DOI: 10.17851/2237-2083.23.3.769-796

# The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary

## *A relevância da ferramenta Sketch Engine para a construção do Field – dicionário de expressões do futebol*

Rove Luiza de Oliveira Chishman
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
rove@unisinos.br

Aline Nardes dos Santos
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
aline.nardes@gmail.com

Diego Spader de Souza
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
dspadersouza@gmail.com

João Gabriel Padilha
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
joaogabrielsl@hotmail.com

**Abstract:** This paper aims at presenting the role of the Sketch Engine concordancer in the development of Field – Football Expressions Dictionary, a trilingual lexicographic resource based on the notion of frame and on linguistic corpora. Here, some considerations are presented concerning not only the resources available in the software, but also the analyses conducted through it. In the second section, Frame Semantics, a theory developed by Fillmore (1982, 1985) and its lexicographic applications are presented; in the third section, the Corpus Linguistics methodological potential is introduced by approaching some of its main principles, and some features of the study corpus. The fourth section describes the analyses procedures to identify (i) polysemic words and (ii) collocations in our corpus, through Sketch Engine. The results indicate that the software can be used for various lexicography-related purposes, once it enables

users not only to visualize in detail the entire productivity of language, especially through the Word Sketch option, but also to define the lexicography policy supported by corpus evidence.

**Keywords:** football; Frame Semantics; Corpus Linguistics.

**Resumo:** O objetivo deste trabalho é apresentar o papel do concordanceador *Sketch Engine* no desenvolvimento do *Field* – Dicionário de Expressões do Futebol, um recurso lexicográfico trilíngue baseado na noção de *Frame* Semântico e em *corpora* linguísticos. Aqui são apresentadas algumas considerações concernentes não apenas aos recursos disponíveis na ferramenta mas também às análises realizadas por meio desse programa. Na segunda seção, a Semântica de *Frames*, teoria proposta por Fillmore (1982, 1985), é apresentada, bem como suas aplicações lexicográficas; na terceira seção, o potencial metodológico da Linguística de *Corpus* é apresentado por meio de alguns de seus princípios basilares, além de se descreverem algumas características dos *corpora* de estudo do projeto. O quarto segmento descreve os procedimentos de análise, visando a identificar (i) as unidades polissêmicas e (ii) as colocações provenientes do *corpus*, por meio do *Sketch Engine*. Os resultados evidenciam que o programa permite aos usuários visualizar detalhadamente toda a produtividade da língua, principalmente por meio do recurso *Word Sketch*.

**Palavras-chave:** futebol; Semântica de Frames; Linguística de *Corpus*.

Recebido em: 31 de julho de 2015.
Aprovado em: 2 de outubro de 2015.

## 1 Introduction

Field – Football expressions dictionary – is a trilingual lexicographic resource organized around the notion of semantic frame (FILLMORE, 1982). Its building process involved, amongst other stages, the compilation of three comparable corpora,[1] which were processed

[1] It is important to highlight that this paper focuses only in Portuguese corpora analyses; therefore, the translation stage is not considered in this study.

using the Sketch Engine tool (KILGARRIF, 2004). Considering the precepts of Corpus Linguistics adopted in this project as a methodological resource, the tool had a fundamental role in the process of identifying frames and treating linguistic phenomena such as polysemy and collocations.

Thus, this paper aims at presenting the functionalities of Sketch Engine in what concerns the development of Field, showing its potential to the process of organizing frames through corpus attestations, as well as treating polysemic units and collocations. In order to illustrate the course of our analysis and demonstrate the results achieved, we approach the use of three Sketch Engine resources: Concordance, Word Sketch, and Collocations.

Having this in mind, we organize this paper in four sections: in the second section, we present the Cognitive Linguistics research program, which includes Frame Semantics and its applications. In the third section, we discuss some important aspects concerning Corpus Linguistics, the Sketch Engine tool and the corpora compiled for this investigation. In the third section, we present the analyses in two distinct moments, focusing on two phenomena: (a) polysemy and (b) collocations. The last section of this paper presents final remarks and perspectives for future researches.

## 2 Theoretical framework

Founded in the second-half of the 1970's, Cognitive Linguistics started as a movement of rupture between a group of linguists and the Chomskyan approach to the study of language. Lakoff (1987), Langacker (1987; 2006), Talmy (1987), Fillmore (1982; 1985) and Fauconnier (1985), among others, are part of this group. The dissatisfaction these researchers with Chomsky's model resided in the minor role semantics and pragmatics were playing along different views towards cognition and its relationship with language and meaning. It is important to point out that the Chomskyan paradigm is, in its fashion, cognitive. The fundamental difference in relation to the cognitivism that emerged in the late seventies is that this notion is broadened by the inclusion of notions

from cognitive psychology, such as the Prototype Theory (ROSCH, 1973; 1975a; 1975b).

Cognitive Linguistics defends that semantics arises from encyclopedic knowledge; meaning is built through our experiences while living and discovering the world around us. Assuming that meaning is encyclopedic, Cognitive Linguistics adopts a usage-based perspective. Such affirmative tells us that within this paradigm, the sharp distinction that separates semantics and pragmatics is no longer accepted: that is to say it is not possible to define strictly what belongs to language, and what belongs to the usage of language. To Fauconnier (2003, p. 1-2),

> Cognitive linguistics recognizes that the study of language is the study of language use and that, when we engage in any language activity, we draw unconsciously on vast cognitive and cultural resources, call up models and frames, set up multiple connections, coordinate large arrays of information, and engage in creative mappings, transfers, and elaborations. Language does not 'represent' meaning; it prompts for the construction of meaning in particular contexts with particular cultural models and cognitive resources.

According to Langacker (1987), the distinction between semantics and pragmatics is quite artificial, and to think of a "viable" approach to semantics means to consider one that rejects false dichotomies – a semantic model that presents encyclopedic nature, for that matter. Frame Semantics is considered to be an innovative theory of this point of view within the realm of Cognitive Linguistics.

## 2.1 Frame Semantics

Frame Semantics is a theory developed by Charles J. Fillmore (1982; 1985) today considered one of the alternatives Cognitive Linguistics presents to the study of meaning. To Miriam Petruck (2001, p. 1), Frame Semantics can be described as "[…] a research program in empirical semantics which emphasizes the continuities between language

and experience […]". In this sense, this theory seeks to investigate the relations between the senses of words of a given language and the experiences speakers go through in life. In other words, Frame Semantics understands that our knowledge of a language depends on the way we perceive the world. In what concerns Frame Semantics, "[…] the meaning dimension is expressed in terms of the cognitive structures (frames) that shape speakers' understanding of linguistic expressions." (FILLMORE; BAKER, 2010, p. 317).

