The TRANSVERB project - An electronic bilingual dictionary for translators: theoretical background and practical perspectives

Beatriz Sánchez Cárdenas and Amalia Todirascu
University of Strasbourg, France

TRANSVERB is a lexicographic resource conceived for novice and professional translators who need assistance when translating texts into a foreign language. It is a semi-bilingual dictionary which can also be used for text production into the users' mother tongue. The case study analyzed in this article pertains to the translation of verbs from French to Spanish. This dictionary is organized onomasiologically in terms of categories. Based on the hypothesis that human cognition organizes concepts in semantic categories (Tranel et al. 2001; Damasio et al. 2004), TRANSVERB is configured in lexical domains (Martín Mingorance 1985, 1987, 1900, 1995; Faber & Mairal 1999). The syntactic information in verb entries includes its combinatory potential, more specifically, its number of arguments as well as their semantic restrictions. This is established through corpus study.

1. Introduction

The project of creating a new dictionary initially seems something of a utopia since any lexicographic endeavour is extremely complex and costly both in terms of economic as well as human resources:

> 'Writing a dictionary is a salutary and humbling experience. It makes you very aware of the extent of your ignorance in almost every field of human activity. It fills your working day with a series of monotonous, humdrum, fascinating, exasperating, frustrating, rewarding and impossible tasks. It goes on for years and years longer than you ever thought it (or you) could. And when it is all over, the fruits of this labour are enshrined forever in a form which allows other people to take it (and you) apart, in print, publicly and permanently' (Atkins 1991: 167).

Nevertheless, despite such difficulties, it is our belief that lexicography is one the most interesting linguistic activities because of its intellectual stimulation and the usefulness of its results.

According to Bergenholtz and Tarp (2003, 2004) and Tarp (2005), dictionaries must be conceived for a specific type of user and user needs taken into account. TRANSVERB is a bilingual dictionary intended for translation and for text production. It is envisaged as semi-bilingual since it does not offer word translation, but rather conceptual linguistic structures translations. As Atkins (2002: 10) suggests, it gives 'its users the opportunity to make their own decisions about equivalence'. Since translators work with electronic resources or tools, TRANSVERB is conceived to be an electronic on-line dictionary.

In this paper, we present here the prototype of this new bilingual dictionary which offers a solution to two of the most frequent problems users have: (1) choosing the correct lexical unit when translating according to the context and (2) using a word in the correct context.

2. TRANSVERB: content

2.1. Lexical domains in TRANSVERB

At a macrostructural level, linguistic content is organized in lexical domains. The lexical domains in TRANSVERB are based on the Functional Lexematic Model (FLM) (1984, 1985, 1987, 1900, 1995), which is based on Dik's (1978) *Functional Grammar* and Coseriu's (1981) *Lexematic Theory*. The FLM postulates that the representation of knowledge in the human mind is linguistically motivated. According to Croft (1993: 337), 'there is no essential

difference between (linguistic) semantic representation and (general) knowledge representation'. In this sense, the FLM claims that lexical units (e.g. verbs) are organised in conceptual categories[1]. Hierarchies play an important role in organising conceptual categories. Many authors such as Lyons (1977), Apresjan (1993) or Nyckees (1998) emphasize that semantically full words (such as verbs and nouns) are stored hierarchically in the human mind. This hypothesis has been corroborated by neurolinguistic studies (Tranel et al. 2001; Damasio et al. 2004).

Thus, a lexical domain can be defined as a hierarchically organised semantic group of lexical units sharing a paradigm. At the uppermost level of this hierarchy is the genus or superordinate term, which in the realm of verbs, often corresponds to one of Wierzbizcka's (1995, 1996) semantic primitives or Apresjan's (1993) near primitives.. The verbs at the lower levels of the hierarchy inherit the semantic and syntactic properties of the superordinate.

