An Intelligent Multi-Dictionary Environment

Gábor Prószéky
MorphoLogic
Késmárki u. 8., H-1118 Budapest, Hungary
proszeky@morphologic.hu

Abstract

An open, extendible multi-dictionary system is introduced in the paper. It supports the translator in accessing adequate entries of various bi- and monolingual dictionaries and translation examples from parallel corpora. Simultaneously an unlimited number of dictionaries can be held open, thus by a single interrogation step, all the dictionaries (translations, explanations, synonyms, etc.) can be surveyed. The implemented system (called MoBiDic) knows morphological rules of the dictionaries' languages. Thus, never the actual (inflected) words, but always their lemmas – that is, the right dictionary entries – are looked up. MoBiDic has an open, multimedial architecture, thus it is suitable for handling not only textual, but speaking or picture dictionaries, as well. The same system is also able to find words and expressions in corpora, dynamically providing the translators with examples from their earlier translations or other translators' works. MoBiDic has been designed for translator workgroups, where the translators' own glossaries (built also with the help of the system) may also be disseminated among the members of the group, with different access rights, if needed. The system has a TCP/IP-based client-server implementation for various platforms and available with a gradually increasing number of dictionaries for numerous language pairs.

Introduction

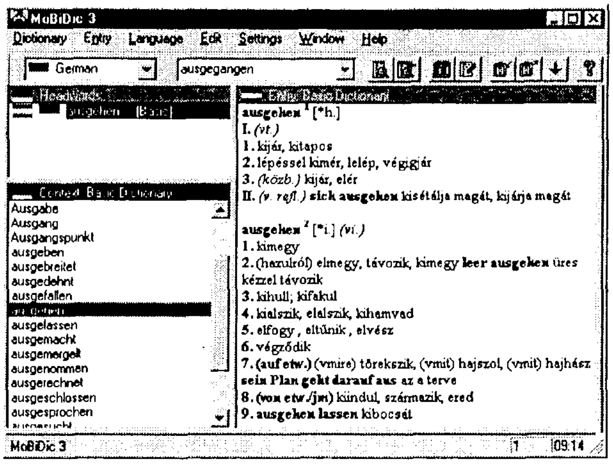
"The whole world of translation is opening up, to new possibilities, and to technological and methodological change" (Kingscott 1993). Some years after the above claim, we see that software tools for translators, even the most recent ones, do not yet guarantee perfect solutions to automatic translation. More and more systems introduce, however, new facilities to the translator working in a computational environment. As Hutchins says, "the best use must be made of those systems that are available, and the producers and developers must be encouraged to improve and introduce new facilities to meet user needs." (Hutchins 1996)

It is almost a commonplace that texts – books, newspapers, letters, official memos, brochures, any type of publications, reports, etc. – in the nineties are written, sent, read and translated with the help of the electronic media. Consequently, traditional information sources, like paper-based dictionaries, and lexicons, are no longer as much a part of the translation environment.

Electronic dictionaries for most developers just mean, however, to make the well-known paper dictionary image appear on the computer screen. It is easy to understand why we say that dictionary computerization does not mean producing machine-readable versions of traditional printed dictionaries, but the combination of the existing lexical resources with up-to-date language technology.

On the other hand, there is a question whether we have to continue in the traditional way of developing new – and different – lexicons for any new application/system, starting from scratch every time and therefore consuming time, money and manpower, or is it new lexicons.

In what follows, timely to think of the possibility of making the effort to converge, trying to avoid unnecessary duplications and – where possible – building on what already exists (Calzolari 1994). Consequently, in the near future we have to combine the two above needs: making existing lexical resources computationally accessible and showing the strategy how to develop we try to argue for changes in development strategies of electronic translation dictionaries. Today's lingware technology can – and must – use dynamic actions, like morpho-syntactic analysis, lemmatization, spell checking, and so on. On the other hand, dictionaries can never be full in any sense, therefore we have to make parallel multi-dictionary access possible. It means that a single dictionary look-up should use an unlimited number of lexical resources that are available for the translator.



