A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

Ana Salgado, Rute Costa

NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa
Avenida de Berna, 26-C 1069-061 Lisboa, Portugal
anasalgado@campus.fcsh.unl.pt
rute.costa@fcsh.unl.pt
https://clunl.fcsh.unl.pt/equipa/ana-salgado/
https://clunl.fcsh.unl.pt/equipa/rute-costa/

Abstract. Terms are a significant part of lexicographical nomenclatures in general language dictionaries. In this paper, we focus on how football terms are treated in three Academy Dictionaries – Portuguese, French, and Spanish – and draw some conclusions about the lexicographical decisions taken in the three languages. After identifying every position football players can have on the field, we verify whether the dictionaries above include these terms. We propose the TEI encoding of the term “defesa” (defence), which designates a position occupied by football players on the field. Bearing in mind concepts such as reusability and interoperability, we intend to present: 1) a comparison of football terms in the three dictionaries; 2) TEI Lex-0 dictionary encoding, a streamlined standard to facilitate interoperability; 3) a consistent TEI modelling and description of the microstructural elements of lexicographical entries. In the end, we draw some conclusions.

1. Introduction

Lexicography is conceived as a field that deals mainly with lexical units (words) but also with specialised lexical units (terms). While Lexicography and Terminology are two different scientific disciplines – with different theoretical-epistemological backgrounds – they have in common the fact that they deal with terms, albeit with different aims more often than not. This means
that working in Terminology and Lexicography requires different approaches since the social, cultural or economic purposes are not the same.

Lexicographers follow mostly a semasiological perspective (from words to senses), and terminologists, mostly concept-oriented, combine conceptual organisation and linguistic analysis where the definition of the concept is central with the view to reduce linguistic ambiguities. In 2013, Costa stated that Terminology and Lexicography should be seen as complementary regarding the methods they use. Bowker (2018, p. 149), arguing for the relation between these fields, sees advantages in the fact that “lexicographers and terminologists continue to work together to tackle new challenges and embrace new opportunities”.

Getting to know the domain and subsequently organising it are two requisite activities for a rapid and systematic identification of the basic concepts, which will result in a better description of the terminology. Bearing in mind that we are working with a specialised domain, the intervention of the expert is necessary to aid in the task of organising knowledge and to validate the descriptions and definitions of terms (Silva & Costa, 2019). This facilitates a more accurate encoding by allowing a tidier classification of the data depending on each element.

The primary purpose of this paper is to show how football terms are treated in the three Academy dictionaries that comprise our lexicographical corpus (i.e. Portuguese, French, and Spanish languages) (section 2). After identifying every position football players can have on the field, we confirmed whether the Academy dictionaries include these football terms (section 3). A contrastive study of the corpus allowed us to observe the data and draw conclusions about the lexicographical work performed in the three languages, gaining some insights into the lexicographical decisions that have been made in the three different dictionaries. Subsequently, we applied and tested the latest version of the TEI Lex-0 guidelines by representing the term “defesa” (defence), which designates a position occupied by players on the field (section 4). Aiming at providing guidelines for a structurally organised and consistent

---


2 This research is part of the PhD project of the first author that intends to design and test a set of methodological guidelines that can facilitate the inclusion of terms in a general language dictionary by systematising information regarding a single domain.
ent processing of lexicographical information, and bearing in mind concepts such as reusability and interoperability, we intend to present: 1) arguments in favour of using TEI Lex-0 dictionary encoding, a streamlined standard to facilitate interoperability; and 2) a consistent encoding and description of the microstructural elements of lexicographical entries, exemplifying the representation of lexicographical content using the term “defesa” (defence).

2. The lexicographical corpus

Recognising the importance of national academies that aim to create a dictionary in order to preserve the language, we decided to build a lexicographical corpus consisting of Academy works. The Academies of Sciences have undertaken dictionaries that are considered official authorities on the usages and the vocabulary of a language.

Our lexicographical corpus is comprised of three dictionaries published by different Academies: the *Diccionario da Língua Portuguesa Contemporânea* by the Academia das Ciências de Lisboa (DLPC), the *Diccionario de la Lengua Española* by the Real Academia Española (DLE), and the *Dictionnaire de l’Académie Française* by the Académie Française (DAF). These are general language contemporary dictionaries with printed editions and a descriptive nature with a normative concern addressed to a vast audience. The reason why we decided to create a contrastive corpus is justified by the fact that although the languages are different and the dictionaries themselves are also different, they share similar problems.

