Archive of Control Sciences 15, No. 4, 637-644 (2005)

Between Understanding and Translating: A Context-Sensitive Comprehension Tool

Gábor Prószéky & András Földes
MorphoLogic
Orbánhegyi út 5, H-1126 Budapest, Hungary
{proszeky, lafoldes}@morphologic.hu

Abstract

This paper introduces an English–Polish/Polish–English comprehension tool. In fact, it is a special electronic dictionary which is sensitive to the context of the input words or expressions. The dictionary program provides translations for any piece of text displayed on a computer screen without requiring user interaction. This functionality is provided by a three-layer process: (1) text acquisition from the screen, (2) morpho-syntactic analysis of the context of the selected word and (3) the dictionary lookup. By dividing dictionary entries into smaller pieces and indexing them individually, the program is able to display a restricted set of information that is as relevant to the context as possible. For this purpose, we utilize automatic and semi-automatic XML tools for processing dictionary content. The construction of such an electronic dictionary involves natural language processing at almost every point of operation. Both dictionary entries and user input require linguistic analysis and intelligent pattern-matching techniques in order to identify multi-word expressions in the context of the input. An on-going research makes the program incorporate more sophisticated language technology: multi-word phrases and sentences are recognized, and translation hints are offered in an intelligent way – by a parser/transformer module matching underspecified patterns of different degrees of abstraction.

Introduction

Most computer users encounter a large number of foreign language texts. They often don't need translations, only to read and understand the text. The foreign language comprehension assistant described in this paper is still not a translation program: it has evolved from a common electronic dictionary engine. Let us emphasize that developing such a tool takes more effort than adapting a bilingual dictionary. Bilingual dictionaries are traditionally composed in a format and structure suitable for assisting with translation. Foreign language comprehension, however, takes a different approach, and it is almost certain that any dictionary needs to be recompiled to some extent before it is incorporated in the comprehension assistant [Feldweg & Breidt 1996].

In this paper, we devise a context-sensitive electronic dictionary we usually label as a 'context-sensitive instant comprehension tool'. It reads text from anywhere on the computer screen, performs linguistic analysis of the context of the input word, and uses an arbitrary number of dictionaries in the background. Finally, it displays context-dependent translations in a bubble—without requiring a single mouse click from the user, and leaving the screen contents intact.

In the following sections, we attempt to give an overview of the problem in general and the requirements of a comprehension tool. Then we describe the program's different phases of operation, with special attention to the linguistic elements and the dictionary development procedures employed throughout the application.

Word comprehension in context

It is clear that the linguistic process of understanding a passage of text is completely different from translating it.. A translator (either human or computational) has to analyse and transform every bit of the source text (and make her/his/its way through the ambiguities and unclear syntactic structures). Users attempting to understand an e-mail message—or any text that is displayed on a computer screen—need only hints about it, but they need it as quickly as possible, and without being disturbed, or even mislead, by ambiguities and other types of non-relevant information.

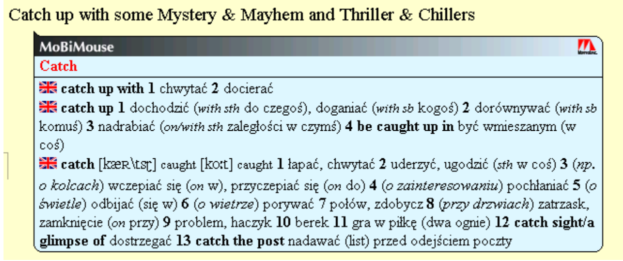
Translation support systems usually comprise of electronic dictionaries, less often aligners, translation memories etc. Such programs present a rather 'intrusive' user interface that draws the user's attention from the text to be understood. Document windows and dialog boxes are too disturbing for the user who needs 'in situ' information about a passage: he needs it exactly where the text appeared and he will not want to start another application, copy the text into it, and click some buttons to see translations. He will want assistance within the application he is currently working with.

Our comprehension assistant fully complies with this requirement: to display the context-sensitive translation of a word or a multi-word expression, all the user has to do is move the mouse pointer over the text in question. Context-sensitivity means that the program indicates if the word is part of a multi-word construction and will select the appropriate translation depending on the syntactic context if possible [Segond & Breidt 1996]. The translation itself is displayed in a bubble over the current screen contents, and disappears when the mouse is moved.

MoBiMouse Plus: a context-sensitive comprehension assistant

An instant comprehension assistant cannot ask for user interaction: it cannot require the user to choose from a list of ambiguous linguistic analyses, and, at the same time, it should keep the number of semantic ambiguities as low as possible.

