Lang Resources & Evaluation (2021) 55:389–430
https://doi.org/10.1007/s10579-020-09508-2

ORIGINAL PAPER

MEmoFC: introducing the Multilingual Emotional Football Corpus

Nadine Braun[1] · **Chris van der Lee**[1] · **Lorenzo Gatti**[2] · **Martijn Goudbeek**[1] · **Emiel Krahmer**[1]

Accepted: 24 September 2020 / Published online: 16 October 2020

Abstract This paper introduces a new corpus of paired football match reports, the Multilingual Emotional Football Corpus, (MEmoFC), which has been manually collected from English, German, and Dutch websites of individual football clubs to investigate the way different emotional states (e.g. happiness for winning and disappointment for losing) are realized in written language. In addition to the reports, it also contains the statistics for the selected matches. MEmoFC is a corpus consisting of comparable subcorpora since the authors of the texts report on the same event from two different perspectives—the winner's and the loser's side, and from an arguably more neutral perspective in tied matches. We demonstrate how the corpus can be used to investigate the influence of affect on the reports through different approaches and illustrate how game outcome influences (1) references to the own team and the opponent, and (2) the use of positive and negative emotion terms in the different languages. The MEmoFC corpus, together with the analyzed aspects of

✉ Nadine Braun
N.Braun@uvt.nl

Chris van der Lee
C.vdrlee@uvt.nl

Lorenzo Gatti
L.gatti@utwente.nl

Martijn Goudbeek
M.B.Goudbeek@uvt.nl

Emiel Krahmer
E.J.Krahmer@uvt.nl

[1] Tilburg Center for Cognition and Communication (TiCC), Tilburg University, 5037 AB Tilburg, The Netherlands

[2] Human Media Interaction Lab, University of Twente, Zilverling 2082, 7500 AE Enschede, The Netherlands

emotional language will open up new approaches for targeted automatic generation of texts.

Keywords Affect · Emotion · Multilingual corpus · Comparable corpora · Natural language generation · Sports · Reportage

1 Introduction

This paper introduces the Multilingual Emotional Football Corpus (MEmoFC),[1] a new corpus consisting of pairs of football reports, which can be used for the study of affective language. We present the text corpus in three languages, English, Dutch, and German, combined with the matching football game statistics, as a resource for investigating how (affective) perspective can change reporting about an event. To the best of our knowledge, this multilingual corpus is the first one where objective data and textual realizations from multiple affective perspectives are systematically combined.

Sports reportage provided by sports clubs themselves is arguably one of the most interesting registers available for linguistic analyses of affect-laden language from different perspectives. It opens up room for creative language, starting already with the headlines of the match reports (Smith and Montgomery 1989). Additionally, 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 football club. So, it seems reasonable to assume that the different possible outcomes of such a match would also produce different match reports in terms of language and affect. 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, MEmoFC). 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, MEmoFC).

Both describe the exact same match and events, but the affective nuances are completely different. The match resulted in a loss for the British club Peterborough United, as evident in the first example, whereas it turned out to be a win for Burton Albion in the second example. This results in very different affective states shining through in the corresponding reports: while all the frustration of Peterborough

[1] After a first presentation of the corpus as MASC (Multilingual Affective Soccer Corpus), the name of our corpus was changed to avoid future confusions with the Manually Annotated Sub-Corpus (also MASC), a subcorpus of the Open American National Corpus (https://www.anc.org/data/masc/corpus/).

seems to be released in a long first sentence (*suffer… a defeat*, *nightmare spell*, *anger*), the winners' text is shorter and much more positive (*entertaining*).

In this paper, we describe how the corpus was collected and preprocessed, we give an overview of properties of the corpus, and we explore it with regard to linguistic differences and similarities related to affect in reports about won, lost, and tied matches in English, German, and Dutch using different tools. In the remainder of this introduction, we position the corpus more broadly in the research field studying the influence of emotion on language, and link it to applications in sentiment analysis and affective natural language generation.

1.1 The psychology of language and emotion

It is a general assumption that a text reflects the affective state of the author. Writing a text involves various cognitive processes, and it is commonly believed that affective states influence these cognitive states, and, hence, that they can have a noticeable effect on the resulting text. This idea has been put forward in psychological theories, such as, for example, Forgas' Affect Infusion Model (1995), which describes how affective states, while seen as different from cognitive processes, "interact with and inform cognition and judgments by influencing the availability of cognitive constructs used in the constructive processing of information" (Forgas 1995, p. 41). Affect infusion is characterized as "the process whereby affectively loaded information exerts an influence on and becomes incorporated into the judgmental process, entering into the judge's deliberations and eventually coloring the judgmental outcome" (Forgas 1995, p. 39). In this study, we aim to investigate whether the influence that affective states (due to winning or losing) exert on cognition extends to language production.

A limited number of psychologists have studied the role of affect on language. Perhaps most notably, Forgas and colleagues found that the affective state influences the politeness of requests, with people in a negative state being more polite (Forgas 1999, 2013; Forgas and East 2008; Koch et al. 2013). In addition, Beukeboom and Semin (2006) found that people in a negative state used more concrete language, in terms of the Linguistic Category Model (Semin and Fiedler 1991), while people in a positive mood used relatively more abstract descriptions.

Many of these psychological studies relied on controlled experiments and small amounts of manually annotated data. To facilitate and speed up these kinds of studies, Pennebaker et al. (2001) developed an automatic tool for assessing texts in terms of different psychological and linguistic categories, including terms related to valence and emotions: the Linguistic Inquiry and Word Count (LIWC). LIWC is a bag-of-words technique that counts words belonging to one or more categories in its dictionary and converts those frequencies to percentages of all relevant words in the text. It has several attractive properties: its emotion word categories and the associated word lists have been validated through human evaluation (Tausczik and Pennebaker 2010), LIWC can be used with arbitrary datasets and requires no preprocessing of the input texts. As a result, LIWC has been used in a large number of psychological studies (Cohn et al. 2004; Pennebaker and Graybeal 2001; Rude et al. 2004; Stirman and Pennebaker 2001) and NLP studies (e.g., Mihalcea and

Strapparava 2009; Nguyen et al. 2011; Strapparava and Mihalcea 2017). For example, in a study on language and depression, Rude et al. (2004) analyzed the language of depressed, formerly-depressed, and never-depressed students and found that, as one would expect, depressed participants used more negatively valenced words, but also, perhaps less expected, used the pronoun "I" more frequently than never- and formerly-depressed students. A similar study was conducted on poems written by suicidal and non-suicidal poets (Stirman and Pennebaker 2001), which confirmed the use of the first person singular as related to negative mood. Text analysis, particularly online, for depression detection has been gaining popularity (see, e.g., Morales et al. 2017, or Losada and Gamallo 2020) with potential applications for mental health, such as early depression detection, treatment, and suicide prevention.

While these studies are indicative of a link between affect and language, most of them focus on less ecologically valid settings (such as the laboratory), use questionnaires or focus on disorders like depression. One can ask how such findings translate to the natural settings outside the laboratory. A study directly addressing this question is Baker-Ward et al. (2005), who analyzed spoken reports of young football players after their final match of the season. They found that the players in a positive state (i.e. winners) produced descriptions of the game that were clearer and more cohesive, while the players in a negative state (the losing players) described the game more interpretatively.

Interestingly, these findings connect to an early study conducted by Hastorf and Cantril (1954), which deals with how different perspectives on a football game between Princeton and Dartmouth influenced viewers' perceptions of the game itself. While Princeton students mostly agreed that the game was played "rough and dirty" by Dartmouth, who ultimately lost the game, and saw more flagrant infractions, the majority of Dartmouth students saw it as "rough and fair" and blamed the roughness on both teams. While this study nicely illustrates how perceptions of events, and, in a way, events themselves may differ according to one's point of view, the precise language used to describe the match was unfortunately not investigated in this study.