The content of these dictionaries is written in the languages of origin of each institution, i.e. Portuguese, Spanish, and French. These three dictionaries are available online³, accessible for free, and are updated continuously.

In Portugal, despite the successive attempts of the Academy, only in 2001, under the coordination of Malaca Casteleiro, did the Academia das Ciências de Lisboa publish a complete dictionary (from A to Z) for the first time in a two-volume paper version: *Diccionário da Língua Portuguesa Contemporânea* (DLPC). A new digital version – a task that has been undertaken by a team working in Natural Language Processing (NLP) at the University of Minho

---

³ The DLPC is currently only available in-house. However, as the coordinator of the new dictionary of the Academia das Ciências de Lisboa, the first author is making steps towards its public availability.
A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

– is the basis of the ongoing review of this dictionary, which now counts with the participation of NOVA CLUNL.4

The *Diccionario de la Lengua Española* is the widest normative dictionary of Spanish, published and created by the RAE. The most recent edition is its 24th, and it has been online since 2005.

The official dictionary of the French language, known as the *Dictionnaire de l’Académie Française* (DAF), served as a model for other dictionaries for many years, and the new version was made available on 7 February 2019, integrating the 9th edition in progress, from letters A to S.

3. Football domain

Football is the domain we have chosen to test the proposal for a set of future methodological guidelines for the lexicographical processing of terms. Our interest in football arises from the fact that it has been the most popular sport on the planet since the end of the 19th century; a sport with worldwide expansion via different societies in every continent. It is estimated that 250 million people are directly involved in football and that 1.4 billion people in the world have some interest in football (Morris, 1985).

We also aim to observe popularisation of the terms, namely the transition of a term into the vocabulary of everyday language. Many football terms have been adopted in everyday language, such as “canto” (corner) or “defesa” (defence) for instance.

This sport is also often referred to as 11-player football because it is played between two teams of 11 players each as seen on each of the definitions of the term in the three dictionaries that comprise our corpus (Fig. 1): “onze jogadores” (DLPC), “onze joueurs” (DAF), “once jugadores” (DLE):

---

4 The Natural Language Processing group of the Computer Science Department of the University of Minho has been developing the technological support of the new digital edition of the DACL, counting on the participation of Alberto Simões from IPCA (Instituto Politécnico do Câvado e do Ave), responsible for the technological support, José João Almeida, and the consultancy of Álvaro Iriarte Sanromán, both from the University of Minho. The participation of NOVA CLUNL (Linguistic Research Center of NOVA University of Lisbon) is related to its transition into the TEI LEX-0 format.
The positions of football players on the field

The 11 football players occupy specific positions on the field which are connected with specific terms (Fig. 2):

![Diagram of football positions](image)

**Fig. 2** – *Football players occupy different positions on the field*
For quick identification of all possible positions of football players on the field, we present Fig. 3:

The positions of the players indicate the specific function that they play on the field; they are typically associated with the tactical scheme used, and can be divided into four fundamental positions: (1) goalkeeper (GR); (2) defender positions (LD, LE, DC, LB); (3) midfielder positions (MD*, MD, ME, MC, MO); (4) attacker positions (AV, SA, PL, ED, EE).

In Tab. 1, and retrieving Fig. 3, we have listed some terms in Portuguese related to positions with their equivalents in Spanish and French. We have marked their presence (✓) or absence (-) in our lexicographical corpus.

---

5 The translation into English is used here only for the purpose of making this communication clearer.
Looking at Tab. 1, we can see that only the term “goalkeeper” is recorded in all the dictionaries. Most terms that designate the positions of the players are not recorded in our dictionaries – e.g., “right-back”, “left-back”, “centre-back”, “right winger”, and “left winger”. We may argue that this happens because we are dealing with polylexical units, such as “left back”, and not just with monolexical units, such as “back” in English, “lateral” in Portuguese, “latéral” in French, and “lateral” in Spanish. In consequence, we decided to search for these units in our lexicographical corpus. The unit “lateral”, when related to football, figures in the DLPC (“Fut. Jogador que actua junto da linha lateral do campo.”) and in the DLE (“Dicho de un futbolista o de un jugador...”)}
de otros deportes: Que actúa junto a las bandas del terreno de juego con funciones generalmente defensivas.”

