TALKING DICTIONARIES

By
Olga Balogh

University of Debrecen
Institute of English and American Studies

Supervisor:
Dr. Béla Hollósi

Debrecen, Hungary
2010

Table of contents

Abstract

The talking dictionaries are very useful for blind people; they have many advantages against the printed ones. For example, if we would like to look up for a word, we need to ask for a sighted person's help, but it is not quite sure that the person near us can speak English. In braille there are only few dictionaries; they would cost a very high price to be printed. They only belong to some English books that are printed. We use the internet and the talking/online dictionaries with the aid of a speech programme named JAWS for Windows.

About The JAWS for Windows

JAWS has more than 15 years experience working with Windows, is distributed in more than 50 countries and is translated into 17 languages. Applications that JAWS supports: Microsoft Office Suite, Internet Explorer, MSN Messenger, Corel, Word Perfect, Adobe, Acrobat Reader. JAWS consists of 2 parts: Eloquence and Solo Realspeak SAPI5 synthesizers. Supported languages: American, Australian, British English, Castilian, Latin American Spanish, French, German, Italian, Brazilian Portuguese, and Finnish.

Source:

http://www.freedomscientific.com/jawsflyer.pdf

Advantages and Disadvantages Of Talking Dictionaries

In this section I would like to view the advantages and disadvantages of talking dictionaries. The Gib dictionary runs under Windows 98, although there are versions that run under XP. (I am going to return to this later). The Dicfor is disadvantaged by the registration page. Without registering the programme cannot be downloaded. Registration page, sometimes, is unmanageable for blind people, because we should have to ask for sighted people’s help, to copy a word or a checking-code, which is below a certain picture, which we – blind people - cannot see.

The WTCD has the same disadvantage, although it has more advantages. WTCd (word training communication and dictionary) is a Hungarian invention. It consists of 3 systems: A word-teaching system, which joins the internet, and it is advantaged for those who do not like opening their dictionaries each day; a talking dictionary, and the most important part of WTCD is a communicating system, which functions like skype or msn. With the aid of this, people can communicate with each other, and the programme can check the vocabulary and the writing.

Data Of GIB

The GIB talking dictionaries are constracted by Scriptum. The first electronic dictionary released in 1992, on a 1.44-MB flopydisc. The name of it was: spt_gib. Expansion was 3.6 mb. The name GIB is an abbreviation for Graphical Interactive Book. As the name shows, these dictionaries are graphical, so it is a disadvantage for blind people. The first addition was the English-Hungarian dictionary by László Országh. Talking dictionaries of Gib: English-Hungarian, Hungarian-English dictionaries. The first version was: Gib 1.0, nowadays: 3.2. The GIBweb: The aim of GIBweb is the availability of those products, which had already appeared on CD-roms. For it’s function need: Microsoft Nt 4.0 and Internet information server 3.0. Advantages are: Gibweb has a searching-card system, which gives opportunity to search in more dictionaries. Disadvantage is that the word-list lacks from it. Communication is slower with gibweb than with CD-rom. Gib4x: The Gib4x is a trial version, which was devoted to help with translation.

The Dictzone dictionary is the best dictionary for blind people; it has many advantages. The first is that we do not have to register for the webpage. The second is, we can put those words down which we searched for. The JAWS also can read them out, and we can make unique notes of words. Only disadvantage is that we need to download the words each by each, because the player does not start at once from the page.

The Role Of Spoken Corpora In Dictionary-making

Definition of corpus: The word “corpus” comes from latin, means “body”. In linguistics and lexicography, a body of texts, utterances, and other speciments considered more or less representative of a language. The plural is corpora. Computer corpora are, especially the bodies of natural language, e.g. whole texts, samples of texts, which are stored in machine-readable forms. Definition of corpus linguistics: Corpus linguistics is the study of language based on examples of real life language use. According to Kennedy, corpus linguistics is not an end in itself but is one source for evidence for improving descriptions of the structure and use of languages, and for various applications, including the processing of natural language by machine and understanding how to learn or teach a language. (Kennedy, 1998 1)

List of pre-electronic corpora

Corpus-based research goes back to the beginning of the 18th century. To make texts machine-readable, there were 5 different ways:

- biblical and literary studies,
- lexicography,
- dialect studies,
- language education studies, and
- grammatical studies.

Creating Of Machine-readable Corpora

Keyboarding: Word-processing became a standard way of creating electronic documents some twenty years ago. Since then, keyboarding has been an obvious option of making written texts machine-readable. Scanning: Although, the keyboarding is a slow process, nowadays scanning is the more popular option. It means two consecutive operations; first, the text is scanned with the aid of a hand-held device (e.g. hand scanner or a pen scanner), a page scanner, a flatbed scanner, or a standalone scanner.

