled, highly-trained, and extremely dedicated lexicographers inevitably attain only a partial knowledge of a language. They also suffer from the general human weakness of a poor or selective memory. And, of course, lexicographers' work is affected by their own prejudices and preferences, however subconsciously.

One way to lend more authority to intuition-based dictionary entries is by adding authentic citations as evidence. Two historical English dictionaries are particularly noted for adopting this policy. Dr. Johnson's Dictionary (1755) deliberately took its citations only from "the best authors" writing in "the golden age of our language", and the citations therefore reflect only the higher culture. Furthermore, Johnson frequently altered the original texts to suit his purposes, for example quoting the same line from Milton's Paradise Lost with "outrageous" at one entry and "outragious" at another, and the same line from the Bible with "**in**discreet" and "**un**discreet", etc. (Kwon 1997). The Oxford English Dictionary (OED, 1879-1928) covered a wider range of authors and texts, but still managed only a piecemeal coverage, because the editors discovered that readers asked to select examples from texts tended to notice the unusual items and overlook the commonplace[1].

3. The corpus as lexicographic resource

John Sinclair has compared the impact of corpora on linguistics with that of telescopes on astronomy. The use of corpora is rapidly changing our ideas about language, and corpus research has already revealed that many of our past intuitions were wrong.

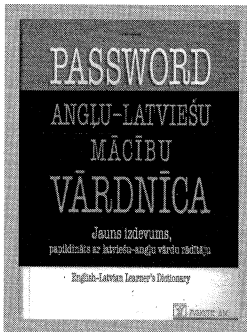
A large computer-held language corpus can overcome many of the limitations of human linguistic intuitions. It can be far more comprehensive and balanced than any individual's language experience. It does not have any memory problems, and can immediately recall all the information that has been input. It does not get distracted by unusual items, but can show us both what is common and typical and what is rare or restricted in use. Ultimately, the corpus can provide more objective evidence.

Further inadequacies of human linguistic informants have come to light: we cannot quantify our knowledge of language[2], we cannot invent natural examples[3], and we are unable (especially since the advent of the Internet) to keep up with language change. Corpora are able to assist us in all these areas: they can give us accurate statistics, a vast number of authentic examples, and (if frequently updated) can reflect even very recent changes in the language.

Another objection to using the intuitions and experiences of one individual is that they can easily be challenged or refuted by others. Corpus data encompasses the language use of many members of the language community, and therefore carries

Ramesh Krishnamurthy has degrees in French and German from Cambridge University, and in Sanskrit and Indian Religions from the School of Oriental and African Studies, London University. He has worked on the COBUILD project at Birmingham University since 1984, contributing to several dictionaries, grammars, and other publications, as well as developing corpora and software. He is an Honorary Research Fellow at Birmingham University and Wolverhampton University, has taught and supervised MA and PhD students, participated in EU linguistic projects, and conducted workshops and courses in several countries.
www.ccl.bham.ac.uk/ramesh

An appendix containing raw corpus data and the author's analysis of *sexy* is available online:
http://kdictioneries.com/newsletter/kdn10-sexy.html



PASSWORD
English-Latvian Learner's Dictionary, Second Edition
Zvaigzne ABC
Riga, 2002
Translation:
Maija Treilona and
Rasma Mozere
Editor: Ieva Tarvida
ISBN 9984-22-075-3
992 pp.
25 x 170 x 55 mm.
Hardcover
http://kdictionaries.com/products/medium/medium-latvian.html

greater authority. Language corpora also represent the democratization of the sources of evidence. We may be able to criticize Johnson's limited range of carefully-vetted sources, and even the wider but only partly used range of OED texts, but it is difficult to argue with evidence of language usage that is repeated by hundreds or even thousands of different speakers and writers in a variety of situations and contexts. In addition to the literary canon, corpora include tabloid newspapers, popular magazines, and recordings of informal conversations.

Finally, every language has its cultural connotations and underlying ideologies, which are difficult for individuals to perceive. The corpus can be invaluable in revealing these[4].

There were some problems with the use of corpora until the 1980s. Very few corpora were available, and they were too small for most purposes (the largest was around 1 million words). They were able to provide only superficial indications about many linguistic features, and were reliable only for the most frequent words in the language (i.e. grammatical words). As larger corpora were built from the 1980s onwards, attention turned to the question of balance: what proportions of texts from which genres should be included? The earlier problems of the non-availability of data, and the technical difficulties of converting printed and spoken texts into digital files had been resolved. But we were now faced with the sudden superabundance of digitalized journalistic texts, especially newspapers.

4. Earlier EFL Dictionaries

The earlier EFL dictionaries for advanced learners (i.e. the 3 editions of Oxford Advanced Learner's Dictionary (OALD), which was the sole example of this genre from 1948 to 1974), developed mainly by language teachers, had a fairly prescriptive attitude to their audience. At that time, most students studied languages at a university, and focussed on literary, historical, and higher-cultural texts. Inclusion policy in EFL dictionaries therefore favoured literary and higher-register items over more colloquial ones.