To avoid such inconstancies, a terminological approach to the domain would be of major help. Building a concept system by identifying the relations between the concepts that embody the positions occupied by football players would allow the lexicographers to compile all the terms designating them. A conceptual approach to domains prevents lexicographers from missing essential terms of a terminology.

A term that is included in all these dictionaries, “goalkeeper”, raises some controversial questions. Although the DLPC uses “Fut.” (football) as a domain label listed in the abbreviation list, in the case of “guarda-redes”, the domain label used is “Desp.” (sports) (“Desp. Jogador que, no jogo do futebol, handebol, hóquei... ocupa o último posto de defesa, entre os postes da baliza, tentando impedir a marcação de golos”). This happens because the ‘definition’ presented above is not only related to the football domain, but includes other sports. In the DLE, “portero” is not identified by any label (“Jugador que en algunos deportes defiende la portería de su bando”). Finally, the DAF use the “Sports” label (“SPORTS. Gardien de but, joueur assurant la défense des buts dans certains jeux de ballon”).

It seems clear that the DLPC and the DAF distance themselves from the DLE by using the domain label to differentiate meanings or contextualize them, merely specifying the domain of the meaning. In fact, any criterion can be validated as long as it is applied uniformly.

4. Text Encoding Initiative (TEI)

In order to make content interoperable and reusable, we decided to follow the Guidelines of the Text Encoding Initiative (TEI), which is a de facto standard in digital edition. These guidelines provide a dictionary module (TEI Consortium) and have been used in numerous dictionary projects for

---

7 Said of a football player or a player of other sports: One that acts alongside the sidelines with generally defensive functions.
8 Player who, in football, handball, hockey..., occupies the last defense position between the goal posts, trying to prevent the scoring of goals.
9 Player who in some sports defends the goal of their side.
10 Goalkeeper, player defending goals in certain ball games.
born-digital lexicographical data (Budin et al., 2012) or retrodigitised projects (Bohbot et al., 2018).

However, since the TEI Guidelines present numerous and flexible encoding possibilities, a new version specifically for dictionary encoding is now being discussed –TEI Lex-0\(^\text{12}\) (Romary and Tasovac, 2018; Bański et al., 2017). We have chosen to test this target format, a streamlined standard to facilitate interoperability and some best-practice guidelines for NLP purposes.

TEI Lex-0 will not replace the “Dictionaries” chapter of the TEI Guidelines; instead it is being discussed as a target format that will standardise the existing heterogeneously encoded lexical resources and is being tested by numerous language dictionaries (Salgado et al., 2019b).

4.1. TEI Lex-0 encoding

The application of TEI Lex-0 will be demonstrated with samples of the term “defence” (Figures 4, 5, and 6) of the DLPC, the DAF, and the DLE. A consistent encoding and description of the microstructural components of the lexicographical article was the first step of this codification. Looking at the three dictionaries, the entries “defesa” (Fig. 4), “défense” (Fig. 5), and “defensa” (Fig. 6) include the same following elements: headword; etymology; part-of-speech (in the DLE this information is given in sense); a series of numbered meanings (polysemy). Depending on the criteria adopted in each of the dictionaries, we can also have domain labelling and other type of usage information, synonyms, collocations, and examples of usage.

---

\(^{12}\) TEI Lex-0 — A baseline encoding for lexicographical data: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

Fig. 4 – Entry “defesa” (DLPC)
1. **DÉFENSE** nom féminin

XIIe siècle. XIXe siècle, comme terme de droit, au sens de « la partie qui se défend ». Emprunté du bas latin *défensa*, « défense », participe passé féminin substantivé de *défendre*, « défendre ».

1. Action de défendre quelqu’un ou de se défendre contre une attaque. *Venir, courir à la défense* [...] 

3. Par analogie. **MÉDECINE.** *Défense musculaire*, contraction douloreuse des muscles (s’emploie surtout à propos de la paroi abdominale, où la défense musculaire témoigne le plus souvent d’une inflammation péritonéale). – **PHYSIOLOGIE.** *Les défenses de l’organisme*, ses moyens de résistance à l’invasion microbienne, virale, parasitaire, etc. L’immunologie étudie le système de défense de l’organisme. – **PSYCHOLOGIE.** Une réaction de défense. L’instinct de défense. Mécanismes de défense, moyens par lesquels le sujet cherche inconsciemment à se protéger. – **JEUX DE BALLON.** Action de défendre les buts d’une équipe, de résister aux attaques de l’équipe adverse et, par extension, en parlant d’un sportif, action de s’opposer aux offensives de l’adversaire. Par métonymie. *La défense*, les lignes arrière d’une équipe, chargées de défendre les buts. – **ÉQUITATION.** *Les défenses d’un cheval*, les mouvements par lesquels un cheval se dérobe à l’action de son cavalier. – **MARINE.** *Défenses*, morceaux de bois, tampons de cordage, pneus usagés, cylindres de matière plastique disposés le long du bord d’un navire afin d’amortir les chocs ou d’empêcher les frottements (on dit aussi *Pare-battage*).

