Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup

Tiago T. Torrent, Maria Margarida M. Salomão, Fernanda C. A. Campos, Regina M. M. Braga, Ely E. S. Matos, Maucha A. Gamonal, Julia A. Gonçalves, Bruno C. P. Souza, Daniela S. Gomes and Simone R. Peron

Federal University of Juiz de Fora – FrameNet Brasil

tiago.torrent@ufjf.edu.br; mm.salomao@uol.com.br
{fernanda.campos|regina.braga|ely.matos}@ufjf.edu.br;
mauchaandrade@gmail.com; julia.goncalves@ifsudeste.edu.br;
brunopereiradesouza@yahoo.com.br;danielasimoesgomes@gmail.com;
speronjf@yahoo.com.br

Abstract

This paper presents the Copa 2014 FrameNet Brasil software (C-14/FN-Br): a frame-based trilingual electronic dictionary covering the domains of Football, Tourism and the World Cup. The dictionary relies on the infrastructure of FrameNet and is meant to be used by tourists, journalists and the staff involved in receiving foreign visitors. Vocabulary from the three domains is made available in English, Spanish and Brazilian Portuguese. Every lexical unit in the dictionary is described against an interlingual background frame.

1 Introduction

The idea of building a frame-based electronic dictionary is not new. Fillmore and Atkins (1992) paper on the semantics of Risk and its neighbors already propose the general guidelines for a lexical resource in which frames would play the key role of organizing lexical units (LUs) – the pairing of a lemma and a frame – and the relations among them. Five years after the Risk paper, Berkeley FrameNet was launched not only to materialize those general guidelines, but also to go beyond them (Fillmore et al. 2003a; 2003b). Currently, Berkeley FrameNet has described 1,164 frames and 12,713 LUs, which, in turn, are attested in 195,590 annotated sentences[1]. Additionally, there are framenets being developed for several languages other than English, such as Chinese (You and Liu, 2005), German (Boas, 2002), Japanese (Ohara et al., 2004), Spanish (Subirats and Petruck, 2003), Swedish (Borin et al., 2010), and Brazilian Portuguese (Salomão, 2009), among others.

However, although FrameNet and its sister projects have made considerable impact in Computational Linguistics, the frame-based approach for creating lexical resources envisioned by Fillmore and his collaborators has not been fully exploited in creating commercial electronic dictionaries, i.e. pieces of software focusing not on an expert user, but in the non-linguist. The Copa 2014 FrameNet Brasil Project (C-14/FN-Br) aims to fill in this blank.

Covering the domains of Football – whose frames were defined by the Semantec group at UNISINOS (Chishman et al., 2013) –, Tourism and the World Cup itself – both modeled by the FrameNet Brasil group at the Federal University of Juiz de Fora –, the dictionary is meant to be used by tourists, journalists and the staff involved both in the organization of the event and in receiving foreign visitors. Vocabulary from the three domains is made available in English, Spanish and Brazilian Portuguese. Every LU in the dictionary is described against a background frame, which is the same for the three languages. Hence, relying on the infrastructure of framenet, C-14/FN-Br serves as a proof of concept that frames – at least the crosscultural ones – can be used as an interlingua.

2 The Database

The C-14/FN-Br database is fully multilingual, in the sense that, besides comprising LUs from the

[1]Data retrieved from http://framenet.icsi.berkeley.edu on December 17th 2013.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 10–14, Dublin, Ireland, August 23-29 2014.

three languages, the frames and Frame Elements (FEs) – the participants and props in a frame – modeled are represented in the three languages as well. This procedure aims to provide the software with a trilingual interface, meaning that the user may choose in which language s/he wants to interact with the software.

To populate the database, C-14/FN-Br uses the same software developed for Berkeley FrameNet, the FrameNet Desktop, with some minor changes, the most important of which being the Translation Relation, the Language Index, the Grammatical Function and Phrase Type Correspondence Tables, and the LU Link.

2.1 Translation Relation

The original FrameNet database works with eight different frame-to-frame relations: Inheritance, Using, Subframe, Perspective on, Precedes, Causative of, Incohative of and See also (Fillmore et al. 2003b; Ruppenhofer et al. 2010). Those relations structure the network of frames, defining which frame is a type – Inheritance – or a part – Subframe – of another, or even if it comes before another frame – Precedes.

For the development of C-14/FN-Br, an additional relation was created: Translation. Since it is a symmetric relation between representations of the same conceptual structure, the Translation relation requires all FE in one representation to be mapped to FEs in the other.

