Gábor Prószéky

How “truly electronic dictionaries” of the 21st Century should look like?

Abstract

The first part of the presentation follows the main points of Atkins (2002) paper on the future of bilingual dictionaries. More than ten years ago she claimed that – besides the still existing printed dictionaries – the future should produce “truly electronic dictionaries” enriched with new types of information. But what sorts of new information have appeared in our dictionaries in the last decade? We try to enumerate the most important features which make today's traditions really different from those of the past. With the help of these features the basic design of our near future's electronic dictionaries is sketched out. Various new aspects must be taken into consideration if the dictionaries of the 21st century are to be better usable than the ones of the previous ages. In the second part of the presentation we show some methods of how new functions of a dictionary appear in the reality.

1. What sorts of new information have appeared in our dictionaries in the last decade?

In 1996, at the EURALEX Congress Sue Atkins gave a speech on the future of bilingual dictionaries. This keynote address was published in a festschrift in her honor six years later, followed by a series of other papers about the same topic (Corréard 2002). One of Atkins's claims formalized a general truth, namely “Change is not something that people tend to associate with dictionaries” (Atkins 1996). Electronic dictionaries, however, are part of the fast changing electronic world, thus, in the last decade several new features have appeared in them.

In the electronic world it is common sense that hypertext functionality eliminates linear text restrictions and opens the way to new types of information by offering new ways of presenting them. In the dictionary world, the first consequence of being electronic is that there are no space constraints, that is, we don't need to follow the well-known dictionary formats, which is a sort of consequence of the Gutenberg-galaxy. Our dictionaries can be disengaged from the shortcomings of being printed. The entries don't need to follow any order, alphabetic or other, the hits for any query can be sorted according to the actual user needs. There can be alternative ways of presenting information: it is not bound to the nature of the paper. There are a lot of opportunities for user customization: for example, if the user likes the more traditional view of dictionary entries with tilde signs, he or she can see the dictionary content according to this, but if other, non-traditional views are preferred, it can be done without changing the dictionary's internal database representation. There are other new options as well: lexicographers are not obliged to insert various examples into the entries, because of the rapid access to large amounts of lexicographical data in available mono- or bilingual corpora showing the actual use of the word or expression in context, with or without translation.

There are other consequences of being electronic: e.g. intelligent dictionary production which is an important branch of computational lexicography of the corpus linguistic age. Starting from corpora (“corpus-based lexicography”) many dictionaries have been

developed in the past years with the help of statistical and linguistic analyses and other extended corpus-linguistic technologies ("corpus-driven lexicography"). Unfortunately, the detailed description of these methods falls outside of the scope of present paper as does dictionary making with the help of sophisticated electronic tools based on current software technology. In the next sections, however, some existing solutions will be shown of how language technology can be combined with dictionaries – mostly along the basic ideas of Sue Atkins.

2. Towards "intelligent" dictionary lookup

Atkins (1996) claims that dictionaries can be used in two different ways: in look-up mode and browsing mode. Look-up mode is where the user is in a quick search of a specific piece of information and browsing mode is where a more relaxed process of reading of dictionary entries takes place. In the traditional dictionary look-up entries are in alphabetic order and the computer search relying on indexing of headwords can be either full or partial string matching. This means that electronic dictionaries using this querying method are very close to the traditional paper dictionaries, only the search in the alphabetic register is much faster in them. As Atkins (1996) says: "The dictionary of the present is at heart little different from the dictionary of the past". Using an analogy: the first electronic dictionaries are similar to the first automobiles where the engine was put in the place of horses, and many decades have been needed to reach the current form of cars which is mostly determined by wind channels. The new, language technology based look-up relies on stemming (that is, the stem of the actual running word is looked up in the dictionary. The new look-up should use some sort of spelling to find information also with misspelled input. There are experimental look-up technologies that use semantic similarity while searching (Segond et al. 2000). A very important difference between paper or paper-like electronic dictionaries and the dictionary look-up using new technologies is that the latter is able to use more than a single source while searching. In other words, parallel lookup in many sources can be done, thus for the end-user it is much less important whether the needed information is found in a single source or in various sources that are currently available for the multi-dictionary look-up system.