ausgehen [1] [*h.]
I. *(vt.)*
1. kijár, kitapos
2. lépéssel kimér, lelép, végigjár
3. *(közb.)* kijár, elér
II. *(v. refl.)* **sich ausgehen** kisétálja magát, kijárja magát

ausgehen [2] [*i.] *(vi.)*
1. kimegy
2. (hazulról) elmegy, távozik, kimegy **leer ausgehen** üres kézzel távozik
3. kihull; kifakul
4. kialszik, elalszik, kihamvad
5. elfogy, eltűnik, elvész
6. végződik
7. **(auf etw.)** (vmire) törekszik, (vmit) hajszol, (vmit) hajhász **sein Plan geht darauf aus** az a terve
8. **(von etw./jm)** kiindul, származik, ered
9. **ausgehen lassen** kibocsát

Figure 1

Look-up of a morphologically complex inflected form: '*ausgegangen*' in a German–Hungarian dictionary.

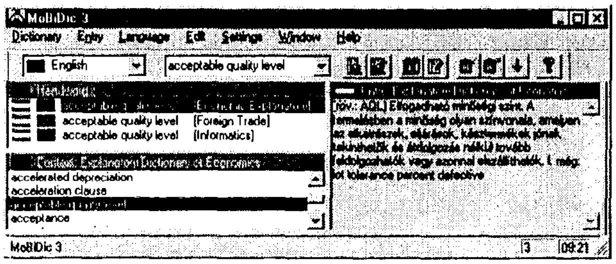
1 The MoBiDic Look-up System

To start with the most natural activity concerning dictionaries is searching them for a single word. There is no problem if it can be found among the headwords of the dictionary, that is, when the input string can match. But sometimes the translator starts the look-up process by clicking an inflected word-form of an open document that cannot be found among the headwords. For the user it is a boring and time-consuming task to type the lexical form, that is, the one accepted letter-by-letter by the dictionary. To make the system able to find the stem of the input word-form automatically, MoBiDic uses a lemmatizer that provides the dictionary look-up module with the stem(s) to be found (Figure 1).

Translators frequently want to find the word as a part of *multi-word expressions* or idioms. If the user does not know whether the actual word is part of some phrasal compound or idiom, the traditional paper dictionaries are very difficult to use. Namely, if the word in question is the so-called headword of a multi-word expression, it can be found easily. In case it is not the headword, one has to know the phrasal compound the word is a part of, but it is a typical "Catch 22" situation: if the expression is known why to search the dictionary for it? MoBiDic helps the user to find all the multi-word expressions containing the actual word's stem, independently whether it is a headword or not. E.g. not only *'lead'* but both *'dog'* and *'life'* provide us (among others) with the multi-word expression '*lead a dog's life*' that can be found under *'lead'* only in a paper dictionary. In other words, users of the traditional dictionaries are supposed to know the expression (what's more: the keyword of the expression) to find it in the lexicon. Search for *'lead a dog's life'* through its components gives the following result in MoBiDic:

lead {lead, leads, leading, led}
  27 occurrences in expressions of the basic dictionary,
dog {dog, dogs, dog's, dogs'}
  21 occurrences in expressions of the basic dictionary,
life {life, lives. life's, lives'}
  77 occurrences in expressions of the basic dictionary,
lead AND life
  5 occurrences in expressions of the basic dictionary,
dog AND life
  2 occurrences in expressions of the basic dictionary,
lead AND dog
  1 occurrence in expressions of the basic dictionary,
lead a dog's life
  1 occurrence as an expression in the basic dictionary.

'Bi' is somewhat misleading in the name MoBiDic. Bilingual in this sense means that the source and the target language are not the same types of object for the program. For MoBiDic, source language is the language the *morphology* of which has to be known, to provide the user with adequate output. The output is expected to be in the target language – the characters, the alphabetic order, etc. of which has to be known to make the hits appear on the screen in adequate format. Of course, the source and target languages can be the same, e.g. in *explanatory or etymological dictionaries* (Figure 2).