A frame, therefore, can be described as an *experience schematization* (EVANS; GREEN, 2006) that structures the information we have regarding the elements that compose a given scene or situation. Fillmore (1982, p. 111) considers the frame to be "[…] any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits […]". When a concept is introduced in a text or in a conversation, all concepts related to it are automatically activated, or, in Fillmore's terms, *evoked*. We can think of the word "waiter", for example: according to Fillmore's concept of frame, to comprehend the concept of "waiter" is to comprehend the whole scene in which this particular concept is inserted, including other related words, such as "menu", "check", "client" etc. One of the purposes of the research undertaken by Frame Semantics refers to investigate the reasons why a community relates a given category to a word. Those reasons are to be included in the description of the meaning of such a word (PETRUCK, 2001). From this perspective, we are able to notice that Fillmore's theory does not share the views of more formal approaches to Semantics when it considers the role of factors such as experience / encyclopedic knowledge in the process of conceptualizing meaning.

Fillmore proposes the concept of frame as a way to cover a set of notions already in existence in the literature, which influenced his work – for instance, schema (BARTLETT, 1932), and script (SCHANK; ABELSON, 1977). It is therefore evident that the idea that underlies the semantic frames suggested by Fillmore was already present in linguistics, though under different denominations. It is worth mentioning that even the word "frame" was already in use by other researchers in different areas of knowledge: Minsky (1974) and Goffman (1975), in Artificial

Intelligence and Sociology respectively. To Minsky, when we are confronted by a new situation, we access a mental structure that provides us information concerning how to act and what to do in the situation – this structure is the frame. Goffman sees the concept through a very similar way, stating that frames are structures used to depict the various ways we behave in different situations. Through these observations, we are able to say that Fillmore's frames are very close to Minsky's and Goffman's.

However, according to Petruck (2001), the concept used in Frame Semantics is influenced by the notion of case frame, used by Fillmore in his Case Grammar, developed during the decade of 1960 – in 1968, more precisely. In the article "The Case for Case", the linguist stablishes the notion of cases, which are similar to semantic roles: in a sentence, the verb selects a determined number of cases, forming a set that Fillmore refers to as case frame.

Another important aspect of Frame Semantics worth mentioning concerns the nature of the frame as responsible for the depiction of a conventional, typical situation. Such characteristic leads us to talk briefly about the concept of prototype. Fillmore (1982) treats this notion using an example with the word *breakfast*, showing us that the understanding of this word is inseparably related to the cultural habit of having a meal during the first part of the day, after a period of sleep, whose menu is made up of a given kind of food and so on (FILLMORE, 1982). It is noticeable, however, that such statement motivates speakers to use the concept of breakfast in unusual situations: one can sleep until noon, eat a toast, drink some juice, and call this a breakfast. At the same time, one can wake up at seven in the morning, eat pizza and call this breakfast as well.

## 2.3 Frame Semantics and the development of lexicographic resources

Because of the possibility of organizing words around the situations in which they occur, Frame Semantics has been used as a basis for lexicographic resources. In this respect, it is essential to mention FrameNet, a pioneer project initiated by Fillmore and a group of linguists and lexicographers that presents semantic information about the lexicon

of the English language based on frames. According to Ruppenhofer *et al*. (2010, p. 5), The Berkeley FrameNet project is "[…] an on-line lexical resource for English, based on Frame Semantics and supported by corpus evidence". FrameNet works essentially with the documentation of the syntactic and semantic possibilities of lexical combination in English, taking into account all the senses words may present (RUPPENHOFER *et al*., 2010): its database has over 10 thousand lexical units. A lexical unit, according to Ruppenhofer *et al*. (2010), is the pairing of a given word and a meaning.[2] In the FrameNet context, it means that, in order to be a lexical unit, the word must be connected to a frame. In cases of polysemy, when a word is related to more than one meaning (thus related to more than one frame), each sense possibility is a lexical unit.

Besides FrameNet, the Kicktionary project (SCHMIDT, 2009) is also worth mentioning. As well as FrameNet, it is a database whose purpose is to describe the language of football in English, French and German, using corpora evidence. Kicktionary groups the football frames in scenes that organize them according to their similarities and differences. Is it important to emphasize, however, that these resources were designed having in mind a specialized user that comprehends the notions concerning Fillmore's theory. In virtue of this aspect, during the compilation of Field, a dictionary for the general public, we discarded the exhibition of some linguistic information in favor of a friendly interface, even though all data associated with semantic frames were mapped and analyzed. For the same reason, Field does not present semantically annotated examples,[3] preferring to occult information related to the elements of each frame.

The next section approaches the structure of Field Dictionary.

### 2.3.1 Field: a user-friendly, frame-based dictionary of football expressions

*Field* is a trilingual dictionary of football language guided by the notion of scenarios, which are an adaptation of Fillmore's concept of

[2] This concept is firstly proposed by Cruse (1986).

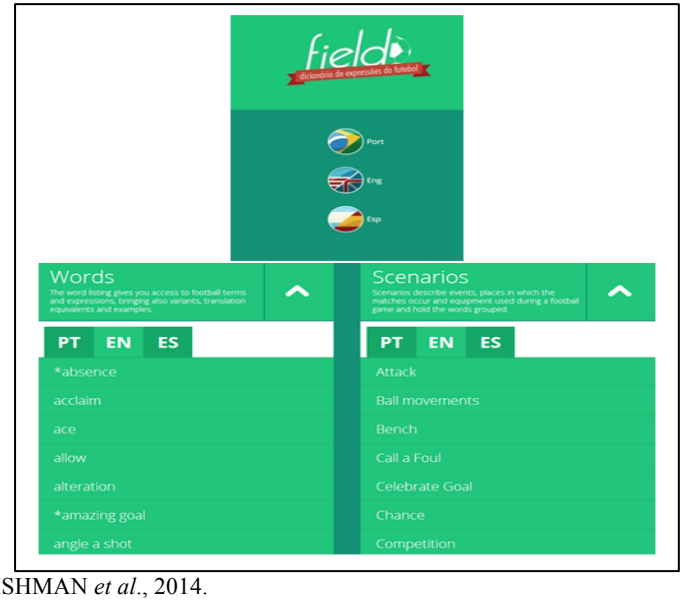
[3] In FrameNet context, semantic annotation means the "assignment of semantic role tags to syntactic constituents." (FILLMORE; PETRUCK, 2003, p. 359).

frame. The main goal of Field creators is not only to offer a word list, but also to organize words regarding their context of use, grouping them according to the situation in which they appear. Field has around 600 word entries, and approximately 40 scenarios; the website has optimized versions for tablets and smartphones.