3. "Intelligent" dictionary visualization

Visualization is another issue which plays a role in the design of future dictionaries. Alternative formats may be needed in various situations. An example of this is when dictionaries use the '~'-sign for the headword in the body of the entries to minimize the size of the paper dictionary, but this is useless in electronic dictionaries. Some publishers, however, insist on the shorter forms. Multiple typographies can be handled easily in current XML/XSLT based dictionary representations, namely, printed dictionaries and their screen-oriented versions come from the same XML source applying different XSL transformations. That's why multiple screen versions do not cause problems for the visualization of today's lexical resources.

As has been mentioned earlier, multiple dictionary look-up is another feature of intelligent dictionary systems. Visualization of entries coming from different sources need a special dynamic combination of search results to produce a virtual single entry. The dy-

namic structure of even a single dictionary entry plays an interesting role in these systems: users may or may not need some parts of the original entries (e.g. optional visualization of non-compulsory information). The original literal context of the running word in question can help the intelligent system to choose the actually not needed information from the dictionary entry. For example, there are many multi-word expressions listed in dictionary entries, and it is the actual context that helps to choose only those entries which play some role in the actual meaning of the word in question. Other contexts may choose other sub-entries from the original entry, so dynamic construction of a larger entry can be done very frequently in the intelligent dictionary systems with context-sensitive dictionary look-up.

Summarizing the above arguments, it can be claimed that today's dictionary structure is a sort of by-product of the Gutenberg-galaxy. The typical lexicographic abbreviations originate from typographic considerations where rules like "use fewer letters" or "save more paper" played a crucial role. In the development of new dictionary solutions, producers of electronic dictionaries play a similar role to the printing houses earlier. In case of traditional dictionaries, linguistic information needed for text comprehension was in the user's mind. Therefore, lexicographers tried to support users to find the needed information effectively. In case of intelligent computer dictionaries a dynamic comprehension module also "sees" what the user sees, therefore the computer is in a situation which was dedicated only for humans earlier. The starting point of the dictionary look-up in contemporary electronic systems is the actual text on the screen and the user triggers the needed dictionary process with some manipulation over this text (e.g. by clicking on the word to start dictionary search, or only leaving the mouse cursor over the actual word for a short while), and the look-up procedure starts with the actual word which is in the actual context. We have to make, however, an important economic consideration here: intelligent electronic dictionaries use language technology knowledge that is always bound to the source language in question, namely, stemming, spelling correction or parsing of the actual context are operations which are not language-independent. Intelligent dictionaries with specific language abilities are consequently bound to those markets where the actual source languages are widely used. In other words: if we have a general dictionary system, its market is the whole world separately from the languages of the dictionaries published with the help of them. In case of applying intelligent, language-specific modules, this market becomes smaller because of the language-dependent modules that guarantee intelligence.

4. Aspects of using electronic dictionaries

If people go to a bookshop, they are able to judge the quality of a paper dictionary with the help of many external features: number of pages (or at least, the thickness of the book), typographical solutions, letter size, etc. It is not too difficult to categorize a dictionary whether it is a reliable one or not. On the other hand, it is rather difficult to measure how up-to-date an electronic dictionary is. The exact size of an electronic dictionary is difficult to check, consequently, the biggest "declared" figures tend to "win" in the marketing competition. Marketing people are always able to "develop" a new way of enumeration which can show that their dictionary is the biggest on the market. Size, of course, is not the only feature which counts, but – again using the "automobile parallel"

– non-experts can be convinced by the speed of the model only. New films or musical productions people can get acquainted with via their reviews and criticisms, but dictionary criticism or particularly, electronic dictionary criticism is not a typical field, so users can only rely mainly on their personal experience.

It is a very interesting experiment how users choose electronic dictionaries. In case of paper dictionaries leafing through the book gives an impression whether it serves the purposes of the customer. The general judgment on the quality of an electronic dictionary is usually based on a simple lookup for a few words: if they are found, the dictionary is generally considered "good", if not, the dictionary is "bad". Generally, non-professional users – mostly non-professional internet users – don't consider the dictionary tools as important as professionals do, they are usually satisfied with a few hits which are provided by even the demo versions as well. Consequently, more and more users don't buy professional dictionaries (where the new look-up technologies can mostly be found) because their needs are already satisfied by lower level free dictionaries on the web. In the paper world, this dichotomy does not exist: users should pay some fee for even the lower level dictionaries. We have arrived to a critical issue here, namely, there is more and more free information on the internet, and this generates a temptation for many potential dictionary users: why to buy an electronic dictionary if an "almost" similar one is available for free. In the electronic world there are dictionaries for fee and dictionaries for free. The question is whether the brand name, the content or the applied services are enough to convince the non-professional end-users to pay for some dictionary content. What is interesting, it is not because of the eventual high price but because of the many freely available dictionaries on the web which also give acceptable results. So, if there are two candidate dictionaries – a professional one for a fee and a less reliable (but generally well-designed) internet dictionary – usually the general message of the web is applied, namely, "who cares whether something is not as professional as the other, but it is for free!"