Lexical domains can be envisaged as a good way of organising the macrostructure of a dictionary. In this way, users have instant access to the whole conceptual system to which a word belongs. This is especially interesting for bilingual lexicography where lexical domains are presented simultaneously in two different languages, as proposed by Faber and Pérez (1997). Since this representation corresponds to the way concepts are stored in the human mind (Tranel et al. 2001; Damasio et al. 2004), users will find this conceptual representation of the lexicon highly useful.

2.2.1. Building a lexical domain

The establishment of an inventory of lexical domains in the FLM is based on the lexicographic information in dictionaries entries, which are factorized according to Dik's (1978) principle of Stepwise Lexical Decomposition. Nevertheless, in our experience, dictionaries entries are always imperfect to some extent, and consequently, they are not sufficient in themselves to ensure the internal configuration of lexical domains. For that reason and in order to increase the reliability of the process, we have added the following two criteria:

1. The FLM claims that the more general verbs situated at the higher levels of the hierarchy have fewer selection restrictions. Gross (1992, 1996) also supports this idea. Accordingly, we studied the semantic restrictions on the arguments of each verb as represented in corpus data. We chose to study contemporary formal written French. Thus, corpora[2] were built from the following on-line databases which reflect this variety: (i) Frantext (novels, essays and novels from 1900 to 2000); (ii) Wortschatz[3] from the Leipzig University; (iii) Corpuseye[4], which includes texts from the European Parliament.

2. The property of troponomy is used in WordNet to represent hyponymy in verbs. Verb entailment is determined by means of the linguistic tests proposed by Fellbaum and Miller (Fellbaum 1990; Miller 1992; Miller et Fellbaum 1991).

[1] This idea is defended by many other authors, such as Lyons 1977; Martín Mingorance 1984; Langacker 1987; Apresjan 1993; Croft 1993; Waxman 1994; Sager 1990; Cabré 1998; Bressé 2000.

[2] Corpora were studied using the semi-automatic analyzer WordSmith tools.

[3] http://corpora.informatik.uni-leipzig.de/?dict=fr site.

[4] http://corp.hum.sdu.dk.

The nine lexical domains obtained corresponded to the inventory in the FLM: EXISTENCE, CHANGE, POSSESION SPEECH, EMOTION, ACTION, COGNITION, MOVEMENT, PHYSICAL PERCEPTION AND MANIPULATION (Mairal & Faber 2007: 7). Each of these domains is subdivided in several lexical sub-domains. For example the domain of PERCEPTION includes the subdomain of physical perception (e.g. *perceive, note, see, observe*). An example of the lexical domain of COGNITION and the lexical subdomain of 'counting' in Spanish and French is shown below:

LEXICAL DOMAIN**: COGNITION**	
Subdomain: **Counting**	
Spanish	**French**
Realizar una operación cognitiva con el objetivo de establecer cuántos elementos componen un grupo **1.** ***contar:*** determinar la cantidad de elementos de un conjunto: 'He contado 150 invitados'. **1.1** ***censar:*** contar los elementos de una población inscribiendo el resultado en el censo: 'En los humedales de León se han censado más de 9.000 aves acuáticas'. **1.2** ***computar***: contar el tiempo según magnitudes numéricas: 'Hemos computado las cifras de participación'. **1.3** ***contabilizar:*** contar aplicando cálculos: 'Sanidad ha contabilizado 2.195 manifestaciones anticipadas'. **2.** ***enumerar:*** nombrar los elementos de un conjunto: 'Ha enumerado el conjunto de medidas aprobadas para mejorar la protección de los trabajadores'.	*Effectuer une opération cognitive visant à déterminer le nombre total d'éléments d'un ensemble* ***compter1:*** déterminer la quantité d'éléments d'un ensemble: 'La police a compté 150 manifestants'. ***dénombrer1:*** compter un à un. Le résultat a un caractère officiel: 'Les enquêteurs ont dénombré une douzaine d'étuis de balles'. ***recenser1:*** dénombrer, en général d'une population, en identifiant qualitativement chaque élément en identifiant chaque élément. Le résultat a un caractère officiel: 'La Fondation Abbé-Pierre recense 200.000 personnes hébergées durablement en hôtel' ***comptabiliser1:*** compter un à un en utilisant des techniques comptables. Le résultat a un caractère officiel: 'Ce logiciel comptabilise plus de 54.000 téléchargements à ce jour'. ***énumérer1***: déterminer la quantité d'éléments d'un ensemble et les identifiant l'un après l'autre: 'Le rapport énumère les cinq critères du transport du bétail'.