Cialdini et al. (1976), however, did investigate language use in relation to success and failure. In three experiments, they demonstrated how individuals involved themselves in victories of (groups of) other people, without having a direct influence on the victory. For example, they suggested that when students where they asked about wins and losses of their own university's team, successful matches were described with significantly more use of the pronoun *we* than lost matches were. This phenomenon of identifying with winners was coined BIRGing ("Basking In Reflected Glory"). Snyder et al. (1986) showed the opposite effect in behavior for failures and coined it CORFing ("Cutting Off Reflected Failure"). While these tendencies of people to bask and distance themselves have been replicated repeatedly (Downs and Sundar 2011; Wann and Branscombe 1990), whether and how these tendencies emerge in language production has not been systematically explored.

1.2 Natural language generation (NLG) and Natural language processing (NLP)

Psychological studies, just as described, have revealed that affective state can influence language production. However, most of these studies only focused on one specific aspect such as politeness or abstractness. Moreover, with the exception of the work done with LIWC, all of these mentioned studies approach the influence of affect on language production experimentally. However, in recent years, there has been a growing interest in more comprehensive studies into emotion and language production, typically using computational approaches. Here, we highlight two: sentiment analysis and affective natural language generation.

Natural language generation (NLG) is the process of converting data into text (Gatt and Krahmer 2018; Reiter and Dale 2000), with applications in, for example, automatic generation of texts for sensitive matters such as neonatal intensive care reports based on medical data (Mahamood and Reiter 2011; Portet et al. 2009), but also automatic generation of photo captions (Chen et al. 2015; Feng and Lapata 2010; Kuznetsova et al. 2012), which can be tailored to the needs of people with visual impairments, or sports commentary (Lee et al. 2014; van der Lee et al. 2017). Bateman and Paris (1989) stress the importance of tailoring machine generated language to the needs of the intended audience. Taking 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 addressees. Several companies worldwide already offer automatically generated narratives based in 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 2011) 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. In general, tailoring automated text to an intended audience especially with regard to sentiment still poses a challenge to whose solution MEmoFC can contribute, for example, in enabling the tailoring of reports specifically to the perspective and affective states fans of specific clubs after specific game outcomes.

Of course, to be able to do this, we need to know how affective state could influence text production, not only concerning the factors studied by psychologists (politeness, abstractness, etc.) but in all aspects of language production. Sentiment analysis can provide valuable clues in this respect. Sentiment analysis, or stance detection, can be characterized as a classification of texts, for example, the labeling of positive versus negative online reviews to capture sentiments and attitudes towards specific topics, brands, and products, which has become a crucial task in recent years (Glorot et al. 2011; Kim 2014; Ravi and Ravi 2015; Socher et al. 2013). Social network sites like Facebook and Twitter have been used to extract opinions and sentiment on a large scale, for example, with a focus on brands or political elections (dos Santos and Gatti 2014; Ghiassi et al. 2013; Isah et al. 2014; Pak and Paroubek 2010; Pang and Lee 2008; Tumasjan et al. 2011).

While most work on stance detection and sentiment analysis has focused on English corpora, there has also been work on other languages, see, e.g., Basile (2013) or Bosco et al. (2013), for work on Italian, more recently, Tsakalidis et al. (2018) for resources in Greek, or on informal and scarce languages (Lo et al. 2017). Increasingly, work is also being done to apply sentiment analysis techniques from English to other, less researched languages automatically, using, for example, machine translation techniques (e.g., Perez-Rosas et al. 2012, or Bautin et al. 2008). These kinds of studies can be informative about which words and phrases are associated with which particular emotional states. Yet, while these approaches are promising, they often still rely on training material in the less-researched languages, for which limited resources are available (at least compared to English).

1.3 The current studies

This paper introduces the MEmoFC corpus, a multilingual, large-scale corpus of soccer reports, which is unique in that it contains pairs of reports for each match, one for each team participating in the match, combined with the original game statistics. In this way, MEmoFC offers controlled (in terms of the source of the events described) yet natural emotionally varied descriptions of the same events. This makes it an attractive resource to study the effect of affect and perspective on language, which, in turn, paves the way for tailoring automatically generated texts to a specific audience.

In this paper, we describe how we constructed and preprocessed the MEmoFC corpus, and we present descriptive statistics for it. MEmoFC can be used to address many different research questions, but to illustrate its potential and evaluate its use as a source for affective science, we perform three example studies:

Example Study 1: Do we see more linguistic indicators of basking behavior in the reports after won matches than after lost ones?

As we have described above, earlier studies have suggested that basking occurs more after winning than after losing (Cialdini et al. 1976). We ask whether this is indeed the case by investigating whether writers in the different languages use the pronoun *we* more often after winning than after tying or losing.

Example Study 2: Which words and phrases are typical for the different game outcomes, and does this differ per language?

We expect the affective states of the authors to be reflected in their lexical choices and possibly also in other linguistic features such as grammar or punctuation (e.g., Stirman and Pennebaker 2001; Hancock et al. 2007). Here, we ask which words and phrases are actually frequently used for specific game outcomes, and whether this differs between the different languages under investigation.

Example Study 3: Can we classify texts as describing a win or loss (and does this vary per language) and which textual elements of the reports are most indicative of the game outcome?

Assuming that different winning and losing reports express different emotions with different language depending on the game results, we ask whether this

Table 1 Example excerpt of the metadata file from the English subcorpus of MEmoFC

Club	FILE	OPPONENT	DATE OF GAME	WIN	LOSS	TIE	WORD COUNT	DATE OF ACCESS	XML FILE	XML DATE OF ACCESS
Leicester City	LC080815	Sunderland	08.08.2015	1	0	0	822	24.11.2016	LC_SL_08082015_goal	23.02.2017
Leicester City	LC150815	West Ham United	15.08.2015	1	0	0	1086	24.11.2016	LC_WHU_15082015_goal	23.02.2017
Leicester City	LC220815	Tottenham Hotspur	22.08.2015	0	0	1	762	24.11.2016	LC_TH_22082015_goal	23.02.2017
Leicester City	LC290815	AFC Bournemouth	29.08.2015	0	0	1	1011	24.11.2016	LC_AFCB_29082015_goal	23.02.2017
Leicester City	LC130915	Aston Villa	13.09.2015	1	0	0	1233	24.11.2016	LC_AV_13092015_goal	23.02.2017
Leicester City	LC190915	Stoke City	19.09.2015	0	0	1	949	24.11.2016	LC_SC_19092015_goal	23.02.2017

Table 2 Example excerpts of matched reports from the English subcorpus of MEmoFC

CU260915 (WIN)	SW260915 (LOSS)
The U's have beaten Swindon Town 2-1 to extend their unbeaten run to five games	SWINDON slipped to a third defeat in as many games, going down 2-1 at home to Colchester United
It was a third win in succession, with George Moncur and Callum Harriott getting the goals to extend	The damage was done in the first half-despite Wes Thomas levelling matter before the break, the U's came away with their third consecutive win
Moncur's came as early as the third minute and Harriott made it 2-1 after thomas had equalised	The major change to Mark Cooper's line-up came in goal, where Tyrell Belford was drafted in as a starter with Lawrence Vigouroux dropping to the substitutes' bench. Skipper, Nathan Thompson missed out with a groin injury so Yaser Kasim was his replacement, with Town lining up in a 3-5-2 formation
The win took the U's up to tenth in the League One table, ahead of Tuesday evening's game against Bradford City	Colchester hit the front with just three minutes on the clock. George Moncur, son of Town legend, John, met a low cross from the left and swept the ball into the back of the net. (…)
The U's were on the offensive from the first whistle and left back Matt Briggs had already got forward a couple of times before he created the opening goal in the second minute. (…)	

Table 3 Information (general, match events, last game, players, substitutes, and managers, last five games, relative strength, match statistics) about MEmoFC statistics stored in the XML files

