Matamala, A. & M. Lorente (2008) "New tools for translators: INTCA, an electronic dictionary of interjections". Chiaro, Delia; Heiss, Christine & Ch. Bucaria (eds.) *Between Text and Image. Updating research in screen translation*. Amsterdam: Benjamins, 63-75.

NEW TOOLS FOR TRANSLATORS: INTCA, AN ELECTRONIC DICTIONARY OF INTERJECTIONS

Anna Matamala, Autonomous University of Barcelona
and
Mercè Lorente, Pompeu Fabra University, Spain

ABSTRACT

This paper presents INTCA (Interjeccions Català-Anglès, i.e. Interjections Catalan-English), a prototype for an electronic dictionary of interjections in English and Catalan, aimed at the needs of language professionals in general and audiovisual translators in particular. This proposal is based on a theoretical framework developed by Cuenca (2004), who includes cognitive postulates to define interjections. The audiovisual corpus used for the analysis of interjections and the collection of lexicographical data for the prototype are also described.

Keywords: audiovisual translation, corpus linguistics, electronic lexicography

1. Introduction

A variety of resources are available for translation problems: lexicographical tools, terminological databases, Internet sites, corpora, etc. Nevertheless, few of these resources take into account the specific needs of translators, in particular those working in the audiovisual and multimedia fields who are confronted with oral language. Matamala (2005)[1] attempts to solve one of these problems by focusing on a specific marker of orality (interjections) and by proposing a prototype version of a dictionary of interjections called INTCA (*Interjeccions Català-Anglès*, i.e. Interjections Catalan-English). The structure of this article begins with the background (2), the definition of interjections (2.1), and the theoretical framework (2.2). The next part is dedicated to the aims and applicability of this prototype (3), the real needs behind an electronic dictionary of interjections (3.1), and its general characteristics (3.2). Finally, some conclusions about the relevance in developing such a tool are discussed (4).

2. Background

This paper is based on previous work (Matamala 2001 and 2005) involving the importance of interjections in audiovisual products and their poor representation in dictionaries. The aims of these investigations included 1) a comprehensive description of interjections based on a corpus and 2) the design of a lexicographical prototype compiling interjections and useful information to language professionals dealing with oral language, and especially for audiovisual and multimedia translators.

A corpus was needed to achieve these goals, but as no adequate sample was found, we created an aligned corpus in 2002. As for genre selectivity, sitcoms were considered an appropriate product in that they represent spontaneous oral language, offering an approximation to real spontaneous speech, and as Fischer (2000: 204) points out, "The dialogues constituted by playwrights or authors of film scripts may display their authors'

competence in communication even more overtly than real dialogue do since the displays can be seen as a kind of typical interaction".

The corpus currently contains two subcorpora: (1) a monolingual subcorpus and (2) a bilingual subcorpus. The monolingual subcorpus (sitcoms originally shot in Catalan) allowed us to analyse interjections from a monolingual perspective and to collect data for the dictionary. The series selected were 25-minute sitcoms broadcast on Catalan Television (TV3): *Plats Bruts* (episodes 2 and 3) and *Jet Lag* (episodes 1 and 2). The selection of the episodes was based on the availability of the pre-production script given to the actors and the final video recording. The subcorpus includes the transcription of the final broadcast version aligned manually with audiovisual excerpts of the sitcoms, as well as extra materials such as the original scripts given to the actors, which were substantially changed during the shooting, as discussed in Matamala (forthcoming *a* and *b*).

The bilingual subcorpus includes sitcoms in English with a dubbed version in Catalan, and provides not only equivalents to be considered in the dictionary, but also a good source for translation analysis. According to the availability of the scripts at different stages of the dubbing process[2] and the recordings in both languages, three sitcoms were chosen (one episode each): the British series *Coupling,* and the American series *Working* and *Normal Ohio,* all broadcast on Catalan television. This subcorpus includes both the transcriptions of the English version and the Catalan dubbing aligned with video excerpts of the sitcoms. The text/video alignment was carried out manually, although software developed by De Yzaguirre *et al*. (2005) was used to align both written transcriptions. The elusive nature of interjections and their rare occurrence in dictionaries made it impossible to apply tagging or annotation procedures to the corpus.