MoBiDic 3
Dictionary Entry Language Edit Settings Window Help
English
acceptable quality level
Headwords
acceptable quality level [Foreign Trade]
acceptable quality level [Informatics]
accelerated depreciation
acceleration clause
acceptance
MoBiDic 3
3
09:21

(röv.: AQL) Elfogadható minőségi szint. A termelésben a minőség olyan színvonala, amelyen az alkatrészek, eljárások, késztermékek jónak tekinthetők és átdolgozás nélkül tovább feldolgozhatók vagy azonnal elszállíthatók. L még: lot tolerance percent defective

Figure 2

Hungarian explanation of '*acceptable quality level*' in the *English–Hungarian Economical Explanatory Dictionary*.

There is an another sort of monolingual dictionary, the *synonym dictionary*. The translator frequently wants to use a synonym (antonym, hypernym, hyponym) of the actual word. An intelligent software tool, like MorphoLogic's Helyette[1], is the combination of a thesaurus (synonym dictionary), a morphological analyzer and a *generator*, because the output is re-inflected according to the morphological information contained by the input word-form. The – so-called inflectional – thesaurus works as follows:

INPUT:	*came*
ANALYSIS :	*came* = *come* + Past
STEM:	*come*
SYNONYM:	*go*
SYNTHESIS:	*go* + Past = *went*
OUTPUT:	*went*

There are special sorts of information in a dictionary. For example, pronunciation is not typically needed for translation, but can be useful for language learners. Pronunciation of the word is, therefore, an information that should be switched on and off, according to the user's needs. In an electronic dictionary it is expected that not only the written phonetic transcription, but also the *spoken* output can be heard. If the dictionary supports multimedia, explanatory *pictures* can help understand the word, even for professionals, not for language learners only (Fig. 3).

If the translator makes a spelling error, first a *speller* starts, and then the corrected word-form is sent to the dictionary look-up system.

Examples do belong to the entries of large, professional paper dictionaries. In electronic dictionaries occurrences of the word in texts of other authors, or wants to see bilingual texts with their aligned translations: monolingual or aligned bilingual *corpus*, a free text search module and a lemmatizer.

2 Dictionaries in MoBiDic

The lexicographic basis for MoBiDic is supplied by various publishing houses. More precisely, MorphoLogic has licenses to almost 50 dictionaries already published in paper format of miscellaneous topics, diverse sizes and many language pairs. The user can choose which dictionary to use in general, and which of them open actually. Currently, if all the available dictionaries are open, MoBiDic handles approximately 1 million lexical entries.

Some of the dictionaries, mainly the terminological ones, have usually a very simple list-based structure. Dictionaries shown by Figure 1 and Figure 2, however, appear on the screen with the traditional paper dictionary image. It is done by using SGML representations and an on-line SGML–RTF conversion. MoBiDic can do exact structural search not influenced by the layout at all.

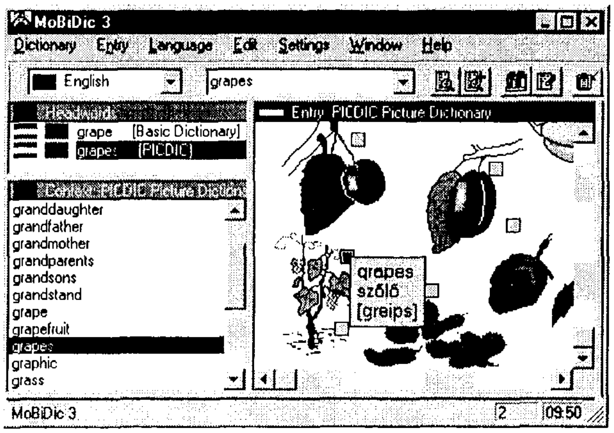
Generally, the original lexical resource – even it has been available in electronic format – did not use SGML. For this reason, a special system for a semi-automatic conversion of some formatted text files containing dictionary data to SGML format has been developed for the MoBiDic environment. This system is not available for the end-users, it serves industrial purposes.[2] First, in order to enable selective access to the information in dictionary entries, a thorough structural analysis is done, while inconsistent and faulty entries are marked. They are corrected later, manually. The resulting SGML-annotated dictionaries are enhanced with the necessary indexes. They are lemma-variants and expanded sub-entries made with the help of existing language technology modules (Prószéky 1994).