Type	Match information
General	League, date, time, stadium, city, referee, attendees, final score, teams, goal scorers
Match events	Assists, regular goals, own goals, penalty goals, penalty misses, yellow cards, red cards (2× yellow), red cards (direct)
Last game	League, date, opponent, final score, played home/away, won/tied/lost, changes in lineup
Players in lineup, substitutes and managers	Name, full name, nickname, birth date, birth place, height, weight, position, kit number, name in *Goal.com*, *Goal.com* player page, youth clubs, senior clubs, national teams represented, current team
Last five games	Opponent, final score, played home/away, won/tied/lost
Relative strength	Wins per team for previous meetings, draws in previous meetings, percentage of people predicting win for the home team/win for the away team/tie, date of previous meetings, which team played home/away in previous meetings, final score previous meetings, most predicted results
Match statistics	Total shots, shots on target, completed passes, passing accuracy, possession, corners, offsides, fouls, total passes, short passes, long passes, forward/left/right/back passes, percentage of forward passes, blocked shots, shots on the left/right/center of the goal, percentage of shots outside the 18-yard box, total crosses, successful crosses, crosses accuracy, crosses inside/outside 18-yard box, left crosses, right crosses, total attempted take-ons, successful take-ons, successful left/right/center/total take-ons in the final third of the match, blocks, interceptions, clearances, recoveries, total tackles, successful tackles, tackle accuracy

knowledge can be used to classify reports; in other words, to what extent can we tell, based on the language, whether a game was won or lost.

The corpus is publicly available for research purposes upon request (https://doi.org/10.34894/07ROT3).

2 Construction of the corpus

2.1 Texts in MEmoFC

The reports in the corpus were manually collected, saved directly from the homepages' archives, and have not been cleaned (typographic errors, wrong grammar, layout etc.). MEmoFC is multilingual in that it contains reports from three languages: English, German, and Dutch. The linguistic subcorpora are further divided into *WIN*, *LOSS* and *TIE*, which are, in turn, distinguished by league (first/second [+ third for the UK]). There are two metadata tables per language: one explaining the abbreviations for the different football clubs and one that allows the identification of the two participating teams of a match, the file name, outcome (win, loss, tie), the date the match took place, and the date the archive of the respective homepage was accessed. An example excerpt from the English metafile can be found in Table 1. Due to the multitude of participating football clubs, possible influences of individual authors' writing styles on the language employed for the text are reduced, which makes it possible to draw more general conclusions for the genre from analyses.

In addition to the written reports, MEmoFC also contains the corresponding match statistics (see Sect. 2.2). The original files are saved in UTF-8 coding and have not been annotated, parsed or PoS-tagged, meaning the texts are exactly how they appeared on the homepages of the clubs right after the matches took place.

With the help of the metadata and the consistently named files as shown in Table 1, the participating clubs and outcomes are easily identifiable, and the matching reports can be aligned and analyzed contrastively. Table 2 illustrates how text excerpts from the corpus are loosely aligned, ending at the same event in the game. Displayed are the two sides of a match that took place on the 26th September 2015 in the British first league. The reports themselves, of course, differ in length and game events described.

2.2 Game statistics

The statistics for the relevant matches were automatically scraped from Goal.com, a website that provides information and content about football. Finding and mining these statistics was done using three modules. First Google queries designed to find pages from *Goal.com* were activated to find the corresponding statistics for each match in MEmoFC. After the corresponding *Goal*.com pages were found, the data that was stored on these pages were mined and, finally, converted to an XML-format. Each XML-file provides data about a football match in MEmoFC. These

Table 4 Overview of football season 2015/2016

League	Season 2015/2016
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
Premier League (UK 1)	08.08.2015–17.05.2016 38 game days 20 clubs
Sky Bet League 1 (UK 2)	08.08.2015–08.05.2016 46 game days 24 clubs
Sky Bet League 2 (UK 3)	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

files contain general-level information as well as more detailed information (see Table 3).

2.3 Descriptive statistics of MEmoFC

The corpus covers between 34 and 46 game days in approximately the same time frame (August 2015 until April/May 2016) in all countries (Table 4). Table 5 shows the difference in text and token numbers: UK1, UK2 and UK3 contain more than twice as many reports as GER and NL. Unfortunately, some of the reports were untraceable on the websites, either because they were removed or never written for individual matches. This concerns 64 reports throughout all leagues and languages, which encompasses just 1.18% of the whole corpus. These matches have been marked *n.a./not available* in the metafiles. Due to the proportion of missing texts being small, these do not cause a significant imbalance in perspective. Hence, we did not treat them as problematic missing data in the exploration of the corpus. Although these missing matches are mentioned in the metafiles, their reports are not counted in Table 5 and Fig. 1. This means that the numbers in Table 5a, b solely result from the texts actually available in MEmoFC, which explains the differences in numbers between wins and losses, as well as the uneven numbers of ties. The corpus now contains 5434 texts, which add up to about 3.5 million tokens, with

Table 5 (a) Number of texts (Txt) and words (W) in MEmoFC by League and Country (1–3 in the UK; 1 and 2 in Germany; 1 and 2 in the Netherlands); (b) Average text length and words per sentence (WPS) in MEmoFC by League and Country (1–3 in the UK; 1 and 2 in Germany; 1 and 2 in the Netherlands)

(a)

	UK1		UK2		UK3		GER1		GER2		NL1		NL2	
	Txt	W	Txt	W	Txt	W	Txt	W	Txt	W	Txt	W	Txt	W
WIN	269	204,669	410	310,727	414	279,166	233	167,125	221	146,123	231	109,269	257	130,833
LOSS	269	182,702	409	289,650	413	261,221	232	164,229	221	125,146	232	102,934	253	115,445
TIE	210	154,581	272	195,256	284	179,422	143	98,622	171	102,548	145	70,111	145	68,443
Total	748	541,952	1091	795,633	1111	719,809	608	429,976	613	373,817	608	282,314	655	314,721
Subcorpus Txt	2950						1221				1263			
Subcorpus W	2,057,394						803,793				597,035			
Overall Txt	5434													
Overall W	3,458,222													

(b)

	UK1		UK2		UK3		GER1		GER2		NL1		NL2	
	Text length (mean)	WPS	Text length (mean)	WPS	Text length (mean)	WPS	Text length (mean)	WPS	Text length (mean)	WPS	Text length (mean)	WPS	Text length (mean)	WPS
WIN	760.85	29.41	757.87	26.70	674.31	28.29	723.48	14.69	658.21	14.81	473.03	17.02	509.08	18.44
LOSS	679.18	28.81	708.19	26.09	632.50	28.47	704.85	15.11	568.85	14.61	456.30	16.84	456.30	17.88
TIE	736.10	29.27	717.85	25.92	631.77	28.32	689.79	14.77	599.69	14.40	472.02	16.63	472.02	18.21

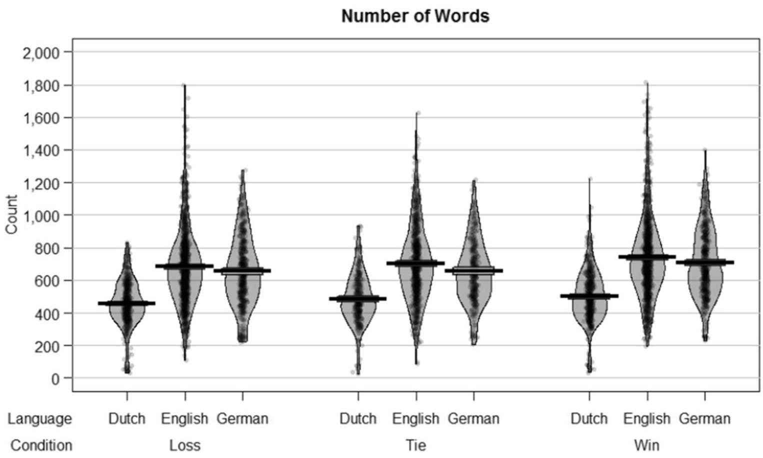


Fig. 1 Distribution of text lengths (words per report) in MEmoFC by language and game outcome

more than half being part of the English subcorpus, 803,793 in the German, and 507,035 in the Dutch subcorpora. The Dutch match reports are the shortest in all conditions, while English and German reports are generally similar in length (see Table 5b). Overall, game outcome seems to have no influence on text length in any of the languages in MEmoFC.

