Football Phraseology: A Bilingual Corpus-Driven study
Sabrina Matuda

Football is the most popular sport in the present century. It has assumed a role beyond that of a national sport by becoming a cultural manifestation. It is now a battleground for several important issues such as economy, the resolution of conflicts, poverty as well as racial and minority awareness (Anchimbe 2008). Football relations across countries have increased significantly over the past decades. In order to regulate these relations, we need to express ourselves through language, that language being, in most cases, English. However, each culture has its own way of playing and supporting its teams, a way that is differently expressed according to each mother tongue. The problem arises when there is an urge to express these particularities in a foreign language. Football constitutes first and foremost a technical domain, though usually not considered as such. Therefore it involves a specialized language. Aiming at understanding the football vocabulary we propose a detailed study of football phraseology.

This presentation aims at presenting a semi-automatic extraction method for the most frequent phraseology in which the term goal is embedded. The study is based on the assumptions that specialized texts are the central object of study of Terminology and that textual type recognition, the culture in which the text is embedded, its purpose as a text, the intended target audience, its systematic traits and the terminological density play an important role in term identification (Krieger e Finatto 2004:81). Therefore, the specialized language shall be seen as an integral part of a culture, representing one of its powerful manifestations in so far as it is not an isolated phenomenon suspended in a vacuum (Snell-Hornby 1988:39 apud Azenha 1999:28).

There are some English-Portuguese bilingual dictionaries of football available in the market. However, all of them describe the term per se. Our study goes beyond the term description. Language patterns for the terms under analysis are identified and through such
patterns typical football phraseologies are brought about, which, according to Krieger and
Finatto (2004) describe up to 70% of a specialized language.

This study is based on the notions of Corpus Linguistics (Bowker & Pearson, 2002; Hunston,
2002; Sardinha, 2004; Sinclair, 1991); corpus-driven translation studies (Tognini-Bonelli,
2001) and terminology (Krieger & Finatto, 2004; Maia, 2002; Temmerman, 2000).

According to Bowker and Pearson (2002:9), a corpus can be described as a large collection of
authentic texts that have been gathered in electronic form according to a specific set of
criteria.

Empiricism, one of the main theoretical aspects of Corpus Linguistics, is described by Berber
Sardinha (2004) as primacy of data deriving from observation of natural language use through
the medium of a corpus.

The probabilistic view of language, another concept of CL, has its roots in Hallydayan
linguistics – a theory that claims that some lexical and grammatical combinations that are
possible are not in fact likely to occur together.

Corpus linguistics has been of great relevance to phraseological studies as it involves
selecting what is worth paying attention to. Many scholars have approached terminology and
phraseology from a corpus linguistic perspective: Bowker and Pearson (2002), Sinclair

As Palumbo (2003) points out while discussing the use of phraseology in translator training
and translation research regarding LSP texts, the identification and definition of phraseology
are highly problematic issues. The problem starts with the numerous labels used to describe
phrases in LSP texts: terminological phrase, LSP phrase, technical phrase, phraseme or
phraseological unit, phraseological term, professionalism or jargon word, multi-word lexical
item, among others.

For the purposes of this corpus-driven study, phraseology is studied from two perspectives as
proposed by Krieger & Finatto (2004): 1) fixed or semi-fixed pluriverbal combinations
constituted by two lexical units – own goal; 2) formulaic sequences in a specialized language -
score an away goal.

Based on these two perspectives, we question the protected status a term usually enjoys and
recognize that it is not likely to be used apart from other lexical because the context within
which it is embedded ends up affecting its meaning.

The adequate use of a term is emphasized by Bowker and Pearson (2002): ‘… How many
times have you found yourself armed with the correct terminology but unable to use it simply
because you do not know which word goes around which?’

Our corpus consists of approximately 1 million words - 557.498 in English and 447.176 in
Portuguese. Each corpus is divided into three subcorpora: laws of the game, media texts about
match results and match reports. The texts were gathered from online newspapers such as
Section 8. Phraseology and Collocation

lancenet, CBFnews, O Globo, Gazeta Esportiva, folha online, The Guardian, The Daily Telegraph, FIFAnews, TalkFootball, eyefootball and BBCsports. Therefore, our corpus is termed comparable. A comparable corpus is a collection of authentic source language texts in more than one language dealing with the same subject matter and compiled according to similar criteria (Vantarola, 2002).

To explore the corpus we used Mike Scott’s WordSmith Tools (version 5, 2007), specifically the tools Concord, Clusters and ConcGrams. We carried out our study in the Portuguese-English direction. Firstly, concordance lines for the word goal were extracted using Concord. Afterwards, the lines were re-sorted to show characteristic lexical patterns. By means of sorting the lines, it is possible to examine the immediate context to the left and to the right. However, due to the large amount of occurrences, identifying the phraseologies through the concordance lines would be very time consuming. Therefore, the tool Cluster was used. The settings were adjusted to find four word clusters, the minimum frequency being four with an MI-score equal or greater than 2.

In order to validate the phraseologies we associated the clusters that seemed to be part of the same unit of meaning, as for instance, ‘an own goal’ and ‘score an own’, which became ‘score an own goal’. By doing so, we reduced the number of clusters from 16 to 10.

A careful analysis of the concordance lines for the word goal re-sorted by the first word to the right and to the left was carried out to find the phraseological equivalents. The next steps greatly depended on whether or not it was possible to come up with equivalents by reading these lines. It should be stressed here that the phraseologies under analysis are often quite different in English and Portuguese. By looking at the lines one can easily identify the equivalents for gol da vitória and marcou um gol goltra, respectively winning goal and score an own goal. However it was not possible to identify equivalents for phraseologies such as melhor chance de gol (best scoring chance). In order to do so, concordance lines for the word chance were extracted:

<table>
<thead>
<tr>
<th>N</th>
<th>Concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>al. It is far easier to build a scoring chance from 40 yards away than from 80. I</td>
</tr>
<tr>
<td>61</td>
<td>len had the Soca Warriors’ best scoring chance, a well-struck shot from the right s</td>
</tr>
<tr>
<td>62</td>
<td>bench. Saha set up France’s best scoring chance, using his chest to drop the ball</td>
</tr>
</tbody>
</table>

Figure 1. concordance lines (fragment) for chance

By looking at the concordance lines one can see that chance is always preceded by scoring which is in turn preceded by best.

I have attempted to show that terms are extended by patterns such as collocation and colligation and, more often than not, become part of a phraseological unit.

Lastly, our findings suggest that a phraseological analysis is a promising approach to terminological extraction because it will identify patterns demonstrating that terms usually occur in context rather than as isolated terms living a life of their own. The phraseological units which were identified are made up of fixed multiword strings, such as marcou um gol contra and scored an own goal, which represent the current use of the term, thus, emphasizing its discursive characteristics.
References


