

The Kicktionary - A Multilingual Resource of the Language of Football

Thomas Schmidt

SFB 538 Mehrsprachigkeit, Max Brauer-Allee 60, D-22765 Hamburg,
thomas.schmidt@uni-hamburg.de

Abstract. This paper presents the Kicktionary, a multilingual (English - German - French) electronic lexical resource of the language of football. It explains how a corpus of football match reports was analysed according to the FrameNet and WordNet approaches and how the result of this analysis is presented to a dictionary user via a website.

1 Overview

The Kicktionary is an electronic resource providing lexicographic information about English, German and French words in the domain of football. It was constructed between September 2005 and July 2006 with the support and advice of the FrameNet team at the International Computer Science Institute (ICSI) in Berkeley¹.

The general aim in the development of the Kicktionary was to explore how linguistic theories about lexical semantics (especially the FrameNet and WordNet approaches to lexicography), corpus linguistic methods and hypermedia technology can help to build lexical resources that are better (or: good in a manner different from) traditional paper dictionaries. Storrer's theses on the use of hyper-text in lexicography [10] were used as a guideline. The focus is thus on questions of computational lexicography for *human* users, rather than on machine-centred fields like natural language processing or artificial intelligence.

The general approach for constructing the Kicktionary was to extract examples for domain specific lexical units from a corpus of football match reports, as described in section 2, and to analyse these items according to the FrameNet and WordNet paradigms. This led to a twofold organisation of the resource: on the one hand, lexical units were structured into a hierarchy of *scenes* and *frames*; on the other hand, they were partitioned into a number of *synsets*, which, in turn, were (partly) organised into a number of *concept hierarchies*. Sections 3 and 4 of this paper explain these notions in more detail.

At present, the Kicktionary contains close to 2000 lexical units and about 8000 example sentences. Table 1 provides some more detailed figures.

¹ The work presented here was carried out with the help of a research grant by the German Academic Exchange Service (DAAD). I am grateful to the FrameNet team and its visitors for their support. The original idea for this project is owing to Dieter Seelbach's and Gaston Gross's work on the lexicography of football language in the lexicon grammar framework ([5], [8], [9]).

	German	English	French	All
Lexical Units	792	599	535	1926
Examples	3551	2374	2239	8164

Table 1. Items in the Kicktionary

Language	Source	# texts	# words (ca.)	Mode
English	uefa.com	535	230000	written
French	uefa.com	482	240000	written
German	uefa.com	486	200000	written
German	kicker.de	1242	700000	written
German	German radio	9	10000	spoken

Table 2. Corpus overview

This paper explains the most important design features of the Kicktionary and briefly discusses some aspects of computational lexicography that were found to be relevant in its construction. A more comprehensive account of this work can be found in [7].

2 Corpus and Method

The Kicktionary was constructed on the basis of a corpus of football match reports from specialised websites. English, French and German texts were taken from the UEFA website (www.uefa.com). For German, additional material was acquired from the online edition of the Kicker journal (www.kicker.de); a small number of transcribed radio commentaries (from the NDR and SWR broadcasting stations) were also added to the corpus. All texts were tokenised and transformed into a TEI conformant XML format. Table 2 gives an overview of the corpus.

Candidates for lexical units were initially selected from a wordlist of the whole corpus without considering their membership in a specific frame or scene. Only in a later stage of the analysis, when a relatively stable scenes-and-frames hierarchy had been established, was the choice of new lexical units guided more directly by the existing structure of the resource. This manner of proceeding was intended to ensure that the scenes-and-frames hierarchy evolves on the basis of an empirical process rather than predetermining the empirical analysis by an "introspective" postulation of frames which are then to be "filled" with lexical material. The assignment of lexical units to synsets and the analysis of semantic relations between synsets were done only after the scenes-and-frames analysis had been more or less completed.

The analysis was carried out with the help of a combined concordancing and annotation tool. For each lexical unit, a KWIC concordance was first created. Suitable example sentences were then selected from this concordance, and the lexical units in these sentences, as well as their arguments, were marked and

annotated with appropriate labels. Example 1 shows an example sentence for the lexical unit *volley* with four arguments.

(1) [Kuijt]SHOOTER **volleyed** [in]TARGET [a Goor cross]MOVING_BALL [from close range]SOURCE

Regarding the cross-lingual part of the analysis, the partly parallel nature of the corpus could be exploited - for about half of the texts from the UEFA website, it was possible to automatically detect that they are direct translations of one another and to establish a cross-lingual alignment of these translations on the paragraph level. During the analysis, this alignment could then be used to discover and compare translation equivalents.

