The Multilingual Affective Soccer Corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch

Nadine Braun (n.braun@uvt.nl)
Martijn Goudbeek (m.b.goudbeek@uvt.nl)
Emiel Krahmer (e.j.krahmer@uvt.nl)

Tilburg center for Cognition and Communication (TiCC), Faculty of Humanities, Tilburg University,
PO Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

The emergence of the internet has led to a whole range of possibilities to not only collect large, but also highly specified text corpora for linguistic research. This paper introduces the Multilingual Affective Soccer Corpus. MASC is a collection of soccer match reports in English, German and Dutch. Parallel texts are collected manually from the involved soccer clubs' homepages with the aim of investigating the role of affect in sports reportage in different languages and cultures, taking into account the different perspectives of the teams and possible outcomes of a match. The analyzed aspects of emotional language will open up new approaches for biased automatic generation of texts.

1 Introduction

Sports reportage provided by sports clubs themselves is one of the most interesting registers available for linguistic analyses of emotionally charged language. It opens up a lot of room for creative language use, starting with the headlines of the match reports (Smith and Montgomery, 1989). Another reason is that the point of view of the author of a match report is clearly definable from the beginning, as it is either a reaction to a tie (that might still be perceived as a net loss or win by the team) or, depending on the perspective, a loss or a win for the soccer club. So, it is easy to assume that the different possible outcomes of such a match would also produce different match reports in terms of language and communicated emotion. Take for example the following introductory sentences:

(1) "Peterborough United suffered a 2-1 defeat at Burton Albion in Sky Bet League One action and lost defender Gabi Zakuani to a straight red card during a nightmare spell at the Pirelli Stadium, but what angered all connected with the club happened in the final moments of the encounter." (PB220815, MASC, 2016)

Compared to:

(2) "If all League One games at the Pirelli Stadium this season are going to be like this it is going to be an entertaining if nerve jangling season." (BA220815, MASC, 2016)

Both describe the exact same match and happenings, but the emotional nuances are completely different. While the match resulted in a loss for the British club Peterborough United, as evident in quote (1), it turned out to be a win for Burton Albion, see quote (2). This results in very different emotions shining through in the corresponding texts: while all the frustration for Peterborough seems to be piled up in a long first sentence already ("suffer... a defeat", "nightmare spell", "anger"), the winners' text is shorter and much more positive ("entertaining").

Knowing about these and other differences that occur in biased sports reporting would be especially valuable for automatic generation of natural language. NLG can be and is currently applied in many different ways, ranging from photo captions (Feng and Lapata, 2010) to neonatal intensive care reports (Portet et al., 2009) and narrative prose (Callaway and Lester, 2001). Bateman and Paris (1989) stress the importance of tailoring machine generated language to the needs of the intended audience. Taking

Proceedings of The 9th International Natural Language Generation conference, pages 74–78,
Edinburgh, UK, September 5-8 2016.

this one step further, Hovy (1990) describes how considering different perspectives on the same event, by taking into account the speaker's emotional state, rhetorical, and communicative goals, is crucial for generating suitable texts for different hearers. Several companies worldwide already offer automatically generated narratives based on databases, e.g. Automated Insights (USA) or Arria NLG (UK). However, the reality of automatic text generation is that not many NLG systems are able to adapt to the mood of the recipients of the produced text (Mahamood and Reiter, 2009) and to convey the mood of the author. While this may not be a problem if simple data-to-text output is the aim of the system, Portet et al.'s (2009) study shows that there are indeed situations that call for a more emotionally informed approach.

To find out more about the emotional language in texts that are produced in negative and positive emotional states, the Multilingual Affective Soccer Corpus (MASC) was compiled and will be analyzed for several aspects of the relation between emotion and written language production in three different languages. To our knowledge, nothing similar to MASC exists at the moment. There is a variety of studies concerned with emotional language (e.g. Stirman and Pennebaker, 2001) and studies that mainly deal with sports reportage (e.g. Müller, 2007), but none of the existing ones includes a complete corpus of parallel texts of the same event from two different perspectives over a whole season in three different languages. This paper introduces this new corpus and highlights possible uses and advantages. MASC is available to interested researchers on request.

2 Building MASC

The corpus includes match reports in (British) English, German and Dutch and was compiled manually, with the texts being copied directly from the individual participating clubs' homepages. This means that the texts are the official reports endorsed by the clubs which are published shortly after the matches have taken place. The overall corpus comprises the 121 different clubs (See Tab.1) which participate in the first and second league in their respective countries. This includes the British Sky Bet League 1 and 2 (UK 1/2), the German Bundesliga 1 and 2 (GER 1/2) as well as the Eredivisie and the Jupiler League (NL 1/2) in the Netherlands (Tab.1).

2.1 Data Collection

Depending on the websites, the match reports are either linked by the clubs themselves as such in the "fixtures and results" tables, in which case those texts were chosen and saved, or the individual reports have to be located in the respective news archives.