2.4 Parsing and lemmatization

In a next step, the corpus was dependency parsed and lemmatized. For the English and German subcorpora, the Spacy Python library (Honnibal and Johnson 2015) was used. The Dutch subcorpus, was lemmatized by Frog (Bosch et al. 2007) because Spacy does not contain a lemmatizer for Dutch. Dutch multiword expressions were automatically conjoined with an underscore by Frog (e.g., *zijn_binnen* [to be in]). For English and German, phrasal verbs and/or separable prefix verbs were "rejoined" (e.g., *climb up* or *ringen_nieder* [wrestle down]). This way it is possible to differentiate, for example, between *kick* and *kick off*. The preprocessed files can be found in a separate folder.

3 Using the MEmoFC

In this section, we will illustrate the potential of the corpus with three exploratory studies, coming from three angles. In the following subsections, we approach the evaluation of the corpus and show its usefulness as an affective linguistic resource with a variety of different techniques in order to demonstrate the diverse ways in which it can be used for research.

Table 6 Proportions of 1PP (EN: We, Our, Ours; NL: We, Wij, Ons, Onze; GER: Wir, Uns, Unser [& variations]) compared to all tokens in Win, Loss, and Tie (Both Perspectives) in English, Dutch, and German

	ENGLISH	DUTCH	GERMAN
Win	0.57	0.84	0.88
Loss	0.33	0.87	0.50
Tie	0.20	0.26	0.22

3.1 Example study 1: Do we see more linguistic indicators of basking behavior in the reports after won matches than after lost ones?

With regard to language reflecting basking tendencies, the focus was on the use of the first person plural pronoun *we* in the aligned match reports. Following the suggestion of Cialdini et al. (1976), we hypothesize more uses of first person plural pronouns (1PP) in reports on won matches compared to reports on ties or losses. We ask whether this is indeed the case, and whether this is the same across languages.

While analyzing and comparing different types of pronouns with NLP tools would also be interesting, in particular the distribution of 1PP compared to *they* (or third person plural pronouns; 3PP), this task proved to be challenging for two reasons. In German and Dutch, some pronouns are ambiguous (e.g., *Sie* in German can be 3rd person plural, formal 2nd person singular and plural, or 3rd person singular female; *zij* in Dutch can be 3rd person plural or singular). This would require a deeper syntactic analysis to detect plural pronouns. However, even after this step, the pronouns' referents would still be ambiguous: whether the more distant 3PP option is indeed used as a reference to the own team (instead of 1PP) cannot be ensured, since 3PP could refer to wide range of referents, such as the opponent, the fans, or a specific group of players—all of which carry no weight for distancing and basking behavior. Currently, no coreference resolution tool for Dutch and German is easily available. Furthermore, Named Entity Recognition is less accurate on the reports of MEmoFC due to the differences with training data (usually annotated newspaper articles) and to identify players' surnames that are often not present in the gazetteer lists of NER tools, and, hence, not recognized. This issue would have had a substantially negative impact on the accuracy of coreference resolution systems, which is why we opted for a different approach. To answer the question guiding ES1, occurrences of 1PP were counted in the tokenized texts and then divided by the overall number of tokens in the review (to account for the fact that longer reviews are more likely to contain more pronouns in general). Afterwards, the results were summarized for the aligned texts in the win, loss, and tie subcorpora in English, Dutch, and German (see Table 6).

In English and German, we find the expected distribution: there are considerably more occurrences of 1PP in reports about won matches than in losses and ties. For Dutch, however, a reverse trend of more 1PP in loss compared to win is apparent, while the proportion of 1PP in ties is lower than in both loss and win. In the reports on ties, we find the overall lowest proportion of 1PP in English, German, and Dutch, with only minor differences between the languages (highest proportion of 1PP in Dutch). The preference for 1PP after lost matches could be a cultural peculiarity that diffuses in language, exemplifying the usefulness of taking into account different languages when constructing language resources for the study of affect. Although English, German, and Dutch are Germanic languages and the subcorpora were collected from Western European cultures, there might still be cultural differences traceable in the language use, e.g., in linguistic distancing behavior. For the aligned reports on ties, it can be assumed that the outcome is perceived differently by the involved clubs: while in some cases the perception might be more similar to a win, in other cases ties can be closer to losses, which might decrease the proportion of 1PP. Examples supporting the different perspectives can be found in the following excerpts, among others, from two aligned reports on tied matches:

3. “Der 1. FC Nürnberg verliert in der Nachspielzeit zwei wichtige Punkte.“ (FCN171015, MEmoFC).
   “1. FC Nürnberg loses two important points during overtime.”
4. “Der FSV Frankfurt sichert sich einen Punkt in Mittelfranken “ (FSV171015, MEmoFC).

“FSV Frankfurt secures one point in Middle Franconia”.

Examples (3) and (4) show that the involved clubs perceive the tie differently—for the FCN it is a lost match because the club *loses points*, while the FSV considers the outcome a victory as they *secure a point*. This means that ties can be perceived as lost or won matches as well, which could also have an influence on the use of the pronoun 1PP in these reports.

Overall, there are generally more uses of 1PP in German and Dutch reports on won and lost matches compared to English. While the pattern is similar in English and German, there is a different, even opposite trend in Dutch, which could be related to cultural differences and should be taken into account in studies on affect and in automatically produced texts.

3.2 Example study 2: Which words and phrases are typical for the different game outcomes, and does this differ per language?

After exploring the distribution of one particular word (1PP), we now ask which words are associated with winning, losing, and tying in the different languages in general. We perform three kinds of analyses: (1) on word frequencies in general (using TF-IDF and concordances, Subsect. 3.2.1), (2) on LIWC categories, and (3) on specific emotion terms (3.2.2). Names of places, players, teams, or managers were filtered out using name entity recognition with Spacy (https://spacy.io/models/) for English and German, and with Frog (Bosch et al. 2007) for Dutch. In addition to individual words, bi- and tri-grams will be inspected.

3.2.1 TF-IDF and concordance

To extract words and n-grams that are especially representative of the conditions and languages, two approaches were used. First, TF-IDF was calculated for each word in each subcorpus. Table 7 shows the extracted most frequent words after lemmatization. While the word lists in reports on wins, losses, and ties differ in all languages and all lists contain interesting frequent words per outcome (e.g., *ecstatic* in English/Win, *embarrassment* in English/Loss, *ringen_nieder* [wrestle down] in German/Win, *entmutigen* [discourage] in German/Loss, *probleemlos* [without problems] in Dutch/Win, *Punt* [point] in Dutch/Tie), there are also various words on these lists that do not appear to be typical for specific game outcomes. Given the relatively large size of the corpus compared to the small number of categories, TF-IDF may be too sensitive to be conveniently used, and other statistics—such as keyness, which compares two corpora instead of calculating word frequencies in a single corpus or document—appear to be better suited for this analysis.

A keyness analysis looks for keywords that are more likely to appear in a target corpus compared to a reference corpus—or, as is this case, in the differences across the conditions in the language subcorpora: win, loss, and tie. As the frequency of a word alone is not an indication of how specific a word is for a corpus, we calculate the keyness of a word with the freely available concordance tool AntConc (Anthony 2004).