2.1. Interjections: a definition

Interjections have been studied from different points of view and with different denominations: traditional grammars include interjections among word classes and highlight their phonological anomalies, syntactic independence, morphological invariance and exclamative nature (Crystal and Quirk 1964, Quirk *et al.* 1972, Leech and Svartvik 1975, Huddleston 1986, Greenbaum and Quirk 1990). Some of these units called interjections in traditional grammars are also found in studies on phraseological units (Sancho 1999, Lorente 2002), and certain monographic articles and theses have considered interjections as an important topic (James 1973, Ameka 1992, Wierzbicka 1992, Wilkins 1992). They also appear in the literature under the label "discourse markers" and are analysed from a pragmatic and discourse perspective (Schourup 1982, Schiffrin 1987, Fraser 1990 and 1999, Redeker 1990, Brinton 1996, Fischer 2000, Andersen 2001, Aijmer 2002) and even within a conversation analysis framework (Kerbrat-Orecchioni 1990, 1996, Calsamiglia and Tusón 1999). Finally, interjections are also found in articles on paralanguage, although with different denominations such as "vocalizations" (Trager 1958, Crystal and Quirk 1964) or "alternants" (Poyatos 1993). Nevertheless, the previous frameworks are not comprehensive and do not embrace all the units that we include under the same term. In fact, interjections are sometimes peripheral and do not comply with the strict conditions inherent in a traditional model of categorization which imposes necessary and sufficient conditions in order to assign a unit to a category.

2.2. Theoretical framework: moulding two perspectives

This article is based on the cognitive perspective adopted by Cuenca (1996, 2000, 2004), which encompasses a wider array of units under a single label by means of a definition

based on prototype theory (Kleiber 1990, Taylor 1995, Ungerer and Schmid 1996) and on the theory of grammaticalization (Traugott and Heine 1991, Lehmann 1995, Hopper and Traugott 2003). The former theory makes it possible to consider interjections as a peripheral class of the category "sentence", since they "correspond to communicative units (utterances) which can be syntactically autonomous and intonationally and semantically complete" (Cuenca 2000), generally expressing pragmatic values. However, unlike prototypical sentences, they do not consist of a subject plus a predicate and are highly context dependent. The latter theory, i.e. the theory of grammaticalization, is used to account for the evolution of secondary interjections through a process of subjectification (Traugott 1995).

Following Cuenca's proposal, and resulting from prototype theory (which states that boundaries are fuzzy and members of a category can share only a small number of attributes with other members of the same category), different units are included under the heading "interjections": paralinguistic features, according to certain authors (Poyatos 1993), such as "mhm" or "tut;" prototypical interjections such as "oh" or "ah" and even expressions close to phraseology such as "damn" and "holy shit".

Depending on their form and function, interjections can be classified into various subcategories. As far as the form is concerned, Cuenca adopts a classification often found in traditional grammars and distinguishes between primary interjections, i.e. simple and fixed units which can only be used interjectionally, and secondary interjections, i.e. units deriving from other grammatical classes which can be used interjectionally due to a process of grammaticalization. Regarding function, Cuenca (2004) adapts Jakobson's (1960) functions and classifies interjections into: expressive, conative, phatic, metalinguistic, and representative.

The framework chosen to support the development of the tool was also based on the concept of applicability (Lorente 1994), derived from the idea of circularity between theoretical and applied linguistics (Slama-Cazacu 1984), which envisages that different expressions of applied linguistics must ensure a certain coherence between theoretical assumptions and the design of applied tools.

3. Aims and applicability of this prototype

INTCA is a multifunctional tool addressed to language professionals whose working language is Catalan and who need a dictionary when producing texts. However, when dealing with interjections, we realized that the problem was not understanding their meaning (evidently easily inferable from the context) but finding an equivalent or their transcription. Thus, considering a future commercial development, the target group of users was widened from translators to other professionals with similar needs. In fact, when the commercial version of INTCA will be fully implemented, a translator will be able to find equivalents for a particular interjection or even interjections that express a particular feeling. Likewise, a Catalan writer will be able to find the transcription of a certain onomatopoeic interjection.