These dictionaries were also more influenced by the native-speaker lexicographic traditions (e.g. OALD claimed that it combined "the traditions of the Oxford Dictionaries" with the "language-teaching skills" of its editor, A.S. Hornby [Preface, 3/e, 1974]). The ordering of senses initially followed native-speaker practice in putting historical and etymological meanings first. The definition style was simpler but still ters rather like the language of telegram and often included abbreviations. Son definitions closely resembled the one-wo or short-phrase synonymic equivalen given in bilingual dictionaries.

The main deviations from native-speak lexicography were the omission of mo etymological information, the marking syllable-division (or hyphenation) poin in headwords, the use of IPA symbols f pronunciation (rather than "respelling"), t inclusion of pictorial illustrations, ar the increase in grammar informatic (mainly concerning noun countability ar details of verb complementation, going f beyond the simple transitive/intransitiv labels in native-speaker dictionaries). EF dictionaries had more examples tha native-speaker dictionaries, but eschewe the use of authentic citations in favour invented pedagogic "model" examples illustrate their definitions.

OALD (1948, 1963, 1974) was belatedl joined by similar dictionaries from oth publishers: Collins (1974), Longma (LDOCE, 1978) and Chambers (1980 Longman introduced some interestin innovations: a controlled defining vocab ulary; examples based on authentic da from London University's Survey English Usage; usage notes to disambiguat near-synonyms; making many embedde items into headwords and thus easie for learners to find; and using academi terminology (e.g. "phrasal verbs" instea of OALD's "verb with a particle preposition", and "collocations" instead "words that the headword usually combine with").

5. Cobuild

The Cobuild project was set up jointly b Collins (now HarperCollins) publishers an the University of Birmingham in 1980, an led by John Sinclair, who had created an analysed the world's first spoken corpu in the 1960s. The project's declared aim were to collect and analyse a large corpu of modern English, and to publish th findings in reference books for learner and teachers of English.

Initial lexicographic analyses wer performed manually on a corpus of million words, using paper printouts frequency lists and concordances, and th analyses were first entered onto pape slips, then keyed into a computer database But computational methods were rapidl introduced into all aspects of Cobuil work. Computer-typesetting was alread established, and Longman had used computer to ensure that words in LDOCE' definitions were part of its controlle

vocabulary.

Cobuild increased its corpus to 20 million words and wrote software to allow online inspection and analysis; results were entered by lexicographers directly into the database; the computer performed various editorial checks, especially to maintain consistency and validate cross-references; progress was automatically monitored; and duplication of effort was reduced, by lexicographers being provided with completed analyses of similar words. Finally, the database entries were extracted automatically into draft dictionary files, edited online, and became input files for typesetting the dictionary. This dictionary, published in 1987, was the first to make use of computers throughout its creation.

The corpus has continued to grow since then: renamed the Bank of English in 1991, it now stands at 450 million words. The corpus retrieval software has also been substantially improved, with more sophisticated search tools, wordclass tagging and syntactic parsing, automatic analyses of collocation, and so on.

6. The impact of corpora on EFL dictionaries

The effects of language corpora were first felt in EFL dictionaries, because the smaller corpora available initially were just about sufficient for the reduced coverage of an EFL dictionary (c. 50,000 entries), but completely inadequate for a large native-speaker dictionary (c. 200,000 entries). The first corpora were exclusively monolingual and mainly English, so bilingual dictionaries benefited only marginally and had to wait until corpora were developed in other languages. The main impact of corpora on bilingual dictionaries required the development of parallel multilingual corpora, which have only started to become available very recently.

A major change had also taken place in the language learning market in the preceding decade or two: students of English were increasingly studying at language schools rather than universities, less interested in literature and high culture, and demanding instead communicative language for much more mundane and practical purposes, such as tourism and commerce. As students are increasingly exposed to unrestricted, unedited and unmediated output via the media and especially the Internet, their dictionaries need to cover much more of the lexicon, at least for decoding purposes. The corpus-based generation of dictionaries therefore became more descriptive[5].

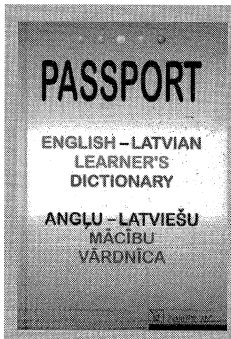
The first impact of corpora can be seen in dictionary inclusion policy. EFL dictionaries began to base their headword lists on corpus frequency, and therefore included many more journalistic and colloquial expressions (e.g. OALD6's new words: *cardboard city, generation X, latchkey child, multiskilling, outsource, innit*), leaving less space to accommodate literary and higher-register items[6]. Later editions (e.g. COBUILD2 1995, LDOCE3 1995) even published the frequency information in the dictionary itself.