Keyness, which we calculate with a word's log-likelihood ratio (Lin and Hovy 2000), is a measure that enables the extraction of keywords based on their probability to appear in the target corpus compared to the reference corpus and, thereby, can identify the words that stand out and define a text most. The log-likelihood ratio of a word is calculated using a contingency table and takes both corpora's sizes into account (based on Rayson and Garside 2000). First, the expected value (E_i) is calculated; N_i being the number of words in the corpora and O_i the observations of the word frequency in both corpora:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

The log-likelihood ratio G2 is then determined as follows:

$$-2ln\lambda = 2\sum_i O_i ln\left(\frac{O_i}{E_i}\right)$$

The higher the log-likelihood value is, the larger the word frequency difference between the corpora and, hence, the more representative a word is for a subcorpus. In contrast to TF-IDF, keyness calculated with log-likelihood (e.g., using a Chi-square distribution) is also an indication of statistical significance since it does not only calculate the frequency of a word within a document or corpus but directly compares the frequencies in two corpora. The critical threshold for a log-likelihood value (or *keyness*) is 3.84 at the level of $p < 0.05$ and 15.13 at a level of $p < 0.0001$.

Table 7 Most relevant unigrams in English, German, and Dutch for Win (W), Loss (L), and Tie (T) extracted with TF-IDF

	English			German			Dutch		
	W	L	T	W	L	T	W	L	T
1	ecstatic	limitation	fight_out	bezwingt	Bittere	trotzen_ab	probleemloos	te_onder	overheersen
2	rojo	divert_behind	flitcroft	Basis	aufholen	Bohl	ll	doen_om	onbeslist
3	climb_up	aarons	coaching	kräftezehrenden	kalte	Affane	smet	jeugdspeler	onvoldoende
4	my	shortliv	bottle	ringen_nieder	Lucien	Verwarnungen	toet	verzuchten	staan_goed
5	crisply	bramall	ng	2000	Wendung	hinausgekommen	zijn_binnen	aspiratie	Koning
6	irresistible	continue_on	behrami	Derbysieg	entmutigen	verwandeln_ direkt	detail	gifbeker	Punt
7	aloft	dion	bobbly	Geduldsspiel	MSVFCH	447	extase	likken	Toekomst
8	purkiss	dispose	cavalry	Halbzeitstatistik	halten_hoch	Angriffswelle	felbegeerd	ondergrens	Van_Bruggen
9	stay_up	embarrassment	chaplain	Spieltagsfakten	liegend	Aufstiegskampf	flow	spelersbus	beduiden
10	blot	forearm	collect_ down	Zugleich	mussten_hinnehmen	Chancenarme	hey	strijden_te_onder	de_toekomst

In our analyses, the 20 most frequent words in the multilingual subcorpora are determined and presented in Table 7, structured according to language, and target condition compared to a reference condition, e.g., win compared to loss (represented as win–loss) or loss compared to tie (loss–tie).

Table 8 illustrates the top 20 words in English. Besides more obvious words like *win* or *victory* in WIN–LOSS, *defeat*, *lose* and *loss* in loss–win and *draw* in tie–win, frequent positive (*superb*, *clean*, *perfect*, *secure, celebration* [win–loss]; *winner*, *opportunity, rescue, chance* [tie–loss]) and negative words (*condemn*, *disappointing*, *cruel*, *suffer, unable,* frustrating [loss–win]; *unable*, *spoil* [tie–win and tie–loss]) are also apparent. In comparison, the English loss–win list consists of mostly negative words. In the reports on tied matches, the unique focus appears to be on the points earned and more neutral (*both*, *share*, *neither*, *settle*, *goalless*). Additionally, the loss–win list contains a preposition (*despite*) and an adversative conjunction (*but*), which likewise occurs unusually frequent in tie compared to win. Upon closer inspection of the context of the occurrences, it becomes apparent that *neither* is more often used as an adjective than as a conjunction. Thus, the negative game outcome affects not only the lexis but also the cohesive structure of the English texts. In addition, the use of the 1PP is more frequent in reports about won matches, hinting at basking tendencies, in line with 3.1.

The patterns for German and English are comparable. As expected, we also find German words describing the outcome of the matches (see Table 9; *Sieg* [*victory*], *Heimsieg* [*home victory*], *Auswärtssieg* [*away victory*], *siegen* [*win*; WIN–LOSS]; *Niederlage* [*defeat*], *verlieren* [*lose*], *unterliegen* [*be defeated*; LOSS–WIN]; *unentschieden* [*tied*], *Remis* [*draw*; TIE–LOSS and TIE–WIN]). However, there are also differences with the English lists. While the number of positive words in the German WIN–LOSS comparison is similar to English ([*Heim-/Auswärts-*] *Sieg* [*home/away victory*], *hochverdient* [*highly deserved*], *gewinnen* [*won*], *besiegen* [*defeat*], *Erfolg* [*success*], *wichtig* [*important*], *perfekt* [*perfect*], *ungeschlagen* [*unbeaten*], *feiern* [*celebrate*], *Tabellenspitze* [*top of the table*], *endlich* [*finally*]), the keyword list of lost matches in relation to won ones seems less negative in comparison. This is especially due to the relative lack of negative adjectives, the only ones being *bitter* (*bitter*) and *unglücklich* (*unlucky*), and to common euphemisms for goals received, such as *kassieren* (*collect*) and *einstecken* (*pocket*). While in tie–win/loss the emphasis is clearly on the fight and the shared points (*unentschieden* [*tied*], *erkämpfen* [*fight for & secure*], *leistungsgerecht* [*performance-based*], *Remis* [*draw*], *Punkteteilung* [*sharing of points*], *torlos* [*goalless*], *beid** [grammatical variations of the word *both*]), the German ties contain also positive keywords (*zufrieden* [*satisfied*], *ungeschlagen* [*unbeaten*], *Punktegewinn* [*winning of points*], *Chance* [*chance*], *gerecht* [*fair*]). Besides such "emotional" words, we again also find other types of words in the lists. In contrast to the English list, the German one contains the simple additive conjunction *und*, which is significantly more frequent in reports about won matches, for example. We again find an adversative preposition (*trotz* [*despite*]) and conjunction/adverb (*jedoch* [*nevertheless*]) in loss–win, hinting at overall differences in text cohesive structure depending on positive and negative game outcome.

Table 8 Keywords across outcomes (Win, Loss, and Tie) compared, respectively, in English after lemmatization

WIN–LOSS	LOSS–WIN	TIE–WIN	TIE–LOSS
#Types Before Cut: 7929	#Types Before Cut: 7299	#Types Before Cut: 6593	#Types Before Cut: 6593
#Types After Cut: 4989	#Types After Cut: 4603	#Types After Cut: 3970	#Types After Cut: 3970
#Search Hits: 0	#Search Hits: 0	#Search Hits: 0	#Search Hits: 0
1 933 313.912 victory	1 1101 681.987 defeat	1 835 496.747 draw	1 835 496.747 draw
2 1980 301.264 win	2 301 193.626 suffer	2 225 242.577 share	2 225 242.577 share
3 189 129.361 sheet	3 87 100.240 condemn	3 140 184.809 spoil	3 140 184.809 spoil
4 389 120.306 secure	4 115 98.662 disappointing	4 892 131.371 both	4 892 131.371 both
5 1161 104.130 season	5 791 81.194 fall	5 1473 123.012 point	5 1473 123.012 point
6 1465 77.610 point	6 602 80.861 despite	6 184 105.002 goalless	6 184 105.002 goalless
7 231 69.870 clean	7 190 80.492 bad	7 563 81.170 level	7 563 81.170 level
8 430 66.628 superb	8 6991 55.773 but	8 1705 67.563 chance	8 1705 67.563 chance
9 263 59.948 record	9 310 48.418 slip	9 228 62.397 settle	9 228 62.397 settle
10 394 58.717 we	10 332 46.032 lose	10 5130 58.106 but	10 5130 58.106 but
11 305 45.876 earn	11 188 43.736 concede	11 162 47.951 neither	11 162 47.951 neither
12 235 42.765 display	12 294 43.015 unable	12 345 37.298 equaliser	12 345 37.298 equaliser
13 378 42.188 since	13 62 38.914 undo	13 102 36.605 snatch	13 102 36.605 snatch
14 1699 41.308 three	14 456 35.388 equaliser	14 94 34.191 rescue	14 94 34.191 rescue
15 280 39.100 performance	15 31 33.073 cruel	15 2479 33.182 side	15 2479 33.182 side
16 63 38.687 ovation	16 34 33.053 mountain	16 492 32.571 opportunity	16 492 32.571 opportunity
17 237 36.996 perfect	17 82 30.865 frustrating	17 314 29.007 earn	17 314 29.007 earn
18 94 35.582 celebration	18 72 30.679 loss	18 198 27.443 unable	18 198 27.443 unable
19 435 35.050 fan	19 20 29.387 hartlepoolunited	19 270 26.247 winner	19 270 26.247 winner
20 221 34.913 table	20 1651 29.217 find	20 1012 25.559 could	20 1012 25.559 could