In the context of Field, frames are called scenarios and are used to structure and organize the entries: football events such as SHOT, INFRACTION and GOAL are scenarios around which words and expressions are organized. This way, if users search for the word *bicicleta* (bicycle kick), they have access to information regarding the corresponding scenario (SHOT), apart from the translation to English and Spanish. By clicking on this scenario, they will find other related words, such as *bomba* (screamer) and *bico* (toe poke).

When accessing the dictionary for the first time, users can choose the desired language, which will also be applied to the lists of entries and scenarios. Field dictionary has two listings: one for words and one for scenarios.

Figure 1 – Field interface: select language screen and lists of words / scenarios

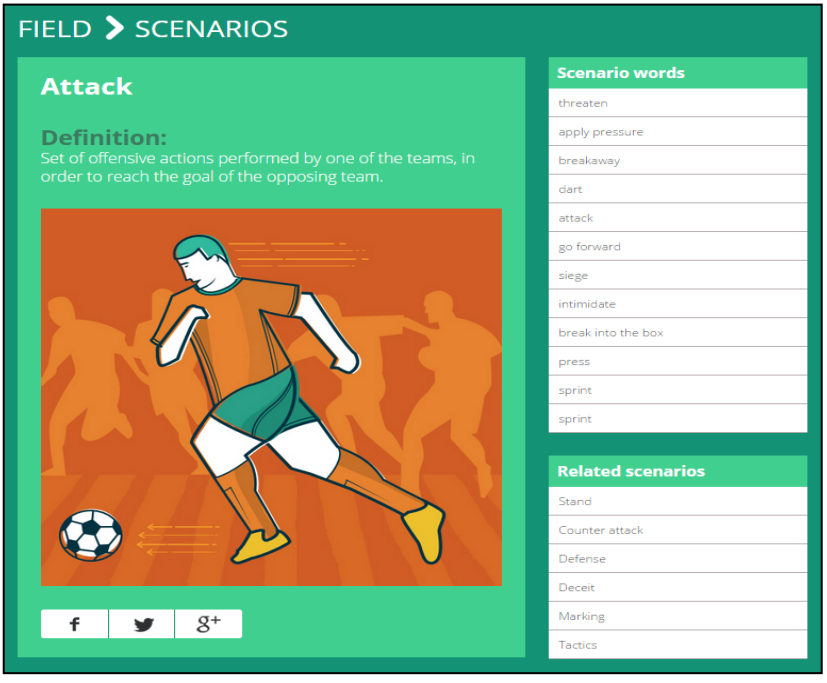



Source: CHISHMAN *et al*., 2014.

When searching for a word entry in Portuguese, users have access to information such as audio, word class and variants (including spelling differences and words with similar meaning), as well as translation to English and Spanish. The work of translation considered three situations: (i) cases of direct equivalence in the target language, (ii) cases of partial equivalence, and (iii) cases with no translation equivalent. Entries that illustrate the second situation (partial or inaccurate equivalence) bring a possible translation or a suggestion, and are marked with an asterisk. Entries to which no equivalence was found are identified with the abbreviation NT and have an explanatory note. This space is also used to explain the use of words that do not have direct equivalents of translation in one of the languages.

Concerning scenario entries, they represent events, match places or equipment used during the game, displaying all the words belonging to them. It is provided a definition and an illustration for each scenario.

Figure 2 – an example of frame / scenario




Source: CHISHMAN *et al*., 2014.

In the next section, we present the methodology adopted in the present investigation.

## 3 The methodologic potential of Corpus Linguistics: the concordance Sketch Engine and the corpora of the Field Dictionary

Corpus Linguistics is a branch in Linguistics that emerged in the second half of the twentieth century (BERBER SARDINHA, 2000). Having an empiricist character, its main objective is studying language through real-usage samples extracted from texts that exist in the world, that is, that were not invented with the purpose of illustrating a linguistic phenomenon. The word *corpus* relates to a "set of textual linguistic data collected properly with the purpose of serving the research of a given language or linguistic variety" (BERBER SARDINHA, 2000, p. 325). This set of data can be handled through computational tools that allow the researcher to deal with significant quantities of data – an example of such tools is the Sketch Engine concordancer.

Therefore, since the arising of personal computers, in 1980, to investigate phenomena such as polysemy and collocations from corpora means organizing corpora in accordance with the purposes of the investigation, as well as to process the data electronically through specific software. However, it is important to consider that electronic storage of corpora had already been created by linguists since the 1950's – despite the limited technologies they had access to –, due to the fact that American structuralists used to collect and analyze real speech data. In accordance to McCarthy and O'Keefe (2010), between the 1980's and the 1990's, Corpus Linguistics took the shape it has today, thanks to the development of computational resources. Nonetheless, is interesting to note that predecessors such as Sinclair *et al*. (1987), the developers of the COBUILD corpus, had to use punched cards technology when it was already considered outdated. The reason why they had to do it was the lack of user-friendly tools for linguists at the time; available resources could only be explored by computer experts (MCCARTHY; O'KEEFE, 2010). This scenario has completely changed since the development of software for linguists such as *Wordsmith* (SCOTT, 1996), *Sketch Engine* (KILGARRIF, 2004), *Unitex* (PAUMIER, 2002) and *AntConc* (ANTHONY, 2014) – these two are free, it is worth to mention.

A corpus linguistics research can be either corpus-driven, via the sole application of statistical measures (GRANGER; PAQUOT, 2008), or corpus-based – as in this study –, when investigations combine corpus data with other theoretical premises, in order to test their assumptions or even to improve them. In addition, Corpus Linguistics as a methodology "[…] does not go to the extreme of rejecting intuition while attaching importance to empirical data." (MCENERY *et al*. *apud* EVISON, 2010, p. 132). Therefore, within the context of Field, the usage of corpora was combined with Frame Semantics premises, as well as with the decisions that constituted what Atkins and Rundell (2008) call Style Guide.[4] This approach is also called "linguistics-with-corpus methodology"[5] (SANTOS, 2008, p. 52).