Hand held devices are usually relatively inexpensive but their performance is modest, the other type of scanners can do a page at a time. Second, as the text scanned in is in some digitized image format, it has to be converted into a flow of characters with softwares. Scanning softwares: Scanning softwares have 2 main types: OCR (Optical Character Recognition), and ICR (Intelligent Character Recognition). Scanning programs: Recognita, Omnipage, Wordscan.

Capturing electronic text

The machine-readable texts are available on the World Wide Web or CD-roms. These texts may serve as the main material of corpus-based studies and database operation.

Spoken Corpora

“Spoken corpora can be created only by transcribing recorded speech first, irrespective of the type of spoken genre (face-to-face conversation, radio quiz programme, inaugural address, telephone conversation, etc) concerned. Transcription is a slow and complex process because of the very conversation from the spoken to the written medium, turn-taking idiosyncrasies, parallel talk, etc. In its handling of speech, converted from a spoken recording into a machine-readable form, discourse analysis is dependent on transcription. There are 2 models in current use: transcriptions which reflect the chronological development of speech based on turn-taking units; and transcriptions which reflect the interpretation of the analyst based on discourse units. Transcription should include as little information about the original recording as possible.

Compilation of spoken corpora

Despite the wide experience gained in the compilation of written corpora, working with spoken language data is not immediately straightforward as spoken language involves many novel aspects that need to be taken care of. The fact that spoken language is transient is sometimes offered as an explanation for why it is more difficult to collect spoken data than to compile corpus of written data. However, it is not just the capturing of data that is anything but trivial. Once the data is stored, the next step is to introduce some transcripts.

Further annotations such as pos tagging, lemmatization, syntactic annotation and prosodic annotation may build upon this transcription. Among the problems encountered in the processing of spoken language data are the following:

1. There is as yet little experience with the large scale transcription of spoken language data. Procedures and guidelines must be developed, and tools implemented.
2. Well-established practices that have originated from working on written language corpora do not hold up when trying to cope with the idiosincracies of the spoken language. This is true for all levels of linguistic annotation. Annotation schemes need to be reconsidered and tools must be adapted.
3. In so far as standards have emerged they need to be adapted in order to be able cater for the needs of spoken corpora.
4. By their very nature, spoken corpora bring together speech and language technologists and linguists from various backgrounds. Ideally, such corpora should address the needs of all these different user groups. Often, however, there is a conflict of interest. For example, the quality of recordings of spontaneous conversations in noisy environments. Although highly interesting and worthwhile from a linguistic perspective will prove too poor to be of any use to someone doing research in speech recognition.

Spoken corpora and their descriptions:

In this section I am going to write about corpora, their description and their roles.

1. Archival Sound Recordings

Archival sound recordings is the development project of the British Library sound archive’s extensive collections. The British Library holds one of the world’s foremost sound archives with a collection of over 3,5 million audio recordings. These come all over the world and cover the entire range of recorded sound from music, drama, literature oral history wildlife and environmental sounds.
Jisc: Joint Information System Committee.

2. Childes: (child Language Data Exchange System)

Childes is the child language component of the Talkbank system. Talkbank is a system for sharing and studying conversational interactions.

Sfb 441 is a link-collection of corpora.
General corpora: ELRA: Europian Language Research Association), LDC (linguistic data consortium), ECI/MCI: European Corpus Initiative Multilingual Corpus.
Electronic Text Center: This is the archive of the library of the University of Virginia.

English Corpora:

1. Penn parsed corpora of historical English
The penn parsed corpora are syntactically annotated corpora of prose text samples of English from the indicated time periods. Their syntactic annotation permits not only searching for words and word sequences, but also for syntactic structure. The corpora are designed for the use of students and scholars of the history of English, especially the historical syntax of the language, and they are publicly available. The penn parsed corpora of historical English are available on CD-rom.

2. The British National Corpus (BNC)
The British National Corpus contains 100 million words and samples of written and spoken language.

3. Icame (international computer archive of modern and medieval English)
Icame is an international organization of linguists and information scientists working with English machine-readable texts. The aim of the organization is to collect and

distribute information of English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and make material available to research institutions.

Cobuild, ICE: international corpus of English

COCA (Corpus of Contemporary American English:
COCA contains 400 + million words (from 1990 to 2009) According to the "word sequency" (a research of words in COCa from 1992 to 2007) there are 20000 words.

Darpa (Defense Advanced Research Projects Agency)

Darpa is the research and development office for the U.S department of defense. Their aim is to teach military techniques.

These projects are aimed at investigating some particular aspects of grammar, vocabulary, discourse, genre or usage usually warrant the creation of some specialized corpus to meet the needs of the project at hand. Some of these are designed to represent narrow fields, such as petroleum industry, doctor-patient interaction or air-trafic control, while others cover some larger segment of language use, such as dialect, regional, non-standard or learners corpora. Leech (1992: 112) has described the development of training corpora and test corpora as specialized corpora to facilitate the building of models processing.