“#Types Before Cut” refers to the overall count of unique types of words. “#Types After Cut” indicates the words are considered for the keyword list. Within columns from left to right: rank, frequency, and keyness of the following word

Table 9 Keywords across outcomes (Win, Loss and Tie) compared, respectively, in German after lemmatization

WIN–LOSS	LOSS–WIN	TIE–WIN	TIE–LOSS
#Types Before Cut: 8780	#Types Before Cut: 8164	#Types Before Cut: 6569	#Types Before Cut: 6569
#Types After Cut: 5679	#Types After Cut: 5782	#Types After Cut: 4347	#Types After Cut: 4416
#Search Hits: 0	#Search Hits: 0	#Search Hits: 0	#Search Hits: 0
1 433 193.744 siegen	1 211 268.160 unterliegen	1 161 214.203 trennen	1 211 195.492 unentschieden
2 311 125.122 gewinnen	2 314 223.904 verlieren	2 211 209.892 unentschieden	2 161 189.114 trennen
3 161 84.483 Heimsieg	3 264 133.777 Niederlage	3 257 150.208 Punkt	3 257 125.965 Punkt
4 235 77.576 feiern	4 118 87.483 bitter	4 121 97.563 Remis	4 121 105.910 Remis
5 98 76.437 Auswärtssieg	5 82 69.992 hinnehmen	5 55 87.776 Punkteteilung	5 55 72.989 Punkteteilung
6 7224 73.028 und	6 135 52.880 leider	6 562 45.162 Chance	6 58 60.001 ungeschlagen
7 283 51.201 verdienen	7 43 49.259 unterlagen	7 317 43.270 enden	7 65 52.227 erkämpfen
8 63 50.577 hochverdient	8 110 46.760 trotzen	8 64 37.626 Ausgleich	8 24 30.781 begnügen
9 56 49.852 besiegen	9 106 46.186 unglücklich	9 410 37.140 jedoch	9 34 28.140 zufrieden
10 66 47.513 Erfolg	10 84 36.837 trotz	10 24 32.882 begnügen	10 34 22.155 fc
11 128 43.342 wichtig	11 431 32.981 musste	11 33 31.730 Punktgewinn	11 43 20.283 torlos
12 134 42.325 Sieg	12 89 31.209 Sicht	12 31 31.284 gerecht	12 14 18.658 Leistungsgerechten
13 503 37.375 drei	13 79 30.828 kassieren	13 476 30.790 bleiben	13 31 18.559 gerecht
14 45 37.121 schwarzgelben	14 29 30.279 einstecken	14 65 29.452 erkämpfen	14 476 18.371 bleiben
15 107 37.109 perfekt	15 1401 29.783 nicht	15 328 27.426 beid	15 170 18.096 gleichen
16 38 32.727 dank	16 61 26.564 reichen	16 34 24.625 fc	16 42 17.999 Haberer
17 54 32.274 ungeschlagen	17 140 26.118 mussten	17 43 24.500 torlos	17 49 16.939 offen
18 306 32.031 Saison	18 183 25.122 Rückstand	18 127 24.145 Gelegenheit	18 328 16.597 beid
19 22 28.829 Tabellenspitze	19 80 22.831 geraten	19 22 22.898 Moral	19 143 16.529 verdienen
20 76 27.449 endlich	20 520 21.027 jedoch	20 71 22.023 leider	20 1788 16.031 sich

“#Types Before Cut” refers to the overall count of unique types of words. “#Types After Cut” indicates the words are considered for the keyword list. Within columns from left to right: rank, frequency, and keyness of the following word

In the Dutch subcorpus, finally, the main keywords distinguishing the conditions are again focused game outcome (see Table 10): win–loss (*thuiszege* [*home victory*], *zege* [*victory*], *gewonnen* [*win*], *overwinning* [*victory*], *winnen* [*win*]), loss–win (*nederlaag* [*defeat*], *verliezen* [*lose*]), tie–win/loss (*gelijkspel* [*tied match*], *gelijk* [*same*], *gelijkspelen* [*tie*], *remise* [*draw*]). Likewise, there are more positive words (*mooi* [*beautiful*], *prachtig* [*magnificent*], *belangrijk* [*important*], *glunderen* [*shine*], *eindelijk* [*finally*]) in win–loss, while in loss–win there are mainly negative words (*pijnlijk* [*embarrassing*], *teleurstellen* [*disappoint*], *balverlies* [*ball loss*], *slecht* [*bad*], *ramp* [*disaster*], *leed* [*sorrow*], *lijden* [*suffer*]). Again, ties seem to be associated more with neutral words.

3.2.2 Emotion words: LIWC, VAD, and emotion analyses

Having looked in general into which words are used to describe the various game outcomes across the different languages, we will now investigate emotion terms more specifically. One way of doing this is using LIWC, originally developed by Pennebaker et al. (2001). For the exploration of the MEmoFC, the original English dictionary, the German dictionary (Wolf et al. 2008), and the Dutch dictionary (Zijlstra et al. 2004) were used. Here, we focus on categories that are arguably most interesting in terms of sentiment and perspective: pronouns, specifically 1PP due to the studies on self-serving and self-preservation (BIRG/CORF; see above), negations, positive and negative emotion words, words relating to anger, sadness, and anxiety, as well as exclamation marks, which indicate positive emotions (Gilbert and Hutto 2014; Hancock et al. 2007). Means and standard deviations for the different LIWC categories are presented in Table 11, and Fig. 2 illustrates the variance in the categories in the corpus by language and outcome/condition in violin plots.

Overall, the LIWC analyses are consistent with the concordance analysis, described above: more positive emotion words are counted in reports about won matches and more negative emotion words, anger, and sadness in reports about lost matches, and this pattern is consistent across languages. Only the level of anxiety does not differ according to game outcome but between languages, with the level overall highest in English and lowest in German. The numbers for tied matches generally fall in between those for wins and losses.

Looking at differences between the languages, we observe that more positive words in English texts and that these are less frequent in German and, especially, in Dutch texts. Negative emotion words are also most frequent in English and occur less in German and Dutch, where the frequencies are similar. We also find a higher proportion of negations in Dutch, specifically in reports on lost and tied matches. The same pattern (more negations in reports about losses and ties) can be observed in English and German as well, although the differences are smaller.