3.1. Is there really a need to create a dictionary of interjections?

Kühn (1989) identifies different reference needs regarding dictionary consultation, as summarised by Hartmann (2001: 88): for checking linguistic competence: text reception, text production, translation, technical knowledge, research, edification and language acquisition. That interjections pose problems concerning both translation and text

production is evident not only from our own experience but also from the opinions expressed by other academics (Cuenca 2004) and professionals such as the Catalan television linguists (Televisió de Catalunya 1997) or the renowned translator Xosé Castro (1999). As stated by Cuenca (2004: 3209),

> *Les imprecisions en els doblatges de pel·lícules i series en anglès són responsables de l'extensió d'aquest calcs. Així, doncs, es fa necessària una descripció acurada y completa de quines són les formes interjectives genuïnes i quins són els usos que tenen. Només d'aquesta manera es podrà aturar l'entrada d'interjeccions alienes, que es produeix amb poc control i amb escassa prevenció, ateses les característiques fonètiques i el significat pragmàtic (poc precís i determinat contextualment) de les interjeccions.*[3]

Likewise, Televisió de Catalunya (1997: 53) points out that,

> *Tot i que podríem trobar algunes coincidències esporàdiques, els recursos expressius vocals i no vocals no són idèntics en totes les llengües. Aquest fet, que pot demostrar qualsevol anàlisi contrastiva, es manifesta especialment en el món del doblatge. Quan es tradueixen els diàlegs pensats originalment en una altra llengua, cal, doncs, tenir presents els recursos propis del català.*

Finally, Castro (1999) considers that,

> *Encuentro cada vez con más frecuencia onomatopeyas mal traducidas (o sin traducir) en las traducciones de películas españolas e incluso en libros.*

Moreover, questions addressed to Internet lists are commonplace and examples of queries concerning interjections have been found in e-groups such as Traducción, Zèfir and TRAG.[4] Users ask for the translation of certain units (for example, "wow" or "tsch" or for the transcription of onomatopoeic interjections (for instance, a kiss).

These are evident needs that led us to design this dictionary of interjections. In fact, current general monolingual and bilingual dictionaries do not cover this grammatical category widely (Matamala 2001), and only a few specific printed dictionaries include this particular unit (Kloe 1976, Arana de Love 1985, Riera-Eures and Sanjaume 2002), although no English-Catalan combination is offered. Furthermore, as shown in Bach and Matamala (2004), printed dictionaries usually offer restrictive access to information, generally by lemma, which is very often what translators are looking for (i.e. translators often look for an entry word which conveys a specific value instead of looking for the meaning of a certain unit). For example, if a translator needs an equivalent for an onomatopoeic interjection which imitates the sound of an animal, he/she will have difficulty finding it in semasiological dictionaries, that is in dictionaries which offer a list of alphabetically organised entry words. On the other hand, certain synonym and thesaurus dictionaries adopt an onomasiological approach and offer alternative ways of accessing information, via synonyms or semantically related units, though interjections are usually discarded for their lack of referential meaning. Therefore, if audiovisual translators were to look for a certain interjection which expresses a particular feeling (for example, surprise), what could they resort to? The solution does not lie in current monolingual or bilingual dictionaries, as inferred from the previous statements, but in new lexicographical works

with innovative proposals for accessing information. The prototype that will be described next attempts to be one of these.

3.2. General characteristics

The general features of the project concept are presented according to a traditional lexicographical classification, i.e. superstructure, macrostructure and microstructure (Hartmann 2001: 59), the notion of "vía de acceso" ('*ways of accessing information,'* Fuentes Morán 1995), similar to Hausmann and Wiegand's (1989) "access structure" and "addressing structure."[6]

INTCA is a descriptive tool and includes both English and Catalan, although it is clearly a single directional dictionary (Kromann 1989: 273) with the direction English>Catalan as its priority. However, its open structure allows it to include not only other languages in future developments, but also prescriptive notes about the presence of certain units in the normative dictionary. Furthermore, this hypermedial tool (i.e. hypertextual tool with multimedia features) contains multiple links and no space restrictions.

