An innovative medical learner's dictionary translated by means of speech recognition

Arnoud van den Eerenbeemt
medical lexicographer
Bohn Stafleu van Loghum, a Springer Media company, Houten, The Netherlands

I will be discussing a medical dictionary based on the Keyword in context (KWiC) concept and speech recognition as a valuable tool for both speeding up data input tasks and preventing overstrained hands and arms while using the keyboard. The latter is vital for me as my daily full-time work consists of compiling medical dictionaries for students and professionals (creating and updating complex and dynamic data) and creating medical spellcheckers.

Non-Anglophone medical students and health care professionals around the globe need an active command of professional English for their career. Yet the lexical tools available for acquiring these skills are few and insufficient: American and British explanatory dictionaries expect readers to be native speakers, while bilingual medical dictionaries are basically glossaries and provide unlabelled translations. The only medical learner's dictionary in the world to date is the excellent Fachwortschatz Medizin, *a comprehensive dictionary compiled by Michael and Ingrid Friedbichler. It helps non-native speakers to acquire language skills step by step, is structured using modular medical concepts and combines various lexical features:*

- *monolingual dictionary: 100,000 medical terms grouped into 1400 sections with key headwords defined in simple English; contextualized with collocations and sample sentences demonstrating correct use, extracted from a 20-million-word corpus of medically authoritative texts;*
- *semi-bilingual dictionary: support in the user's native language (German, Dutch) in the form of 42,000 translated keywords;*
- *thesaurus: synonyms, antonyms and related terms;*
- *domain-specific glossary: readers from all medical fields can focus on content relevant to their specialization;*
- *PC edition: full-context search, customizable display (pronunciation, definition, translations, collocations), cross-references etc.*

After acquiring the Dutch rights I realised out that farming out the translation work would require me to extensively monitor translators. I decided to translate the 42,000 medical terms myself instead, using speech recognition and a large monitor to display my database, a web browser, a word processor and two medical dictionaries. I developed voice-driven macros for automating 600,000 Google searches, creating 2000 records, searching dictionaries and my 52,000-record medical database etc. This allowed me to translate up to 400 terms per working day and complete this immense task within reasonable time.

1. Introduction

Medical lexicography involves working on highly changeable data, subject to frequent changes in spelling, nomenclature, classifications, laws and regulations etcetera. Terms and their description may differ from country to country, depending on medical opinions and culture (e.g. homeotherapy and phytotherapy in Germany). My work consists mainly of compiling mono- and bilingual medical dictionaries for medical students and professionals: updating on symptoms, diagnosis, therapy, adding new terms, labelling or even deleting obsolete terms and applying nomenclature and spelling. Besides I build medical spellcheckers and even write medical dictates for sponsored spelling contests for medical doctors.

2. English, a medical lingua franca

English has become the leading language of international communication in science and medicine. Non-Anglophone medical students and health care professionals around the globe are increasingly faced with the need to have not only a passive but also an active command of professional English, an essential skill for their career. They need lexical tools to acquire these language skills step by step, but these are few and insufficient. Monolingual defining dictionaries don't serve well medical students and professionals writing in medical English.

Bilingual medical dictionaries are little more than glossaries and tend to provide numerous unlabelled translations.

Fachwortschatz Medizin - KWiC-Web is the sole learner's dictionary for medical English. This 900-page dictionary was written by Michael and Ingrid Friedbichler, teachers of English for medical purposes (EMP) at Innsbruck University and other educational institutes in Austria and Germany. During their career they have been dealing with the needs and challenges of health care professionals whose first language is not English. In 1997 they decided to develop a dictionary that is comprehensive and, at the same time, specific enough to help non-native speakers from different medical fields acquire the English language skills they need step by step. On the basis of their experience teaching EMP, countless face-to-face editing sessions with medical writers and research in language acquisition, they developed in the past decade the KWiC-Web approach with the support of medical professionals and students and the help of computer-based lexicographic methods.