With this relation, the software is able to link the set of LUs evoking a frame in a given language to the set of LUs evoking the same frame in another language. Also, the Translation relation provides the dictionary user with the possibility of accessing the frame description in all the three interface languages, which may or may not be the same as the language of the word being looked up.

2.2 Language Index

A framenet-like database is composed of both a set of frames, FEs and the relations among them, and a set of annotations. In the annotation sets, sentences are labeled not only in regards to the FEs occurring in them, but also in regards to the Grammatical Functions (GFs) and Phrase Types (PTs), i.e. the linguistic material, in which they are instantiated.

Each language requires its own set of GFs and PTs, which were defined by the correspondent framenet projects (see Ruppenhofer et al., 2010; Spanish FrameNet, 2013; Torrent and Ellsworth, 2013). In C-14/FN-Br, GF and PT labels for English, Spanish and Brazilian Portuguese are included in the database and made available for the lexicographer when annotating a sentence. A Language Index (*en*, for English; *es*, for Spanish, and *br*, for Brazilian Portuguese) was created and ascribed to the layers containing those labels and also to the frames. During the process of importing aanotation corpora into the Desktop, the Language Index in the annotation set is unified with that of the layers, thus providing the relevant GF and PT labels for each sentence.

2.3 Grammatical Function and Phrase Type Correspondence Tables

Another change made to the database was the creation of a relational table in which GF and PT labels in the three languages are associated to each other. The purpose of this table is to provide the software with a base for comparing valence patterns of the lexical units in the dictionary. Through such a comparison, C-14/FN-Br is able to suggest the best translation equivalents for a given word in the other two languages covered by the dictionary (see Torrent et al., forthcoming).

2.4 LU Link

The last change incorporated to the C-14/FN-Br database is the LU Link. Through this relation, LUs evoking the same frame in different languages can be linked to each other as translation equivalents. The process of assigning those equivalences is based on the analysis of the lexicographer. The LU Link tool is specially important for defining equivalence relations between nouns evoking entities, such as city and stadiums names, since their valence patterns are not informative enough for the automatic suggestion of translation equivalents.

3 The Application

C-14/FN-Br was developed as a web app and can be accessed at http://dicionariodacopa.com.br. On the client side, the web interface is built using HTML5 and Javascript. Data is stored in a MySQL database and is delivered to clients through a PHP5 web service. Data transfer uses AJAX techniques.

Every piece of information in the multilingual framenet-like database of C-14/FN-Br is, to some extent, used for presenting information to the final user. Thus, besides being able to search for LUs and their equivalents in the other languages, users can also see the frame and the FEs evoked by the lexical item, as well as selected annotated sentences, images and videos, which illustrate the meaning evoked by the word.

Since the dictionary is aimed at a non-specialist audience, adaptations in the Frame Semantics terminology were made. Frames are called scenes (or *esceneas* and *cenas*, in Spanish and Portuguese) and Frame Elements are called participants (or *participantes*). Frame-to-frame relations also had their names changed into more transparent concepts. For instance, instead of presenting the names Inheritance and Subframe, the software shows these relations as nominal predicates, respectivelly, "is a type of" and "is a part of".

3.1 Query Modes

Users may access the information in the dictionary in four different ways: by searching a word, by typing a sentence, by browsing the list of frames grouped by cognitive domains, and by exploring the frame grapher.

The first mode is similar to the one found in the majority of electronic dictionaries (Pastor and Alcina, 2010): users are presented a list of words in alphabetical order and a search box. Icons with the Brazilian, British and Spanish flags provide the possibility of choosing the language to which the words being looked up belong.

When a given lemma is paired with more than one frame in the dictionary, that is, it is polysemous, C-14/FN-Br shows a disambiguation screen to the user, in which s/he can choose in which scene s/he wants to explore the meaning of the word. Alternatively, if the user types in a sentence containing the LU s/he wants to look up, the dictionary may infer the most appropriate scene, using the frame disambiguation tool.