Regarding the building of our corpora, considering there was not a football language corpus in Brazilian Portuguese available, it was necessary to compile our own. The texts that form the Field Dictionary corpus were collected from the official websites of Brazilian football teams, from news websites and from Twitter profiles. The texts that interested our purposes were the ones that described how the matches had occurred. Known in the sports domain as *match report*, this genre resumes the main moments in a match: who scored the goals, how the goals were scored, which team won, what were the main moments of the match etc. These features ensured that we could verify how the football frames are organized by the occurrences of typical evocators as *shot* and *goal*. Among the texts available on the sites, there were some that did not fit our research, once they do not evoke football frames, and were, thus, ignored – their content was based on institutional news of the clubs, like the facilities, the promotions, the products related to the teams etc.

By following these features, our corpus can be considered representative of the football events in brazilian Portuguese, concerning our research purposes. Our collection of texts can be considered *big*, once it totalizes one million words to each language – Portugese, English and Spanish. Size of corpora is not a consensus amongst the authors, so it is important to say that we consider ours a big corpus based on the sum proposed by Berber Sardinha (2000). Once collected, the texts were

[4] A Style Guide is composed by all the instructions that guide the building of a dictionary (ATKINS; RUNDELL, 2008).

[5] In Portuguese: "metodologia da linguística com *corpora*".

converted to the .txt format (UTF-8), then processed by the *parser* PALAVRAS (which labelled the texts with morpho-syntactic information) and, finally, processed by the software Maestro, a pre-requisite for uploading the corpus to the Sketch Engine tool.

The Sketch Engine is a tool that allows the creation, the manipulation and the study of corpora. Through the search options, the user is taken to the *concordances*, which consist of lines based on fraction of texts in which the queried word or expression (the so-called *node word*), appear highlighted, as well as its co-texts (portions of texts that surround the node word). Other useful resources present in Sketch Engine are *Word Sketch* – which presents schematically the syntactic realizations of the lexical items in the corpus – and *Collocations*, where the user can visualize the words that tend to occur together in the corpus. These functionalities are shown in more detail in the next chapter.

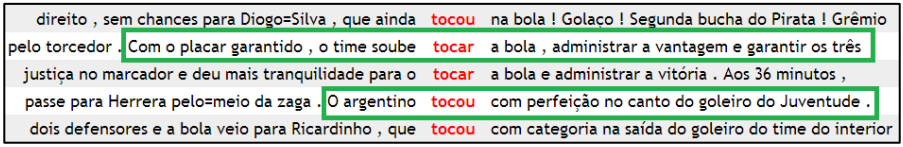
## 4 Analysis and results

### 4.1 The identification of polysemic lexical units using Sketch Engine

The Sketch Engine tool made it possible to identify polysemic lexical units in the corpus. In general lines, polysemy is the phenomenon whereby a linguistic form can have more than one related meanings, and it is a fundamental feature of human language (ULLMANN, 1961). It is important to mention that the considerations concerning polysemy here presented are part of a masters dissertation (PADILHA, 2015) written in the context of FIELD – Football Expressions Dictionary. Its main objective was to verify how polysemy presents itself in the football language. One of the main findings in this paper is that the most frequent polysemous items analysed represent *complex categories* radially shaped, confirming Langacker's (1987) and Lakoff's (1987) hypothesis, respectively. That is to say: the polysemous senses of words are stored in cognition in respect to a *prototypical sense*, the first one that comes to the mind of speakers when they hear or think of a word, through which all the other related senses are generated, in this case, by extension, as we intend to show.

Through a simple query in the *concordance* menu, we realized that the verb *tocar*, for example, presents polysemic behaviour, judging by the co-text that surrounds the realizations of this verb:

Figure 3 – KWIC-list of the verb *tocar* (part 1)[6]

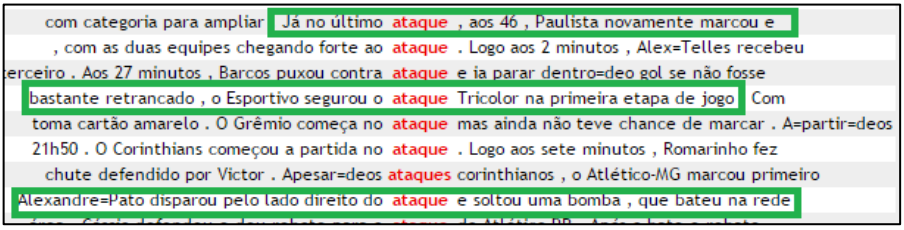



Source: KILGARRIF *et al*., 2004.

Right above, highlighted in green, it is possible to see two senses related to *tocar*: in [...] *o time soube tocar a bola* [...], the frame Pass is evoked, once it concerns the transfer of the ball possession between two players from the same team. In the second case, [...] *o argentino tocou com perfeição no canto do goleiro do Juventude*, Shot is the frame evoked by *tocar*: in this sense, this verb conceptualizes the action in which a player kicks the ball against the adversary team's goal.

Another case of polysemy that we find interesting to mention relates to the noun *ataque*, for which there were accounted three relates senses:

Figure 4 – KWIC-list of the verb *tocar* (part 2)[7]




Source: KILGARRIF *et al*., 2004.

---

[6] Free translation of the highlighted examples: 1) Possessing a fair scoreboard, the team could *pass* the ball around, hold the lead and guarantee the three…; 2) The Argentinian *placed* the ball perfectly into Juventude's goalkeeper corner.

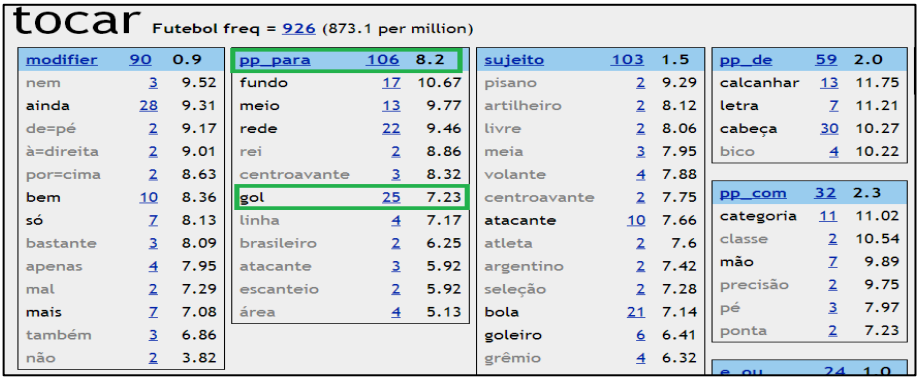
[7] Free translation of the highlighted examples:
1) In the last *attack move*, at the 46th minute, Paulista scored another goal and…;
2) …very defensive, Esportivo held back the Tricolor's attack in the first half of the match;
3) Alexandre Pato bolted to the right wing of the *attack* and fired a screamer, which hit the net.