Users like to work with their own little vocabularies, glossaries, and the professional translator is usually asked to use official translation

[1] To be combined with **MoBiDic** in the near future.

[2] See http://www.morphologic.hu/e_sgml.htm

equivalents provided by the employer. These glossaries are generally never published, but there is a need to us them in the same environment. MoBiDic is able to treat user dictionaries containing any type of information sources (lexicons, encyclopedias and dictionaries).

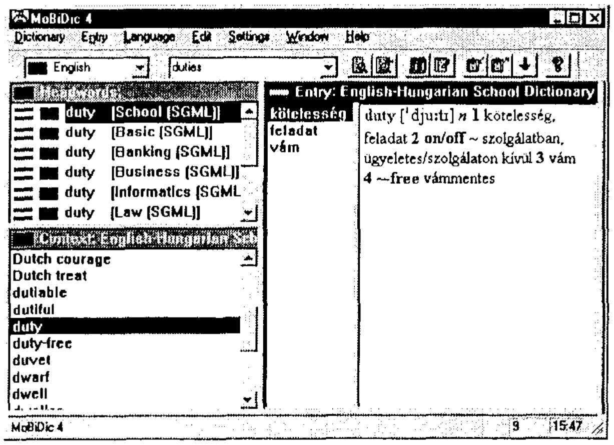

MoBiDic 3
Dictionary Entry Language Edit Settings Window Help
English
grapes
Headwords
Context: PICDIC Picture Diction
Entry: PICDIC Picture Dictionary
grapes
szőlő
[greips]
MoBiDic 3
2
09:50

grape [Basic Dictionary]
grapes [PICDIC]

granddaughter
grandfather
grandmother
grandparents
grandsons
grandstand
grape
grapefruit
grapes
graphic
grass

Figure 3

'*grapes*' (from the PicDIC picture dictionary) with pronunciation in **MoBiDic**


MoBiDic 4
Dictionary Entry Language Edit Settings Window Help
English
duties
Headwords:
Context: English-Hungarian Sch
Entry: English-Hungarian School Dictionary
MoBiDic 4
9
15:47

duty [School (SGML)]
duty [Basic (SGML)]
duty [Banking (SGML)]
duty [Business (SGML)]
duty [Informatics (SGML
duty [Law (SGML)]

Dutch courage
Dutch treat
dutiable
dutiful
duty
duty-free
duvet
dwarf
dwell

kötelesség
feladat
vám

duty ['dju:tı] *n* **1** kötelesség, feladat **2** on/off ~ szolgálatban, ügyeletes/szolgálaton kívül **3** vám **4** ~free vámmentes

Figure 4

Search for the (lemma of) '*duties*' in a set of English–Hungarian dictionaries

The strength of this method is that user dictionaries are looked up for a word exactly when other dictionaries, thus translator's remarks can also be read when other dictionaries provide the user with their translation equivalents. Here we have to emphasize again that MoBiDic is not yet another electronic dictionary, but a multi-dictionary environment where a single word is sent to every open dictionary by a single mouse-click. In Figure 4 the user started from the word-form '*duties*', and eight dictionaries (that are open and contain English either on the source or the target side) send translations to the screen.

3 Implementation Features

The most recent development is MoBiDic's client–server implementation. Its server side (Windows NT, Unix and Novell) consists, in fact, of two servers: the linguistic server and the dictionary server. The user interface and screen handling modules will take place on the (Win, Mac, Linux, Java, etc.) client side.

There are many software modules of other vendors on the market that can also be combined with MoBiDic through its well-defined *application programming interface* (API). With the help of this API the user can communicate to the other modules from MoBiDic without leaving it. Because of technical and legal reasons, it can, of course, be done in collaboration with the developer of the product in question. The picture dictionary shown by Figure 4 is a working example: the vocabulary part of the (also commercial) CALL program called PicDIC is available for MoBiDic users from the familiar environment.

Translators who generally use their favorite word-processor while translating can use MoBiDic from their word-processing tools with the help of the included macros. Another important issue is that users can use their CD-ROM drive for other purposes while translating. Namely, MoBiDic has minimal space requirement because of its compression method[3], therefore the full dictionary system can be copied to the hard disk: thus the CD drive is freed and can be used for other purposes.

