Some issues concerning a corpus-based English-Arabic dictionary of hotel promotion

MIGUEL FUSTER-MÁRQUEZ &
FRANCISCA SUAU-JIMÉNEZ
IULMA-UV

First of all …

This contribution discusses a project *within* a project.

It is part of a larger research project:

"***Análisis léxico y discursivo de corpus paralelos y comparables (español-inglés-francés) de páginas electrónicas de promoción turística.***" Ref. FFI2011 (2012-2014] [Awarded by the Ministerio de Economía y Competitividad]

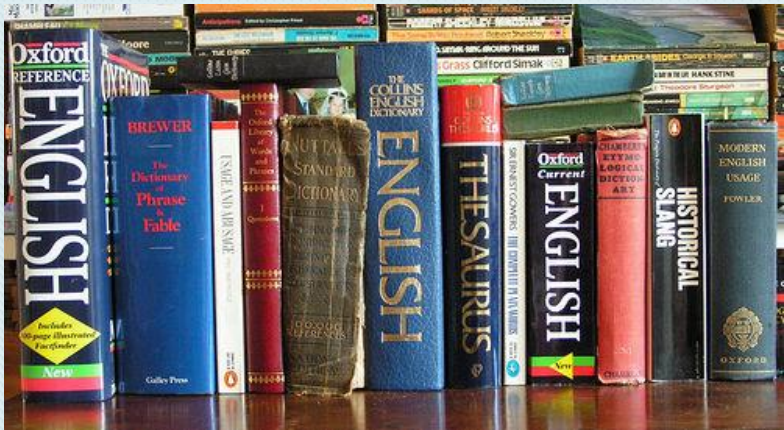
COMETVAL Research Group, IULMA, University of Valencia: http://www.uv.es/cometval

Some remarks about modern lexicography

- It is well-known that current monolingual dictionaries -most notably those designed for learners- are **corpus-based.**
- This is part of a trend which started in the 1960s whose significance is seen in the number of dictionaries and grammars which are being published with principles which originate in CL research (see Fuster & Clavel 2010).
- Moreover, the corpora owned by leading publishing houses has grown exponentially, in number and size.

Some remarks about modern lexicography

Thus, it is hard to imagine the existence of modern monolingual dictionaries where corpora do not have a central role, at least in the English context.



And what about other dictionaries?

The presence of CL in lexicography is not a matter of *either or.* There are various degrees which have to do with two factors:

To what extent the corpus provides the basis for the items selected or described;

To what extent CL informs how the entry is designed.

And now our project …

- Our goal is to make a small specialized corpus-based dictionary of interest to the tourist industry.

- The corpus is made up of texts extracted from high, medium and low budget British hotel websites which were collected during the years 2010 and 2011.

Other features

At this stage, our corpus contains around

150,000 words

Is the corpus large enough for this type of dictionary? *Representativeness.*

This size refers just to *the English side*, not to Arabic.

But some decisions about the final size of such corpus have not been made.

Corpus size is determined by…

the language (genre(s), text types and national variety) of the texts collected that should inform the dictionary.

Consider this site, which has been one of our sources.

Hotel Web SITES as sources

Websites are more complex than ***Webpages***:

- **A website** is a collection of one or more web pages designed to convey information on a particular subject or theme to a web user.
- A **web page** is one screen full of information (from a web site) that may contain links to other pages in the web site or links to external information.
- [http://wiki.answers.com/Q/The_difference_between_a_website_and_a_webpage]

Hotel Web SITES as sources

Decisions have to be made about which text(s) will be part of our corpus.

Hotel websites typically host webpages which inform clients about various issues related to their hospitality business.

Since the hosted information may be too diverse, a selective process is required (fields/categories selected).

Hotel Web SITES … not just informative

- It is misleading to think that these hospitality websites have a purely informative goal.
- Hotel websites are carefully designed to inform potential guests about the qualities of services offered by the product.
- The whole presentation (images, videos, etc.) and the language choices contribute to a single goal: attract customers.

Other languages …

- Similar translation corpora and/or comparable corpora are being gathered for the Spanish and French bilingual dictionary versions. [See Granger 2010]
- Let us return to the hotel website and observe the presence of languages.

Other languages …

- As a next step, we are considering the creation of a multilingual corpus of translation, but their number is affected by

1. their availability: different websites offer different languages.
2. the SL and TLs are not always certain.
3. but Arabic is an uncommon choice on the websites visited [to be discussed later].

…concordancing tools and word lists

- So far, these small-sized comparable corpora are being explored and word lists obtained individually by means of the free AntConc software.
- The latest version can be obtained from Lawrence Anthony's website, together with other instructions. [http://www.antlab.sci.waseda.ac.jp/antconc_index.html]

…from corpus design to dictionary design…

Not all corpus-based based dictionaries are corpus-based in the same way or to an equal degree (see Atkins and Rundell 2008).

In our case, following IULMA's lexicographical experience, it was decided that the corpus should have great weight in relation to:

(1) *collecting data*

(2) *selection of entry words*

(3) *selection of data within the entry*.