The first highlighted sentence shows the sense of *ataque* that conceptualizes the *action of attacking the adversary team in order to score a goal*, evoking the Attack frame. The second sense related to *ataque* observed through the *concordance* query conceptualizes not the action, but the *group of players responsible for attacking the other team*. This comes to the fore by the excerpt *ataque tricolor*, which conceptualizes these players from Grêmio, referred to like that because it is one of the brazilian clubs whose uniforms have three colors. In this sense, *ataque* evokes the Actors frame. The third and last sense recorded for *ataque* relates to the *place in which the players responsible for scoring goals act.* This sense evokes the Field frame.

Though the *concordance* query allowed us to identify the senses of polysemous items, as we tried to show above, it would take us an endless amount of time to analyse all the instances for the senses of *tocar* and *ataque*. It would be so due to two basic reasons: the first one is that the tool shows *all* the sentences grouped, in a way that the user cannot separate one sense from the others; the second reason is that the sentences are not shown completely, just part of them (even considering that the tool allows the user to extend the co-texts to the left and right sides, it would still take a considerable amount of time from the user). Considering these issues, we relied on another Sketch Engine resource: the *Word Sketch*.

The *Word Sketch* option provides a more detailed search by allowing the user to access all the syntactic realizations of the queried word, as shown in the box below:

Figure 5 – Word Sketch for the verb *tocar* (part 1)




Source: KILGARRIF *et al*., 2004.

In the upper image, we have the sketch obtained for the verb *tocar*. Within each box it is possible to see a syntactic relation and its occurrences in the corpus. Highlighted in green, we can observe the realization of *tocar* when it is complemented by the PP *para*, totalizing 106 occurrences, of which 25 involve the noun *goal* as an object – *tocar para o gol*, sense that evokes the Shot frame, as we said before. Other senses related to *tocar* emerge through the objects *bola* and *mão*:

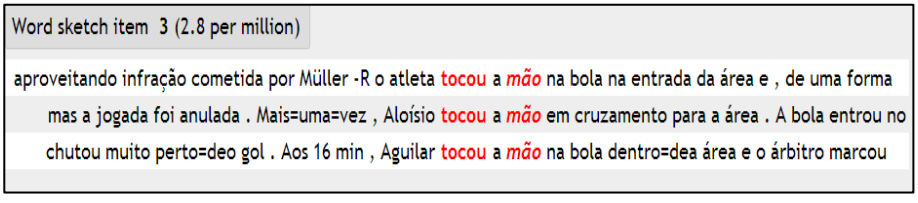
Figure 6 – Word Sketch for the verb *tocar* (part 2)

| objeto | 100 | 1.1 |
|---|---|---|
| bola | 75 | 8.98 |
| mão | 3 | 8.31 |
| 10 | 2 | 8.09 |
| travessão | 3 | 7.92 |
| trave | 2 | 6.4 |
| camisa | 2 | 6.36 |

Source: KILGARRIF *et al*., 2004.

*Tocar a bola* can relate not only to a pass, but also to a shot, as we have stated. On the other hand, the object *mão* involves another sense, and, consequently, another frame. By clicking on the number right beside the word, we have access to the concordances that show this sense:

Figure 7 – KWIC-list of the verb *tocar* (part 3)



Word sketch item 3 (2.8 per million)

aproveitando infração cometida por Müller -R o atleta tocou a *mão* na bola na entrada da área e , de uma forma

mas a jogada foi anulada . Mais=uma=vez , Aloísio tocou a *mão* em cruzamento para a área . A bola entrou no

chutou muito perto=deo gol . Aos 16 min , Aguilar tocou a *mão* na bola dentro=dea área e o árbitro marcou

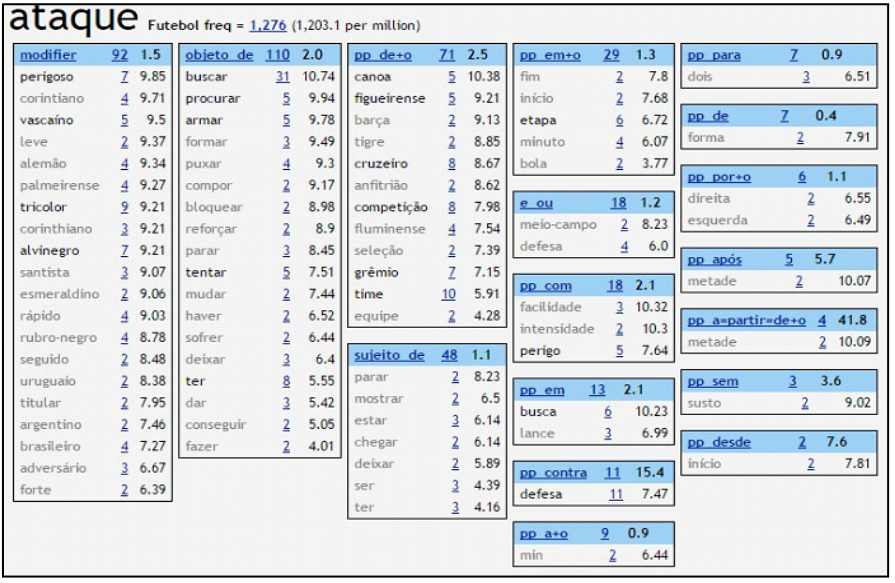
Source: KILGARRIF *et al*., 2004.

By checking the three occurrences in which *mão* occurs as a complement for *tocar*, we could observe that there is a third sense for this verb: this sense of *tocar* conceptualizes the moment a player touches the ball with his hand, which is an infraction, unless this player is the

goalkeeper. The frame evoked by the third sense, thus, is the Infraction one.

Right below, the word sketch for *ataque* is shown:

Figure 8 – Word Sketch for the noun *ataque* (part 1)




Source: KILGARRIF *et al*., 2004.

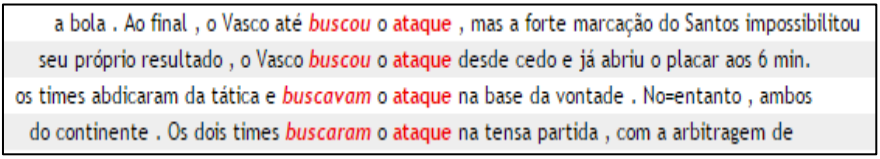
The relation *objeto de* (object_of) shows us all verbs that take *ataque* as an object in our corpus and their occurrence – 110 times. The verb *buscar* (search for) is the more prominent in this list, followed by *procurar* (look for) and *armar* (set) as the image below shows:

Figure 9 – Word Sketch for the noun *ataque* (part 2)


Source: KILGARRIF *et al*., 2004.