The types of corpora mentioned so far are finite in size and for electronic corpora one million running words became a kind of unofficial standard size from 1964 until the early 1990s. More recently, however, some corpora have been compiled containing vast amounts of text which are added to, and which are not necessarily balanced and structured, so text does not systematically and proportionately come from particular genres or registers.

Size vs. Balance in Corpus Building

A general corpus is usually designed and created in a way that it will be balanced. (E.g. objective sampling techniques will have been used to ensure that it will be a representative corpus of a language as a whole and will not be biased or skewed in any way. Although, this representativeness is only a qualified one in most cases as even when major publishers channel a lot of funds into creating large balanced corpora of a language (for example, for lexicographic purposes), the end product is likely to show disbalance. In most of cases size solves the problem of the representativeness of a corpus.

There are 4 reasons for focusing merely on size.

> “A collection of machine-readable text does not make a corpus.
>
> 1. The vast growth in resources of machine-readable text has taken place exclusively in the medium of written language. Until speech-recognition devices have developed the automatic input of spoken language to the level of present ocr devices for written languages, the collection of spoken discourse on the same scale as written text will remain a dream of the future.
> 2. While technology advances quickly, human institutions evolve slowly. This platitude has particular relevance to the collection and distribution of computer corpora, where that most slowly evolving human institutions retards the full availability of the resources which technology makes possible. The law forbids copying files without the express permission of copyright holders.
> 3. In the present context, software may be taken to include not the raw or pure corpus, which can be collected more or less automatically in its original machine-readable form, but the annotations which linguists others may add to it, and the computer programs designed to process corpora linguistically.

(leech, 1991: 10-12

Description of other corpora

In this section I am going to write about the COBUILD project and other corpora.

Cobuild Project

"Cobuild is the first major machine-readable corpus-based lexicographical project since the American Heritage Project of the 1970s. This project was made use of the Birmingham Collection of English test and was joint venture between a major commercial publisher, Collins, and a research teambased in the English Department of the University of Birmingham.

Cobuild: (Collins Birmingham University International Language Database)

This corpus has usually been referred as the Cobuild corpus, but also known as The Birmingham Corpus.

The task of compiling the corpus, which was to be the basis initially for the production of a new English dictionary, began in 1980.

The Main Cobuild Corpus had about 7.3 million words by August 1982. A quarter of the corpus was constituted by spoken texts, it set out to represent broadly general rather than technical language, authorship was 75% of male and 25% of female, and the passages included mostly standard British English producted by adults aged 16 or over (70%) The Cobuild Dictionary was published in 1987. The Main Corpus supplemented, by a 13-million-word Reserve Corpus. It was for fiction and non-fiction.

Cobuild project was not only constituted the largest storehouse of authentic English for lexicography, but they also served as empirical grammars, concordance lists and language teaching corpora. In 1990, John Sinclair was the director of Cobuild; this year the project became a corpus-building, named Bank of English, a corpus of a hundreds of million of words. By 1997 the size was about 300 million of words and growing.

The Role of Computational Tools In Corpus Linguistics

Nowadays, it is evident that a corpus needs the support of computational environment, providing software tools both to retrieve data from the corpus and to process linguistically the corpus itself. (LDC users have the opportunity to create new data). In spite of wide availability of tools such as concordancer packages, grammatical tagging systems, most of them exist only on prototype forms without any documentations or public availability. These tools are the following types:

1. General-purposed corpus data retrieval tools: These are existing in concordance facilities, in being able to handle corpora in complex formats, in being able to sort and search in various ways, and to derive from an annotated corpus various kinds of data.
2. Annotation tools: These might be used for automatic processing or semi-automatic interactive use or accelerated manual analyses and input. Interactive windowing facilities have much unrealized potential in this field. One tool for which has strong demand is a robust corpus parser: something that will be able to provide a reliable but shallow parse for large quantities of text.
3. Tools that provide interchange of information between corpora and grammatical databases: At the most simplest level, a program which derives lexical frequency lists from raw corpora, is a device for deriving lexical information from a corpus. From a tagged corpus, a lemmatized frequency may be derived.

These are examples how corpora can create or augment linguistic databases. A testing algorithm can also use observed corpus data as a mean of evaluating the coverage of a grammar or the performance of a parser.

From these and similar instances we can see that between corpora and linguistic databases or linguistic models there is an important and rather complex channel of information transfer, for which special tools are required.

The mere existence of a large corpus will not satisfy demand for linguistic data. A set of the general tools for processing the corpus will be essential. Many of such tools are already exist and are in use, but they are often designed to meet very specific local needs and there is work to be done on agreeing on standard formats for data derived from a corpus.