In recent years, various alternative methods to assess the emotional nature of a text have been developed. For example, one can assess the valence (positive–negative), Arousal (calm–excited) and Dominance (low–high) of the words in a text. To measure how texts in the MEmoFC differ in terms of these VAD dimensions,

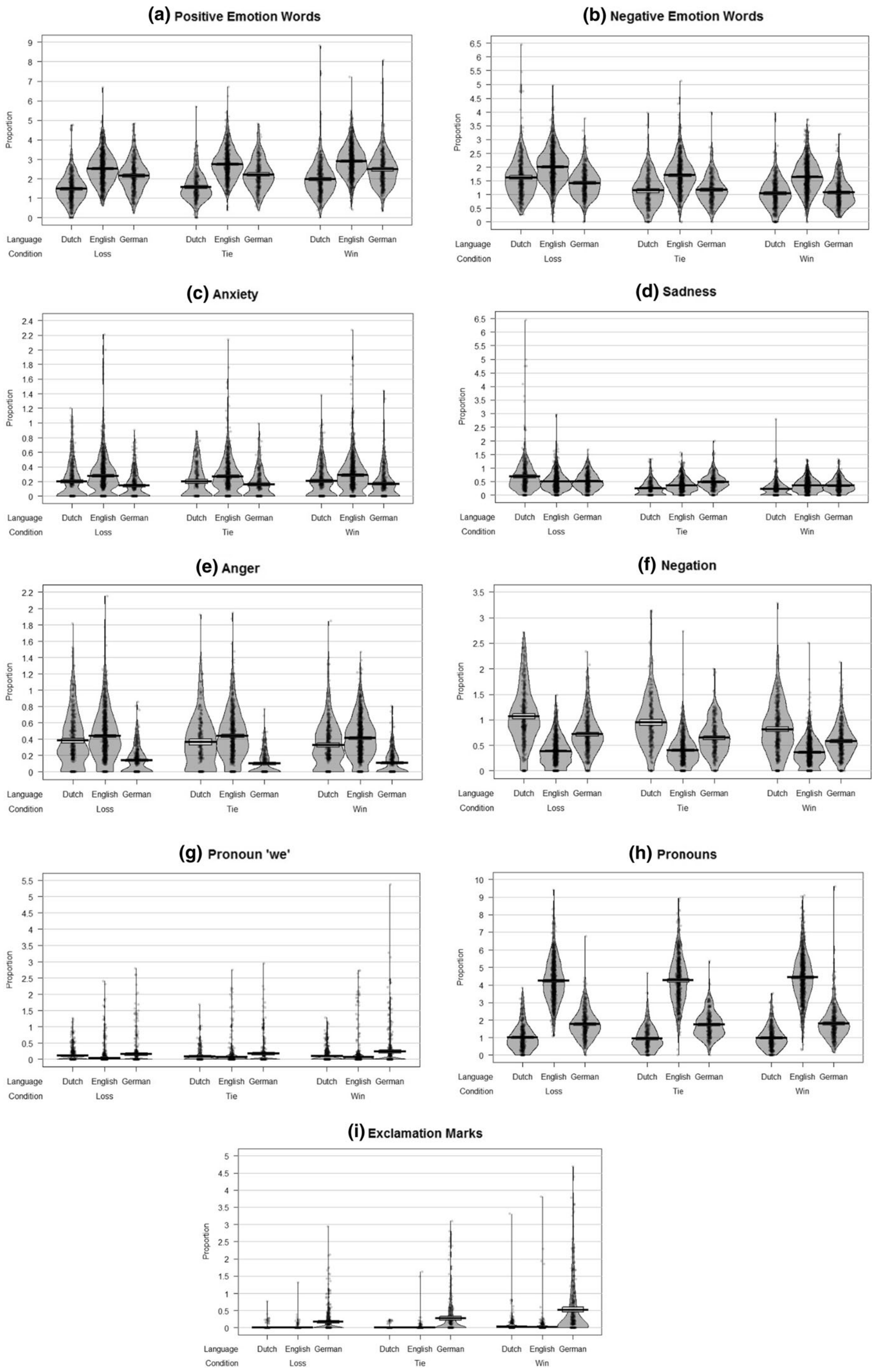
Table 10 Keywords across outcomes (Win, Loss and Tie) compared, respectively, in Dutch after lemmatization

WIN–LOSS	LOSS–WIN	TIE–WIN	TIE–LOSS
#Types Before Cut: 6354	#Types Before Cut: 5783	#Types Before Cut: 5099	#Types Before Cut: 5099
#Types After Cut: 4104	#Types After Cut: 3921	#Types After Cut: 3341	#Types After Cut: 3498
#Search Hits: 0	#Search Hits: 0	#Search Hits: 0	#Search Hits: 0
1 362 182.782 zege	1 350 240.884 verliezen	1 194 251.174 gelijkspel	1 194 196.324 gelijkspel
2 158 102.258 gewinnen	2 324 230.284 nederlaag	2 172 84.015 gelijk	2 393 132.938 punt
3 304 63.628 overwinning	3 116 67.251 onderuit	3 51 82.267 gelijkspelen	3 51 81.081 gelijkspelen
4 96 55.418 boeken	4 1657 65.668 niet	4 393 63.599 punt	4 30 43.564 puntendeling
5 288 40.347 winnen	5 74 37.074 helaas	5 30 47.275 puntendeling	5 172 37.738 gelijk
6 613 36.355 drie	6 43 32.679 lijden	6 194 41.042 gelijkmaker	6 25 31.021 remise
7 25 32.136 bevrijden	7 19 28.357 overmaat	7 224 39.268 beide	7 104 21.772 pakken
8 67 30.481 geweldig	8 19 28.357 ramp	8 25 34.009 remise	8 40 21.328 tevreden
9 187 30.213 mooi	9 29 27.226 pijnlijk	9 27 28.539 overhouden	9 14 20.030 veerkracht
10 26 26.360 thuiszege	10 17 25.372 leed	10 109 24.685 eindigen	10 10 18.851 zwaarbevochten
11 37 24.362 stijgen	11 48 25.148 teleurstellen	11 969 24.096 niet	11 20 17.618 genoegen
12 48 23.302 eindelijk	12 248 22.820 gelijkmaker	12 695 22.181 kans	12 9 16.966 overheersen
13 97 23.093 belangrijk	13 993 21.555 er	13 615 19.313 tegen	13 27 16.031 overhouden
14 115 21.874 prachtig	14 912 21.215 krijgen	14 38 18.802 helaas	14 20 15.535 doelpuntloos
15 75 21.254 venlonaren	15 121 21.056 resultaat	15 53 18.476 bedrijf	15 109 13.912 eindigen
16 529 21.238 doelpunt	16 89 19.633 aansluitingstreffer	16 9 18.126 overheersen	16 54 13.807 openen
17 25 20.863 buit	17 2625 19.341 te	17 587 17.800 krijgen	17 146 13.418 slotfase
18 27 19.680 klimmen	18 218 19.169 proberen	18 20 17.644 doelpuntloos	18 423 13.355 over
19 15 19.282 glunderen	19 51 18.686 balverlies	19 26 17.332 steken	19 7 13.195 bevrijden
20 15 19.282 ontlading	20 57 18.654 slecht	20 20 15.741 genoegen	20 7 13.195 klassiek

“#Types Before Cut” refers to the overall count of unique types of words. “#Types After Cut” indicates the words are considered for the keyword list. Within columns from left to right: rank, frequency, and keyness of the following word

Table 11 Means and standard deviations for LIWC categories across game outcomes (Win, Loss, and Tie) and languages

	Win			Loss			Tie		
	EN	NL	GER	EN	NL	GER	EN	NL	GER
Positive emotion	2.89 (0.78)	1.97 (0.86)	2.45 (0.89)	2.52 (0.78)	1.48 (0.72)	2.17 (0.79)	2.76 (0.79)	1.58 (0.73)	2.22 (0.79)
Negative emotion	1.64 (0.57)	1.05 (0.53)	1.07 (0.50)	2.00 (0.65)	1.62 (0.73)	1.42 (0.51)	1.70 (0.63)	1.15 (0.36)	1.17 (0.51)
Anxiety	0.29 (0.26)	0.21 (0.22)	0.17 (0.19)	0.28 (0.25)	0.20 (0.24)	0.14 (0.17)	0.27 (0.25)	0.20 (0.22)	0.16 (0.18)
Anger	0.41 (0.26)	0.33 (0.28)	0.11 (0.14)	0.44 (0.30)	0.38 (0.32)	0.14 (0.16)	0.44 (0.29)	0.37 (0.32)	0.10 (0.14)
Sadness	0.35 (0.25)	0.24 (0.26)	0.35 (0.24)	0.50 (0.33)	0.69 (0.64)	0.51 (0.30)	0.37 (0.26)	0.25 (0.27)	0.49 (0.29)
Negation	0.36 (0.27)	0.81 (0.49)	0.59 (0.33)	0.38 (0.28)	1.02 (0.56)	0.72 (0.38)	0.40 (0.28)	0.95 (0.52)	0.65 (0.35)
Pronouns	4.44 (1.30)	0.99 (0.64)	1.81 (0.90)	4.23 (1.19)	1.01 (0.73)	1.78 (0.79)	4.27 (1.29)	0.94 (0.68)	1.74 (0.74)
We	0.06 (0.31)	0.09 (0.22)	0.24 (0.55)	0.04 (0.20)	0.11 (0.23)	0.16 (0.41)	0.07 (0.29)	0.09 (0.24)	0.18 (0.41)
Exclamation marks	0.02 (0.16)	0.04 (0.18)	0.52 (0.78)	0.005 (0.04)	0.01 (0.06)	0.17 (0.35)	0.008 (0.08)	0.01 (0.04)	0.28 (0.53)