Appearing as an object of *buscar* 31 times, the high occurrence of this pattern showed us the first sense of *ataque*, according to our corpus: *action of attacking the adversary team*, as the concordances below show:

Figure 10 – KWIC-list of the verb *tocar* (part 3)




Source: KILGARRIF *et al*., 2004.

In all the concordance lines in the image, *ataque* conceptualizes an *action,* it is possible to observe: in the first two lines, this action is performed by a team – *Vasco* – that tried to score a goal in the match, as expected. In the following lines, the subjects of *buscar* are *os times* (the teams) and *os dois times* (the two teams), respectively, conceptualizing the act of attacking performed by the players responsible for this function in the game. This sense occurred 124 times in our corpus, being the most recurrent one. The second sense related to *ataque*, *group of players that perform the action of attacking*, according to our analysis, appears 52 times, being followed by the third sense – *part of the field* – that was counted 25 times in our data.

It is quite significant to mention that, even though the *Word Sketch* filter shows *all* the combinations for the nouns and verbs queried, the analysis must rely on the researcher proficiency when identifying the senses registered for such a language. To put it simply, the tool presents the combinations and the exemplars of such combinations, but it is the researcher that judges what counts or not as a sense, once the sense is not in the form, as the data reveal.

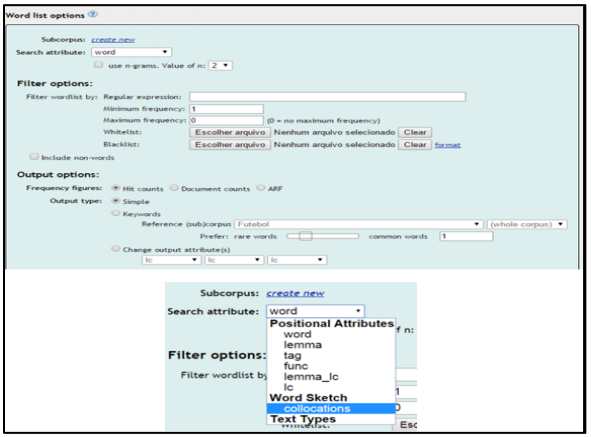
In the next section, we present the study of the collocations in the football language through the Sketch Engine concordancer.

## 4.2 *The* Collocations *tool and the identification and treatment of collocations*

Sketch Engine was also useful when it comes to the identification of collocations – groupings of words that usually appear together, such as *close friend* and *key question* (TAGNIN, 2013)[8] – in our *corpus*. It is important to say that this section is based on the results of a master's degree dissertation (SOUZA, 2015) developed in the context of *Field: Football Expressions Dictionary*. The objective of the study was to evaluate the role of the FrameNet methodology in the lexicographic treatment of collocations in the aforementioned dictionary. The hypothesis is that the FrameNet concept of *lexical unit* (a pairing between a linguist form and a meaning / frame, as we have previously said) allows collocations to be in the main list of headwords of the dictionary. Therefore, this section provides information regarding the methodological steps taken in the thesis dissertation and reports some of its results.

First, it is fair to start by addressing some characteristics of the *Collocations* tool of Sketch Engine, used in the study. Under the *Word list options* menu (FIGURE 9), *Collocations* allows the search to generate a list of all lexical combinations in a corpus, organizing the results based on a frequency criterion.

Figure 11 – The *Word list options* menu and *Collocations* filter




Source: KILGARRIF *et al*., 2004.

[8] For more information concerning collocations, cf. Tagnin (2013) and Hausmann (1989).

By clicking the options in *Search attribute*, the resource *Collocations* appears available. It is important to emphasize that in *Frequency figures* (below *Output options*), we maintain the option *Hit counts* selected, once we want the software to show how many times each one of the collocations appear in the corpus. For this particular study, we used the Brazilian Portuguese corpus *Futebol*, compiled for *Field*.

The search for collocations resulted in a vast list of combinations, and not all of them were structures we could consider football collocations. For the ongoing research, we selected the 500 most frequent results for a manual analysis, which consisted in eliminating the structures that did not have place in the study (non-collocations such as *que não* [that don't] for example). After this first scrutiny of the data provided by Sketch Engine, 74 collocations of football language were identified. The first highlight we were able to see in the data corresponds to the very high incidence of verbal structures: out of 74 collocations, 44 have a verb as one of its components, therefore meaning 59,4% of this subcorpus. Beforehand, we can consider this aspect a reflection of the language, as football is a dynamic sport, made of actions and events.

From this final list, in this section we present an overall quantitative and qualitative analysis of six verbal collocations related to the event of scoring a goal and three concerning the event of passing the ball.

The most frequent collocation in the corpus fits in this category: *abrir (o) placar* [open (the) score] with 784 occurrences. See the example below:

(1) Com o jogo bastante concentrado no meio-campo, o Grêmio TIME conseguiu **abrir o placar** aos 31 minutos TEMPO.
*With the game highly centered in the midfield, Grêmio* TEAM *was able to* ***open the score*** *at 31 minutes* TIME.

In the sentence above, we can see elements such as the *team* that scored a goal and *time*, fitting the collocation within the Score Goal frame in *Field*.

The next two collocations are *marcar gol* [score goal] and *fazer gol* [make goal] with 441 and 401 occurrences respectively. One

interesting semantic aspect of these structures is that they can work as substitutes of one another, as we can see in the examples:

(2) Juanfran JOGADOR **marcou gol** e deixou tudo igual no estádio El Madrigal.
*Juanfran* PLAYER ***scored a goal*** *and left it all the same in the El Madrigal stadium.*

(3) Zagueiro JOGADOR volta e **faz gol**.
*The defender* PLAYER *comes back and* ***makes a goal****.*

Having identified the element of a player who scores, these collocations can also integrate the Score Goal frame.

It is also interesting to notice the presence of the collocation *marcar pênalti* [sign penalty] in our data. Unlike what happens in the previous structures, in which the verbs refer to an action performed by a player or team, the verb here is related to the referee, who defines the penalty for a given offense made by some of the players in the game to the other team. See the example:

(4) O árbitro ÁRBITRO **marcou pênalti** e o veterano Blanco JOGADOR converteu a cobrança com categoria, ampliando o marcador PLACAR.
*The referee REFEREE* ***signaled penalty*** *and the veteran Blanco PLAYER converted the charging with category, expanding the marker MARKER.*

Differently from the other collocations, which are all part of the Score Goal frame, this particular structure, even though related to the same big event, belongs to the Referee Decisions frame as it is, in fact, depicting one of the actions done by referees in the match.

