TwicPen : Hand-held Scanner and Translation Software for non-Native Readers

Eric Wehrli
LATL-Dept. of Linguistics
University of Geneva
Eric.Wehrli@lettres.unige.ch

Abstract

TwicPen is a terminology-assistance system for readers of printed (ie. off-line) material in foreign languages. It consists of a hand-held scanner and sophisticated parsing and translation software to 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 system exists for the following language-pairs: English-French, French-English, German-French and Italian-French.

1 Introduction

As a consequence of globalization, a large and increasing number of people must cope with documents in a language other than their own. While readers who do not know the language can find help with machine translation services, people who have a basic fluency in the language while still experiencing some terminological difficulties do not want a full translation but rather more specific help for an unknown term or an opaque expression. Such typical users are the huge crowd of students and scientists around the world who routinely browse documents in English on the Internet or elsewhere. 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) or TWiC (Wehrli, 2003, 2004).

Similar needs are less easy to satisfy when it comes to more traditional documents such as books and other printed material. 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 multiword 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 help as is available 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 all, the number and the complexity of its compounds, as exemplified in figure 1 below.[2]

[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).

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

Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 61–64,
Sydney, July 2006.

This paper will describe the TwicPen system, 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.

2 Overview of TwicPen

The TwicPen system is a natural follow-up of TWiC (Translation of Words in Context), (see Wehrli, 2003, 2004), which is a system for online terminological help based on a full linguistic analysis of the source material. TwicPen uses a very similar technology, but is available on personal computers (or even PDAs) and uses 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.

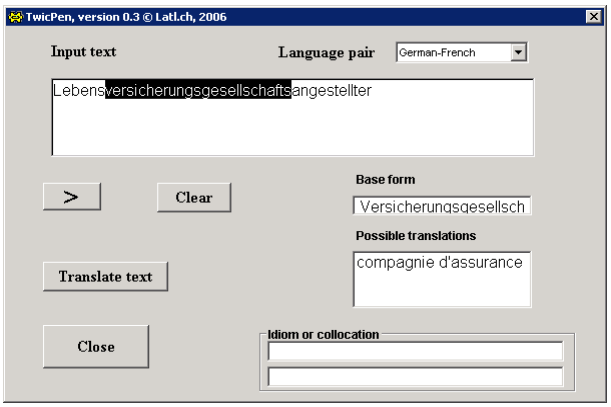

Figure 1: TwicPen user interface with a German compound

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.

3 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).

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.

3.1 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$_i$]]

3.2 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.

3.3 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]$_i$]]]

4 Multi-word expressions

Perhaps the most advanced feature of TwicPen 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 3.

(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 3 shows, TwicPen is capable of identifying the collocation.

The screenshot given in Figure 3 shows that the user selected the word *battu*, which is a form of

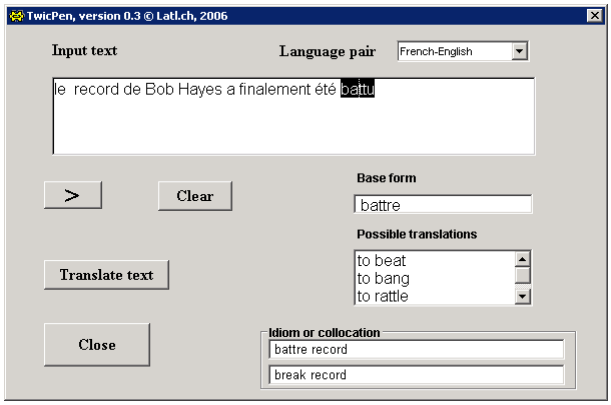

Figure 2: Example of a collocation

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 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). 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*.

5 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 TwicPen system presented in this paper has been developed for readers of printed material. They scan the sentence (or a fragment of it) containing a word that they don't understand and the system will display (on their laptop) a short list of translations. 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 multiwords expressions even when their constituents are not adjacent to each other.

Acknowledgement

Thanks to Luka Nerima and Antonio Leoni de Len for their suggestions and comments. The research described in this paper has been supported in part by a grant for the Swiss National Science Foundation (No 101412-103999).

6 References

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

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

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

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

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.

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

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