3 Scenes and Frames

Based on Fillmore's work on scenes and frames semantics [3, 4] and on the FrameNet methodology for constructing a lexical resource on the basis of frame semantics [4, 6], scenes and frames, as understood in the kicktionary, can be defined as follows: A frame is a structural entity used to group linguistic expression which share a common perspective on a given conceptual scene. A scene, in that sense is a superordinate construct to a frame. It is defined in terms of pieces of abstract, and possibly non-linguistic, knowledge, whereas the subordinate notion of a frame is concerned with the properties of concrete linguistic means of expressing this kind of knowledge.

One example for a scene in the domain of football is a one-on-one situation, i.e. an occasion in which the player in possession of the ball (PLAYER) is attacked by an opponent (OPPONENT) at some location (AREA) on the field. There are numerous ways of linguistically referring to such a scene, and they can be differentiated according to the perspective they impose on it. Thus, a speaker can choose to take the point of view of either the PLAYER (examples 2b and 2c) or of the OPPONENT (2a and 2d). Likewise, he can choose to relate the event in situ (2a and 2b) or describe it from the perspective of its outcome (2c and 2d).

(2a) [Zahovaiko]OPPONENT **challenged** [Manou Schauls]PLAYER [in the penalty area]AREA.

(2b) [He]PLAYER turned inside to **take on** [Roma]OPPONENT and finish with his left foot from close range.

(2c) [Hector Font]PLAYER tried to **nutmeg** [Ioannis Skopelitis]OPPONENT.

(2d) [Ronaldo]OPPONENT **dispossessed** [Wisla goalkeeper Radoslaw Majdan]PLAYER [on the edge of the box]AREA.

According to this differentiation by perspective, there are at least four frames associated with the one-on-one-scene. Frames are usually named after their semantically least specific English member; in this case the names are "Challenge"

(2a), "Take.On" (2b), "Beat" (2c) and "Deny" (2d). Each of these frames contains several lexical units. For instance, verbs like *beat*, *outstrip*, *round* or *sidestep* have similar properties with respect to a scenes-and-frames analysis to the verb *nutmeg* and are therefore all assigned to the same frame "Beat". Note that assigning lexical units to one and the same frame does not necessarily postulate a specific semantic relation (like, e.g., synonymy) between these items; conversely, however, synonymous expressions will inevitably end up in the same frame.

Since this kind of analysis is independent of the part of speech of lexical units, frames can contain verbal, nominal and adjectival items side-by-side. For example, both the verb *tackle* and the nominal expression *sliding tackle*, like the verb *challenge*, are part of the frame "Challenge". This is especially important for the multilingual analysis, because it frequently happens that a translation equivalent for a given lexical unit can only be found in a different part of speech. For instance, the idea of *nutmegging* (*an opponent*) is usually expressed in French with the help of the nominal multi-word-expression (*faire un*) *petit pont* (*sur un adversaire*), and both these lexical units are accommodated by the frame "Beat".

As Boas [1] argues, a scenes-and-frames analysis carried out for one language is usually transferable to other languages. For the Kicktionary, this means that scenes, which are language-independent by definition, remain valid across languages, and that frames can accommodate lexical material from an arbitrary number of languages. Thus, the afore-mentioned frame "Beat" contains, among others, the English, German and French lexical units listed in (3a) to (3c).

- (3a) beat, outstrip, nutmeg, shake off, sidefoot
- (3b) Beinschuss, düpieren, stehen lassen, tunneln, umdribbeln
- (3c) coup du sombrero, dribbler, échapper, mystifier, petit pont

A total of 16 scenes were defined for the football domain, consisting of altogether 104 frames.

4 Synsets and Concept Hierarchies

While a scenes-and-frames analysis of the vocabulary reveals many regularities and relationships between lexical units which are not covered by traditional dictionaries, it does not explicitly state some more basic associations between words, like synonymy, hypernymy or holonymy. In addition to the scenes-and-frames structuring of the resource, a second analysis was therefore carried out using the WordNet [3] approach of partitioning the vocabulary into sets of synonyms and establishing semantic relations between such "synsets".

As an example, consider the frame "Shot" (part of a scene of the same name) which contains, among many other lexical units, the English, German and French nouns listed in (4a) to (4c).

- (4a) shot, drive, volley, header, diving header
- (4b) Schuss, Torschuss, Volley, Direktabnahme, Kopfball, Kopfstoß, Flugkopfball, Kopfballtorpedo
- (4c) tir, frappe, vollée, tête, coup de tête, tête plongeante

Assigning all of these lexical units to the same frame is justified by the fact that they all impose the same perspective (the shooter's) on the same prototypical scene (a shot), but it does not provide any more specific information about commonalities and differences between the meaning of these words. As a first step towards adding this kind of information, lexical units with identical meanings were subsumed in synsets. As the examples (5a) and (5b) show, the notion of a synset was extended in the Kicktionary to include translation equivalence between lexical units of different languages as well as synonymy within one language.