Though the information used for the prototype derives from the previously described corpus, the future dictionary should have this data complemented with a real oral language corpus, or reproductions of the same, such as comic strips or theatre plays, as "there is hardly any alternative to corpora as the primary and main resource for lexicographers now" (Cermák 2003: 20). The great number of interjections found in comic strips make this kind of product an interesting candidate to be digitally included in the corpus as scripts (in GIF or JPEG formats) along with audiovisual video-clips, both aligned with their transcriptions to facilitate their textual processing. Nevertheless, we must take into account that enlarging the corpus is a time-consuming task, and the extensive corpus needed to develop a complete dictionary with a wide selection of interjections and communicative situations would involve high costs. Consequently, whereas the audiovisual corpus has proven to be sufficient for the prototype, the commercial version will probably need to (a) resort to different corpora (lexicographical, textual, audiovisual) to select the entries, (b) use this nomenclature to develop a list of hypothetical functions and communicative situations, and (c) search for those functions and situations in textual and audiovisual excerpts which document each linguistic instance and provide examples for the dictionary.

3.2.1 Superstructure

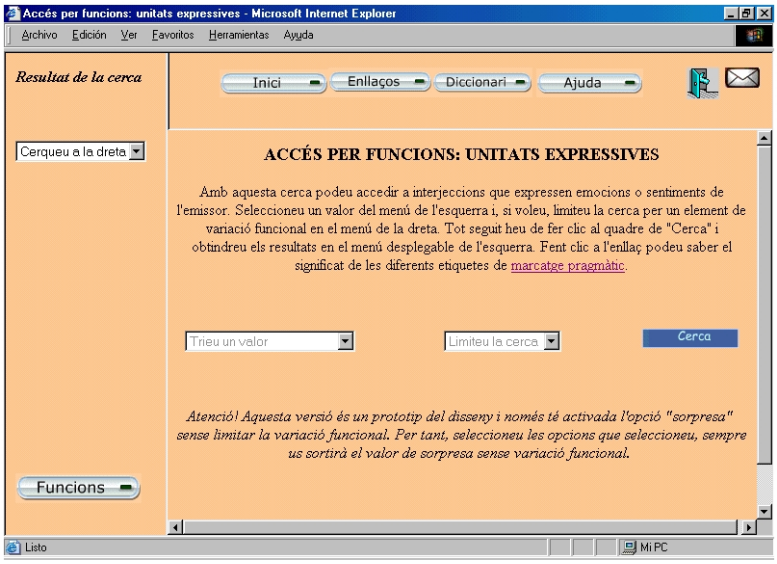
Superstructure embraces the global structure of the dictionary, and traditionally contains introductory documents, a main body and back matter. However, this view is biased by the linear structure traditionally followed by printed dictionaries and does not make sense in a prototype created as an electronic tool. Consequently, these items will be described from a hypertextual point of view, accentuating the links established between its various sections.

Having considered two languages, the first key issue to be solved was whether two parallel superstructures, both in English and Catalan, or a single superstructure was needed. According to the aim of the project – the development of a monodirectional prototype - the Catalan superstructure was favoured along with the main page links to: (a) an introduction with the objectives and sources of the dictionary aimed at intended users who are able to send suggestions through a feedback function, (b) a help page, (c) the dictionary itself, and (d) a page with links to monolingual dictionaries.