When the mouse pointer is left over a word for more than a given time interval (200 ms in our case), it indicates that the user needs information about that word and its context The program determines what is the largest possible context, analyzes it, and gives any available information on it – based on the dictionaries behind the system. The minimum requirement is that the program should recognize all obvious multi-word expressions and idioms, and provide appropriate translations. (See Figure 1.)



Catch up with some Mystery & Mayhem and Thriller & Chillers

Catch

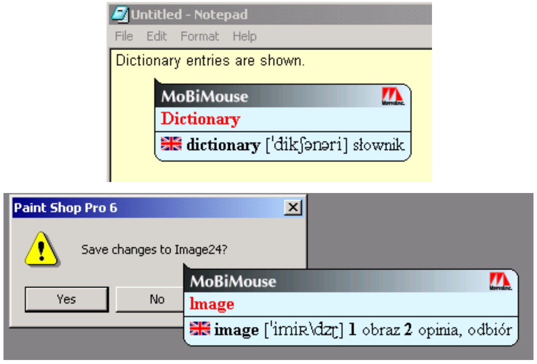
catch up with 1 chwytać **2** docierać

catch up 1 dochodzić (*with sth* do czegoś), doganiać (*with sb* kogoś) **2** dorównywać (*with sb* komuś) **3** nadrabiać (*on/with sth* zaległości w czymś) **4 be caught up in** być wmieszanym (w coś)

catch [kæʀ\tsʈ] caught [kɔːt] caught **1** łapać, chwytać **2** uderzyć, ugodzić (*sth* w coś) **3** (*np. o kolcach*) wczepiać się (*on* w), przyczepiać się (*on* do) **4** (*o zainteresowaniu*) pochłaniać **5** (*o świetle*) odbijać (się w) **6** (*o wietrze*) porywać **7** połów, zdobycz **8** (*przy drzwiach*) zatrzask, zamknięcie (*on* przy) **9** problem, haczyk **10** berek **11** gra w piłkę (dwa ognie) **12 catch sight/a glimpse of** dostrzegać **13 catch the post** nadawać (list) przed odejściem poczty

1. Dictionary information displayed by the comprehension tool, using an English-Polish dictionary

The recognition process can be divided into three main phases. The program watches the motion of the mouse pointer and acquires the text from the screen position where the mouse pointer stops. Our implementation either performs an OCR-like procedure, or uses special application dependent API-s to get the screen content.



Dictionary entries are shown.

Dictionary

dictionary ['dikʃənəri] słownik

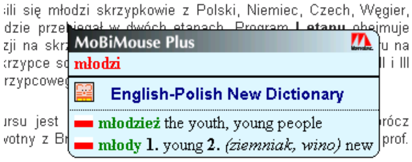
Save changes to Image24?

Image

image ['imiʀ\dʒ] **1** obraz **2** opinia, odbiór

2. Different text recognition situations

The next step – the linguistic analysis – is supposed to identify the word that was pointed at, and perform a morpho-syntactic analysis of its context to determine the headwords to look up in the dictionaries. It is important to note that the initial data are results of an OCR-like process whose ambiguities require correction during subsequent linguistic analysis. The linguistic analyzer module then performs morpho-syntactic analysis for the selected word in context—by means of the HUMOR engine [Prószéky & Kis 1999]. At this point, morphological analysis has three main purposes: linguistic stemming for accurate dictionary lookups (see Figure 3.), spelling correction and preparation of shallow parsing of the context to identify candidates for multi-word expressions.



młodzi

English-Polish New Dictionary

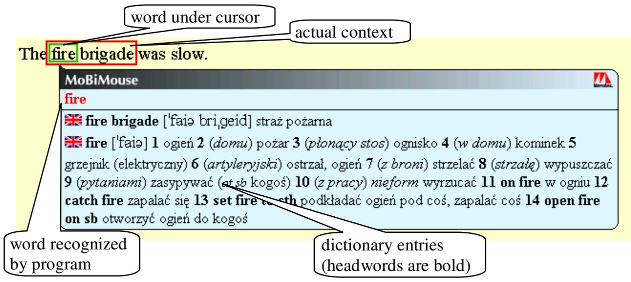
młodzież the youth, young people

młody 1. young **2.** *(ziemniak, wino)* new

3. Stemming in action

If linguistic analysis fails to recognize any multi-word expressions, words from the context are still passed on to the dictionary lookup phase as the dictionaries may contain idiomatic phrases that cannot be recognized on a linguistic basis.