In addition, end-users can easily see if something is not perfect in a paper dictionary, because the full version is in their hands even in the bookshop. In case of electronic dictionaries, you can meet the full version only if you have already bought it. Additionally, in the electronic world, you can meet special – in most cases: "non-official" – variants of well-known dictionaries which are not as easy to identify. For example, if dictionary entries are split up into pairs of source-target expressions, that is, entries without internal structure, then much less typography is required, consequently, identification of the original lexicographical source is not easy. What is more, in electronic dictionaries there is no "basic" order of entries, which is a crucial issue in comparing two paper dictionaries. Order of entries is again a feature which cannot be used for identifying the original source of an electronic wordlist because there can be various indices providing different ordering among the same entries. In some cases, special technical terms or occurrence of seldom used expressions can be clear signs of the origin of the dictionary. In case of general dictionaries there are not too many very specific entries, and it would be difficult to argue, for example, that *table*'s German equivalent is not *Tisch* in an English-German dictionary, or one of *horse*'s equivalent is not *Pferd*. The question is always the same: where is the border between common (bilingual) knowledge which is generally available for free and special lexicographical information that should be paid for?

5. "Truly electronic dictionaries" enriched with new types of information

As we have seen, dictionary content and/or reputation of a dictionary publisher may be crucial, but in the case of electronic dictionaries, technology is also be important. People generally don't have time enough for anything, and if the new technology helps to spare some seconds or minutes for them, they are getting interested. Users prefer less typing and they like if they don't need to open new applications which eventually partly cover the screen, mostly those parts where the text to be understood takes place. For example, if the mouse pointer is left over a word for more than one second indicates automatically that the user would like to have information about that word and its context, the hits to this query should be suddenly shown in a bubble on the screen. Users don't have much time for interaction, so the tool should rely on its own linguistic knowledge only: the actual context should be taken into consideration to identify all possible multi-word expressions in the context, even if word forms are inflected or the actual word order is different from the one of the lexical form of the expression. If no multi-word expressions are found, the tool displays a simple dictionary entry for the selected word (Figure 1). What is important is that the tool provides a list all possible translations found in all active dictionaries for the given language as source language.

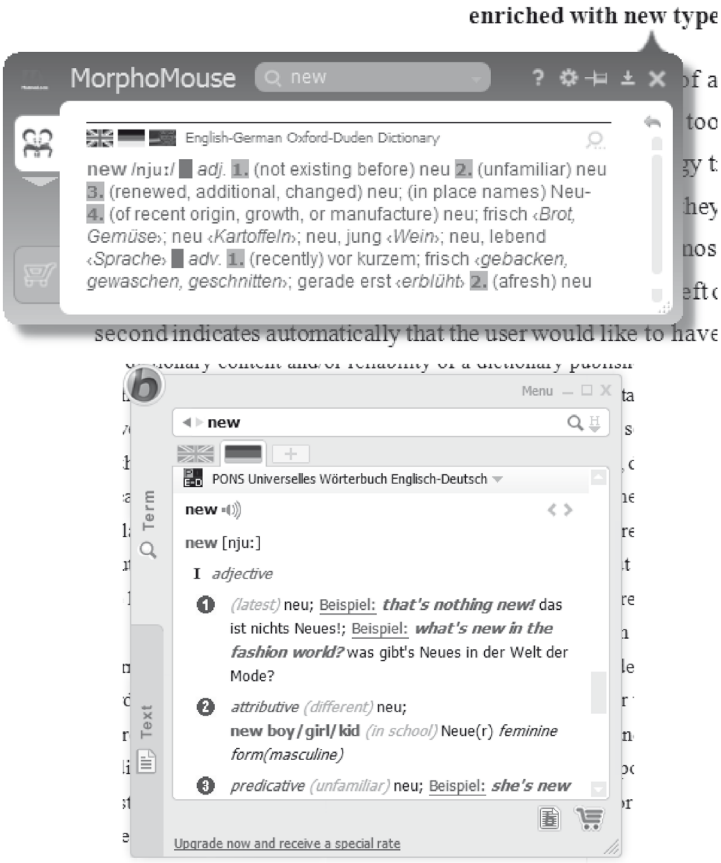

Figure 1: Look-up for a simple dictionary entry (MorphoMouse and Babylon)