3.2 Frame Disambiguation

The frame disambiguation tool chooses, among the frames evoked by the various LUs of a lexeme, the best fit for a given context. First, a text-matching script identifies lemmas in the sentence the user types in. Second, all frames evoked by the LUs grouped under those lemmas are retrieved and each possible combination of frames forms a cluster. Since it is very unlikely that two or more LUs evoke the same frame in one sentence, and since frames in FrameNet are related to each other via frame-to-frame relations, cluster formation is extended so as to include related frames. The similarity/proximity between frames is assessed within each cluster, based on the weights ascribed to each frame-to-frame relation. The cluster in which relations between frames are stronger receive higher grades, enabling C-14/FN-br to present to the user the most probable meaning of a polysemous word given the sentential context provided in the query. Currently, the frame disambiguation tool works only for verbal lexical units.

3.3 Frame Grapher

The two other query modes focus on frames, not on LUs. Users can either browse a list of the frames defined for each domain, accessing their definitions and FEs, or explore the frame grapher, a dynamic graph in which frames are linked to each other via frame-to-frame relations. The purpose of the grapher is to show to the user, in an interactive tool, how the domains of the World Cup, Football and Tourism are organized. Through this tool it is possible not only to access the vocabulary evoking a specific stage of the World Cup, such as the Playoffs, for example, but also to see that it is a part of the World Cup that precedes the Final Match and comes after the Group Stage.

3.4 Multimedia Features and External Links

As an additional resource for understanding the frame evoked by each lexical unit, the dictionary provides the users with pictures and annotated pieces of video illustrating some of the frames. Currently, video annotation is made using a free web tool from Mozilla Corporation: PopcornMaker (https://popcorn.webmaker.org).

Also, users may follow a link to the websites from which the example sentences for each LU were collected, thus being able to read travel blogs, news and other pieces of text related to the domains coverd by the dictionary.

4 Preliminary Analytics

The C-14/FN-Br web app was launched on June 4th 2014, one week prior to the World Cup. During the first month, from 6/4/2014 to 7/4/2014, the app was accessed 1,883 times by 1,367 different users, with a total of 12,322 pages visited, i.e. an average of 6.54 pages per session. Bounce rate during the first month was of 22.68%, and the average session duration was 3 minutes and 52 seconds.

The web app was accessed from 59 different coutries. Table 1 lists the top ten countries in number of sessions. For each country, the percentage of new sessions, the bounce rate, the average number of pages visited and the average session duration is also presented.

Country	Sessions	New users	Bounce rate	Pages per session	Session duration
Brazil	1,352(71.80%)	924(68.39%)	22.26%	6.99	00:04:21
United States	75(3.98%)	57(4.22%)	29.33%	4.49	00:02:23
Spain	71(3,77%)	65(4.81%)	14.08%	6.11	00:02:39
Colombia	31(1.65%)	16(1.18%)	41.94%	3.32	00:02:13
Peru	31(1.65%)	27(2.00%)	12.90%	6.45	00:03:47
United Kingdom	30(1.59%)	18(1.33%)	36.67%	6.17	00:02:49
Venezuela	27(1.43%)	24(1.78%)	29.63%	4.89	00:03:05
Germany	24(1.27%)	22(1.63%)	16.67%	4.62	00:01:23
France	19(1.01%)	18(1.33%)	10.53%	5.89	00:02:04
Canada	18(0.96%)	10(0.74%)	11.11%	6.56	00:02:09
Other	205(10.88%)	170(12.58%)	---	---	---
Total/Average	**1,883(100%)**	**1,351(100%)**	**22.68%**	**6.54**	**00:03:52**

Table 1. Sessions per country.

Although 71.80% of the sessions originated from Brazil, only 65.00% (1,224) of them were opened from devices in which the default language was Brazilian Portuguese. Devices in which English is the default language represent 18.00% of the total (339 sessions), while those whose default language is Spanish stand for 10.20% (192 sessions). Those data might point to the fact that the app has been used by foreigners visitng Brazil during the World Cup, however, the information on the default language retrieved from the analytics software is not competely reliable, since it is set by the user while configuring the device for the first time.

Nevertheless, stronger data supporting the claim that the app has been used by foreigners while visiting Brazil come from the analysis of the number of clicks received by each link in the first page of the app, in which users choose the interface language in their first visit – this page is not shown after the second time the same device accesses C-14/FN-Br. In the 1,367 sessions opened in this page, Brazilian Portuguese was chosen as the interface language 794 times, English was chosen 202 times and Spanish 160 times. 211 sessions were closed before any clicks were recorded.

Regarding the query modes offered in the main menu, Table 2 presents the users' behavior in regards to the interface language chosen. Totals presented in the last row of the table include visits by both first-time and returning users and do not include sessions in which no query mode was selected and users simply accessed information about the project or quit the app. Data in Table 2 show that the word search was the most used query mode in all three languages, representing at least half of the

users'choices. Neverthless, the other query modes oferred by C-14/FN-Br were tested by a substantial number of users.

