Web Corpora for Web Translation:
Discovering Translation Equivalents in Comparable Tourism Corpora

© Patrizia Pierini 2010

Pre-print text written for the Challenges for the 21st Century: Dilemmas, Ambiguities, Directions, edited by Di Martino G. - Lombardo L. - Nuccorini S.

1. Introduction

Web translation is the transfer of the content of websites to another language/culture within the process of localisation, which also includes non-linguistic activities such as software engineering (Esselink 2003). The website as hypertext is a multisemiotic genre resulting from the interplay of various semiotic resources — written and spoken language, images, graphics and acoustic elements (sounds, music). The transfer of the verbal component concerns both connected discourse and paratext — text links, menus, headings, headers, footers, etc.

Because of the peculiarities of the website as genre, Web translation calls for the exploitation of Web corpora providing samples of natural-occurring language in hypertexts that can contribute to improving translation quality. ‘Web corpus’ is used here to refer to a corpus compiled by downloading full hypertexts, i.e. all web pages accessible under a common Web address, including paratextual elements.

It is well-known that the translation quality of printed material containing tourism discourses is often not very high, while websites for tourism are sometimes no better (Pierini 2007). One of the reasons for low quality translations is that the complexity of tourism discourse is often underestimated: it is a specialised discourse characterised by specific features (Dann 1996; Gotti 2006) and a culturally constructed phraseology (Gerbig – Shek 2007). One aspect to consider is the seriousness of translation errors in hypertexts; even a trivial error in a web page is to be regarded as serious because of its ‘visibility’: it appears in a message available to a potentially global audience (Scarpa 2008, 237), and in the case of promotional and commercial messages, it may affect sales.

This paper deals with the issue of ‘translation equivalence’, understood as the semantic, pragmatic and cultural relationship between the source text and the target text, as studied by means of the tools of corpus linguistics. Two comparable Web corpora were built by downloading the pages of British and Italian hotel websites and saving them in text format. The aim is twofold: to present a new perspective on Web corpora and to discover translation equivalents through a corpus-driven approach. Following Tognini Bonelli (2001), the methodology combines frequency and
collocational information. Designed for connected discourse, it will also be tentatively adopted for paratext. The analysis will focus on hotel as link label in paratext, more precisely in the main navigation bar, and on has/have as main verb in connected discourse.

2. Why Web corpora

There are three reasons why there is a need for the compilation of a Web corpus when translating websites. The first reason is that Web writing differs in some aspects from traditional writing in print: Web style is characterised by the use of simple sentences, conciseness, discourse organisation in short lexias, one idea per paragraph, the inverted pyramid style (from main information to details), typographic features facilitating scanning. The second reason is that specialised discourse, when used on the Web, exhibits particular features, as illustrated by Pierini (2008) on promotional tourism discourse. The third reason is that a Web corpus can provide paratextual data. Paratext is not an accessory element, but a constitutive part of hypertexts whose communicative relevance is underlined in studies on Web usability (see microcontent in Nielsen 1999) and Web communication (Garzone 2002). Paratextual elements give the receiver an idea of the content of a website, and in the case of links, they also represent the instruments for accessing information.

In the conversion of web pages, paratextual elements are flattened into text format. They constitute a significant amount of (often repetitive) material, usually considered as fragments ‘disturbing’ connected discourse (Fletcher 2004). Depending on their aims, researchers can post-edit the corpus by removing paratextual elements (Scarpa 2008, 316). Our working hypothesis is that, leaving the corpus unedited, it is possible to explore paratextual elements, in particular, the use of lexical items as link labels. Such exploration poses some methodological questions: how can data be extracted from them? Can a corpus-driven approach work when dealing with paratext? We shall try to answer these questions in Section 5, by tentatively analysing hotel as link label.

3. Corpus compilation and methodology

Two comparable Web corpora were assembled at the University of “Roma Tre” by downloading full British and Italian websites of 3- and 4-star hotels, and saving them in text format. The corpora were put together on the basis of similarity in their contextual features: medium (the Web); genre (Web advert); communicative purpose (persuading the receiver to purchase a similar tourist product); sender (individual establishment) and encoder (tourism professional); intended audience (a market segment defined by the product).

The British corpus amounts to 196,500 running words, and the Italian one to 118,600. They are small and untagged, but, as pointed out by Bowker – Pearson (2002, 48), even small corpora,
that are anywhere from about ten thousand to several hundreds of thousands of words, can be reliable aids in the study of domain-specific English. The analysis of the data was assisted by the use of *Wordsmith Tools 5.0* (Scott 2007).