The first attempt in the direction of new dictionaries was COMPASS (Feldweg/Breidt 1996). The system we try to illustrate the above features of intelligent dictionaries with is MorphoMouse, a technology that relies on the earlier MoBiMouse (Prószéky/Kis

2002). Atkins' "look-up mode" and "browsing mode" are supported: hits are shown either in a bubble-shaped pop-up window on the screen, or the same small bubble-like window which can be used traditionally via typing (Figure 2).

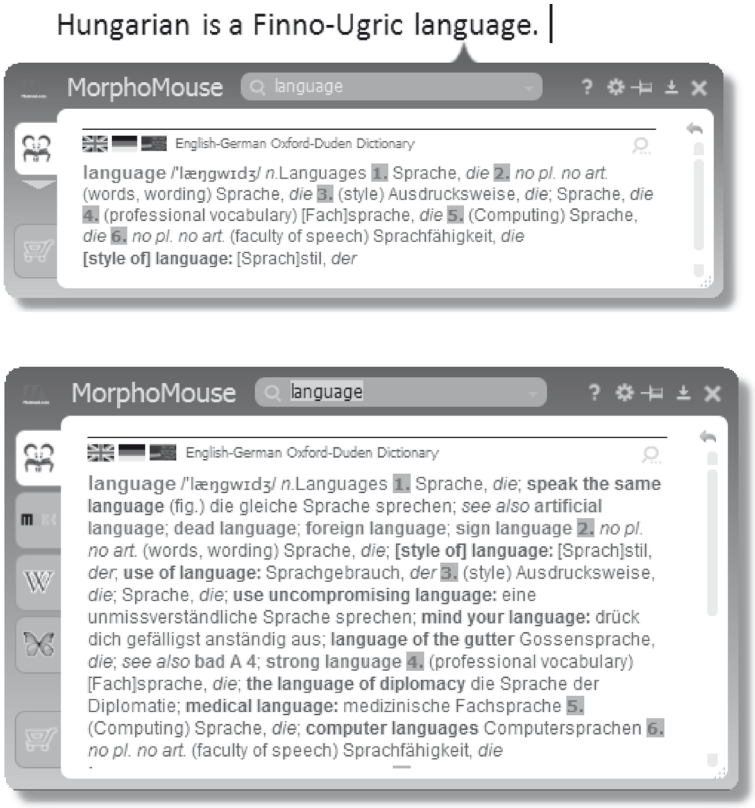

Figure 2: Look-up mode (partial entry) and browsing mode (full entry)

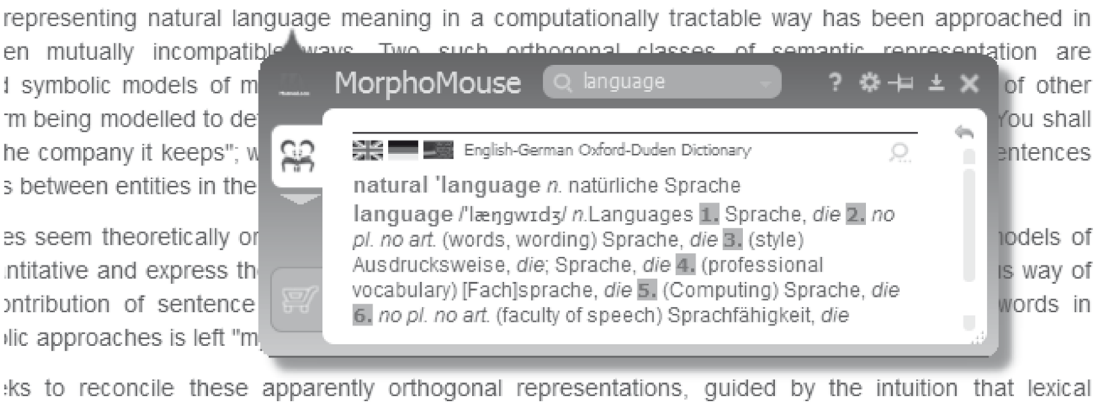

Figure 3: Contextual look-up

Instant dictionary look-up means that the mouse cursor is left over a word on the screen and dictionary look-up is triggered by either no mouse movement being identified in the next second or by the pressing of some special button (e.g. press Ctrl button twice). Look-up relies on the actual context: the potential stems of the actual word (under the cursor) are identified with the help of language technology modules: a language identifier, a morphological analyzer for the language in question and a sort of multi-word analysis is also done to check whether the actual entry and some surrounding words can form multi-word expressions or not (Figure 3).