Table 1. Representation of 'COGNITION-counting' in TRANSVERB

2.2.2. Advangages of lexical domains for bilingual lexicography

Regarding bilingual verb entries, one of the advantages of representing verbs related to its lexical domains in both A and B language is that it allows quick and easy access to the conceptual schema of the language, which is helpful for translators, who deal with conceptual transfers between two languages.

compter1	*COGNITION*	J'ai compté 7 enfants. J'ai compté les enfants.
compter2	*EXISTENCE*	La Crète compte 25 millions d'oliviers.
compter3	*SPEECH*	Marie compte: 1, 2, 3, 4…

Table 2. Domain membership of *compter*

Another benefit of this kind of macrostructural lexicographic conception is that when verbs are represented within their lexical domain, this makes polysemy disambiguation much simpler. For example, through dictionaries entries and corpora analysis, we have isolated three different uses of the verb *compter* as shown in table 2 (Sánchez Cárdenas, in press):

Although these three verbs are all counting verbs, they belong to subdomains within different lexical domains, as can be observed in their lexical decomposition. For example, *compter1* is related to COGNITION (this structure implies an agent who gets to a numeric result thanks to

his/her cognitive capacities); *compter2* belongs to the domain of EXISTENCE (since it does not imply any activity and describes the localisation of a number of elements in a specific place) and *compter3* is classified in the domain of SPEECH (the verb describes the verbalisation of the cardinal numbers). Users can select the sense of the verb they looking for, depending on text needs.

3. Verb entries

Macrorole	Thematic role	Type of syntagma	Gram. function	Syntactic structure	Semantic Class	Example	Example in context
ACTOR	**Cognizer:** **Cognitive entity who accomplishes an action.**	**NP (SN1)**	**Sub-ject**	**SN1** V SN2	**N person**	***Isabelle, M. Foulon***	***M. Foulon énumère les quatre facteurs de sa réussite.***
					N institutional	***Observatoire, Andra, Commission***	***La commission énumère 25 produits phytosanitaires dont l'utilisation est admise dans l'UE.***
					N profession	***Magistrat, président***	***Le Président énumère les principales décisions adoptées.***
	False Cognizer				**N speech**	***Rapport, quotidien, loi***	***Le rapport énumère de nombreuses violations flagrantes et massives des droits de l'homme.***
UNDER-GOER	**Result: Result of the ACTOR'S action**	**NP**	**Direct Ob-ject**	SN1 V **SN2**	**N intellectual construction**	***faits, causes de la récession, points importants, propositions, raisons***	***M. Bayrou énumère les trois circonstances de la défaite du parti.***
					N event	***réformes, mesures, modifications des programmes***	***Le secrétaire général a énuméré les cinq mesures déjà prises.***
					N artifact	***transports, produits phytosanitaires***	***La commission énumère dans son avis 25 produits phytosanitaires dont l'utilisation est admise dans l'UE.***
					N speech	***mots, chapitres, liste, questions***	***Le dictionnaire de Jacques Attali énumère plus de 400 mots clés du futur.***
					N category	***noms, catégories, variétés***	***Le procureur du roi avait énuméré 26 variétés de poires.***
					N inanimate natural object	**matériaux**	***Le texte énumère les matériaux précieux ou semi-précieux utilisés pour la décoration du bâtiment.***

Table 3. Entry for *énumerer*

TRANSVERB'S entries are mainly based on Faber and Mairal (1999) and Atkins (2002). Verb entries thus contain information pertaining to the valence of the verb as well as information about frame elements, grammatical function, phrase type, and sortal features of each verb argument. Since no complete inventory of frame elements exists, we have preferred to describe the semantic nature of each argument in terms of macroroles and thematic roles, following Role and Reference Grammar (Van Valin 1993, 2000, 2001, 2004, 2005; Van Valin and LaPolla 1997). To this information about the syntactic structure of the sentence, we have added the semantic class of the nouns in the arguments as well as examples of the type of nouns on each argument and examples in context.