Query Modes	Brazilian Portuguese	English	Spanish
Search a word	543 (54.08%)	206 (60.41%)	142 (66.98%)
Type in a sentence	141 (14.04%)	52 (15,25%)	26 (12.26%)
See the meaning	174 (17.33%)	42 (12.31%)	27 (12.73%)
Explore the network	146 (14.54%)	41 (12.02%)	17 (8.01%)
Total	**1,004 (100%)**	**341 (100%)**	**212 (100%)**

Table 2. Query modes chosen by users according to each of the possible interface languages.

5 Conclusion

This paper presented C-14/FN-Br, a frame-based trilingual electronic dictionary covering the domains of football, tourism and the Word Cup. Our aim with this research was both to extend the use of FrameNet to the construction of lexical resources for non-linguists and to explore the use of frames as interlingual representations.

Acknowledgments

The authors thank the National Council for Scientific and Technological Development – CNPq – (grant #474270/2011-4), the Minas Gerais State Research Foundation – FAPEMIG – (grant #CHE-APQ-00567-12), and the Federal University of Juiz de Fora (grant #PAGP-24975/12-14) for funding this project. The authors are grateful to the Semantec Group at UNISINOS, for providing the Football frames, and to Berkeley FrameNet and Spanish FrameNet, for the helpful insights on the English and Spanish frames and lexical units.

References

Carlos Subirats and Miriam R. L. Petruck. 2003. Surprise: Spanish FrameNet! *Proceedings of the 17th International Conference of Linguists*. Prague, Czech Republic.

Charles J. Fillmore and Beryl T. Atkins. 1992. Toward a Frame-Based Lexicon: the semantics of RISK and its neighbors. In Adrienne Lehrer and Eva F. Kittay (eds.). *Frames, Fields and Contrasts*. Routledge, New York. 75-102.

Charles J. Fillmore, Christopher R. Johnson and Miriam R. L. Petruck. 2003a. Background to FrameNet. *International Journal of Lexicography*, 16(3): 235-250.

Charles J. Fillmore, Miriam R. L. Petruck, Joseph Ruppenhofer and Abby Wright. 2003b. FrameNet in action: the case of attaching. *International Journal of Lexicography*, 16(3):297-332.

Hans C. Boas. 2002. Bilingual FrameNet Dictionaries for Machine Translation. *Proceedings of The Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain.

Joseph Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley, USA.

Kyoko H. Ohara, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito and Shun Ishizaki. 2004. The Japanese FrameNet Project: An Introduction. *Proceedings of the LREC 2004 SatelliteWorkshop: Building Lexical Resources from Semantically Annotated Corpora*. Lisboa, Portugal.

Lars Borin, Danna Dannélls, Markus Forsberg, Maria Toporowska Gronostaj and Dimitrios Kokkinakis. 2010. Swedish FrameNet++. *Proceedings of the Swedish Language Technology Conference*. Linköping, Sweden.

Liping You and Kaiying Liu. 2005. Building Chinese FrameNet Database. *Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*. ICCS Shangai Normal University, Shangai, China.

Maria Margarida M. Salomão. 2009. FrameNet Brasil: um trabalho em progresso. *Calidoscópio*, 7(2): 171-182.

Rove L. O. Chishman, Diego S. Souza and João G. Padilha. 2013. Kicktionary_Br: um relato sobre a anotação semântica de um corpus voltado ao domínio do futebol. *Veredas,* 17(1): 101-116.

Spanish FrameNet. 2013. Spanish FrameNet web site. http://sfn.uab.es:8080/SFN/.

Thomas Schmidt. 2007. The Kicktionary: A Multilingual Resource of the Language of Football. In Georg Rehm, Andreas Witt and Lothar Lemnitzer (eds.). *Data Structures for Linguistic Resources and Applications*. Gunter Narr, Tübingen, Germany. 189-196.

Tiago T. Torrent and Michael Ellsworth. 2013. Behind the Labels: criteria for defining analytical categories in FrameNet Brasil. *Veredas,* 17(1): 44-65.

Verónica Pastor and Amparo Alcina. 2010. Search Techniques in Electronic Dictionaries: a classification for translators. *International Journal of Lexicography*, 23(3): 333-357.