In some instances, reports were missing for individual matches. Those cases are marked as "not available (n.a.)" in the metadata files. As the perspectives on those unavailable matches cannot be compared later on, they might be disregarded in the actual analysis. In the affected matches, the counterparts to the missing texts are still included in the dataset.

League	Time Frame 2015/16
Bundesliga 1 (GER 1)	14.08.2015 – 14.05.2016 34 game days 18 clubs
Bundesliga 2 (GER 2)	14.08.2015 – 14.05.2016 34 game days 18 clubs
Sky Bet League 1 (UK 1)	08.08.2015 – 08.05.2016 46 game days 24 clubs
Sky Bet League 2 (UK 2)	08.08.2015 – 07.05.2016 46 game days 24 clubs
Eredivisie (NL 1)	07.08.2015 – 08.05.2016 34 game days 18 clubs
Jupiler (NL 2)	07.08.2015 – 29.04.2016 38 game days 19 clubs

Table 1: Overview: soccer season 2015/16 (UK, GER, NL)

The reports are saved as plain text files in UTF-8 coding in separate folders according to which subcorpus and category (WIN, LOSS, TIE) they belong to. The metadata for the three main subcorpora is split into three separate files. These tables contain the names of the text files, the clubs' and the opponents' names, the dates the matches actually took place, the outcomes from the respective clubs' perspectives and the date the club homepages were accessed. They also include basic information about the subcorpus, like average lengths or number of texts in the conditions.

As of now, MASC includes the written reports themselves, meaning that (elementary) statistics on the match, match photos etc. are not part of the corpus.

3 Descriptive Statistics

This description will present observations about the completed corpus, including the whole season 2015/16 in the three aforementioned countries. It contains an overall of 2,916,265 tokens (Tab.2). MASC can be divided into different subcorpora, either according to language, league or outcome. Differentiating between the three languages, 1,515,442 tokens are part of the British subcorpus, while 803,793 belong to the German and 597,030 to the Dutch part (Tab.3).

	UK 1	**UK 2**	**GER 1**	**GER 2**	**NL 1**	**NL 2**
WIN	410	414	233	221	231	257
LOSS	409	413	232	221	232	253
TIE	272	284	143	171	145	145
Texts	**4,686**					
Tokens	**2,916,265**					

Table 2: Number of texts and tokens

In general, the corpus includes 4,686 reports (Tab. 2). The difference in numbers between WINs and LOSSes as well as the uneven number of TIEs is caused by not available texts, which could not be collected and are therefore left aside in the final calculations. The substantially greater numbers of participating clubs and game days result in almost twice as many texts in the British leagues compared to the Dutch or German ones (Tab.2). This is also one reason for the significantly higher number of tokens in the English subcorpus.

Table 3 provides a first impression of the average lengths of the match reports, which might be an interesting factor for NLG. There are clear differences (or preferences) not only between the three languages, but also the competitions themselves and the outcomes. The shortest texts throughout all languages and leagues by far are the Dutch match reports, which fall short of the English and German ones by about 200 tokens on average. The shortest Dutch report comprises only 24 tokens (Tab.3, *NL 1*) in total. Compared to this, the shortest texts in the other first leagues of the other countries are at least about four times as long. Furthermore, reports describing WINs are, on average, longer than reports describing LOSSes or TIEs throughout all languages and leagues. The length of the reportage on tied or lost matches, on the other hand, varies slightly across leagues and languages (Tab.3).

Besides text length and emotion words, which have already been mentioned in examples (1) and (2) in the introductory part of this paper, shift of focus is another interesting aspect that we observe in the texts in the different conditions. For example, consider the following excerpts that have been selected from several possible alternatives in the corpus:

(3) "Pijnlijke nederlaag Ajax bij FC Utrecht (…) Ajax kreeg de bal niet uit het eigen strafschopgebied, waarop de middenvelder venijnig uithaalde: 1-0." (AX131215, MASC, 2016)

(4) "FC Utrecht wint van Ajax (…) Het is dat ene balletje waarvan je 86 minuten lang hoopt dat-ie valt. En drie minuten voor tijd gebeurt dat." (FCU131215, MASC, 2016)

The texts again describe the same match, but they stress different details. While the loss is an "embarrassing defeat" for league leader Ajax ("pijnlijke nederlaag"), the win for Utrecht triggers pride and happiness ("the one thing you've been hoping for all 86 minutes long"). Following example (4), we can find a detailed account of the winning goal. For Ajax, on the other hand, the short mention of the deciding goal in example (3) is preceded by a detailed account of the teams' (unsuccessful) defense. So, the focus shifts according to the author's affiliation. However, emotions and focus shift do not only show in reports of decided

	UK 1	**UK 2**	**GER 1**	**GER 2**	**NL 1**	**NL 2**
Shortest	290	87	294	201	24	39
Longest	1,798	1,634	1,261	1,350	986	1208
TOTAL	**1,516,876**		**803,793**		**597,035**	
WIN	757.87	674.31	723.48	658.21	473.03	509.08
LOSS	708.19	632.50	704.85	568.85	443.68	456.30
TIE	717.85	631.77	689.66	599.69	483.52	477.78
MEAN	**688.21**		**658.31**		**472.71**	

Table 3: (Average) text lengths in the MASC subcorpora and conditions

matches. Examples (5) and (6) are taken from texts about an, again, randomly selected tied match.