**Fig. 5 – Entry “défense” (DAF)**

defensa

Del lat. tardío *défensa*.

1. f. Acción y efecto de defender o defenderse.
2. f. Arma, instrumento u otra cosa con que alguien se defiende en un peligro.
3. f. Amparo, protección, socorro.
4. f. Obra de fortificación que sirve para defender una plaza, un campamento, etc. U. m. en pl.
5. f. Jugada del trespaso en la que un jugador sustituye en sus derechos y deberes al hombre que rinde la jugada.
6. f. Mecanismo natural por el que un organismo se protege de agresiones externas. U. t. en pl. con el mismo significado que en sing.
7. f. *línea defensiva*.
8. f. En ajedrez, conjunto de jugadas previstas que constituyen una determinada estrategia.
9. f. Razón o motivo que se alega en juicio para contradecir o desvirtuar la pretensión del demandante.
10. f. Abogado defensor del litigante o del reo. U. m. en pl. cuando hay varios reos en el mismo juicio.
11. f. *Cuba, Méx., R. Dom. y Ur. parachoques*.
12. f. pl. Colmillos del elefante, cuernos del toro, etc.
13. f. pl. *Mar. Conjunto de objetos que se cuelgan del costado de la embarcación para que esta no roce o golpee contra el muelle u otra embarcación.
14. m. y f. Cada uno de los jugadores que forman la línea defensiva.

**Fig. 6 – Entry “defensa” (DLE)**
A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

Focusing on the form element that contains the orthographic, phonetic, and grammatical information of the entry (headword), all the elements of the entries are distinguished as clearly as possible:

```
<entry xml:lang="pt" xml:id="defesa_1">
  <form type="lemma">
    <orth>defesa</orth>
    <pron>dɨfézɐ</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">f.</gram>
  </gramGrp>
  <etym>[…]</etym>
</entry>

<entry xml:lang="fr" xml:id="défense">
  <form type="lemma">
    <orth>défense</orth>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">nom</gram>
  </gramGrp>
  <etym>[…]</etym>
</entry>

<entry xml:lang="es" xml:id="defensa">
  <form type="lemma">
    <orth>defensa</orth>
  </form>
  <gramGrp>
    <gram type="gen">feminin</gram>
  </gramGrp>
  <etym>[…]</etym>
</entry>
```

**Fig. 7 – Entry “defesa/défense/defensa” (DLPC, DAF, DLE) encoded in TEI Lex-0**

From these encoding examples, we highlight that the TEI Lex-0 schema only uses entry, the basic element of the dictionary’s microstructure, once we have constrained the general structure of a lexical entry – in our schema, entryFree, superEntry and re (related entry) of the current Guidelines are not used. In Fig. 7, entry contains the following elements – form, etym, and gramGrp – and requires the attributes @xml:id and @xml:lang, the appropriate language code (BCP 47) – “pt” (DLPC), “fr” (DAF), or “es” (DLE). Dictionary entries always start with a lemma (headword), a canonical form. The lemma form is encoded as form @type="lemma".

To specify the morpho-syntactic properties of the entry, we use gramGrp, which includes the part-of-speech of the entry (pos) and further specifications, such as the morphological gender (gen) of a lexical item. In
the example, for part-of-speech we use the universal dependencies tagset. In addition, we highlight the fact that the DLPC has also phonetic transcription (pron) and a numerical index (number), and, in the DLE, the part of speech (a noun) is inferred from the attribution of the female gender and is marked in sense.