Diccionario de Términos de la Piedra Natural e Industrias Afines (2005), de Alcaraz, Hughes, Mateo, Vargas y Gómez.

slope[1] *n:* GRAL/EXPLOT pendiente, talud. [Exp: **slope**[2] (CONST vertiente, agua, faldón; V. *hip; rafter; roof; pitch; valley*), **slope angle** (EXPLOT talud de banco; ángulo delimitado entre el plano horizontal y la línea de máxima pendiente de la cara del banco; V. *bench; berm*[2]; *face*[2]; *stope*), **slope stability** (EXPLOT estabilidad del talud), **sloping** (GRAL inclinado, oblicuo ◊ *Natural slate shingles were used on sloping roofs*), **sloping ground** (EXPLOT terreno en pendiente)].

…selecting entry words from word lists…

Data collection and compilation have been briefly mentioned. Now, let us draw our attention to a sample of the English word list which stems from these websites, starting from letter ***p***:

plan, player, plus, policy, pool, port, portal, portals, porter, position, room, services, suite, price, products, program, professional, …

What kind of wordlist is this!!??

…selecting entry words from word lists…

After discarding stop words (if, you, are, in, etc.) we were left with a dictionary-worth list of content words which had these problems:

- **Some lexical words appeared with a single word form (see Fuster-Márquez 2012). [size?]**
- **For some words, we had the impression that the corpus had given us poor information. [size?]**
- **Most of the words did not appear to be terminological enough for a specialised dictionary. [General English?]**

…not sufficiently specialised?

The third issue led us to think that when words were examined in isolation they seemed quite ordinary. They did not provide us with the clue to their relevance. But when combined they appeared to be more revealing

plan	***open plan***
	floor plans
tv	***plasma (screen) tv***
bed	***floating bed***
	platform bed
policy	***private policies***
pool	***swimming pool, rooftop***
	swimming pool, indoor pool
port	***data port***

…potential user

It has been noted that specialised dictionaries may contain words which show different levels of specialisation according to the user profile/point of view:

A) Highly specific vocabulary ← **for experts**
B) Less specialised ← **experts and non-experts**
C) Not very specialised ← **non-experts/laymen**.

Most of the words qualify as B and C

[see Vargas Sierra (2008) http://rua.ua.es/dspace/bitstream/10045/13212/1/1453%20Vargas.pdf]

And, of course, the user in mind

No dictionary is conceived without a *user profile*. This matter is decided by lexicographers (or terminographers) as soon as the project begins.

The potential users of a dictionary like this is someone:

- who works in the hospitality business
- or works in the tourist industry business.
- or needs to translate a text of this kind.
- and students, teachers of ESP (tourism-related businesses), etc.

the user in mind

In other words:

these dictionaries are conceived as part of research which seeks to investigate the language of an emerging internet genre or subgenre: hospitality promotion on websites.

Language is here a central component, there is an intentional use of a multimodal presentation with very selective language content.

A corpus-based approach to content

Both common and less common words reflect actual use in the websites and are described as part and parcel of typical discourse patterns which emerge when the corpus is analysed and collocational patterns revealed (see Suau-Jiménez & Dolón-Herrero 2007, Suau-Jiménez 2011a, 2011b, and Mapelli 2008).

Not just a corpus-based approach

Create a *lexico-grammatical profile*:

1) Collocational patterns
2) Colligation
3) Semantic preference
4) Semantic profile

A contextual description starts with the examination of a word's lexico-grammatical profile is the basis for the entry word. Not different from a corpus-driven approach with attention to detail is a necessary step.

[see Altenberg and Granger 2002: 32; Sinclair 2004, O'Keefe, McCarthy and Carter 2007: 14-15; etc.]

think about the relevance of context…

Collocates, semantic preference and semantic profile play a crucial role in lexical choice:

- [WIDE/HUGE/FULL] ***RANGE***
- [BEST, LOWER, ONLINE] ***RATE(S)***
- [STAR] ***RATED***
- [WITHIN EASY] ***REACH***
- [EASILY, CONVENIENTLY] ***REACHED***

Work in progress: example of entry word (English)

- **EXCHANGE:** a) *n.*

  Conversion or interchange of sums of money of equivalent value, as between different currencies or different issues of the same currency.

  Collocations/Multiwords units:

  Currency exchange: *The role will be to conduct guest check-ins and check-outs, including rooming of our guests, as well as carrying out accurate currency exchange and handling of cash.* Exchange rates: *Exchange rates and reference to currency converters will not be considered by Customer Service at the time of review.*

Arabic and directionality

To be discussed at a later stage:

1. The selection of code for Arabic in a diglossic context where there is variation (Moroccan).
2. The lack of an adequate corpus in Arabic.

Hence: bilingual unidirectional dictionary:

the source text is provided by English

the target language is Arabic.

Some final remarks

- All these choices will necessarily be reflected in the concept of an initial product, a small English-Arabic hospitality dictionary, that could be useful to translators and hospitality students as well as for professionals in Spanish and Moroccan travel agencies and hotels.

And thanks for being here …

Francisca Suau Jiménez *francisca.suau@uv.es*
Miguel Fuster Márquez *miguel.fuster@uv.es*

IULMA
Institut Interuniversitari de Llengües Modernes Aplicades de la Comunitat Valenciana
Instituto Interuniversitario de Lenguas Modernas Aplicadas de la Comunidad Valenciana