Ordering of senses within entries also changed substantially, reflecting the relative frequency of the senses in corpus data, and especially the importance of hitherto largely overlooked areas of meaning. For example, COBUILD was the first dictionary to give the "homosexual" sense of *gay* first (and the "lively and cheerful" sense was labelled as old-fashioned), and the first to bring to our attention the frequent use of the verb *see* in discoursal expressions such as *I see*, and *You see*, meaning "understand", rather than in semantic meanings relating to the faculty of vision. Common verbs like *have*, *take* and *make* were seen to function in a semantically depleted way, as "delexical" verbs which merely provided the syntactic link with the following noun objects which carried the major semantic component, in phrases like *have a bath*, *take a nap* and *make a decision.*

EFL dictionaries were now able to give much better information on collocation[7], because of improved corpus software. Grammar coding became simpler, but more extensive. Wordclasses were subdivided into more subclasses, and detailed grammar patterns were given for all wordclasses, not just for verbs. And, of course, more authentic examples were supplied from the corpus data.

COBUILD in particular introduced several other major innovations: all the main forms of a headword were given in full (not abbreviated); definitions were expressed in full sentences showing typical linguistic patterns and contexts (cf "When a horse gallops, it runs very fast" with the traditional "(of a horse) to run very fast"); examples were taken straight from the corpus, with minimal editing; and, grammar and semantic relations were printed in a separate column to the right of the main text. However, unlike most of the other dictionaries, COBUILD did not use syllable markers or pictorial illustrations.

Although all of the current EFL dictionaries make some claim to the use of corpora in their compilation, they vary considerably in the extent to which they take the corpus evidence seriously. There is still heated lexicographic debate on



PASSPORT
English-Latvian Learner's Dictionary
Zvaigzne ABC
Riga, 2000
Translation:
Maija Treilona and
Rasma Mozere
Editor: Ieva Tarvida
ISBN 9984-17-232-5
650 pp.
22 x 150 x 40 mm.
Hardcover
http://kdictionaries.com/products/junior/junior-latvian.html

various issues: What is the ideal corpus? To what extent should the corpus evidence affect lexicographers' decisions[8]? Should encyclopaedic items, abundant in corpora, be included in EFL dictionaries[9]? Which aspects of the descriptive apparatus are pedagogically relevant to the student? Should authentic corpus examples be edited for pedagogic purposes?

7. Subsequent developments and the future

The influence of the innovations in EFL dictionaries is also evident in corpus-based native-speaker and bilingual dictionaries. Collins started using Cobuild's Bank of English corpus and also built its own Language Databanks for other languages (French, German, Spanish, etc). Oxford used the British National Corpus and a corpus of French in the Oxford-Hachette French Dictionary (1994), and produced the corpus-based New Oxford Dictionary of English (1998).

EFL dictionaries have seen a shorter time gap between editions (cf OALD 1948, 1963, 1974, 1989, 1995, 2000; LDOCE 1978, 1987, 1995, 2001; COBUILD 1987, 1995, 2001). This is partly due to market considerations, and partly due to the greater ease of producing new editions of computer-held texts.

Developments in computer technology have also led to the release of EFL dictionaries on CD-Rom, increasingly in simultaneous publication with the paper edition. One important benefit of CD-Roms is that pronunciations can now be heard, and do not have to be interpreted from phonetic symbols. Many dictionaries are now online; indeed, the OED has ceased paper publication and updates will only be available online from now on.

Corpus data has also become publicly available. Cobuild first released corpus data in printed form in its Concordance Samplers series, then on CD-Rom (a 5-million-word Word Bank forms part of 'Cobuild on CD-Rom'; the Collocations CD-Rom contains 2.6 million corpus examples). The availability of corpora online (Bank of English, British National Corpus, and many others, in many languages) has allowed teachers and students permanent access to native-speaker data (whereas native-speaker informants may not always be available for consultation). Researchers in CALL (computer-assisted-language-learning) are creating new software to make more use of corpus data.

As multilingual corpora become more available, and increase in size, genre variety and degree of automation, the impact on bilingual dictionaries will be immense. As monolingual corpora increase in size they will underpin most native-speaker dictionaries. Enhanced annotations of text corpora will facilitate the study of semantics and pragmatics. Improved software will deepen our understanding of collocation. Frequent corpus updates will improve our ability to identify important trends in language change. Audio and video corpora are being developed and will contribute to better information on pronunciation and intonation, on body language and gesture.

Dictionary design is one area that has been somewhat resistant to change. Even corpus-based dictionaries on CD-Rom are still rather dependent on the paper product design. Considering the exciting developments in interactive computer games, it must be feasible to create more user-stimulating language reference resources.

PASSPORT
English-Chinese Learner's Dictionary
Bookman Books
Taipei, 2001
ISBN: 957-586-945-1
685 pp.
216 x 157 x 37 mm.
Softcover
http://kdictionaries.com/products/junior/junior-chin.html

Notes

1. James Murray, the first editor of the Oxford English Dictionary, made the following complaint (1879): "The editor or his assistants have to search for precious hours for examples of common words which readers passed by... Thus, of *abusion*, we found in the slips about 50 instances: of *abuse* not five...There was not a single quotation for *imaginable*...".
2. Stubbs (1995): "Native speakers can often give a few examples of the collocates of a word ... But they certainly cannot document collocations with any thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations."
3. 'Naturalness' is a concept put forward by John Sinclair (1984), which goes beyond the earlier purely formal concepts of 'grammaticality' and 'well-formedness'. An instance of what I would consider to be a non-natural example is "Never hold a gun by the business end" (OALD 6, 2000). Apart from its pragmatic oddity (it is difficult – though not impossible – to imagine who would actually say this, and in what situation), there are no examples of *"**by** the business end"* in the current 450-million-word Bank of English corpus.
4. See, for example, Krishnamurthy (1996).
5. Even the corpus-based dictionaries could not be purely descriptive, however because of their student target audience so they still attached warning labels to non-recommended usages (e.g. in COBUILD, for the sentence-adverbial use of *hopefully*: "Some careful speakers of English think that this use of *hopefully*

is not correct, but it is very frequently used.").

