Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 299–303
ISBN 978-83-60810-22-4
ISSN 1896-7094

On-line and off-line translation aids for non-native readers

Eric Wehrli, Luka Nerima, Violeta Seretan, Yves Scherrer
LATL-Dept. of Linguistics
University of Geneva
{Eric.Wehrli, Luka.Nerima, Violeta.Seretan, Yves.Scherrer}@unige.ch

***Abstract*—Twic and TwicPen are reading aid systems for readers of material in foreign languages. Although they include a sentence translation engine, both systems are primarily conceived to give word and expression translation to readers with a basic knowledge of the language they read. Twic has been designed for on-line material and consists of a plug-in for internet browsers communicating with our server. TwicPen offers a similar assistance for readers of printed material. It consists of a hand-held scanner connected to a lap-top (or desk-top) computer running our parsing and translation software. Both systems provide readers a limited number of translations selected on the basis of a linguistic analysis of the whole scanned text fragment (a phrase, part of the sentence, etc.). The use of a morphological and syntactic parser makes it possible (i) to disambiguate to a large extent the word selected by the user (and hence to drastically reduce the noise in the response), and (ii) to handle expressions (compounds, collocations, idioms), often a major source of difficulty for non-native readers. The systems are available for the following language-pairs: English-French, French-English, German-French, German-English, Italian-French, Spanish-French. Several other pairs are under development.**

I. Introduction

Technological developments, such as the Internet, the globalization trend of recent years, the developments of plurilingual political and entities (such as the EU), not to mention international tourism, have all lead to a huge increase of situations in which people read documents in a language other than their own. For instance, scientists around the world are expected to read documents in English (or other languages), tourists in foreign countries may not always find written information in their own language. Other examples include people interested in reading foreign press, or reading novels in their original version. In such situations, if the language is unknown to the reader, machine translation is probably the key answer. Readers who have some knowledge of the language they read, however, would usually prefer a more specific aid, such as bilingual terminological aid for an unknown term or an opaque expression.

In this paper, we would like to describe the Twic and TwicPen systems developed at LATL to provide terminological aid to readers of non-native documents. Both systems rely on the same simple concepts, that is: first submit the sentence containing the word for which the reader requests help to a full morphological and syntactic analysis. The lemmatized and largely disambiguated output of that analysis is then used to retrieve possible translations from a bilingual dictionary. The main advantages of such a system over a standard on-line bilingual dictionary are (i) lemmatization – since the parser includes a morphological analyzer, any word form is recognized, i.e. there is no need to enter (and know) the relevant base form; (ii) better precision – the disambiguation process enforced by the parser reduces the bilingual lookup to words and readings syntactically compatible with the grammatical context of the sentence, thereby greatly reducing the noise inherent to a simple dictionary lookup; (iii) multiword expressions – in our opinion (and experience) the capability of Twic to identify multiword expressions (such as idioms and collocations), even when the componants are non adjacent is by far the most remarkable advantage of this system.

For on-line documents, a variety of terminological tools are available, some of them commercially, such as the ones provided by Google (word translation services) or Babylon Ltd. More advanced, research-oriented systems based on computational linguistics technologies have also been developed, such as GLOSSER-RuG (Nerbonne et al, 1996, 1999), Compass (Breidt et al., 1997), but to date none have been commercialized or made available to a wide audience.

When it comes to more traditional documents such as books and other printed material, terminological aid is much harder to satisfy. Multilingual scanning devices have been commercialized[1], but they lack the computational linguistic resources to make them truly useful. The shortcomings of such systems are particularly blatant with inflected languages, or with compound-rich languages such as German, while the inadequate treatment of multi-word expressions is obvious for all languages.

TwicPen has been designed to overcome these shortcomings and intends to provide readers of printed material with the same kind and quality of terminological aid as Twic provides for on-line documents. For concreteness, we will take our typical user to be a French-speaking reader with knowledge of English and German reading printed material, for instance a novel or a technical document, in English or in German.