The nouns of the arguments have been assigned to semantic classes such as ‘N human’, ‘N institutional’. Following Hanks (2000) and Béjoint (2007), we envisage frequency of use in corpora as relevant data that should be included in dictionaries, and in verb entries. As an example, the entry for the verb ‘*énumérer*’ is shown in Table 3.

This kind of information is very useful when users must produce a text in a language, whether it is their own language or a foreign language. The dictionary is implemented as a website. We use a database to represent lexical units, lexical domains, and the properties describing each paradigm. The database is structured in the following way.

One table was defined to represent data for each language. For each lexical unit, the table provides information about domains and subdomains, about the macrorole and thematic role of arguments, and about the syntactic context. It also provides examples. Furthermore, the database can be easily expanded to include other languages. The database structure has been defined and data is available. We use a MySql server to store the database, and are in the process of developing the various interfaces for querying and visualisation.

4. Web site content

Since TRANSVERB will eventually host a multiplicity of languages combinations, the first thing users need to do is to define the ‘language A’ – or source language – and the ‘language B’ – or target language. Once this is done, users can consult the dictionary.

To access the linguistic information, users need to decide whether to search for a word or for a lexical domain. The first of these two options does not need any further explanation, since most electronic dictionaries follow this trend. All users have to do is to type the word they are looking for, and this will take them to a screen in which they will find the linguistic information they are looking for. Of course, the process is not entirely that simple since there are other things that must be considered. In TRANSVERB, this process first gives users access to the different lexical domains a given verb can belong to.

4.1. Searching by lexical domains

When searching by lexical domain in TRANSVERB, the home page displays a pop-up menu with all the different domains and subdomains available (for example PERCEPTION-physical perception or COGNITION-counting). The interface then searches the database for available domains and subdomains, and builds the choice list. Users then choose a domain from the menu; the database is queried for all verbs from the selected domain; and the results are displayed on page showing the bilingual lexical subdomain.

Each verb of the lexical domain is presented as a hyperlink leading to the verb entry. The idea behind this conceptual structure is that translators can use bilingual lexical domains to find the most suitable term available in the target language by viewing approximate correspondences between the members of entire lexical domains.

4.2. Searching by lexical unit

Users can also perform a more conventional word search by entering the verb in the search box. The database is queried to search the domains of the word. If the verb only belongs to one domain, users will go directly to the screen with the contrastive lexical domain representation, shown in Figure 1. When verbs belong to more than one domain, users must decide which domain they are interested in. For example, when in the case of *compter*, the user must choose one of the following three options:

Compter1: COGNITION-counting: 'J'ai compté 10 enfants dans la salle'.
Compter2: EXISTENCE-counting. 'Cette université compte de célèbres chercheurs'.
Compter3: SPEECH-counting. 'Je compte: 1, 2, 3…'.

Both the verb and the DOMAIN-subdomain are hyperlinks. The domain hyperlink connects to the representation of the lexical domain, whereas the verb hyperlink takes users the lexical entry.

5. Conclusion and Future Work

From a macrostructural point of view, TRANSVERB is organized in lexical domains, which gives the user easly acces to concept structure. At the microstructural level, this dictionary gives information about the syntactic structure of the arguments of the verb and their semantic classes. Current research provides a remarkable theoretical framework and methodologies that can be used to create innovative lexicographic tools. It is a matter of taking full advantage of them.

Bibliography

Apresjan, J. (1993). 'Systemic lexicography as a basis of dictionary-making'. In *Journal of the Dictionary Society of North America* 14. 79-87.

Atkins, B.T.S. (1991). 'Building a Lexicon. The Contribution of Lexicography'. In *International Journal of Lexicography* 4 (3). 167-204.