6. I remember an academic review of COBUILD (1987) decrying the omission of "mizzenmast", because it occurred frequently in Herman Melville's Moby Dick, and EFL students would therefore need it.

7. However, collocations are still inaccurately reflected, even in the latest editions: for the headword *overshoot* LDOCE omits both *target* and *runway* (the most significant collocates in the corpus) and instead gives *turning* (1 example in 450m words); OALD gives *runway*, but not *target*; only COBUILD gives both *target* and *runway*.

8. Of course, dictionaries that already had pre-corpus editions faced the problem of 'text inertia': a lot of money had been invested in creating a satisfactory text from intuition, so they were unwilling to make wholesale editorial changes just because of corpus evidence; pedagogical conservatism in the teachers and students also contributed to this inertia.

9. Longman (LDELC 1992) and Oxford (OALED 1992) produced separate editions with copious encyclopaedic entries in an attempt to resolve this issue.

References

Chambers Universal Learners' Dictionary. 1980. Edinburgh: W&R Chambers.

Collins English Learner's Dictionary. 1974. London: Collins.

COBUILD: *Collins Cobuild English Dictionary for Advanced Learners*, Third edition. 2001. Glasgow: HarperCollins.

Cowie, A. P. 2000. 'The EFL Dictionary Pioneers and their Legacies.' In *Kernerman Dictionary News,* 8. (http://kdictionaries.com/newsletter/kdn8-1.html)

Johnson, S. 1755. *A Dictionary of the English Language.* London: Longman.

Krishnamurthy, R. 1996. 'Ethnic, Racial and Tribal: The Language of Racism?' In *Texts and Practices.* C. Caldas-Coulthard and M. Coulthard, (eds.). London: Routledge.

Kwon, H. K. 1997. *English Negative Prefixation: Past, Present, and Future.* Unpublished PhD thesis, University of Birmingham.

LDELC: *Longman Dictionary of English Language and Culture.* 1992. Harlow: Longman.

LDOCE: *Longman Dictionary of Contemporary English,* Third edition. 1995. Harlow: Pearson.

Lewis, M. 2001. 'Is anyone in EFL actually awake and thinking?' In *ELGazette*, 261.

Murray, J. 1879. *Address to the Philological Society.* Oxford: Clarendon Press.

New Oxford Dictionary of English. 1998. Oxford: Oxford University Press.

OALD: *Oxford Advanced Learner's Dictionary of Current English,* Sixth edition. 2000. Oxford: Oxford University Press.

OALED: *Oxford Advanced Learner's Encyclopedic Dictionary.* 1992. Oxford: Oxford University Press.

OED: *Oxford English Dictionary.* 1928. Oxford: Oxford University Press.

Sinclair, J. 1984. 'Naturalness in Language.' In *Corpus Linguistics: Recent developments in the use of computer corpora in English language research.* J. Aarts and W. Meijs, (eds.). Amsterdam: Rodopi.

Sinclair, J. 1987. *Introduction to the Collins Cobuild English Language Dictionary.* London: Collins.

Sinclair, J. 1995. *Introduction to the Collins Cobuild English Dictionary,* Second edition. London: HarperCollins.

Stubbs, M. 1995. 'Collocations and Semantic Profiles: On the cause of the trouble with quantitative studies.' In *Functions of Language,* 2, 1.

PASSPORT – Second Edition

The PASSPORT English Learner's Dictionary has undergone its first thorough editorial revision since publication in 1996. The new Second Edition now serves as the basis for new language versions appearing by the end of the year.

Created by Yaakov Levy and Raphael Gefen, PASSPORT is designed for young pre-intermediate learners, in particular pupils in junior high (lower secondary) and in the upper grades of primary school, and weaker learners in high (upper secondary) school.

The language versions published so far include Bulgarian, Chinese (Traditional), Czech, Estonian, Hebrew, Italian, Latvian, Lithuanian and Thai. In preparation are Chinese (Simplified), French, Greek, Malay, Romanian and Slovak.

The recent revision has profited and incorporated the feedback received from publishing partners in different countries, and can be divided as follows:

- corrections, changes and additions to the existing entries, with an emphasis on more synonyms and references to American/British English;
- the introduction of new entries including (a) new words widely used at all levels, especially but not only in the field of hi-tech and computers (e.g. email, e-learning, Internet), (b) existing everyday words which were not included in the earlier version but are more common now (e.g. ethnic), and (c) new meanings to existing entries (e.g. cool);
- in a number of cases, entries specifically relevant to different languages and societies have been entered at the request of the local editor.