For such a user, German vocabulary is likely to be a major source of difficulty due in part to its opacity (for non-Germanic language speakers), the richness of its inflection and, above

[1]The three main text scanner manufacturers are Whizcom Technologies (http://www.whizcomtech.com), C-Pen (http://www.cpen.com) and Iris Pen (http://www.irislink.com).

all, the number and the complexity of its compounds, as exemplified in figure 1 below.[2]

This paper will describe the Twic and TwicPen systems, showing how an in-depth linguistic analysis of the sentence in which a problematic word occurs helps to provide a relevant answer to the reader. We will show, in particular, that the advantage of such an approach over a more traditional bilingual terminology system is (i) to reduce the noise with a better selection (disambiguation) of the source word, (ii) to provide in-depth morphological analysis and (iii) to handle multi-word expressions (compounds, collocations, idioms), even when the terms of the expression are not adjacent.

II. Overview of the Twic and TwicPen systems

Twic and TwicPen are both terminological aid systems based on a full linguistic analysis of the source material. Twic is design for on-line documents, and TwicPen for printed material, using a hand-held scanner to get the input material. In other words, TwicPen consists of (i) a simple hand-held scanner and (ii) parsing and translation software. TwicPen functions as follows:

- The user scans a fragment of text, which can be as short as one word or as long as a whole sentence or even a whole paragraph.
- The text appears in the user interface of the TwicPen system and is immediately parsed and tagged by the Fips parser described in the next section.
- The user can either position the cursor on the specific word for which help is requested, or navigate word by word in the sentence.
- For each word, the system retrieves from the tagged information the relevant lexeme and consults a bilingual dictionary to get one or several translations, which are then displayed in the user interface.

Figure 1 shows the user interface. The input text is the well-known German compound discussed by Kay et al. (1994) reproduced in (1):

(1) Lebensversicherungsgesellschaftsangestellter
Leben(s)-versicherung(s)-gesellschaft(s)- angestellter
life-insurance-company-employee

Such examples are not at all uncommon in German, in particular in administrative or technical documents.

Notice that the word *Versicherungsgesellschaft* (English *insurance company* and French *compagnie d'assurance*), which is a compound, has not been analyzed. This is due to the fact that, like many common compounds, it has been lexicalized.

III. The Fips parser

Fips is a robust multilingual parser which is based on generative grammar concepts for its linguistic component and object-oriented design for its implementation. It uses a bottom-up parsing algorithm with parallel treatment of alternatives, as well as heuristics to rank alternatives (and cut their numbers when necessary).

[2]See the discussion on "The Longest German Word" on http://german.about.com/library/blwort_long.htm

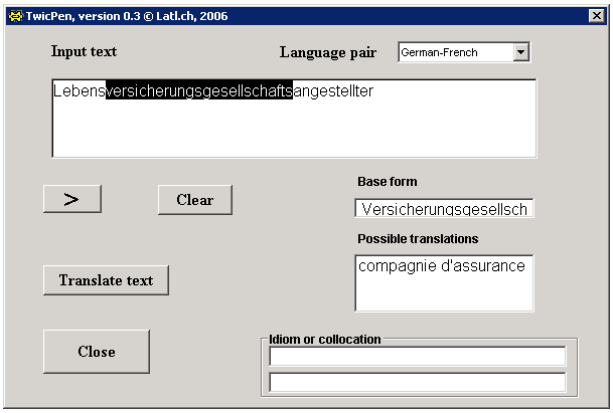

Fig. 1. TwicPen user interface with a German compound

The syntactic structures built by Fips are all of the same pattern, that is: $[_{XP}$ L X R], where L stands for the possibly empty list of left constituents, X for the (possibly empty) head of the phrase and R for the (possibly empty) list of right constituents. The possible values for X are the usual part of speech **Adv**erb, **A**djective, **N**oun, **D**eterminer, **V**erb, **T**ense, **P**reposition, **C**omplementizer, **Inter**jection.

The parser makes use of 3 fundamental mechanisms: projection, merge and move.