3. *Fachwortschatz Medizin* - KWiC-Web

The acronym 'KWiC-Web' stands for 'keywords [placed] in context and interconnected in a semantic web'. *Fachwortschatz Medizin - KWiC-Web* and its first foreign edition, *Pinkhof Medisch Engels*, is essentially the world's first learner's dictionary of medicine. It has been designed to help non-Anglophone health professionals use medical English terms in practice, unlike American or British medical dictionaries which assume the user is a native speaker practicing medicine. This semi-bilingual dictionary is structured along medical concepts in a modular fashion. It combines various functions:

- a monolingual medical dictionary, presenting 42,000 medical terms grouped into 1400 sections with key headwords defined in simple English;
- these key words are extensively contextualized with their typical collocations commonly used in medicine;
- authentic sample sentences demonstrate how medical terms are correctly used in English; for this purpose data have been extracted from a 20-million-word corpus of authentic and professionally edited texts authored by leading medical authors who are native speakers of English;
- medical terminology, jargon, and lay expressions are differentiated (register tags);
- thus all materials are evidence-based, i.e. they are up-to-date, authentic and reflect the way English is actually used among native-speaking medical professionals;
- help is given with correct pronunciation of difficult words;
- there is a domain-specific glossary, making professional English used in medicine accessible to health care professionals from all medical fields, who can familiarize themselves with terms and expressions relevant to their fields of specialization;
- the thesaurus provides synonyms, antonyms and related terms for each headword;
- the bilingual medical dictionary, acts as a learning tool, promoting efficient language learning in meaningful medical context; terms and phrases are indexed alphabetically in both languages, enabling users to look up search words in either direction;
- support in Dutch is offered for difficult terms and phrases by translating keywords and problem words in the user's language and giving language tips on pitfalls.

The main purpose is to enable the English-speaking reader to understand a medical term or expression. The modular structure enables medical professionals to train the special language

of selected medical fields step by step and according to their individual needs – both in EMP courses as well as by self-study.

The Dutch paper edition is an attractive 900-page book with blue support colour. A sample PDF can be downloaded for free on the product web site www.pinkhof.nl. The PC application is based on the Xaver interface (Doctronic GmbH). It features full-context search capability, a possibility to display or hide selected components (pronunciation, definition, translations, sample sentences, etc.), hyperlinked cross-references facilitate hopping to related entries in other modules and scalable screen text. Its full version can be downloaded and tried for free during 10 days, after which a validation code has to be purchased.

4. A lexicographer challenged to translate this dictionary on his own

I acquired the translation rights for a Dutch edition in 2007, to find out soon that farming out the substantial translating work required was both costly and inefficient, with a disproportionate amount of time needed for preparing data, monitoring external translators and proofreading their work result. Experienced medical translators were discouraged by the highly condensed lexical content and refused to accept the well-paid job. Institutions teaching medical translating suggested to have their students do the job as a training. So I decided to carry out the translation task myself: translating 42,000 medical terms and adding as many Dutch synonyms as possible.

A 24-inch monitor used in HD resolution allowed me to work with five applications without switching: database, browser, word processor and two medical dictionaries. The German source content in xml was reformatted, allowing me to work on it in a single file in TextPad, where with search/replace I performed repetitive partial translations, e.g. replacing unconditionally all occurrences of *Harn* into *urine* and all *-ung* suffixes into *-ing*, in many cases already yielding correct translations. I had my speech recognition software, Dragon Naturally Speaking, configured by a specialist with custom-built macros for pasting a candidate translation of a source-text transeme selected in Microsoft Word to the Windows clipboard, subsequently having my dictionary workbench Uniterm Pro (Acolada GmbH) look it up in 52,000 records, if needed having Uniterm create and fill out new records with the clipboard's content, having my browser search the internet on the clipboard's content etc.

With these automated, voice-driven tasks I was able to translate the book's full content through some 600,000 Google searches, scrutinizing search results and matching fitting results with my medicolexical data. After some fourteen months of brain straining the task was accomplished successfully: 42,000 Dutch valid translations had been found for an equal number of English terms. Besides, some 2400 Dutch medical synonyms had been found and added by speech commands to my database, to become headwords with a *see* reference.

5. Marketing

After the launch of the PC edition in January 2010 annual subscriptions have already been sold to several medical faculties. The semi-bilingual structure allows for adapting to the needs of medical professionals in other countries by replacing the original translations with equivalents in other languages. The book is currently being translated into Japanese by a team of six translators who expect to complete the work mid-2010, after some ten months' work.

http://www.pinkhof.nl/medisch-engels: full PC (Windows) edition 10-day trial period, free download of 60-page sample PDF.