The English-L1 section of PASSPORT has 12,000 entries and 16,000 meanings. Each version includes an extensive L1-English part, appendices on the English language and grammar, dictionary exercises, and illustrations.

The PASSPORT Electronic Dictionary is being adapted in line with the recent editorial revision and its software upgraded. The first new versions to incorporate the PASSPORT revision in electronic form will be Traditional Chinese and Greek.

Translation, the Key or the Equivalent?

a study of the dictionary use strategies of Finnish senior secondary school students

Seppo Raudaskoski

Seppo Raudaskoski holds a masters degree in English translation and interpreting, and is a project researcher in the School of Modern Languages and Translation Studies at the University of Tampere, Finland. This article is a summary of his MA thesis.
seppo.raudaskoski@uta.fi

1. Introduction

It has long been the desire of lexicographers and publishers alike to find out as much as they can about the needs, wishes and skills of dictionary users, so that they can customize their products accordingly. To this end, numerous questionnaire and test studies have been conducted over the past few decades. The study under discussion was based on a test aimed to compare how well Finnish senior secondary school students make use of the information available to them in the representatives of two dictionary archetypes, the bilingual dictionary (represented in the test by *English-Finnish General Dictionary* and *Finnish-English General Dictionary*) and the bilingualized dictionary (represented by *Englannin opiskelijan sanakirja,* a bilingualized version of the *Collins Cobuild New Student's Dictionary*). The author was part of the editorial team responsible for the bilingualization of the Cobuild dictionary. A similar type of bilingualized dictionary was first published in Finland as part of Kernerman Semi-Bilingual Dictionaries in 1993 (*Password English Dictionary for Speakers of Finnish*).

2. Different dictionary types

It is well documented that the EFL dictionary market is still characterized by a rigid dichotomy, that is, the battle between the monolingual and the bilingual dictionary. The monolingual dictionary is favored by language teachers, who feel that monolinguals contain more information about the foreign language (L2) than bilinguals (see, for example, Atkins 1985). More importantly, monolinguals present their L2 information in L2. With their definitions and examples, they make every dictionary search a useful experience in more ways than the one perhaps originally intended; besides pinpointing the meaning of a headword, the user finds out about its collocations, learns how to paraphrase it, and receives several good examples of how to use it in a sentence. In addition, the user learns to think in L2 instead of relating every new word he or she comes across to his or her own mother tongue (L1). The drawback of monolinguals is that they are often difficult to use for a beginner. With their L2 definitions, grammar codes and lengthy entries, they may leave the user confused and unsatisfied, but their main problem is that they are inherently circular; the L2 definitions may send the user searching all over the dictionary for the meanings of the words contained by the definition. What is more, if the user wants to express something in L2 but does not know the necessary words, he/she is unable to start searching from among the L2 headwords of the monolingual dictionary.

Bilingual dictionaries, on the other hand, are reviled by EFL teachers because they help students maintain a 'translation barrier': by concentrating on isolated headwords and their equivalents, they keep up the students' habit of relating every new word they learn to their L1. The listing of equivalents is particularly harmful because of the anisomorphic nature of languages (Zgusta 1971). All languages have a unique way of naming and organizing reality, which means that full equivalence is, in fact, quite rare outside of terminology. Neat juxtaposition of headwords and equivalents may keep the student under the illusion that there is always full equivalence between the lexemes of two different languages and, what is worse, that the equivalent can be inserted to all contexts the student might come across. The illusion is made all the more dangerous by the fact that bilinguals rarely provide enough information on how to use the headwords or the equivalents in an actual textual environment. Bilinguals are nevertheless easier to use than monolinguals and they provide instant answers. For these reasons, bilingual dictionaries are the popular choice among students, especially in the beginner and intermediate levels.

The bilingualized dictionary is, of course, the supposedly happy marriage of the two above-mentioned paradigms. It contains the L2 definitions and examples of the monolingual dictionary and the easy-to-use L1 equivalents of the bilingual dictionary. (This type of dictionary is often based on an existing monolingual learner's dictionary.) The emphasis in the entries is on the L2 material, and for this reason the equivalents are often called 'keys', as they are rather aids for understanding than stand-alone translations of the headword. The user is supposed to turn to the definitions and examples first, and if the meaning of the headword still remains somewhat unclear, the key is there to provide clarification and reassurance (cf. Reif 1987). If the bilingualized dictionary is equipped with an index of all the keys used, the user

also has handles by which to access the L2 headwords when in need of an L1-L2 translation. In short, the bilingualized dictionary can be seen as an all-in-one solution to the needs of a learner's dictionary user.

The bilingualized paradigm, however, does not escape all criticism. The concept of the key is slightly problematic, as the key should be a competent L1 translation, but simultaneously draw as little attention to itself as possible. There is a danger that the user may skip definitions and examples altogether and only pick up the instant translation proffered by the key (Nakamoto 1995). Furthermore, the index is a double-edged sword in the hands of an inexperienced user. Since it contains only the keys used in the entries, it is by no means a representative sample of the L1. It merely puts on display the reactions of the dictionary editors to a series of L2 situations, that is, the entries of the original monolingual dictionary. At worst, the index could be used as a misleading and incomplete L1-L2 dictionary.