Multi-entry look-up means that even in a normal dictionary there can be different headwords containing each word of the original input. Different multi-word entries can contain common parts, e.g. the dictionary entry of *ability* contains a sub-headword which contains *natural* and *for: she has a natural ability for teaching*. In the same dictionary we can find the entry of *natural* which also contains an expression with *for*, namely, *he was a natural for the job*. Both hits should be provided if the query consists of two words, *natural* and *for* (Figure 4).

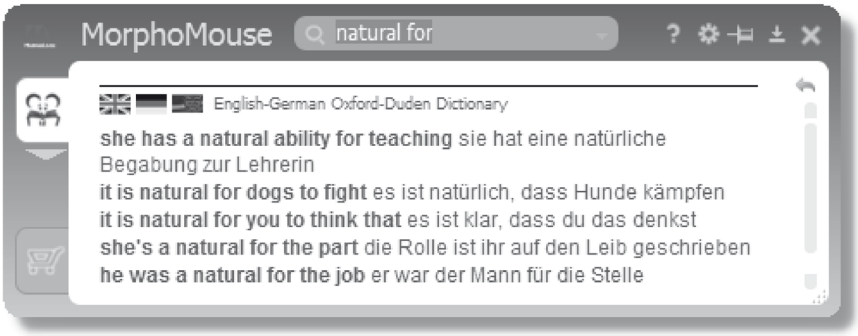

Figure 4: Parallel look-up for multi-word expressions in a single dictionary

Parallel multi-dictionary look-up is a procedure which goes through each open dictionary and the actual query is searched for all of them. Users are not generally interested in the details how to do this, or whether some dictionaries don't have equivalents for the actual input but some others do. The user would like to have as many hits as the available dictionaries are able to offer (Figure 5). Even multi-language look-up can help the user if he/she is able to use more than a single target language (Figure 6). Talking about true multilingual dictionaries Atkins says that in the past lack of space and commercial pressures made a true multilingual dictionary impossible (Atkins 1996) and continues: "If a multilingual dictionary is to be compiled, we have to devise an analysis technique common to all languages involved".