A. Projection

The projection mechanism assigns a fully developed structure to each incoming word, based on their category and other inherent properties. Thus, a common noun is directly projected to an NP structure, with the noun as its head, an adjective to an AP structure, a preposition to a PP structure, and so on. We assume that pronouns and, in some languages proper nouns, project to a DP structure (as illustrated in (2a). Furthermore, the occurrence of a tensed verb triggers a more elaborate projection, since a whole TP-VP structure will be assigned. For instance, in French, tensed verbs occur in T position, as illustrated in (2b):

(2)a. $[_{DP}$ Paul], $[_{DP}$ elle]
b. $[_{TP}$ manges$_i$ $[_{VP}$ e$_i$]]

B. Merge

The merge mechanism combines two adjacent constituents, A and B, either by attaching constituent A as a left constituent of B, or by attaching B as a right constituent of any active node of A (an active node is one that can still accept subconstituents).

Merge operations are constrained by various, mostly language-specific, conditions which can be described by means of procedural rules. Those rules are stated in a pseudo formalism which attempts to be both intuitive for linguists and relatively straightforward to code (for the time being, this

is done manually). The conditions take the form of boolean functions, as described in (3) for left attachments and in (4) for right attachments, where **a** and **b** refer, respectively, to the first and to the second constituent of a merge operation.

(3) **D + T**
a.AgreeWith(b,{number,person})
a.IsArgumentOf(b,subject)

Rule 3 states that a DP constituent (ie. a traditional noun phrase) can (left-)merge with a TP constituent (ie. an inflected verb phrase constituent) if (i) both constituents agree in number and person and (ii) the DP constituent can be interpreted as the subject of the TP constituent.

(4)a. **D + N**
a.HasSelectionFeature(Ncomplement)
b.HasFeature(commonNoun)
a.AgreeWith(b,{number,gender})

b. **V + D**
a.HasFeature(mainVerb)
b.IsArgumentOf(a, directObject)

Rule (4a) states that a common noun can be (right-)attached to a determiner phrase, under the conditions (i) that the head of the DP bears the selectional feature `[+Ncomplement]` (ie. the determiner selects a noun), and (ii) the determiner and the noun agree in gender and number. Finally, rule (4b) allows the attachment of a DP as a right subconstituent of a verb (i) if the verb is not an auxiliary or modal (ie. it is a main verb) and (ii) if the DP can be interpreted as a direct object argument of the verb.

C. Move

Although the general architecture of surface structures results from the combination of projection and merge operations, an additional mechanism is necessary to handle so-called extraposed elements and link them to empty constituents (noted **e** in the structural representation below) in canonical positions, thereby creating a chain between the base (canonical) position and the surface (extraposed) position of the "moved" constituent as illustrated in the following example:

(5)a. who did you invite ?
b. $[_{CP}\ [_{DP}$ who$]_i$did$_j$ $[_{TP}\ [_{DP}$ you $]$ e$_j$ $[_{VP}$ invite$[_{DP}$ e$]_i$ $]\ ]\ ]$

IV. Multi-word expressions

Perhaps the most advanced feature of the Twic/TwicPen system is its ability to handle multiword expressions (idioms, collocations), including those in which the elements of the expression are not immediately adjacent to each other. Consider the French verb-object collocation *battre-record* (*break-record*), illustrated in (6a, b), as well as in the figure 2.

(6)a. Paul a battu le record national.
Paul broke the national record
b. L'ancien record de Bob Hayes a finalement été battu.
Bob Hayes' old record was finally broken.

The collocation is relatively easy to identify in (6a), where the verb and the direct object noun are almost adjacent and occur in the expected order. It is of course much harder to spot in the (6b) sentence, where the order is reversed (due to passivization) and the distance between the two elements of the collocation is seven words. Nevertheless, as Figure 2 shows, TwicPen is capable of identifying the collocation.

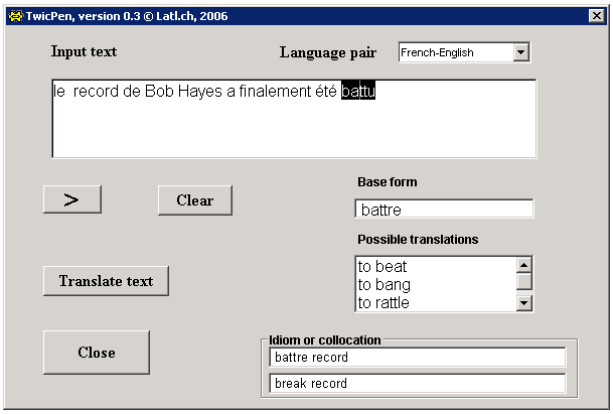

Fig. 2. Example of a collocation

The screenshot given in Figure 2 shows that the user selected the word *battu*, which is a form of the transitive verb *battre*, as indicated in the base form field of the user interface. This lexeme is commonly translated into English as *to beat, to bang, to rattle, etc.*. However, the collocation field shows that *battu* in that sentence is part of the collocation *battre-record* which is translated as *break-record*.