The third step – dictionary lookup phase receives lexical stems of the selected words and words in the context, which are then matched against all dictionaries. We utilize an intelligent dictionary engine capable of handling multiple dictionaries at the same time [Prószéky 1998]. Dictionaries are often adapted to comprehension needs by filtering out non-relevant information [Feldweg & Breidt 1996]. The output of the comprehension assistant is then shown in a bubble-shaped pop-up window on the screen that disappears if the user moves the mouse cursor again. Figure 4. describes the layout of a dictionary entry bubble displayed by the comprehension assistant.



word under cursor
actual context
MoBiMouse
word recognized by program
dictionary entries (headwords are bold)

The fire brigade was slow.

fire

fire brigade ['faiə bri,geid] straż pożarna

fire ['faiə] **1** ogień **2** (*domu*) pożar **3** (*płonący stos*) ognisko **4** (*w domu*) kominek **5** grzejnik (elektryczny) **6** (*artyleryjski*) ostrzał, ogień **7** (*z broni*) strzelać **8** (*strzałę*) wypuszczać **9** (*pytaniami*) zasypywać (*at sb* kogoś) **10** (*z pracy*) *nieform* wyrzucać **11 on fire** w ogniu **12 catch fire** zapalać się **13 set fire to sth** podkładać ogień pod coś, zapalać coś **14 open fire on sb** otworzyć ogień do kogoś

4. Parts of a dictionary entry as displayed by the comprehension assistant in the bubble.

Dictionaries

Dictionaries in our system are represented as lexical databases where the structure of each dictionary is strictly preserved. This is achieved through using XML as the single dictionary format. Dictionaries are either originally written in XML or transformed from a printed or another electronic format by means of automatic and semi-automatic tools. Any dictionary must be restructured in order to be used with a comprehension assistant. In the examples of the present paper, we use an English-Polish dictionary that was first converted from TeX to XML format.

As a first approach toward intelligent treatment of multi-word expressions, we needed to recognize that the sample phrases and sentences in a large dictionary provide enough multi-word expressions for the program to be useful. This fact is very important for providing translations—as the translations are there in the dictionary—, because as long as we have nothing more than a dictionary program, the only way to construct translations is to read them from dictionary databases and display them as they are stored.

As a basic rule, it is also clear that we cannot display entire dictionary entries. The main reason for this is not that they tend to be too lengthy (and could easily cover the entire screen) but the reply of the program must be relevant to the actual context. For this reason, original dictionary entries were split up into smaller parts: all examples and otherwise included multi-word expressions were treated as separate dictionary entries. After the original dictionary was properly tagged into XML, the recompilation process could be largely a mechanic conversion. The above step is the most important in restructuring entire dictionaries. Thus all multi-word expressions in the dictionary are treated by the program as headwords. This is important because headwords are directly indexed and can be found in the database by a single lookup operation.

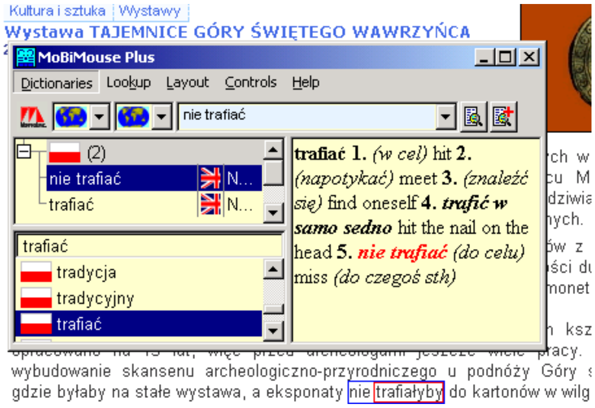
The hit in the dictionary is rarely a single entry. As a general rule, the comprehension assistant finds all suitable entries for the context of the selected word, and displays them in the same bubble. It is also important to index each multi-word expression word by word as they must be found by any of their words. The text acquisition module returns a whole line of text with one single word marked: it is, in a non-linguistic sense, in the 'focus' of the user query. The marked word can be any one of a multi-word expression. Thus, for instance, the expression *Commonwealth of Independent States* is found by pointing at either *Commonwealth* or *Independent*. (Figure 5.)