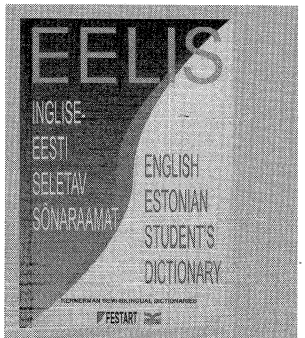
3. The test

The bilingualized dictionary used in the test, *Englannin opiskelijan sanakirja*, is aimed specifically at senior secondary school students. As its Cobuild background implies, all its definitions are in simple English consisting of complete sentences ("if you X something, you Y it"; "an X is a Y"), and its examples are culled from the Bank of English, a corpus of newspaper, literary and spoken texts. There are few symbols or abbreviations in the entries, and each headword is complemented with at least one key in Finnish. There is only one key per headword whenever possible, as it is crucial that the user not get bogged down in the Finnish part of the entry, but concentrate on the information in English instead. With 35,000 headwords, it displays only the essential vocabulary of the English language.

The bilingual dictionaries that were used, *English-Finnish General Dictionary* and *Finnish-English General Dictionary*, are much more comprehensive (90,000 and 160,000 headwords, respectively) than *Englannin opiskelijan sanakirja.* The information contained in them is packed very densely with the help of abbreviations, symbols, parentheses, tildes and other space-saving methods. In addition, there can be dozens of headwords in a single entry, which sometimes makes finding the necessary information a time-consuming task. The dictionaries contain some made-up examples of how to use the equivalents, but these are often short phrases lacking vital collocations.

The test described in this study was devised to determine which type of dictionary, bilingual or bilingualized, would be more helpful to a completely untutored user working in an actual textual environment. The test consisted of sixteen translation assignments, eight from English into Finnish and eight from Finnish into English. The study was decided to be conducted in the form of a test, because observation studies would have required too much time and manpower, and surveys can be a rather unreliable source of information: the subject might give answers that he/she thinks are appropriate, or he/she might misunderstand the questions. A translation test was chosen over a reading comprehension test on the grounds that in a reading comprehension test, the subjects could use guessing techniques to deduce the correct answer from the textual context. Finally, open-form assignments were chosen over a multiple-choice study so that the subjects could not reach the correct answer by way of eliminating the least plausible options. As the subjects were confronted with English source text words they did not know, or with Finnish source text words they did not know how to translate into English, they resorted to dictionaries in a natural, unforced manner. In other words, dictionary use was dictated by the situation, not the test form. More than one word was usually required in the translation, which made it possible to reach an acceptable answer in more ways than one. To avoid the pitfalls that have proved to be the undoing of many dictionary tests and surveys in the past, the work of Nesi (2000) proved to be a useful guide.

The test group comprised of twenty Finnish senior secondary school students, all of whom had at least 9 (out of 10) as their previous English module grade number. The point in choosing apt students was to prevent the test from deteriorating into a cavalcade of simple grammar mistakes, which would have undermined the original intent of testing language learners for their dictionary use skills rather than their elementary language skills. The students were divided into two groups of ten, one using the bilingualized dictionary and the other using the bilingual dictionaries. The students had had no prior guidance in the use of dictionaries apart from the exhortations of their teachers to use monolinguals and distrust bilinguals. They had 105 minutes to fill sixteen blank spots with the help of the Finnish and English source texts. The texts were fairly long, so there was little time to contemplate proper search strategies. Hopefully, this made it



EELIS
English Estonian Student's Dictionary
Festart
Tallinn, 2001
Translation: Ants Pihlak, Kadri Soon, Tiia Rinne
Editor: Helju Jüssi
ISBN 9985-9298-7-X
1043 pp.
221 x 149 x 50 mm.
Hardcover
http://kdictionaries.com/products/medium/medium-estonian2.html

possible to record the students' instinctive reaction to the information on offer in the dictionaries.

The test was completed twice, once without any dictionary and once with a chance to make use of the dictionaries. In the first round, the students were asked under every blank spot whether they were satisfied with the translation themselves. In the dictionary round, two new questions were asked in addition to the satisfaction question: on what page(s) the student had found information useful for the translation, and how many searches he/she had made in the dictionary.

To illustrate, one English-Finnish blank spot in the dictionary round looked like this (the English passage requiring translation is *is likely to be diluted or shelved*):

> ...a ban on snowmobiles in the park, due to come into effect in two years' time, is likely to be diluted or shelved.
>
> ...puiston moottorikelkkakielto, jonka pitäisi astua voimaan kahden vuoden
>
> päästä, _____.
> page(s): _____
> no. of searches: _____
> satisfied (y/n)? _____

4. The results

When the answers to the dictionary use questions were analyzed and compared to the actual translations, it was possible to "triangulate" quite reliably, whether the students had used their dictionary, where they had gone to search for information, what kind of information they had found there, and what they thought of its usefulness.