4 Comparison with other methods

There are several dictionary programs both in laboratories and on the market, but only some of them share the so-called "intelligent" features with MoBiDic. Rank Xerox developed in the COMPASS and Locolex projects a prototype that accesses enhanced and structurally elaborated dictionaries with an intelligent, context-sensitive

[3] Average 1–2 Mb/dictionary.

look-up procedure, presenting the information to the user through an attractive graphical interface. (Feldweg and Breidt 1996) Unlike MoBiDic, it does not have access to more than one dictionary at the same time. Consequently, user dictionaries are not supported. SGML is, however, used both in the dictionary and the corpus modules. There is a focus on the intelligent treatment of multi-word units in the IDAREX formalism (Breidt et al 1996). Another project with similar aims is GLOSSER. Its prototype (Nerbonne et al. 1997) carries out a morphological analysis of the sentence in which the selected word occurs and a stochastic disambiguation of the word class information. This information is then matched against a (single, but SGML) dictionary and corpora. The GLOSSER prototype displays context dependent translations and on request, examples from the available corpora. Neither of the above developments nor other web dictionary services (e.g. WordBot) share all the important features with MoBiDic: client–server architecture, multi-dictionary access, user dictionary handling, parallel (and intelligent) dictionary and corpus look-up. What's more, MoBiDic is commercially also available, that is tested by thousands of "real" end-users.

Conclusion

MoBiDic is a multi-dictionary translation environment based on a client–server architecture. It consists of the following main parts: linguistic server, dictionary server and the client with the graphical user interface. There are several benefits:

(1) the linguistic server is dictionary independent and language dependent[4];

(2) the dictionary server has intelligent access to various sorts of dictionaries (from SGML to multimedia) and bilingual corpora;

(3) simultaneously an unlimited number of dictionaries can be held open, thus by a single interrogation step, all the dictionaries (with translations, explanations, synonyms, etc.) can be surveyed;

(4) the translators' own glossaries built with the help of the system may also be disseminated (as new dictionaries, with the needed copyrights) among other users, if needed;

(5) it has an open architecture and a well-defined API;.

(6) it has been implemented and is available with a gradually increasing number of dictionaries for numerous language pairs.

MoBiDic is, therefore, not a research project only, but a set of translation tools for a wider public.

References

Breidt. E., F. Segond and G. Valetto (1994) Local Grammars for the Description of Multi-Word Lexemes and Their Automatic Recognition in Texts. *Papers in Computational Lexicography,* Linguistics Institute, HAS, Budapest, pp. 19–28.

Calzolari, N. (1994) Issues for Lexicon Building. In: A. Zampolli, N. Calzolari & M. Palmer (eds.) *Current Issues in Computational Linguistics: In Honour of Don Walker.* Kluwer / Giardini Editori, Pisa, pp. 267–281.

Feldweg, H. and E. Breidt. (1996) COMPASS – An Intelligent Dictionary System for Reading Text in a Foreign Language. *Papers in Computational Lexicography,* Linguistics Institute, HAS, Budapest, pp. 53–62.

Hutchins, J. (1996) Introduction. *Proceediings of the EAMT Machine Translation Workshop,* Vienna, pp. 7-8.

Kingscott, G. (1993) Applications of Machine Translation. In: *Transferre necesse est... (Current Issues of Translation Theory),* Szombathely, pp. 239–248.

Nerbonne, L. Karttunen, E. Paskaleva, G. Prószéky and T. Roosmaa (1997) Reading More into Foreign Languages. *Proceedings of the Fifth Conference on Applied Natural Language Processing,* Washington..

Prószéky, G. (1994) Industrial Applications of Unification Morphology. *Proceedings of the 4th Conference on Applied Natural Language Processing,* Stuttgart, pp. 157–159.

[4] Recently, English, German, Hungarian, Polish, Czech and Romanian morphological components are available for the MoBiDic users. Descriptions for further languages are under development, see the web site http://www.morphologic.hu for the actual list of languages.