The ability of Twic/TwicPen to handle expressions comes from the quality of the linguistic analysis provided by the multilingual Fips parser and of the collocation knowledge base (Seretan et al., 2004, 2008). A sample analysis is given in (7b), showing how extraposed elements are connected with canonical empty positions, as assumed by generative linguists.

(7)a. The record that John broke was old.
b. $[_{TP}\ [_{DP}$ the $[_{NP}$ record$_i$ $[_{CP}$ that$_i$ $[_{TP}\ [_{DP}$ John $]\ [_{VP}$ broke $[_{DP}$ e$]_i$ $]\ ]\ ]\ ]\ ]\ [_{VP}$ was $[_{AP}$ old $]\ ]\ ]$

In this analysis, notice that the noun *record* is coindexed with the relative pronoun *that*, which in turn is coindexed with the empty direct object of the verb *broke*. Given this antecedent-trace chain, it is relatively easy for the system to identify the verb-object collocation *break-record*.

Consider now slightly more complex examples, with the same *battre-record* collocation.

(8)a. Ce **record** ne semble pas facile à **battre**.
This record doesn't seem easy to break.

b. [$_{TP}$ [$_{DP}$ ce [$_{NP}$ record]$_i$] ne semble pas [$_{FP}$ [$_{AP}$ facile [$_{CP}$ à [$_{TP}$ [$_{VP}$ battre [$_{DP}$ e]$_i$]]]]]]

With sentence (8), we have an example of the so-called *tough*-movement construction, in which the subject of a predicative adjective is also construed as the direct object of the infinitival verb complement of that adjective. So, in our example, *record* which occurs (and functions) as the main clause subject is connected, as shown by the chain of indices, to the direct object position of the verb *battre* (break). Again, thanks to that analysis, Twic/TwicPen recognize an occurrence of the *battre-record* (break-record) collocation.

(9)a. Ce **record**, Paul espérait le **battre** cette saison.
This record, Paul was hoping to break [it] this season.
b. [$_{CP}$ [$_{DP}$ ce record]$_i$ [$_{TP}$ [$_{DP}$ Paul] espérait [$_{VP}$ e [$_{TP}$ [$_{VP}$ le$_i$ battre [$_{DP}$ e]$_i$ [$_{Adv}$ [$_{DP}$ cette saison]]]]]]

Example (9) gives an illustration of the French left-dislocation structure, in which the topicalized DP *ce record* (this record) is necessarily interpreted as corefential with the direct object clitic pronoun *le* (it). Again, with such an analysis, it is easy for the system to recognize the collocation *battre-record* in such a sentence.

Our last example (10) shows a combination of *tough*-movement and left dislocation. We will not give the full analysis for that sentence, but simply observe that the the left dislocated element *ce record* is coindexed with the direct object clitic *le* (it) of the verb *considérer* (consider), which takes a predicative adjective controlled by the direct object. Since that adjective is of the *tough* variety, its subject is interpreted as the direct object of its verbal complement.

(10) Ce **record**, Paul pensait que Marie le considérait difficile à **battre**.
This record, Paul thought that Mary would consider it difficult to break.

V. Demo

Twic & TwicPen are translation aids for non-native users reading, respectively, on-line documents or printed material (books or other off-line material). They have been designed for users who have a basic knowledge of the language they read, while still experiencing some terminological problems.

Twic consists of a plug-in for common web-browsers (Mozilla Fox and Internet Explorer) which communicates with a server equipped with the lexical databases as well as the parsing and translation software. TwicPen consists of a hand-held scanner to copy the relevant material to a laptop computer on which the parser and translation software as well as the relevant databases have been copied. For both systems, the sentence containing the selected word is first analyzed by the parser in order to select the most relevant reading of the selected word given the grammatical context and then return a small number of translations (at most 5) for the selected word. In addition, when the parser identifies the selected word as a component of a multiword expression, a translation for that expression is also returned. On request, both systems are also equipped with a sentence translation facility, to provide a full translation of the sentence containing the selected word. The latest version of the Mozilla Fox Twic plug-in provides two translation options, with our own (and still under development) Its system, and with the Google API.