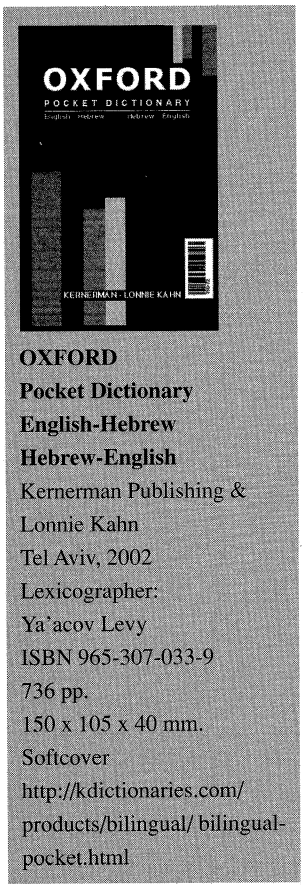
The test translations were marked according to two criteria: they had to fit in to the sentence around them and they had to represent the meaning of the source text accurately, leaving nothing out. The open-endedness of the translations may have left the test scores somewhat open for debate, as there were quite a few translation proposals that were not clear-cut *correct* or *incorrect* cases; in future tests, it would be advisable to have more than one marker available in order to reach some sort of consensus in such cases. The blank spots were chosen so that the students could not simply copy a key or an equivalent from the dictionary. Instead, they were often forced to adapt the translations provided in the entry to make the translation adequate. It was important to gauge the adaptability of the students; should a user fail to do any thinking on her/his own and simply accept the key or the equivalent at face value, any extra information present in the entries, such as the definitions and examples of the bilingualized dictionary, is rendered useless.

Content analysis of the students' translations revealed that the students often used their dictionaries uncritically. A prime example would be the translation of the word *seething* in the context *seething sulphur spring*. The bilingualized dictionary offered the key *kuhiseva*, an adjective used to describe a place full of something that is animate. Most students went for this key, with the resulting translation, *kuhiseva rikkilähde*, being something of an absurdity, since a seething sulphur spring can hardly sustain much life. In another case, the text *natural assets* was translated word-for-ford with help of the dictionary, a strategy which resulted in a nonsense concept *luonnolliset varat*, "nature-like assets". The most serious problem was, however, that the students used the Finnish-English index as a dictionary of its own and seldom bothered to consult the actual dictionary after they had located an English word in the index. This resulted in errors when there was something unusual about the inflection of the word or there were several headwords to choose from, but nothing to give clues about the suitability of each word for the context at hand.

The bilingual dictionary, with its dense entries full of symbols and abbreviations, caused difficulties for many students, especially when the necessary headword or equivalent was concealed inside a long entry. Lack of entry navigation skills was a problem especially during the Finnish-English test, as the students could not simply deduce the correct English equivalent from a long string like they could when faced with Finnish equivalents.

One major source of translation errors in both groups was the students' inability to make some translations fit in with the surrounding text. A missing definite or indefinite article or a wrong case ending could make all the difference between a successful and an unsuccessful translation. In such cases, there was little any dictionary could do to help, and in some instances the test measured the language skills and translator's instincts of the students as much as their dictionary use skills.

Examination of the test scores revealed that the bilingualized dictionary users improved their performance from the first (non-dictionary) round more than the bilingual dictionary users. Due to the small size of the sample, it is impossible to make any universal statements. It can be said, however, that despite all the translation errors caused by poor use of



OXFORD
Pocket Dictionary
English-Hebrew
Hebrew-English
Kernerman Publishing & Lonnie Kahn
Tel Aviv, 2002
Lexicographer:
Ya'acov Levy
ISBN 965-307-033-9
736 pp.
150 x 105 x 40 mm.
Softcover
http://kdictionaries.com/products/bilingual/ bilingual-pocket.html

the bilingualized dictionary and its index, the test group made better use of it than the bilingual dictionary. Regardless of the dictionary used, the English-Finnish test scores improved markedly in the dictionary round, whereas the influence of dictionaries was only marginal in the Finnish-English test. This could have been partly because the students only made roughly half the number of searches while translating from Finnish into English when compared to the English-Finnish test.

In the non-dictionary round, the students scored significantly better in the Finnish-English test than the English-Finnish test. In a way, it may have been easier to translate into L2, since students knew the exact semantic content of the source text and could attempt to paraphrase it in any number of ways. When translating from L2 into L1, however, students were stuck with the source text words they did not know; one cannot paraphrase something one does not quite grasp in the first place.

Both dictionaries eroded the students' ability to evaluate the adequateness of their translations, the effect being more pronounced in Finnish-English translations. This could be detected by comparing the adequateness of translations with the answers to questions concerning satisfaction. The comparison showed that the number of correct evaluations (correct-satisfied, incorrect-unsatisfied) decreased in the dictionary round. This is a slightly alarming trend, as finding a vaguely appropriate-feeling word during a dictionary search should not be a universal seal of approval that makes the user oblivious to all errors in her or his text.

5. Conclusion

The main lesson learned from the test results is that effective dictionary use requires some rudimentary skills and a healthy attitude towards dictionaries. The two contradictory traits of the users, that is, not bothering to find out about the proper uses of a dictionary while simultaneously accepting dictionary information as final truth, often defeat the best efforts of the lexicographer.