Atkins, S. (2002). 'Bilingual Dictionaries, Past, Present and Future'. In *Proceedings from the Euralex Conference*. UK. 1-29.

Béjoint, H. (2007). 'Informatique et lexicographie de corpus: les nouveaux dictionnaires'. In *Revue française de linguistique appliquée*. 12 (1). 7-23.

Bergenholtz, H.; Tarp, S. (2004). 'The concept of dictionary usage'. In *Nordic Journal of English Studies* 3 (1). 23-36.

Bergenholtz, H.; Tarp, S. (2003). 'Two opposting theories: On H.E Wiegand's recent discovery of lexicographic functions'. In *Hermes: Journal of Linguistics* 31. 171-196.

Coseriu, E. (1981). *Lecciones de lingüística general*. Madrid: Gredos.

Croft, W. (1993). 'The role of domains in the interpretation of metaphors and metonymies'. In *Cognitive Linguistics* 4 (4). 335-370.

Dik, S. (1978). *Functional Grammar*. Dordrecht: Foris Publications.

Faber, P.; Pérez, Ch. (1997). 'Image Schemata and Light: a study in contrastive lexical domains in English and Spanish'. In *Folia Linguistica* 36. 63-107.

Faber, P.; Mairal, R. (1999). *Constructing a Lexicon of English Verbs*. Berlin: Mouton de Gruyter.

Fellbaum, C. (1990). 'English verbs as a semantic net'. In *International Journal of Lexicography* 3 (4). 278-301.

Gross, G. (1992). 'Forme d'un dictionnaire électronique'. In Clas, A.; Safar, H. (dir.). *L'environnement traductionnel, la station de travail du traducteur de l'an 2001*. Sillery: Presses de l'Université du Québec et AUPLEF-UREF. 255-271.

Gross, G. (1996). *Les expressions figées en français: noms composés et autres locutions*. Paris, Ophrys.

Hanks, P. (2000). 'Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance'. In *Euralex Proceedings* 1. 3-13.

Lyons, J. ([1977] 1990). *Sémantique linguistique*, Paris: Larousse.

Mairal, R.; Faber, P. (2007). 'Lexical templates within a functional cognitive theory of meaning'. In *Annual Review of Cognitive Linguistics* 5. 137-172.

Martín Mingorance, L. (1987). *Classematics in a Functional-Lexematic grammar of English, Actas del X Congreso de AEDEAN*. Zaragoza.

Martín Mingorance, L. (1990). 'Functional Grammar and Lexematics'. In Tomaszczyk, J.; Lewandowska-Tomaszczyk, B. (eds.). *Meaning and Lexicography*. Amsterdam: Benjamins. 227-253.

Martín Mingorance, L. (1995). 'Lexical logic and structural semantics: methodological underpinnings in the structuring of a lexical database for natural language processing'. In Hoinkes, U. (ed.). *Panorama der Lexikalischen Semantik*. Tübingen: Gunter Narr. 461-474.

Martín Mingorance, L. (1998 [1985]). 'Bases metodológicas para un estudio contrastivo del léxico derivado'. In Martín Rubiales, A. (ed.). *El Modelo Lexemático Funcional*. Granada: Editorial Universidad de Granada. 61-82.

Miller, G. A.; Fellbaum, Ch. (1991). 'Semantic networks of English'. In *Cognition* 41. 197-229.

Miller, G. A. (1992). 'WordNet and the Organization of Lexical Memory'. In *Intelligent tutoring systems for foreign language learning; the bridge to international communication*. Berlin/New York: Springer-Verlag. 89-102.

Nyckees, V. (1998). *La sémantique*. Paris: Belin.

Sánchez Cárdenas, B. (pending publication). 'Lorsque compter compte: vers une représentation lexicographique'. In Dchicha, S.; Grass, T.; Wallaert, I. (eds.). *Outils de traduction, outils du traducteur? - Translation tools, Tools for the Translator?* Strasbourg: Université de Strasbourg.

Tarp, S. (2005). 'The pedagogical dimension of the well-conceived specialised dictionary'. In *Iberica* 7. 7-21.