The demo will focus primarily on the TwicPen system, with the following topics:

1) the whole process of scanning a word, a fragment of the text or a whole sentence to get terminological help on selected words;
2) the underlying linguistic analysis and show how this analysis helps
   a) by determining the base form (morphological analysis)
   b) by reducing the noise in the bilingual retrieval
   c) with an in-depth treatment of multi-word expressions
3) the treatment of multi-word expressions, ie. compounds, collocations and idioms;
4) the Fips multilingual parser will be presented with English, French, German and Italian data;
5) time permitting, the sentence translation component can be discussed

Twic & TwicPen have been developed for the language pairs indicated in the table below (with the current number of bilingual correspondences and number of multi-word expressions):

language pair	number of correspondences	number of expressions
English-French	79’000	7’500
English-German	77’000	3’500
French-English	79’000	9’900
French-German	48’000	6’300
French-Italian	39’000	3’000
French-Spanish	25’000	1’800
German-French	48’000	2’400
German-English	77’000	2’000
Italian-French	39’000	2’900
Spanish-French	25’000	1’900
Greek-French	8’200	350
French-Japanese	18’000	

Other language pairs are under development. By the end of 2009 we expect to have all the pairs for the following set of languages English, French, German, Italian, Spanish. In addition, we are in considering the addition of Japanese (first as target language) and Greek.

VI. Conclusion

Demand for terminological tools for readers of material in a foreign language, either on-line or off-line, is likely to increase with the development of global, multilingual societies. The Twic and TwicPen systems presented in this paper have been developed for readers of both on-line and printed material.

Using Twic with on-line documents, the user selects a word for which he wants a translation. Using TwicPen with printed material, users scan the sentence (or a fragment of it) containing a word that they don't understand. For both systems, a short list of translations will be provided, as well as the translation of an expression the word might be part of. We have argued that the use of a linguistic parser in such a system brings several major benefits for the word translation task, such as (i) determining the citation form of the word, (ii) drastically reducing word ambiguities, and (iii) identifying multi-words expressions even when their constituents are not adjacent to each other.

ACKNOWLEDGEMENT

The research described in this paper has been supported in part by a grant for the Swiss National Science Foundation (No 101412-103999).

REFERENCES

[1] Breidt, E. and H. Feldweg, 1997. "Accessing Foreign Languages with COMPASS", *Machine Translation*, 12:1-2, 153–174.

[2] Kay, M., M. Gawron and P. Norvig, 1994. *Verbmobil: A Translation System for Face-to-Face Dialog*, Lecture Notes 33, Stanford, CSLI.

[3] Nerbonne, J. and P. Smit, 1996. "GLOSSER-RuG: in Support of Reading" in *Proceedings of COLING-1996*, 830-835.

[4] erbonne, J. and D. Dokter, 1999. "An Intelligent Word-Based Language Learning Assistant" in *TAL* 40:1, 125–142.

[5] Seretan V., Nerima L. and E. Wehrli, 2004. "Multi-word collocation extraction by syntactic composition of collocation bigrams", in Nicolas Nicolov et al. (eds), Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, Amsterdam, John Benjamins, 91–100.

[6] Seretan, V. and E. Wehrli, 2008. "Multilingual collocation extraction with a syntactic parser", in Language Resources and Evaluation. [available at www.springerlink.com]

[7] Wehrli, E. 2003. "Translation of Words in Context", *Proceedings of MT-Summit IX*, New Orleans, 502–504.

[8] Wehrli, E. 2004. "Traduction, traduction de mots, traduction de phrases", in B. Bel et I. Marlien (eds.), *Proceedings of TALN XI*, Fes, 483–491.

[9] Wehrli, E. 2007. "Fips, a Deep Linguistic Multilingual Parser", in *Proceedings of the ACL 2007 Workshop on Deep Linguistic Parsing*, Prag, 120–127.