Before any further studies are conducted utilizing the method described here, its biggest flaw needs to be addressed: a test group of twenty is far too small to make any sweeping statements about any dictionary type. Considerably larger groups, with more than one person available to mark the translations, would be the way to go. Ideally, future studies would dovetail with courses on dictionary use; the test could be used to compare the scores attained by an untutored group with the scores of a group that has been given dictionary training. The results of such a study might well be helpful to anyone devising a short course module on dictionary use.

References

Atkins, S. 1985. 'Monolingual and bilingual learners' dictionaries: a comparison.' In *ELT Documents 120: Dictionaries, Lexicography and Language Learning,* Robert Ilson (ed.). Oxford: Pergamon Press, 15-24.

Englannin opiskelijan sanakirja. 2001. Helsinki: Otava.

English-Finnish General Dictionary. 1990. Porvoo: WSOY.

Finnish-English General Dictionary. 1984. Porvoo: WSOY.

Nakamoto, K. 1995. 'Monolingual or Bilingual, that is not the Question: the 'Bilingualised' Dictionary.' In *Kernerman Dictionary News,* 2. http://kdictionies.com/newsletter/kdn2-2.html

Nesi, H. 2000. *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries.* Tübingen: Max Niemeyer Verlag.

Password English Dictionary for Speakers of Finnish. 1995. Helsinki: WSOY.

Reif, J. A. 1987. 'The development of a dictionary concept: an English learner's dictionary and an exotic alphabet.' In *The Dictionary and the Language Learner: Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985,* Anthony Cowie, (ed.). Tübingen: Max Niemeyer Verlag, 146–158.

Zgusta, L 1971. *Manual of Lexicography.* Praha: Academia.

The Kernerman Dictionary Research Grants

Kernerman Publishing and K Dictionaries are launching a project aimed at encouraging research in selected areas of lexicography. This project, named Kernerman Dictionary Research Grants, will have a 3-year trial run. At the end of that period the project will be reviewed, and the next 3-year period will be planned.

In the first period (2002-2004) the sum of USD 9,000 will be available. The grants will be administered by the associations for lexicography in Africa, Asia and Europe (AFRILEX, ASIALEX and EURALEX). Each association will appoint an Assessment Committee consisting of its president and two officers, who will review and approve applications. The Committees are independent, and their decisions are final.

One grant will be awarded annually by each Committee. The maximun sum of each grant is $1,000, but a grant may be renewed fully or partially at the discretion of the Committee, or money held over for the following year. The grants are open to candidates anywhere in the world, who may apply to any one of the three Committees.

Applicants should submit a 500-word outline of their proposal, which, if accepted, will be published in Kernerman Dictionary News, as well as a 2,000-word summary on completion of their project. Five areas of lexicography have been selected for consideration for grants during the first three years:

1. **The study of the dictionary-using behaviour of language learners at the elementary school level, the junior-high school level and the high school level, and of non-academic adult language learners at the beginning and intermediate levels, as well as the design of these dictionaries.**
   Lower and intermediate level students constitute the vast majority of foreign language learners and dictionary users. Until now, their interests have been vastly neglected, as most dictionary research is carried out in universities and colleges, and have as subjects university and college students. It is hoped that by encouraging research at the pre-tertiary levels lexicographers can gain much-needed information about the dictionary needs of pre-academic language learners.

2. **Specialized corpora for foreign language learners**
   More and more dictionaries for L2 are being based on general word corpora. However, these corpora do not meet the linguistic and lexical needs of students who are learning another language. Basing learner dictionaries on corpora that reflect more closely the needs of learners of that particular language, not the level of a native speaker, might enhance pedagogic lexicography, making modern learner dictionaries more relevant and user-friendly. Work done toward the creation of specialized corpora would be very beneficial for learners of those languages.

3. **The function of lexicography in the process of vocabulary acquisition**
   Where new vocabulary means both new words and phrases, and new meanings of familiar words and phrases, and acquisition means storing in the reader's long-term memory, the function of the learner dictionary in vocabulary acquisition takes on an important dimension. Studies are needed not simply of how the dictionary helps the learner understand new meanings and uses, but also of how it facilitates their retention in the long-term memory. How can dictionaries assist users in remembering what they read?

4. **Trilingual and multilingual lexicography**
   A surprisingly large number of persons are bilingual. These include members of such groups as national and ethnic minorities who speak the language of the majority, refugees, emigrants, foreign students, transient workers and their families, and numerous others who may be on the move. Many of these learn a third language (such as English as the global lingua franca), becoming trilingual. The theory and design of trilingual and multilingual dictionaries, yet in its infancy, is now more viable with the development of computational lexicography.

5. **Lexicography programs concerning language preservation and survival**
   The genuine fears about the extinction of small languages in the face of globalization, together with the dissemination of a few favoured languages, are stimulating efforts to preserve small languages, or even to revive them in cases where the number of speakers has declined almost to the vanishing point. Dictionaries, whether historical, monolingual, bilingual or descriptive, can help in the preservation and revival of endangered languages. The construction of oral and written corpora is required for such languages, as well as critical research into available material.

These are general guidelines, and the Assessment Committees may use their own discretion in selecting awardees for innovative research in other areas.