The next collocation, *balançar a rede* [swing the net], whose frequency is 245 occurrences, presents a change of focus, a change in the perspective through which the moment of goal is perceived. See the examples below:

(5) Quando Vagner Love JOGADOR **balançou a rede** do Palmeiras, neste domingo (18.11), em Volta Redonda, não foram só os torcedores do Flamengo que vibraram muito.

*When Vagner Love* PLAYER ***swung*** *Palmeiras'* ***net****, this Sunday (18.11), in Volta Redonda, it was not only Flamengo's crowd that celebrated.*

It seems that while the previous collocations focused more on the ball and the player, the collocation *balançar a rede* focuses on the result of the action in which the ball takes part, performed by the player towards the net. We highlight the fact that collocations need to be analyzed taking into consideration the category it is expressing. In a football match, what else could swing the net if not the ball? Maybe something else, like an object or even a player. However, even though that is possible, this particular collocation would not be used as it is related to a very specific event: scoring a goal.

The last collocation that conceptualizes the moment of goal is *sofrer gol* [concede goal], with 90 cases in the corpus. We notice that this collocation is on the same level of *marcar gol* and *fazer gol*, but referring to the opposing team's point of view, as we are able to see in (6):

(6) Mas Antunes JOGADOR não conseguiu acabar com a sina de **sofrer gols** logo no início.

*But Antunes* PLAYER *failed to end the fate of* ***conceding goals*** *early on.*

Another case in which verbal collocations designate two contrary points of view of a given action happens with the following structures: *dar (o) passe* [give pass], *fazer (o) passe* [make pass] e *receber (o) passe* [receive pass], with 118, 53 and 198 occurrences respectively. The first two collocations conceptualize the perspective of the player who has the ball and is passing it on to another player from their team. The third, however, depicts the point of view of the player to whom the ball is passed. The interesting feature about these collocations is how the frame elements are organized in the sentences, considering that different perspectives will possibly conceive the arrangement of such elements under different ways. See the examples from the corpus:

(7) Ivanildo JOGADOR QUE PASSA desceu pela meia direita, **fez o passe** para Magique JOGADOR QUE RECEBE, que apareceu na cara do goleiro e tocou para o fundo das redes.

*Ivanildo PASSER went down by the right wing, **made the pass** to Magique RECEIVER, who appeared in the front of the goalkeeper and kicked to the net.*

(8) Leandro JOGADOR QUE PASSA **faz o passe** para o centroavante JOGADOR QUE RECEBE.

*Leandro PASSER **does the pass** to the center-forward RECEIVER.*

(9) O jovem JOGADOR QUE PASSA fez um gol e **deu passe** para outro JOGADOR QUE RECEBE.

*The young PASSER scored a goal and **gave a pass** to another RECEIVER.*

(10) Ganso JOGADOR QUE PASSA **deu passes** precisos que criaram boas chances.

*Ganso PASSER **gave** precise **passes** that created good chances.*

(11) Com 25 minutos, Edenílson JOGADOR QUE RECEBE **recebeu passe**, invadiu a área e quando se preparava para tentar fazer o gol, foi derrubado por Élton.

*With 25 minutes, Edenível RECEIVER **received the pass**, invaded the area and while he was preparing to attempt a goal, was knocked down by Élton.*

(12) Primeiro, ele JOGADOR QUE RECEBE **recebeu passe** de Dadá JOGADOR QUE PASSA e tocou sem chances para o camisa 1 do Coxa.

*First, he RECEIVER received the pass from Dadá PASSER and shot with no chances to Coxa's no. 1.*

The first feature we can see through these examples is that, even though a pass situation will always involve at least two players, not always both of them will appear in the sentence, as we are able to notice in (10) and (11). Another interesting aspect also concerns the example

(10): in many cases in which we have the plural form *passes*, it is followed by an adjective. The data indicates that the singular form is often used to describe a single moment, a specific pass situation, while the plural is more directed to the *performance* the player had along the play, his style and competence. It is an evaluation not only of the pass or passes, but of the player as well.

## 5 Final Remarks

The present paper shows how the Sketch Engine tool contributed to the development of Field. It is highly valid to emphasize the treatment of the corpora through this software, considering not only the size of these collections, but also the possibility that it offers to explore the linguistic properties of these texts. The tool makes it possible not only to identify the most frequent words in the corpora, but also the way these words relate to each other. In addition, this study demonstrates the fundamental role of Corpus Linguistics in this corpus-based approach, providing a precise methodology for verifying empirically polysemic units and collocations.

One of the main advantages of using Sketch Engine is that its functionalities work in an integrated form, offering different levels of analysis: users can either perform a basic, fast query through the concordance option, or a more detailed one through Word Sketch. What is worth emphasizing is that, in our understanding, these resources do not superpose each other, but work together, once the user is always taken to the concordances in the end, as we demonstrated above.

Beyond its validity concerning the studies of polysemy and collocations, we emphasize the relevance of the Sketch Engine concordancer to the frame construction, considering that the *Word Sketches* show the syntactic-semantic behaviour of the frame evokers through empiric evidences.

Moreover, we would like to stress that, if on the one hand Sketch Engine provides several kinds and quantity of information, allowing different levels of analysis, on the other hand, it also poses a challenge to users, once they have to deal with large amount of data, not considering the task of compiling a corpus of their own – an activity that demands a considerable knowledge not only of the tool, but also of programing.

Finally, the results indicate that the software can be used for various lexicography-related purposes, once it enables users not only to visualize in detail the entire productivity of language, especially through the Word Sketch option, but also to define the lexicography policy supported by corpus evidence, as we intended to demonstrate.

To the next stage of the research, we foresee the construction of new corpora, having in mind the continuity of Field and also the creation of a dictionary devoted to the Olympic games.

## References

ANTHONY, L. *AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University, 2014. Retrieved on: September 15th, 2015, from: <http://www.laurenceanthony.net/>.

ATKINS, S.; RUNDELL, M. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press, 2008.

BARTLETT, F. Remembering: A study in Experimental and Social Psychology. Cambridge: Cambridge University Press, 1932.

BERBER SARDINHA, T. *Linguística de Corpus: histórico e problemática. D.E.L.T.A.*, [S. l.], v. 16, n. 2, p. 323-367, 2000. Retrieved on: March 4th, 2015, from: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso>.