Kwic concordances, word frequency lists, collocation statistics will be basic requirements and a software to perform these basic tasks can be available as part of the corpus package. Tagging software and parsers will be more difficult to implement and their design more difficult, but these tools should also be available.

Basic Tools

These tools might include:

1. Word frequency: A software to produce lists of word types their frequency in the corpus and also some statistical profile of the relation of types to tokens in the corpus. (e.g. the word frequency of COCA)
2. Concordancing: Concordancing is a text retrieval and indexing software with features appropriate for linguistic analysis.
3. Interactive searching: Flexibility in search and display/presentation.

Advanced Text Handling

In order to allow more sophisticated statistical analyses to be carried out on a large corpus, it may be useful to implement a number of more advanced text processing tools which can automatically process linguistic information in a corpus. Advanced text handling might include the following software:

1. Lemmatization: to relate a particular inflected form of a word to its base form or lemma, whereby enabling the production of frequency and distribution figures which are less sensitive to the incidence of surface strings
2. Part-of-speech labeling: To assign a word class or part-of-speech label to every word. This allows simple syntactic searches to be carried out.
3. Parsing: to assign a fully labeled syntactic tree or bracketing constituents to sentences of the corpus.
4. Collocation: to compute the statistical association of word forms in the sentence.
5. Sense disambiguation: to distinguish which several possible senses of a given word is being employed at each occurrence. This area shows results for look-up in a machine-readable dictionary or identifying collocational patterns.

Comparison of the work of John Newman and Steven Greenberg

In this section I will examine the aspects of a corpus linguist and a theoretical linguist through “Rationale and analysis” by John Newman and a paper from 2003 by Steven Greenberg.

John Newman is the professor of the University of Alberta. He is a corpus linguist. Steven Greenberg is a theoretical (or armchair) linguist. Newman focuses on practice, relies on data of corpus. He works with parsers. His researches rely on databases of corpora.

Despite of John Newman, Steven Greenberg treats corpora as theories. He deals with the future, and his research relies on the forms of speech. (for example, on breaks, etc).

Project work

The Sound Archive of the British Library contains 25300 digitized texts. Unfortunately, the files are difficult to download for the blind, because, in one hand, there are graphics which JAWS cannot read, or the programmes need special treatment to download. On the other hand, the files are copyrighted, and only the members can get to them. The third mean reason why the files cannot be downloaded, is the collection of broken links. The best example for this the BASS (Bavarian Archive for Speech Signals.) corpus.

There are corpora, which have special softwares; for example, the SLAAP (Sociolinguistic Archive and Analysis Project) has such complicated softwares, which are impossible to download for the blind people. It has, for example, a vowel normalizer. But the files cannot be downloaded.

The COCA contains 400 million words (1990-2009); according to the word frequency, it contains 385 million texts (1990-2007) and these contain 2000 words/year. The COCA (corpus of contemporary American English) serves for making frequency-based dictionaries.

The Darpa is aimed for military uses. It has foundation also, and we have opportunity to make small business.

The Elisa corpus (English Language Interview Corpus As a second-language learning application) is developed at the Tuebingen and the University of Surrey as a resource for learning and teaching and interpreter training. It contains interviews with the native speakers of English. They talk about their professional career in the media or tourism.

Elisa currently contains 28 interviews and about 6000 words.

The ICE (international corpus of English) began in 1990 with the primary aim of collecting materials for comparative use of English worldwide. 20 research teams around the world are working on new corpora of regional variety of English. Each Ice corpus contains 1 million words.

The LDC (linguistic Data Consortium) is an open consortium for universities, companies and research laboratories. It researches and distributes databases, lexicons. LDC was founded in 1992 by the ARPA, and was supported by the IRI.

The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription. It consists of 100 texts and 5000 running words.

The SOHP is the documenting of oral history of South America. This corpus gives opportunity to download the files with several formats, for instance: mp3, xls; and we can read them also. My aspects will be focused on the MP3 files, because of the theme of my thesis are talking dictionaries. I found lots of interesting interviews in this page, and heard old words. This webpage contains for example political interviews, interviews about family lives, school life etc. It was very interesting for me to hear elder peoples' voices who born in the 19th century.

Glossary

American researchers say that we use more than 100 million words a day during our work and everyday communication. This human activity would overburden an average computer because it demands complex abilities and huge amount of energy. As modern people live in the fast lane, when they want to learn foreign language they have no time to use printed dictionaries, but they need modern and easy-to-use devices. Talking dictionaries are ideal solutions for this purpose especially for blind people who are unable to use the printed materials.

In my thesis I compared some talking dictionaries which are available nowadays. I found the Dictzone dictionary the best because blind people can take notes in Braille and it can help a lot when they learn languages and prepare for exams. It can be used on the internet for free for everybody.

Bibliography

Leech, G. 1992. Computer Corpora, Lanchester: Lanchester University Press.