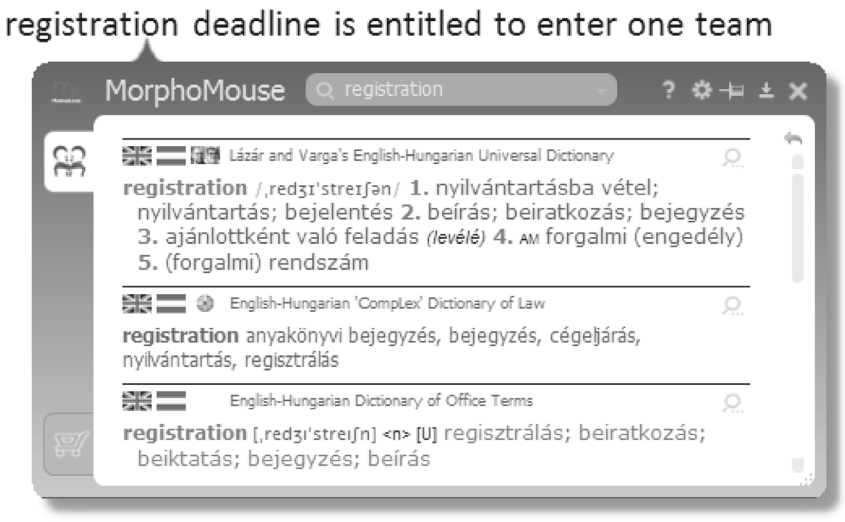

Figure 5: Combined parallel look-up in general and specific dictionaries

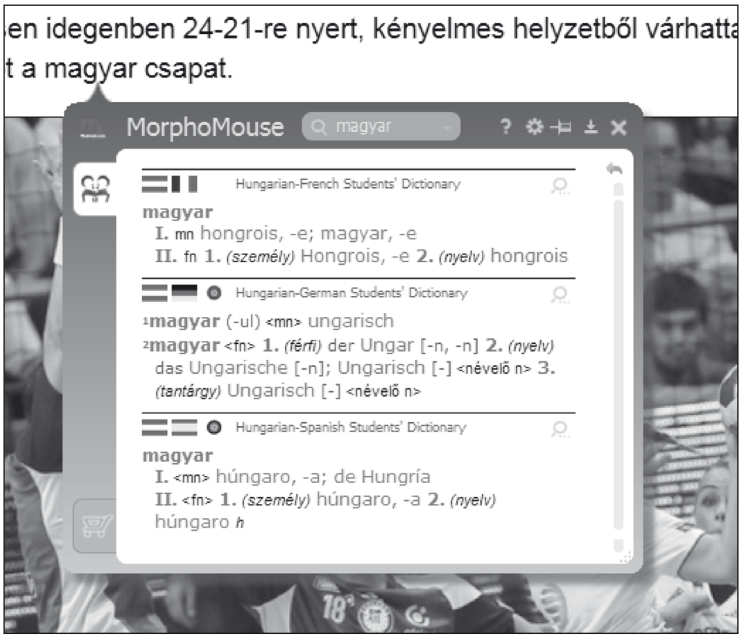

Figure 6: Multi-language look-up

Important combination of traditional dictionaries and other linguistic sources can be guaranteed by language technology solutions. Let's quote Atkins (1996) again: "research described in Atkins and Varantola shows that people often turn to a monolingual dictionary during a bilingual search. The ideal dictionary should offer monolingual functions (definitions, etymologies, usage notes) to the bilingual dictionary user". For example, a set of dictionary hits can be combined with hits from other, monolingual lexicographical sources or encyclopaedias: an example is shown from Wikipedia on Figure 7. "The ideal dictionary should allow the user to browse through genuine attested examples of the foreign expression in use in various types of texts" (Atkins 1996). Translation memories or just parallel corpora are available for today's electronic dictionaries: Figure 8 shows some hits (for the English verb *prefer*) in Hunglish, a web-based Hungarian-English bilingual corpus (Varga et al. 2005). Electronic dictionaries are frequently used in combination with machine translation applications, thus, not only the word and its local context are processed by linguistic technologies but the full sentence containing the word in question (Figure 9).

Figure 7: Combined instant look-up in other internet sources (Wikipedia)