The methodology adopted follows the approach suggested in Tognini Bonelli (2001, 131-36): it sees meaning as ‘function in context’ and considers the lexico-grammatical information provided by the co-text as being of fundamental importance. It consists of three steps: given an initial word that we want to translate, we shall examine its collocational profile in order to find its most frequent collocates; then, for each collocate we shall posit a *prima facie* translation equivalent that will be searched for; finally, within the collocational range of the translation equivalent, we shall try to identify an equivalent to the initial node word.

### 4. Starting the investigation

To begin with, we created a frequency wordlist for both corpora. At the top of the list of lexical items, we find *hotel* in both of them; in the Italian corpus, there are other items denoting ‘hotel’ as well (Table 1).

<table>
<thead>
<tr>
<th>British Corpus</th>
<th>Italian Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>hotel</em> 4231 (2.17%)</td>
<td><em>hotel</em> 2131 (1.79%)</td>
</tr>
<tr>
<td><em>albergo</em> 312 (0.26%)</td>
<td><em>struttura</em> 52 (0.04%)</td>
</tr>
</tbody>
</table>

**Table 1. Frequency of words denoting ‘hotel’**

The high frequency of *hotel* in both corpora is due to its frequent use both in connected discourse and paratext. For this reason, we decided to analyse *hotel* as our link label.

From the frequency list of the British corpus, it emerges that some verbal items are very frequent (Table 2).

<table>
<thead>
<tr>
<th>word</th>
<th>frequency</th>
<th>percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>is</em></td>
<td>1428</td>
<td>0.73</td>
</tr>
<tr>
<td><em>are</em></td>
<td>1085</td>
<td>0.56</td>
</tr>
<tr>
<td><em>contact</em></td>
<td>624</td>
<td>0.32</td>
</tr>
<tr>
<td><em>be</em></td>
<td>581</td>
<td>0.30</td>
</tr>
<tr>
<td><em>can</em></td>
<td>465</td>
<td>0.24</td>
</tr>
<tr>
<td><em>will</em></td>
<td>461</td>
<td>0.24</td>
</tr>
<tr>
<td><em>have</em></td>
<td>453</td>
<td>0.23</td>
</tr>
<tr>
<td><em>has</em></td>
<td>316</td>
<td>0.16</td>
</tr>
<tr>
<td><em>book</em></td>
<td>311</td>
<td>0.16</td>
</tr>
<tr>
<td><em>click</em></td>
<td>298</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Table 2. Top verbal items in the frequency list from the British corpus**
The verbs in Table 2 can be grouped into three classes: a) basic verbs (be, have), frequently serving as main verb in propositions describing the hotel. For example, is occurs in propositions aimed at defining the identity of the hotel (e.g. The hotel is ideal for business guests). b) activity verbs (contact, book, click), mainly in the imperative form, that call the receiver to action by interacting with the computer, and whose frequent presence is one of the main features of Web communication. c) the modals can and will, typical of promotional tourism discourse (Pierini 2008).

With regard to connected discourse, we decided to concentrate on have as main verb when used to describe the structure of a hotel. In order to explore the usage of has and have, we considered all concordances of each of them and sorted one word to the right and one to the left to spot the phrases co-occurring with each form.

5. **Paratext: discovering the translation equivalents of hotel as link label**

In hotel websites, general information about the hotel is provided on the home page or on a page usually labelled hotel in British sites. When we search for this node word, we find concordances showing two uses, i.e. its use in connected discourse or as link label in paratext. The label of a link can be easily identified: the item is found in a sequence of formally unconnected phrases occurring several times in the corpus. Consider the following:

This non smoking hotel offers guests all modern comforts expected of a three star

<table>
<thead>
<tr>
<th>Home</th>
<th>Hotel</th>
<th>Reservations</th>
<th>Dining &amp; Events</th>
<th>Weddings</th>
<th>Contact us</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hotel</td>
<td>Tariff</td>
<td>Food and Drink</td>
<td>Conferences</td>
<td>How to contact us</td>
</tr>
<tr>
<td></td>
<td>The Hotel</td>
<td>Rooms and suites</td>
<td>Restaurant</td>
<td>Special Events</td>
<td>Special Offers</td>
</tr>
</tbody>
</table>

The first concordance shows the use of the item in connected discourse; the others illustrate its use as label, showing the co-text of hotel: on the left, home is the only collocate of hotel; on the right, the most frequent labels are rooms, reservations, tariff.