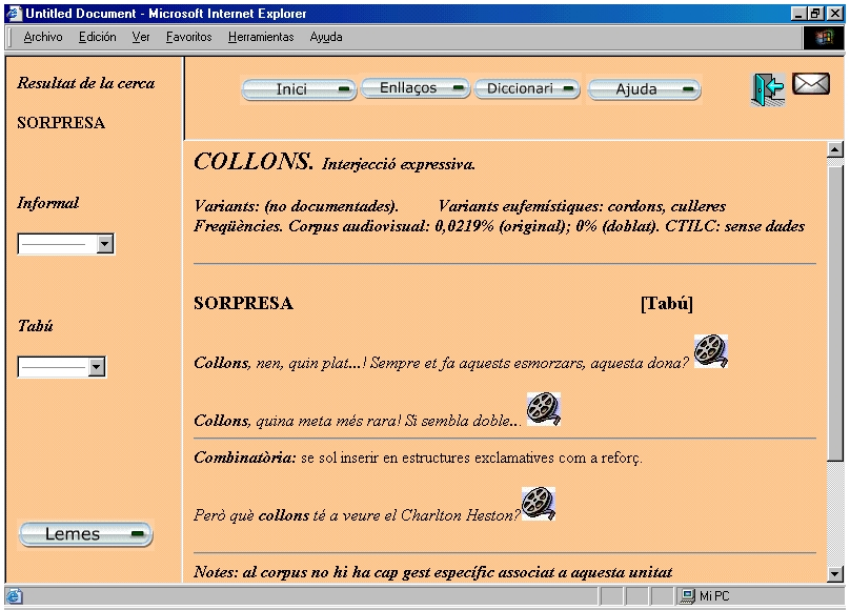
3.2.2. Macrostructure

The macrostructure refers to the way in which the entries are organised (Fuentes Morán 1995: 51) and two elements concerning macrostructures must be clearly defined: the selection, delimitation and lemmatisation of the units, and their presentation. In this prototype, the criteria used for unit selection was its inclusion in the word class "interjections", following Cuenca's proposal, although its structure can be opened up to other pragmatic and phraseological units in future versions. Four access ways are proposed for the presentation of information: (a) direct access, (b) access by English lemma, (c) access by Catalan lemma, and (d) access by function.

- *Direct access* allows users to search for a specific interjection by typing it directly into a dialogue box that pops up after clicking on the "direct access" button, and the system automatically retrieves the lexicographical entry.
- *Access by English* and *Access by Catalan lemma* involve accessing the information knowing beforehand what interjection is being looked for. The proposed interface includes a column with the list of interjections in the left frame. Once a unit is selected from the list, the microstructural information appears in the right frame. At the top of the page there is always a group of links giving access to other sections of the tool.
- *Access by function* enables users to look for those units expressing a given function, without having a certain form in mind, in a wide range of registers. First, the user selects whether he or she is looking for a unit which expresses a feeling, encourages somebody to do something, and interacts with or imitates a sound or a movement. Once a general function has been selected, the user can choose a subfunction and limit the search by marking a variation constraint (e.g., taboo, slang, etc.), as shown in Figure 1.



For instance, a user can search for a unit which expresses a feeling like "surprise", without restrictions, and the final result will be a list of proposed units in the left frame organized according to their connotations, whereas the microstructural information will come up in the right frame (Figure 2).



3.2.3. Microstructure

The microstructure includes all the items within the article, and the challenge in this case is integrating an exclusively monolingual interface (search by function or by Catalan lemma) with a complementary search starting from the English unit, resulting in two slightly different microstructures which will be described next.

3.2.3.1. Access to a Catalan unit. If the user searches for a particular unit in Catalan, either by lemma or by function, the final result includes the following microstructural information:

- *Word class and type of interjection:* the label "interjection" in a dictionary of interjections seems to make no sense. However, as space is no longer a problem in an electronic dictionary, it seems quite useful to use the label. In Figure 2 this item has been included with the label *"interjecció expressiva"* (expressive interjection) next to the entry word.
- *Formal variations of the entry word*: sometimes, interjections do not have an established form, and formal variations of the entry (such as *"cordons"* or *"culleres"* in Figure 2) are included.
- *Links:* regardless of the page consulted, there is always a series of links at the top (button *"Enllaços"* in Figure 2) which not only gives access to other sections of the tool, but also to external links, such as dictionary web pages. This allows users to consult whether a certain interjection occurs in descriptive and normative dictionaries. Moreover, an asterisk is added to those units not compiled by the normative dictionary (*Diccionari de la llengua catalana,* by the Institut d'Estudis Catalans), so that users know that the interjection is not considered correct from a prescriptive point of view although found in corpora or descriptive works and, therefore, used by speakers and writers.
- *Frequency data:* this information is essential for various reasons. First, different units might convey the same meaning as an English interjection and frequencies are crucial in selecting the most adequate one. Second, if a writer is looking for the transcription