(5) "Der 1. FC Nürnberg verliert in der Nachspielzeit zwei wichtige Punkte." (FCN171015, MASC, 2016)

(6) "Der FSV Frankfurt sichert sich einen Punkt in Mittelfranken" (FSV171015, MASC, 2016)

As we can see, both clubs perceive the tie differently – for the FCN in example (5), it is a lost match because the club "loses points ("verliert… Punkte"), while the FSV in (6) thinks of the outcome as a WIN ("sichert sich einen Punkt") as they "secure a point". This means that TIEs are usually also perceived as lost or won matches and might even trigger the same emotional response in both teams (LOSS/LOSS or WIN/WIN). So far, the mentioned aspects of match reportage seem to appear in all three languages.

4 Discussion

In this paper, we introduced MASC as a new text collection for linguistic research aimed at improving biased output of NLG systems across different languages. English, German and Dutch might be similar and from the same language family, but the realization of emotions in a text is not only a matter of linguistic preferences, but also rooted in the respective soccer culture. This is why – even though close in geographic and linguistic proximity – the way emotions are expressed and the emotions themselves (e.g. excitement, disappointment, shame, happiness etc.) in the conditions may vary more than the similarity in languages would imply.

As a first step towards analyzing the corpus for emotional language, we will use the text analysis program LIWC (Pennebaker et al., 2001). For example, LIWC can help to determine the proportions of negative and positive emotion words, such as "defeat" in example (1) or "entertaining" in example (2). It can even be expected that the soccer culture differences in the three countries in question are significant enough to also shine through in the language of the match reports. The corpus will help to contribute to the understanding of how different emotional states influence and change written language production. After MASC has been completed, we are planning a detailed descriptive analysis on surface features, such as already indicated text lengths and emotion words, as well as a more in-depth analysis of, for example, referential expressions and pronouns. Further, an analysis of the preferred pronouns or referential items in general can be carried out. By analyzing the pronouns, it is possible to ascertain the focus of the author in the respective outcome of the game. If the match results in a WIN, does the report focus on the own team's great performance or on the opponent's failure ("us vs. them")? Does even the perspective on one's own team change ("we vs. they")? Or, in case of a LOSS, are the positive aspects of the game for the own team highlighted or rather the superiority of the other team? Additionally, we plan to investigate whether there are linguistic features that are related to the affect present in the texts – for example, whether certain grammatical constructions occur more in positive or negative contexts. For instance, Beukeboom and Semin (2006) suggest that abstract language correlates with positive affect.

Besides looking at potential effects of emotional state on language production, we also want to investigate how authors select game events for their reportage. For this purpose, we plan to collect game statistics for all games in MASC, to see which events are realized in the respective reports, and whether there is any bias in this selection procedure. This could also provide useful information about how game events are generally expressed in language, which is helpful for the development of new NLG applications.

These are some of the research questions that we seek to answer with MASC. As indicated before, the corpus is available on request.

Acknowledgements

We received financial support for this work from The Netherlands Organization for Scientific Research (NWO), via Grant PR-14-87 (Producing Affective Language: Content Selection, Message Formulation and Computational Modelling), which is gratefully acknowledged. We benefitted from discussions with the members of the Tilburg Language Production group, and with Charlotte Out in particular.

References

Bateman, J. A., & Paris, C. (1989, August). Phrasing a text in terms the user can understand. In *IJCAI* (pp. 1511-1517).

Beukeboom, C. J., & Semin, G. R. (2006). How mood turns on language. *Journal of experimental social psychology*, *42*(5), 553-566.

Callaway, C. B., & Lester, J. C. (2002). Narrative prose generation. *Artificial Intelligence*, *139*(2), 213-252.

Feng, Y., & Lapata, M. (2010, July). How many words is a picture worth? Automatic caption generation for news images. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1239-1249). Association for Computational Linguistics.

Hovy, E. H. (1990). Pragmatics and natural language generation. *Artificial Intelligence*, *43*(2), 153-197.

Mahamood, S., & Reiter, E. (2011, September). Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 12-21). Association for Computational Linguistics.

Müller, T. (2007). *Football, language and linguistics: time-critical utterances in unplanned spoken language, their structures and their relation to non-linguistic situations and events* (Doctoral dissertation, The University of Sheffield).

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7), 789-816.

Smith, M. K., & Montgomery, M. B. (1989). The semantics of winning and losing. *Language in Society*, *18*(01), 31-57.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*(4), 517-522.

MASC. (2016). AX131215.
BA220815.
FCU131215.
FCN171015.
FSV171015.
PB220815.