Although we thoroughly surveyed future users when starting the project, it has been clear from the beginning that existing dictionaries will be far from perfect at any stage of the project life cycle. Therefore, dictionaries are continuously reviewed and updated. The speciality of this process is that it is built largely on user feedback. For dictionary development (and even for linguistic research, in general), the comprehension assistant is an ideal source because it reaches a potentially large number of users. Based on this insight, we have implemented an instant feedback feature, which comprises of two processes: logging and contacting the developers. Logging means that the comprehension assistant continuously logs words and multi-word expressions it was unable to analyze or failed to find in the dictionaries. Contacting the developers means that if the user agrees the program automatically sends e-mails containing the current logs to the development team.

This process effectively reveals errors and deficiencies in the dictionaries, and, at the same time, it helps defining directions of further improvements.

Related research

There are two categories where our context-sensitive comprehension assistant tool, MoBiMouse Plus, can be compared to other systems: functionality and linguistic accuracy. There are only various pop-up dictionaries on the market: among the best-known ones there are Babylon, WordPoint or Langenscheidt's Pop-up Dictionary, but none of them have as many language technology features as MoBiMouse Plus. The character identification techniques applied in our comprehension assistant are independent of both the language and the writing system: it is rather different from all known applications that work with English characters only. There is a version that recognizes complex scripts such as characters of Far East languages—which enables users to read and understand texts that they cannot even type in.



trafiać 1. *(w cel)* hit **2.** *(napotykać)* meet **3.** *(znaleźć się)* find oneself **4.** ***trafić w samo sedno*** hit the nail on the head **5.** ***nie trafiać*** *(do celu)* miss *(do czegoś sth)*

wybudowanie skansenu archeologiczno-przyrodniczego u podnóży Góry
gdzie byłaby na stałe wystawa, a eksponaty nie trafiałyby do kartonów w wilg

5. Combination of pop-up and fix-window dictionaries

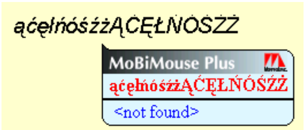
Pop-up dictionary applications mostly start by pressing a button or clicking the mouse. MoBiMouse Plus is the only known application that starts without clicking; therefore it can be used to acquire any text from the screen without affecting other running applications.

The speed of the text acquisition module is 1000 characters/s, stemming is performed at 0.002 s/word-form; an average dictionary lookup takes 0.02 s. MoBiMouse Plus, unlike many other applications, can be used in both language directions of the dictionary.

Issues specific to Polish language

As we noted the system described here is the first Polish–English implementation of our already existing system.

The Polish language is rich in accented characters, some of which exist only on the Windows' Eastern-European code page. Therefore we decided to reimplement our application on the basis of the Unicode character representation. In the new implementation theoretically all characters of the Unicode code pages can be used on the screen, so similar French–Polish or even Chinese–Polish dictionaries are supported too.



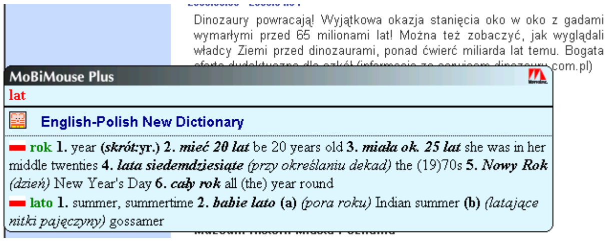
ąćęłńóśźżĄĆĘŁŃÓŚŹŻ

ąćęłńóśźżĄĆĘŁŃÓŚŹŻ

<not found>

6. Recognition of 'Polish' characters

Our intelligent dictionary works in both English–Polish and Polish–English language direction, so the linguistic features described above must work for the Polish language as well. Our Polish stemming module is based on the Humor morphologic analyzer developed by MorphoLogic earlier. [Prószéky & Kis 1999] The tool works with Hungarian and other agglutinative languages as well. Polish is known for its rather complex morphology but it is no challenge for a HLT tool that is able to handle languages with a potential of billions of separate word forms. Hence, it is able to handle several millions of inflected Polish entries, generated from over 100 000 Polish stems [Wołosz 2000].



Dinozaury powracają! Wyjątkowa okazja stanięcia oko w oko z gadami wymarłymi przed 65 milionami lat! Można też zobaczyć, jak wyglądali władcy Ziemi przed dinozaurami, ponad ćwierć miliarda lat temu. Bogata

lat

English-Polish New Dictionary

rok **1.** year **(*skrót:*yr.) 2. *mieć 20 lat*** be 20 years old **3. *miała ok. 25 lat*** she was in her middle twenties **4. *lata siedemdziesiąte*** *(przy określaniu dekad)* the (19)70s **5. *Nowy Rok*** *(dzień)* New Year's Day **6. *cały rok*** all (the) year round

lato **1.** summer, summertime **2. *babie lato* (a)** *(pora roku)* Indian summer **(b)** *(latające nitki pajęczyny)* gossamer

7. Stemming and dictionary lookup in Polish–English direction

In the present implementation we use a medium size (44 000 entries) dictionary [Piotrowski & Saloni 1998]. The Hungarian version was able to handle much larger dictionaries with nearly the same performance, so we are eager to publish similar Polish implementations as well.