of an onomatopoeic interjection and finds it in our dictionary, it is important to know whether it is an established form found both in dictionaries and corpora, or a rare unit with a low usage frequency. At this initial stage, the percentages based on such a limited corpus are not representative, but data from the CTILC (*Corpus Textual Informatitzat de la Llengua Catalana*), a reference corpus in Catalan,[5] have been included, when available, in order to overcome this initial setback.

- *Definition:* defining interjections by means of the traditional definition based on *genus* plus *differentiae* (Lara 1994), i.e. based on the first metalanguage (Rey-Debove 1967), is not effective because they lack referential meaning. A definition based on a second metalanguage (Rey-Debove 1967) whereby the functional features of these units are elucidated (Kipfer 1984: 103-104) is the only alternative to overcome the definitions currently found in dictionaries, which "are not of the kind that could help anyone to learn how to use them" (Wierzbicka 1992: 160). Our proposal is a special type of definition that embraces "core" and "additional" elements. The core element is a keyword expressing the value transmitted by the unit, and additional elements would be subspecifications, i.e. specific constraints or peculiarities to this general idea. For instance, the noun "*porta"* (literally "door") can be used interjectionally to make someone go. In this case, the core meaning would be to "incite somebody to an action", and the additional element would be "to go out".
- *Variation elements:* these are what are traditionally known as pragmatic labels and following the fourth edition of the *Oxford Advanced Learner's Dictionary* (1989), a five-point scale of rethoric, formal, informal, slang, and taboo label for the units is proposed (see the label *"tabú"* in Figure 2).
- *Examples and pronunciation:* current examples derive from the corpus presented in section 2. As shown in figures 2 and 3, the examples have the icon of a reel next to the transcription, a link to an audiovisual clip from the corpus.
- *Collocations:* although interjections are independent units equivalent to a sentence, they tend to appear next to certain units in semi-fixed collocations, so this information is also provided by the interface. For example, it is quite common to find the interjection *"d'acord"* ("ok", "alright") preceded by the interjection *"ah,"* and this is clearly stated in the prototype, as well as the usual position, lexical combinations, and syntactic relationships of interjections (Matamala and Lorente, forthcoming).
- *Remarks:* this field is used to integrate complementary items which might be relevant, such as the body language related to a certain interjection.

3.2.3.2. Access to a Catalan unit through an English interjection. When searching for a unit through an English interjection, information is presented slightly differently. The user selects an interjection from different folding menus (see Figure 3), so that a lexicographical article appears on the right of the screen. This article provides the same microstructural information plus a list of possible translations with pragmatic labels (*"propostes de traducció"* in Figure 3) and clicking on these equivalents (*mira, escolta, aviam, a veure, daixò, escolti* and *miri* are proposed equivalents for the interjection *'look'* in Figure 3) a lexicographical article in Catalan appears and the same microstructural information is found.