- (5a) {shot, drive / Schuss, Torschuss / tir, frappe}
- (5b) {header / Kopfball, Kopfstoß / tête, coup de tête}

In a second step, semantic relations between synsets were analysed. For instance, lexical units in the synsets (6a) and (6b) were found to be hyponyms of those in (5a) and (5b), respectively.

- (6a) {volley / Volley, Direktabnahme / vollée}
- (6b) {diving header / Flugkopfball, Kopfballtorpedo / tête plongeante}

Besides the hyponymy/hypernymy relation, nominal synsets were also linked via a part-whole (holonymy/meronymy) relation as demonstrated in (7).

- (7) {goal / Tor, Kasten, Gehuse / but, cage} *holonym of* {crossbar, bar / Latte, Querbalken / barre, transversale}


Verbal synsets were connected via the troponymy ("to X is to Y in some way") relation as demonstrated in (8).

- (8) {beat, defeat / schlagen, bezwingen / batter, s'imposer} *troponym of* {thrash / deklassieren / balayer}

Since all of these relations are transitive, they can be used to build hierarchies of synsets. Altogether, 36 such concept hierarchies were built for a total of 552 synsets. In contrast to all other structural assignments, the mapping of synsets to concept hierarchies is neither complete nor unique - i.e. whereas each lexical unit belongs to exactly one frame and exactly one synset, and each frame to exactly one scene, some synsets are not assigned to a concept hierarchy at all, while others are part of two or more such hierarchies. Further semantic relations which play a role in WordNet, e.g. antonymy, have not yet been explored for the Kicktionary.

5 Presentation

Since the Kicktionary is mainly intended as a lexicographic resource for human users, great attention was paid to an adequate, human-readable presentation of lexical units and their structural organisation. The resource is presented as a website on www.kicktionary.de². Figure 1 depicts an exemplary entry for the lexical unit *bicycle kick*. The entry indicates the lexical unit's scene and frame assignment and lists the annotated example sentences in two different forms - once as full text and once in a schematic overview. Synonyms and superordinate synsets are also provided. Furthermore, each component of the presentation is hyperlinked to corresponding other parts of the resource. For instance, clicking on the name of the scene will take the user to a description of that scene. Likewise, examples are linked to the corpus text from which they were taken, and the synsets are linked to a presentation of the corresponding concept hierarchies.

bicycle-kick.n  Scenario Shot Frame Shot

SHOOTER [Player]

1. Not content with that, [Crespo]_{SHOOTER} then attempted a **bicycle kick** only for Laštuvka to produce a reflex save to deny him a second goal. [1077219 / p9]
2. Cazorla shot narrowly wide from distance on the half-hour mark and Luciano saw [his]_{SHOOTER} **bicycle-kick** saved by Vasili Khomutovski five minutes later before José Mari shot wide. [80107 / p3]
3. The Danish forward headed Pirlo's long pass into the path of Shevchenko who latched on to the ball but saw his shot cleared by [Celtic defender Dianbobo Balde's]_{SHOOTER} spectacular **bicycle-kick**. [1077172 / p6]

Support	LU	SHOOTER
<i>attempted</i>	bicycle kick	Crespo
	bicycle-kick	his
	bicycle-kick	Celtic defender Dianbobo ...

Synonyms Fallrückzieher.n
overhead_kick.n bicycle-kick.n
retourné.n

Hypernyms Torschuss.n Schuss.n
shot.n drive.n strike.n
tir.n frappe.n

[Moving_Balls]

Fig. 1. Kicktionary presentation of the lexical unit *bicycle kick*

The Kicktionary offers several points of entry to a user for navigating and exploring the resource. For a simple bottom-up access, an alphabetic list of lexical units, separated by language, is provided. For top-down access, a list of scenes or an index of concept hierarchies can be used. A fourth point of entry is given in the form of an annotated parallel text with links into the resource.

² The site is password protected. Interested users can request a free account

6 Discussion

In a discussion of the Kicktionary’s contribution to current research in computational lexicography, three points seem especially important. Firstly, the Kicktionary is one of the first attempts to construct a domain-specific resource using a frame-semantic approach. Secondly, it is also one of the first examples of a multilingual resource on the basis of this theory. And thirdly, the Kicktionary has explored new ground by trying to combine a FrameNet-like approach with elements taken from WordNet-style lexicographic analyses. Against this background, the most important findings in the work on the kicktionary can be summarised as follows:

- A frame semantic approach is very well suited for the construction of domain-specific lexical resources. Even more than general language dictionaries, such resources need to relate detailed linguistic information with knowledge about the world, and the notion of scenes and frames provides a systematic method for fulfilling that task.
- Owing to the language-independent nature of a scene and to the possibility to populate frames with lexical units from different languages, the scenes-and-frames approach also lends itself very well to the construction of a multilingual resource. The resulting organization of the multilingual dictionary can be helpful in various translation tasks.
- A scenes-and-frames analysis and a WordNet style analysis of the vocabulary can be utilized in a complementary manner. Many of the more basic semantic relationships between lexical units are not covered by the scenes-and-frames hierarchy, and a separate organisation of the vocabulary into hierarchies of synsets is one practicable way of providing this missing information.