First, we read the concordances of the equivalent of hotel, i.e. hotel, albergo, struttura. Then, we searched for the left collocate of hotel, i.e. home, commonly used in Italian websites, and the translation equivalents of its right collocates, i.e. Camere, Prenotazione/i, Tariffe. Analysing the sets of concordances, we identified the navigation bars where the page devoted to the hotel, when present, is labelled in various ways. Here follow some selected concordances:

<table>
<thead>
<tr>
<th>Home</th>
<th>Chi siamo</th>
<th>Dove siamo</th>
<th>Camere e servizi</th>
<th>Bar</th>
<th>Giardino</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>Descrizione</td>
<td>Camere e Prezzi</td>
<td>Posizione</td>
<td>Eventi</td>
<td></td>
</tr>
<tr>
<td>L’Hotel</td>
<td>Tariffe</td>
<td>Offerte speciali</td>
<td>Dove Siamo</td>
<td>Prenotazioni</td>
<td></td>
</tr>
<tr>
<td>Hotel</td>
<td>Prenotazione</td>
<td>Pacchetti</td>
<td>Promozioni</td>
<td>Camere</td>
<td>Ristorante</td>
</tr>
</tbody>
</table>
The labels most frequently used, constituting the functional translation equivalents of hotel, are: hotel (10), descrizione (6), l’hotel (5), chi siamo (4). The data suggest a variation in link labelling that in Italian hotel websites appears to be less standardised than in British websites. For example, Chi siamo is usually found in corporate and institutional websites; Descrizione, referring to a communicative purpose, is not communicatively efficient: a link label should present relevant information about the content of its associated web page (Nielsen 2009).

6. Connected discourse: discovering the translation equivalents of has and have

In the British corpus, has and have are used to describe the hotel and its components. Has occurs as main verb 180 times out of 316. On the left, its most frequent collocate is hotel; on the right, its collocates are words referring to rooms and facilities:

ever since. The hotel now has 23 suites and 35 bedrooms - each
Seckford Hall Hotel has two restaurants for you to enjoy
City Centre. The hotel has a large private car park at the rear

Other frequent left collocates of has are room/bedroom-suite; in the right co-text, we find words denoting their facilities. More rarely, its left collocates denote other hotel components:

Bedrooms. Each bedroom has a private shower, colour TV
in gracious living. The Marryat Suite has air-conditioning, mini-bar, satellite
the luminous, 350 square foot lounge has a comfortable seating area with

When collocating on the left with hotel or one of its components, has also co-occurs with views on the right:

East corner of Cornwall. The hotel has the most spectacular views over
not try our bar menu? The bar also has stunning views and sandwiches

Then, we scanned the concordances of have, that occurs as main verb 306 times out of 453. On the left, its most frequent collocates are rooms/bedrooms/suites, and the exclusive-we pronoun (referring to the sender). The pattern rooms/bedrooms/suites + have collocate with words denoting facilities and views:

plenty of storage space. Most rooms have a balcony or a terrace. All
private garden. All our luxurious suites have large bedrooms with separate sitting
DOGS ARE WELCOME Most bedrooms have sea views, are tastefully decorated and

The patterns we + have and we+ also/even+ have occur 57 times and 25 times, respectively; they frequently collocate with words denoting some additional, non-standard facility offered by the hotel:
We have a limited supply of cots and rollaway beds which are available weather permitting. In addition we have a golf practice area. It is a three-quarters drive away. We even have our own helipad, so there's no excuse

With regard to the Italian corpus, we searched for the prima facie equivalents of has and have, i.e. the inflected forms of avere: the frequencies of ha, hanno, abbiamo are low, respectively 83, 53, 17. By analysing the concordances of ha and hanno, we note that they are used mainly as auxiliary with a percentage of about 70%. When main verbs, their left collocates are varied; when co-occurring with hotel, ha does not collocate on the right with words denoting its components:

San Marco. L’Hotel Giorgione ha il vantaggio incomparabile che gli deriva
Il Ristorante ha una capienza di 80 persone ma con l’utilizzo
I nostri Clienti hanno diritto a una riduzione sul costo del parcheggio

The next step was to search for the prima facie equivalent of hotel as left collocate of has, i.e. hotel, albergo, struttura. The following occurrences were found:

confortevole ed elegante struttura dispone di 95 camere e 5 junior suite
il modulo di richiesta. L’Hotel Delfino dispone di un lussuoso ristorante
nelle nostre strutture attrezzate. L’Hotel dispone, inoltre, di bar, di una saletta

Less frequently, we also found:

i banchetti di nozze, l’Hotel Delfino mette a disposizione una grande sala con ampio parcheggio interno che l’albergo mette a disposizione dei clienti gratuitamente
Da oggi l’Hotel mette a disposizione dei suoi ospiti tre nuovi Internet Point

Comparing the two sets of concordances, we note two different patterns: the right collocates of dispone di are both camere and facilities; the right collocates of mette a disposizione (di + client/ospiti) are words denoting facilities only. So, the pattern hotel + has has two different functional equivalents according to its right collocates.