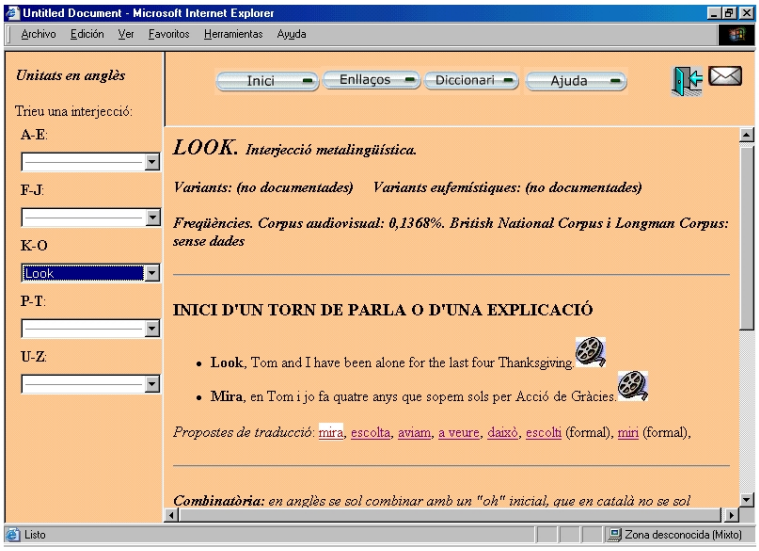


Figure 3. English lexicographical article.

4. Conclusions

To sum up, interjections (Cuenca 1996, 2000, 2004) are key items in audiovisual products such as films, cartoons and TV series which often try to imitate spontaneous colloquial oral language. However, dictionaries do not always fulfil the needs of language professionals working with these types of texts. Thus, the creation of an electronic dictionary of interjections from English into Catalan, covering a wide range of information seemed a useful and practical idea. As the initial data come from a collection of aligned sitcoms which include video files, this corpus might serve not only to develop a lexicographical prototype, but also to carry out systematic studies on screen translation. The cognitive approach by Cuenca has shown to be an adequate framework for explaining the categories and functions of interjections, since it encompasses a wide variety of units under a single label and accepts fuzzy boundaries in categorizations. The theoretical idea behind this dictionary is that linguistic tools should address the needs of the user and the development of friendly interfaces that both represent and give access to relevant information, making the most of hypertext and audiovisual possibilities.

What has been described is a prototype. Prototypes are characterized by their ability to offer a whole range of innovative features which prove their eventual feasibility on the marketplace. Nonetheless, in order to turn INTCA into a commercial dictionary, two basic preparatory actions are required: (a) the information available for each interjection should be evaluated, giving priority to the most relevant data for translation, and (b) the hypertextual navigational systems should also be evaluated by professional translators in real or simulated working situations in order to improve information access, taking into account the most common search strategies used.

To conclude, it should be pointed out that this is just a hint of what we hope will be an important field of research in Translation Studies, a line which develops, or at least proposes, tools which translators will be able to use to their advantage.

Acknowledgments

We would like to express our gratitude to Patrick Zabalbeascoa and Marcela Rivadeneira who kindly revised the article. Needless to say, any mistakes should be attributed solely to the authors.

Notes

1. This essay is based on Matamala's doctoral thesis, supervised by Mercè Lorente and developed within the PhD in Applied Linguistics at the Universitat Pompeu Fabra (Barcelona), published online in 2005.
2. The scripts available correspond to three key stages in the dubbing process: non-synchronised translation, synchronised translation, and the linguistically revised final version.
3. We provide our own translation for this and the following quotations.
Cuenca 2004: "Inaccuracies in the dubbing of English films and series are responsible for the spread of loan translations. Therefore, an accurate and complete description of genuine interjective forms and their uses is needed. This is the only way to avoid the introduction of foreign interjections, which is taking place with little control and insufficient prevention, due to the phonetic features and pragmatic meaning (rather vague and context-bound) of interjections."
TVC 1997: "Despite sporadic coincidences, expressive resources – both vocal and non vocal – are not the same in all languages. This is provable by any contrastive analysis, and is especially obvious in dubbing. When translating dialogues originally conceived in another language, translators should bear in mind expressions specific to Catalan."
Castro 1999: "I find more and more badly translated (or even non-translated) onomatopoeias in films translated into Spanish or even in books."
4. *Traducción* is a Spanish list devoted to translation in general (http://www.rediris.es/list/info/traduccion.es.html); *Zèfir* is a list addressed to Catalan language professionals (http://www.llengcat.com/zefir/), and TRAG is the only list for audiovisual translators in Spain (http://xcastro.com/trag/)
5. This corpus can be found at http://pdl.iec.es.
6. The previously described corpus was used to obtain information for the prototype and both the corpus and the prototype are accessible from a single interface in a DVD which is included in the thesis. Hopefully, they will be available on the Internet in the future.