CHISHMAN, R. L. O. *et al*. *Field – Dicionário de Expressões do Futebol*. São Leopoldo: Unisinos, 2014. Retrieved on: Apr. 25th, 2015, from: < http://dicionariofield.com.br/>.

CRUSE, D. A. *Lexical Semantics*. New York: Cambridge University Press, 1986.

EVANS, V.; GREEN, M. *Cognitive linguistics*: an introduction. Edinburgh: Edinburgh University Press, 2006.

EVISON, J. What are the basics of analyzing corpus? In: MCCARTHY, M.; O'KEEFE, A. (Ed.) *The Routledge handbook of Corpus Linguistics*. London/New York: Routledge, 2010.
DOI: <http://dx.doi.org/10.4324/9780203856949.ch10>.

FAUCONNIER, G. *Mental Spaces*. Cambridge: MIT Press, 1985.

_____. Cognitive Linguistics. In: NADEL, L. *Encyclopedia of Cognitive Science*. London: Macmillan, 2003.

FILLMORE, C. Frame Semantics. *Linguistics in the Morning Calm*. Ed.: The Linguistic Society of Korea, Seoul: Hansinh Publishing Co., p.111-137, 1982.

_____. Frames and the semantics of understanding. *Quaderni di Semantica*, vol. 6, n. 2, 1985. p. 222-254.

_____.; BAKER, C. A frames approach to semantic analysis. In: HEINE, B.; NARROG, H. (Ed.). *The Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, 2010. p. 313-339.

FILLMORE, C. J.; PETRUCK, M. R. L. FrameNet Glossary. *International Journal of Lexicography*, Oxford, v.16, n.3, p. 359-361, 2003. Retrieved on: June 20th, 2015, from: <http://ijl.oxfordjournals.org/content/16/3/359.full.pdf>.

GOFFMAN, E. *Frame Analysis*: An Essay on the Organization of Experience. Cambridge, MA (U.S.): Harvard University Press, 1975.

GRANGER, S.; PAQUOT, M. Disentangling the phraseological web. In: GRANGER, S.; MEUNIER, F. (Ed.). *Phraseology*: an interdisciplinary perspective. Amsterdam: John Benjamins, 2008. p. 27-49. DOI: <http://dx.doi.org/10.1075/z.139.07gra>.

HAUSMANN, F. J. Le dictionnaire de collocations. In: HAUSMANN, F.J.; REICHMANN, O.; WIEGAND, H. E.; ZGUSTA, L. (Org.). *Wörterbücher, Dictionaries, Dictionnaires*. Ein Internationales Handbuch zur Lexikographie, v. 1. Berlin: Walter de Grujter, 1989, p. 1010-1019.

KILGARRIFF, A. *et al*. *The Sketch Engine*. Lorient: Euralex, 2004. Retrieved on: 28 jun. 2014, from: <http://www.sketchengine.co.uk/>.

LAKOFF, G. *Women, fire, and dangerous things*: what categories reveal about mind. Chicago: The University of Chicago, 1987.
DOI: <http://dx.doi.org/10.7208/chicago/9780226471013.001.0001>.

LANGACKER, R. *Foundations of Cognitive Grammar*. V. 1. Stanford: Stanford University Press, 1987.

_____. Cognitive Grammar. Introduction to Concept, Image, and Symbol. In: GEERAERTS, D. (Ed.) *Cognitive Linguistics*: basic readings. Berlin/New York: Mouton de Gruyter, 2006.

LYONS, L. *Semantics*. V. 2. Cambridge: Cambridge University Press, 1977.

MCCARTHY; M.; O'KEEFE, A. Historical perspective: what are corpora and how have they evolved? In: MCCARTHY, M.; O'KEEFE, A. (Ed.) *The Routledge handbook of Corpus Linguistics*. London/New York: Routledge, 2010.

MCENERY, A.; XIAO, R; TONO, Y. *Corpus-based Language Studies*. London: Routledge, 2006.

MINSKY, M. A framework for representing knowledge. *Artificial Intelligence Memo*, n. 306, Cambridge: Massachusetts Institute of Technology, 1974.

PADILHA, J. G. M. *A polissemia na linguagem do futebol*: uma proposta de aproximação entre redes lexicais e frames semânticos. 2015. 134 f. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, 2015. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/3774>. Acesso em: 25 abr. 2015.

PAUMIER, S. *Manuel d'utilisation du logiciel Unitex*. IGM, Université de Marne-la Vallée, 2002. Retrieved on: September 15th, 2015, from: <http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf>.

PETRUCK, M. R. L. *Frame semantics*. Berkeley: University of California, 2001.

RUPPENHOFER, J. *et al*. *FrameNet II*: Extended Theory and Practice. Berkeley, California: International Computer Science Institute, 2010.

ROSCH, E. Natural categories. *Cognitive Psychology*, n. 4, p. 328-350, 1973. DOI: <http://dx.doi.org/10.1016/0010-0285(73)90017-0>.

_____. Cognitive Reference Points. *Cognitive Psychology*, n. 7, p. 532-547, 1975a.

_____. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology*: General, v. 104, n. 3, p. 192-233, 1975b. DOI: <http://dx.doi.org/10.1037/0096-3445.104.3.192>.

SANTOS, D**.** Corporizando algumas questões. In: TAGNIN, S. E. O.; VALE, O. A. (Org.). *Avanços da linguística de corpus do Brasil.* São Paulo: Humanitas, 2008.

SCHANK, R. C.; ABELSON, R. P. *Scripts, plans, goals, and understanding*: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum, 1977.

SCHMIDT, T. The Kicktionary: a multilingual lexical resource of football language. In: BOAS, H. C. (Ed.) *Multilingual FrameNets in computational lexicography*: Methods and applications. Berlin/New York: Mouton de Gruyter, 2009, p.102-132.

SCOTT, M. *Wordsmith Tools.* Oxford: Oxford University Press, 1996.

SINCLAIR, J. M. *et al*. *Cobuild English Dictionary*. London/Birmingham: Collins Cobuild, 1987.

SOUZA, D. S. *Jogada de letra*: um estudo sobre colocações à luz da Semântica de Frames. 2015. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, 2015. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/3924>. Acesso em: 25 abr. 2015.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. São Paulo: Disal Editora, 2013.

TALMY, G. Beyond foreground and background. In: TOMLIN, E. R. *Coherence and grounding in discourse* (Typological studies in language, 11), Amsterdam: John Benjamins, 1987.

ULLMAN, S. *Semântica*: a ciência do significado. Lisboa: Fundação Calouste Gulbenkian, 1961.