Then, we searched for the Italian equivalents of the other left collocates of has/have, i.e. stanza*/camer*/suite/sal*, etc.:

molte ville venete. Ogni camera è dotata di cassaforte, mini-bar, aria
da luce artificiale. Le sale sono dotate delle più sofisticate attrezzature

The most frequent verb form appears to be essere dotato di.

Our next step was to search for the Italian equivalents of view(s), one of the right collocates of has/have, i.e. vista and panorama. Here are some concordances:

allo stile toscano, e molte offrono la vista su caratteristici scorci
ristorante, situato al sesto piano, gode di una vista eccezionale che sovrasta
Luna Wellness Hotel gode di un meraviglioso panorama intorno a sé.
residenziale e commerciale, Offre anche uno splendido panorama

So the functional equivalents of the pattern has/have + views are the verbs offrire/godere di + vista/panorama.

In order to find the equivalents of the pattern we have, we analysed the concordances of abbiamo, that occurs as main verb 12 times out of 17:

Ombrelli. In caso di pioggia, abbiamo ombrelli da poter dare in prestito.
Libri. In albergo abbiamo diversi scaffali con libri per i nostri clienti.
Al ricevimento abbiamo a disposizione dei ns. clienti cartoline, francobolli

Since the right collocates of we have are varied and the additional facilities available in Italian hotels may differ from those in UK hotels, we consulted the frequency list. Here, we found other verb forms inflected by the first person plural with no subject. Exploring their concordances, we identified the expressions mettiamo a disposizione (4 occurrences) and disponiamo di (3):

la spiaggia e il centro di Taormina, mettiamo gratuitamente a disposizione la navetta check-out dalla camera è alle ore 11.00. Disponiamo di deposito bagagli al ricevimento.

The verbal expressions abbiamo (a disposizione), mettiamo a disposizione, disponiamo di collocate with words denoting some particular facility (ombrelli, libri, cartoline, navetta, deposito bagagli). They appear to highlight something ‘extra’, similarly to the pattern we have.

We also searched for anche/persino, the prima facie equivalents of also/even in the English pattern. Anche, occurring 174 times, co-occurs with words denoting facilities in already identified patterns:

permesso di parcheggio “Pass”. Disponiamo anche di un garage privato (prenotabile).
al ricevimento. Mettiamo a disposizione anche posate e piatti. Su disponibilità e con libri già letti. Il ricevimento dispone anche di libri e guide sulla Sicilia

Noting that a disposizione co-occurs with two verbs (avere, mettere), we searched for this expression to find other collocates. We found 24 occurrences of the pattern è/sono + a disposizione, that take as subject words denoting specific facilities:

residenza Azzurra. Sale giochi a disposizione dei nostri ospiti.
Sulla nostra terrazza sul tetto sono a disposizione dei nostri clienti sedie e sdraio. che invece desiderano fare da soli, è a disposizione una piccola stireria.

The higher frequency of this pattern, when compared to the patterns above, suggests that Italian prefers an impersonal presentation to verbal expressions such as abbiamo and disponiamo.

We summarise our findings in Table 3, showing English patterns and their equivalents in Italian.
<table>
<thead>
<tr>
<th>English pattern</th>
<th>Italian equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotel + has + rooms/facilities</td>
<td>• dispone di + rooms/facilities</td>
</tr>
<tr>
<td></td>
<td>• mette a disposizione (di + ospiti/clienti) + facilities</td>
</tr>
<tr>
<td>bedroom(s)/room(s)/suite(s)/component of hotel + has/have + facilities</td>
<td>essere dotato di</td>
</tr>
<tr>
<td>hotel/component + has/have + views</td>
<td>offrire /godere di + vista/panorama</td>
</tr>
<tr>
<td>we + have+ additional facility</td>
<td>facility/ies + essere a disposizione</td>
</tr>
</tbody>
</table>

Table 3. Has/have and their functional translation equivalents in Italian

7. Conclusions

This study has shown that a corpus-driven approach can work when dealing with connected discourse as well as paratext: we have been able to find the Italian translation equivalents of hotel when used as link label in the navigation bar. In connected discourse, we have seen that the Italian equivalents of has/have are different verbal expressions depending on the collocational patterns in which they occur. The equivalents belong to a more formal register, and impersonal constructions are preferred to personal ones. We can conclude by saying that a study of the co-text of a word allows to find functional translation equivalents not provided by dictionaries.

References