Conclusion

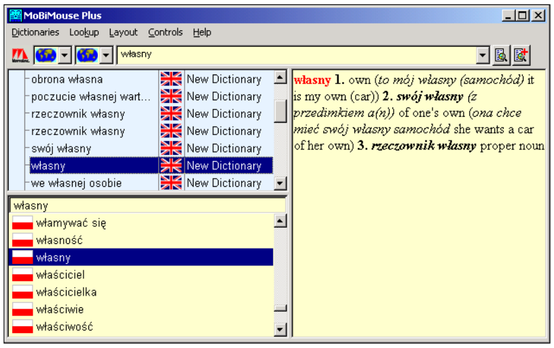
The context-sensitive instant comprehension assistant named MoBiMouse Plus offers translations for words and expressions on the computer screen. The program can be activated by moving the mouse pointer over a particular word: the translation is displayed in a bubble-like fashion. The dictionary content can be displayed in a classical window based user interface as well.

MoBiMouse Plus has been written in C^{++} and can be used anywhere in a Windows (98/ME/2000/XP/2003) environment. It is especially useful when browsing the net: it helps reading web pages fast and without interrupts. We adapted the original dictionary software to the specific needs of the Polish language: the accented Polish characters are fully supported, and our stemmer recognises the inflected Polish word forms as well. Therefore, MoBiMouse Plus performs extensive linguistic analysis of the selected Polish and English words and their contexts, and provides intelligent replies for multi-word expressions where all translations are relevant to the context.

Acknowledgements

We would like to thank the editors, Tadeusz Piotrowski and Zygmunt Saloni for their co-operation, and Robert Wołosz, without whom the Polish version of our software tool could not have been possible.

8. MoBiMouse Plus window with expressions containing any form of *własny*



własny

obrona własna	New Dictionary
poczucie własnej wart...	New Dictionary
rzeczownik własny	New Dictionary
rzeczownik własny	New Dictionary
swój własny	New Dictionary
własny	New Dictionary
we własnej osobie	New Dictionary

własny

- włamywać się
- własność
- własny
- właściciel
- właścicielka
- właściwie
- właściwość

własny **1.** own (*to mój własny (samochód)* it is my own (car)) **2. *swój własny*** *(z przedimkiem a(n))* of one's own (*ona chce mieć swój własny samochód* she wants a car of her own) **3. *rzeczownik własny*** proper noun

References

[Feldweg & Breidt 1996]
Feldweg, H. and E. Breidt.. COMPASS – An Intelligent Dictionary System for Reading Text in a Foreign Language. *Papers in Computational Lexicography (COMPLEX 96),* Linguistics Institute, HAS, Budapest, pp. 53–62. (1996)

[Piotrowski & Saloni 1998]
Piotrowski, T., Z. Saloni. *Słownik Angielsko–Polski Polsko–Angielski,* WILGA, Warszawa (1998)

[Poznanski et al. 1998]
Poznanski, V., P. Whitelock, J. Udens and S. Corley. Practical Glossing by Prioritised Tiling. *Proceedings of the COLING-98*, Montreal, pp. 1060–1066. (1998)

[Prószéky 1998]
Prószéky, G.. Intelligent Multi-Dictionary Environment. *Proceedings of the COLING-98*, Montreal, pp. 1067–1071. (1998)

[Prószéky & Kis 1999]
Prószéky, G. and B. Kis. A Unification-based Approach to Morpho-Syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. *Proceedings of the37th Annual Meeting of ACL, College Park*, pp. 261–268. (1999)

[Segond & Breidt 1996]
Segond, F. and E. Breidt. IDAREX: description formelle des expression à mots multiples en français et en allemand. In: A. Clas, Ph. Thoiron and H. Béjoint (eds.) *Lexicomatique et dictionnairiques,* Montreal, Aupelf-Uref. (1996)

[Wołosz 2000]
Wołosz, R.: Efektywna metoda analizy i syntezy morfologicznej w języku polskim. *Unpublished doctoral thesis.* Warszawa (2000)