A more detailed evaluation of the resource will be carried out once there is a sufficient amount of user feedback from the website presentation.

7 Outlook

The Kicktionary in its present form is complete in the sense that a reasonably large³ list of vocabulary items from the football domain has been analysed and integrated into the described architecture. It is also complete in the sense that this architecture has been made fully accessible to the user via the presentation

³ "Reasonably large" meaning that a) the number of lexical units in the kicktionary is considerably higher than in comparable printed dictionaries (e.g. [2, 11]) and that b) a further analysis of the corpus would turn up no or very few additional lexical units.

of the resource on a website. There are, however, various ways in which the kicktionary could be improved and extended in the future.

Firstly, an extension of the corpus is likely to uncover lexical units that have been overlooked so far. A larger corpus could also be used to increase the number of annotated examples for the existing lexical units. In both cases, the additional material may make it necessary to remodel parts of the scenes-and-frames hierarchy and of the concept hierarchies. Further text material from the UEFA website (again, about 250,000 tokens for English, French and German) has been acquired for this purpose and is presently being processed.

Secondly, user feedback for the kicktionary website should make it possible to evaluate the quality of the resource and its presentation.

Thirdly, the existing architecture, together with the concordancing and annotation tool developed for the analysis, should make it relatively easy to supplement the kicktionary with lexical units and examples from other languages. There are plans for cooperations to produce a Polish and an Icelandic version of the kicktionary. Furthermore, corpus material in Italian, Portuguese, Spanish, Russian and Japanese is available for lexicographers interested in producing versions for these languages.

Lastly and more generally, the Kicktionary may be a promising test case for the development and application of methods for collaborative creation of specialized multilingual lexical resources. This is so because, on the one hand, football is a well-delimited special domain with a large, but manageably-sized vocabulary. On the other hand, and contrary to many other specialized areas, it is not too difficult to find "experts" who are competent users of that vocabulary (in different languages) and who may be able and willing to contribute to such a collaborative effort either as lexicographers or as evaluators of the resulting resource. First steps towards an architecture in which dictionary creators and dictionary users can work together to construct an improved version of the kicktionary have already been taken.

References

1. Hans C. Boas. Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography*, 4(18):445–478, 2006.
2. Roberta Colombo, Klaus Heimeroth, Olivier Humbert, Michael Jackson, Frank Kohl, and Josep Ràfols. *PONS Fußballwörterbuch*. Ernst Klett Verlag, Stuttgart, 2006.
3. Charles J. Fillmore. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, 1982.
4. Charles J. Fillmore, Christopher Johnson, and Miriam R.L. Petruck. Background to framenet. *International Journal of Lexicography*, 3(16):235–250, 2003.
5. Gaston Gross. Comment décrire une langue de spécialité? *Cahiers de lexicologie: Revue Internationale de Lexicologie et Lexicographie*, (80):179–200, 2002.

6. Josef Ruppenhofer, Michael J. Ellsworth, Miriam R.L. Petruck, and Christopher Johnson. *FrameNet: Theory and Practice*, 2005.
7. Thomas Schmidt. The kicktionary - a multilingual lexical resource of football language. In Hans C. Boas, editor, *Multilingual Framenets*. De Gruyter, New York, 2007.
8. Dieter Seelbach. Das kleine multilinguale Fußballlexikon. In Walter Bisang and Gabriela Schmidt, editors, *Philologica et Linguistica. Historia, Pluralitas, Universitas. Festschrift für Helmut Humbach zum 80. Geburtstag am 4. Dezember 2001*, pages 323–350. Wissenschaftlicher Verlag, Trier, 2001.
9. Dieter Seelbach. Separable Partikelverben und Verben mit typischen Adverbialen. Systematische Kontraste Deutsch-Französisch /Französisch-Deutsch. In Uta Seewald-Heeg, editor, *Sprachtechnologie für die multilinguale Kommunikation. Beiträge der GLDV-Frhjahrstagung*, pages 103–115. Gardez!, St. Augustin, 2003.
10. Angelika Storrer. Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In Ingrid Lemberg, Bernhard Schröder, and Angelika Storrer, editors, *Chancen und Perspektiven computergestzter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*, pages 88–104. Niemeyer, Tübingen, 2001.
11. Kaya Yildirim. *Fußballwörterbuch in 7 Sprachen*. Kauderwelsch. Reise-Know-How Verlag Peter Rump GmbH, Bielefeld, 2006.