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

WEHRLI, Eric, et al.

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 […]

Reference

Available at:
http://archive-ouverte.unige.ch/unige:22825

Disclaimer: layout of this document may differ from the published version.
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

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 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 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
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 ad-

tantage 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 materia-

l. 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)
replicated in (1):

\[
\text{(1) Lebensversicherungsgesellschaftsangestellter} \\
\text{Leben(s)-versicherung(s)-gesellschaft(s)-} \\
\text{angestellter} \\
\text{life-insurance-company-employee}
\]

Such examples are not at all uncommon in German, in

particular in administrative or technical documents.

Notice that the word Versicherungsgesellschaft (English in-

surance 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

\(^2\)See the discussion on “The Longest German Word” on http://german.
about.com/library/blwort_long.htm

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: \([\_L \_ \_X \_R\_]\), 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 Adverb, Adjective, Noun, Determiner, Verb, Tense,
Preposition, Complementizer, Interjection.

The parser makes use of 3 fundamental mechanisms: pro-
jection, merge and move.

A. Projection

The projection mechanism assigns a fully developed struc-
ture 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):

\[
\text{(2a) } \left[\_\_DP \text{ Paul } \_\_DP \text{ elle } \_\_\right] \\
\text{(2b) } \left[\_\_TP \text{ manges } \_\_TP \text{ e } \_\_\right]
\]

B. Merge

The merge mechanism combines two adjacent constituents,
A and B, either by attaching constituent A as a left constituen-
t 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 sub-

constituents).

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 for-
malism 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.

\[ (3) \quad D + T \]
\[ \begin{align*}
& a. \text{AgreeWith}(b, \{ \text{number,person} \}) \\
& a. \text{IsArgumentOf}(b, \text{subject})
\end{align*} \]

Rule 3 states that if 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.

\[ (4a) \quad D + N \]
\[ \begin{align*}
& a. \text{HasSelectionFeature} \{ \text{Ncomplement} \} \\
& b. \text{HasFeature} \{ \text{commonNoun} \} \\
& a. \text{AgreeWith}(b, \{ \text{number,gender} \})
\end{align*} \]

\[ (4b) \quad V + D \]
\[ \begin{align*}
& a. \text{HasFeature} \{ \text{mainVerb} \} \\
& b. \text{IsArgumentOf}(a, \text{directObject})
\end{align*} \]

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 \( \{ + \text{Ncomplement} \} \) (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:

\[ (5a) \]
\[ \begin{align*}
& \text{who did you invite ?} \\
& \text{b. \{ CP \{ DP \{ who \}\}, did, \{ TP \{ DP \{ you \}\}, \{ \text{ invite} \{ DP \{ e \}\}, \} \]\} }
\end{align*} \]

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 \textit{battre-record} (\textit{break-record}), illustrated in (6a, b), as well as in the figure 2.

\[ (6a) \]
\[ \text{Paul a battu le record national.} \]

Paul broke the national record

\[ (6b) \]
\[ \text{L’ancien record de Bob Hayes a finalement \text{\textit{battu}.}} \]

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.

\[ \text{Fig. 2. Example of a collocation} \]

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

\[ (7b) \]
\[ \text{b. \{ TP \{ DP \{ \text{ the \{ record }\{ AP \{ old \} \text{ }\}\}, \{ that, \{ TP \{ DP \text{ John }\}\}, \{ VP \{ \text{broke }\{ DP \text{ e }\}, \} \}\}\}\text{ }\}\}\]

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

Consider now slightly more complex examples, with the same \textit{battre-record} collocation.

\[ (8a) \]
\[ \text{Ce \textit{record} ne semble pas facile à \textit{battre}.} \]

This record doesn’t seem easy to break.
b. $[\text{TP} [\text{NP} \text{ce record}]_i] ne \text{semble pas} [\text{FP} [\text{AP} \text{facile} \ [\text{CP} [\text{VP} \text{battere (DP e)}_i] \text{ce saison}]]]]]

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 battere (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. $[\text{CP} [\text{DP} \text{ce record}]_i] [\text{TP} [\text{DP} \text{Paul}] \text{espérait} [\text{VP} \text{e} [\text{TP} \text{le} \text{battere (DP e)}_i] \text{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 anal-
ysis)
b) by reducing the noise in the bilingual retrieval
c) with an in-depth treatment of multi-word expres-
sions
3) the treatment of multi-word expressions, ie. compounds,
colloctions 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 bilin-
gual correspondences and number of multi-word expressions):

<table>
<thead>
<tr>
<th>language pair</th>
<th>number of correspondences</th>
<th>number of expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-French</td>
<td>79'000</td>
<td>7'500</td>
</tr>
<tr>
<td>English-German</td>
<td>77'000</td>
<td>3'500</td>
</tr>
<tr>
<td>French-English</td>
<td>79'000</td>
<td>9'900</td>
</tr>
<tr>
<td>French-German</td>
<td>48'000</td>
<td>6'300</td>
</tr>
<tr>
<td>French-Italian</td>
<td>39'000</td>
<td>3'000</td>
</tr>
<tr>
<td>French-Spanish</td>
<td>25'000</td>
<td>1'800</td>
</tr>
<tr>
<td>German-French</td>
<td>48'000</td>
<td>2'400</td>
</tr>
<tr>
<td>German-English</td>
<td>77'000</td>
<td>2'000</td>
</tr>
<tr>
<td>Italian-French</td>
<td>39'000</td>
<td>2'900</td>
</tr>
<tr>
<td>Spanish-French</td>
<td>25'000</td>
<td>1'900</td>
</tr>
<tr>
<td>Greek-French</td>
<td>8'200</td>
<td>350</td>
</tr>
<tr>
<td>French-Japanese</td>
<td>18'000</td>
<td></td>
</tr>
</tbody>
</table>

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