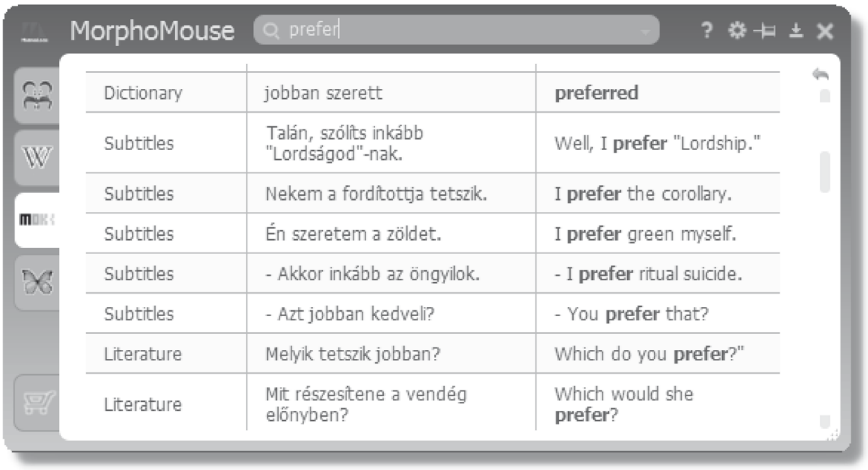

Figure 8: Parallel corpus as special dictionary

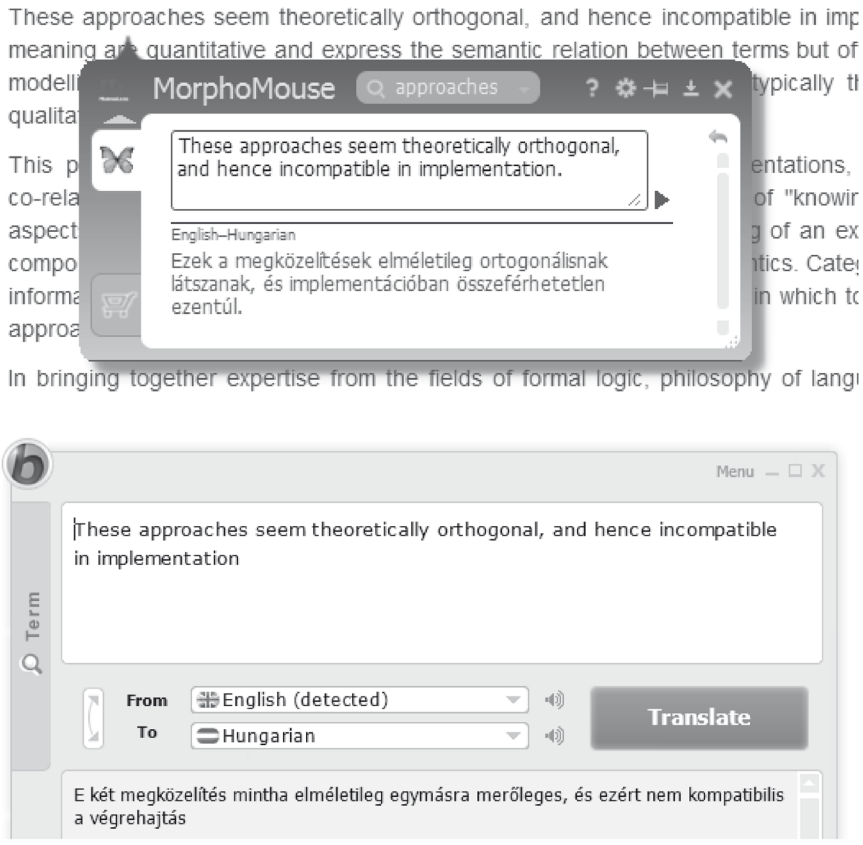

Figure 9: Combined instant look-up for sentence translation (MorphoMouse and Babylon)

6. Conclusion

Atkins (1996) says: "Many of the obstacles to the creation of tomorrow's improved bilingual dictionary have been removed in the past few decades by the advent of the computer (computer-assisted lexicography, rich electronic text corpora as sources of lexicographical evidence, computational searches of dictionaries, and so on)". She describes how new-age bilingual dictionaries must exploit the electronic medium. She lists a lot of claims which have become well-known since then, like "no space constraints", "flex-

ible compiling liberated from alphabetical order", "alternative ways of presenting information", "rapid access to large amounts of lexicographical evidence in corpora" and many others. According to Atkins (1996) the electronic dictionaries of the mid-nineties were little more than reincarnation of print dictionaries. In our paper we have tried to show (via examples of publicly available tools) that around fifteen years later there are electronic dictionary systems which use technologies carrying out the plan sketched first by Sue Atkins. These new tools even use multi-dictionary look-up methods, language identification, stemming, linguistic treatment of multi-word expressions with the help of language technology solutions. Unfortunately, there are many features Atkins mentioned that no solutions have been made for since then. It's high time to read again about those challenging tasks of today's computational lexicography.

7. References

Atkins, B.T.S. (1996) Bilingual dictionaries – past, present and future. In: Gellerstam, M./ Järborg, J./Malmgren, S.-G./Norén, K./Rogström, L./Papmehl, C.R. (eds.): *Euralex '96 Proceedings*. Gothenburg: Gothenburg University, 515-590.

Corréard, M.-H. (ed.) (2002): *Lexicography and Natural Language Processing* (*a festschrift in honour of B.T.S. Atkins*). Grenoble: Euralex.

Feldweg, H./Breidt, E. (1996): COMPASS – an intelligent dictionary system for reading text in a foreign language. In: Kiefer, F./Kiss, G./ Pajzs, J. (eds.): *Papers in Computational Lexicography*. Budapest: Linguistics Institute, 53-62.

Prószéky, G./Kis, B. (2002): Context-sensitive dictionaries. In: Tseng, Shu-Chuan (ed.): *Proceedings of the 19th International Conference on Computational Linguistics*, Vol. II. Taipei: Howard International House, 1268-1272.

Segond, F./Aimelet, E./Lux, V./Jean, C. (2000): Dictionary-driven semantic look-up. In: *Computers and the Humanities* Vol. 34, No. 1/2, 193-197.

Varga, D./Németh, L./Halácsy, P./Kornai, A./Trón, V./Nagy, V. (2005): Parallel corpora for medium density languages. In: *Proceedings of the RANLP 2005*, 590-596.