13 In the ELEXIS context, a @norm attribute is mandatory to specify a normalised (UD) part of speech value. For more details: https://universaldependencies.org/u/pos/
A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

Fig. 8 – Entry “defesa/défense/defensa” (DLPC, DAF, DLE) encoded in TEI Lex-0
When dealing with football terms, the associated lexical items belong to a specific domain. This usage information is usually identified in dictionaries through domain labelling (Salgado and Costa, 2019; Salgado et al., 2019a). In Fig. 4, senses 11, 12, and 13 are marked with “Desp.” (the abbreviated form of “Desporto” [Sports]). In Fig. 5, the sense related to football is marked with the expanded form “JEUX DE BALLON”. Lastly, in Fig. 6, senses number 7 and 14, which refer to football, are not annotated. In TEI-style encoding, domain is a marker that identifies the specialised field of a lexical item – in TEI Lex-0 we use the element usg to mark usage information such as this (Fig. 8). Looking at other examples from Tab. 1, you can detect inconsistencies, since, for example, in the DLPC, both sports and football domains are used (e.g. “defesa-direito” is marked with the football domain, while “ponta-direita”, a synonym of “extremo-direito”, again a position occupied by a football player on the field, is marked as sports). Finally, we will also highlight multiword expressions, such as “jogar à defesa” (to play defence), which are treated as entries. We argue that this information should be standardised to a single value.

5. Concluding remarks

We want to conclude with four final remarks:

1. Our research has strictly lexicographical purposes, using Terminology’s methodologies as a contribution to the definition of guidelines and a methodology for the selection, inclusion, and processing of terms in general language dictionaries, namely the Academy Dictionaries we have analysed here, proposing a new dictionary model that, in a harmonised and balanced way, combines lexicographical methodologies and terminological methodologies.

2. Combining conceptual organisation and linguistic analysis is a necessity. Getting to know the domain and subsequently organising it are two requisite activities for a rapid and systematic identification of the basic concepts, which will result in a better description of the lexicon. This facilitates encoding by allowing a tidier classification of the data depending on each element, such as entry or sense, or @type “domain”> attribute. Bearing in mind that we are working with specialised domains, the intervention of the expert is necessary to aid in the task of organising knowledge, which will result in more accurate encoding.

3. By examining the codification of the term “defesa”, we have confirmed that the Lex-0 TEI meets our research needs. After encoding the
A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0

microstructure of the dictionary, we can guarantee the interoperability and reusability of the information. The advantage of applying TEI Lex-0 lies in the fact that lexicographers and terminologists alike are currently making efforts to apply TEI to the ongoing revision of ISO LMF (Romary, 2015). Given its (still) non-standard nature, it can be changed to accommodate relevant dictionary structures. We intend to demonstrate that the results obtained are useful for computational lexical encoding and may serve the purpose of NLP.

4. Finding that the dictionaries that make up our lexicographical corpus share common problems regarding the example we worked with (“defesa”) has led us to suggest that it would be interesting to present identical solutions for all of them. The reason why we decided to create a contrastive corpus is justified by the fact that although the languages are different and the dictionaries themselves are also different, they share similar problems. A contrastive analysis of the three dictionaries allowed us to confirm that. The solutions we have presented for the Portuguese example (DLPC) would be replicable in dictionaries of other languages since TEI is a recommendation for standardisation. An agreement between academies and other institutions would be desirable to systematise and optimise a new type of Lexicography that can provide a better representation of the entire European lexicographical heritage.

Acknowledgements

Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020, and by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS).

References


Larousse Illustré. In GLOBALEX 2018 – Globalex workshop at LREC2018, May 2018, Miyazaki, Japan, 1-6. <hal-01728328>.


A good TACTIC for lexicographical work: football terms encoded in TEI Lex-0


**Dictionaries**


**Résumé**

La complémentarité entre lexicographie et terminologie est le point de départ de cet article. Travaillant dans le domaine du football, cet article examine les termes du football dans trois dictionnaires de l’Académies de Sciences — portugais, français et espagnol — et présente l’encodage TEI du terme « defesa » (défense), qui désigne une position occupée par les joueurs sur le terrain. Après avoir identifié toutes les positions que les joueurs de football peuvent occuper sur le terrain, nous avons vérifié si les dictionnaires incluent ces termes. Une étude contrastive du corpus nous a permis d’observer les données et de tirer des conclusions sur le travail lexicographique réalisé dans les trois langues. En gardant à l’esprit des concepts tels que la réutilisabilité et l’interopérabilité, nous présentons : 1) une comparaison des termes dans les trois dictionnaires en analyse ; 2) le codage du dictionnaire TEI Lex-0, une norme *de facto* pour faciliter l’interopérabilité ; 3) une modélisation TEI cohérente et une description des éléments microstructuraux des entrées lexicographiques.