(a) Positive Emotion Words
(b) Negative Emotion Words
(c) Anxiety
(d) Sadness
(e) Anger
(f) Negation
(g) Pronoun 'we'
(h) Pronouns
(i) Exclamation Marks
Proportion
Language
Condition
Dutch
English
German
Loss
Tie
Win
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7
8
9
10
0.2
0.4
0.6
0.8
1.2
1.4
1.6
1.8
2.2
2.4

◀**Fig. 2** Percentages of LIWC categories (positive and negative emotion words, anger, anxiety, sadness, negation, pronouns, “we”, exclamation marks) per total words per text by condition (Win, Loss, and Tie) and language (English, German, and Dutch). Points are raw data points, the line shows central tendencies of the data, the bean is the smoothed density, and the rectangle around the line represents the inference interval

and whether this differs across languages, we use normative lexicons of English (Warriner et al. 2013), German (Vo et al. 2009) and Dutch (Moors et al. 2013), which have been developed by relying on a large number of native speakers rating thousands of words on these dimensions (except for German, where dominance is not reported).

The three lexicons differ not only in size (about 14,000 words for English, 4,000 for Dutch, and 3,000 for German), but also in the rating scales that were used during the data collection (1–9 for English, 1–7 for Dutch, and between ± 3 for German). To obtain more consistent results across languages, we used min–max normalization to rescale all dictionaries between − 4 and + 4, with 0 indicating neutral valence/arousal/dominance.

After the scale adjustment, the average VAD scores for each report were calculated by summing each dimension’s scores of all words in the report and then averaging them for all matches. A similar approach has been used in, for example, Gatti et al. (2016), and has been shown to be useful when there is no sufficient annotated data for supervised classification (Jurafsky and Martin 2009; Taboada et al. 2006), or when pre-trained sentiment or emotion analysis tools are not available (as it is the case for Dutch and German).

As can be seen in Table 12, reports of winning matches have a higher, positive valence across all languages compared to losses. In a similar vein, reports on losses have a more negative valence than those on ties, which in turn are more negative than wins. We observe no difference in arousal between reports, while dominance is slightly higher for wins. Ties are consistently ranked between wins and losses.

Given that large parts of the dictionaries consist of moderate/neutral words, this might “flatten” the differences between conditions. New dictionaries with only words that have (normalized) valence, arousal, and dominance scores of more than 2

Table 12 Valence (V), arousal (A), and dominance (D) for aligned reports in English (Win–Loss: 1037 matches compared; Tie: 365 matches compared; Dictionary: 13,915 entries), Dutch (Win–Loss: 468 matches compared; Tie: 135 matches compared; Dictionary: 4299 entries), and German (Win–Loss: 404 matches compared; Tie: 146 matches compared; Dictionary: 2903 entries)

	English			Dutch			German		
	V	A	D	V	A	D	V	A	D
Win	0.70	− 1.01	0.59	0.56	0.28	0.43	0.65	− 0.80	N.A
Loss	0.64	− 1.02	0.54	0.48	0.25	0.39	0.54	− 0.83	N.A
Tie	0.66	− 1.01	0.56	0.50	0.25	0.42	0.59	− 0.83	N.A

Table 13 Valence (V), arousal, and dominance for extreme values in aligned reports in English (Dictionary entries left: 1750 [valence]; 919 [arousal]; 403 [dominance]), Dutch (Dictionary entries left: 801 [valence]; 317 [arousal]; 145 [dominance]), and German (Dictionary entries left: 748 [valence]; 466 [arousal]; n.a. [dominance])

	English			Dutch			German		
	V	A	D	V	A	D	V	A	D
Win	1.28	− 2.20	2.13	1.68	1.32	1.29	1.45	− 1.77	N.A
Loss	1.04	− 2.20	2.00	1.28	1.05	1.26	1.08	− 1.82	N.A
Tie	1.12	− 2.17	2.00	1.48	1.22	1.30	1.34	− 1.85	N.A

or less than − 2 were created and the analysis rerun using only the more extreme values.

In Table 13, the differences for the extreme values between the matched reports about winning and losing matches are even bigger, which confirms the trend that winners use more positive and more strongly positive language. Again, more negative affect is expressed in texts about losses than in reports about won matches.

A final exploratory analysis of the emotion words in the different texts zooms in on discrete emotion terms using the EmoMap technique of Buechel and Hahn (2018), which maps the VAD lexicons onto Ekman's set of basic emotions (1992). The resulting lexicons are then scaled between 0 and 10, with 0 representing absence of a particular emotion and 10 representing the maximum intensity. Note that 10 is a theoretical maximum, while in fact no word in the resulting dictionary has a value higher than 8.5. Table 14 shows the distribution of emotions across languages and game outcomes. Although the numeric differences are relatively small, the pattern is broadly consistent with the earlier LIWC and VAD analyses, with joy scores being higher for wins, and sadness, fear, and disgust higher for losses.

3.3 Example study 3: Can we classify texts as describing a win or loss (and does this vary per language) and which textual elements of the reports are most indicative of the game outcome?

The analyses so far suggest that there are systematic differences in words and phrases used for different outcomes, whether we look at personal pronouns (we), LIWC categories, VAD scores, or discrete emotion terms. This raises the question whether we can automatically predict whether a text reports about, say, a win or a loss, taking into account words, but also other textual features. To investigate this, we conducted a multiclass classification task to further explore possible differences in the language used to report on a win, tie, or loss. Our classifier is based on the classification framework of Van der Lee and Van den Bosch (2017). Similar to that framework, a distinction was made between word statistics, syntactic, and content-specific text features (see Table 15). The word statistic features are measures on the word or character level, such as sentence length and word length distributions.

Table 14 Discrete emotions (joy [J], anger [A], sadness [S], fear [F], and disgust [D]) in aligned reports in English (Win–Loss: 1037 matches compared; 365 Ties compared), Dutch (Win–Loss: 468 matches compared; 135 Ties compared), and German (Win–Loss: 404 matches compared; 146 Ties compared)

	English					Dutch					German				
	J	A	S	F	D	J	A	S	F	D	J	A	S	F	D
Win	3.63	1.06	1.07	1.18	0.99	1.64	1.66	1.99	1.85	1.58	3.81	1.62	0.96	1.40	0.91
Loss	3.54	1.10	1.11	1.21	1.02	1.56	1.69	2.05	1.88	1.63	3.80	1.62	0.97	1.41	0.93
Tie	3.57	1.08	1.09	1.19	1.01	1.57	1.68	2.01	1.86	1.61	3.78	1.62	0.94	1.39	0.92

Table 15 Text features by word statistics, syntax, and content adopted for the classification of MEmoFC texts

Group	Category	Description
Word	Average word length	
Statistics	Average sentence length in terms of words	
	Average sentence length in terms of characters	
	Type/token ratio	Ratio of different words to the total number of words
	Hapax legomena ratio	Ratio of once-occurring words to the total number of words
	Dis legomena ratio	Ratio of twice-occurring words to the total number of words
	Short words ratio	Words < 4 characters to the total number of words
	Long words ratio	Words > 6 characters to the total number of words
	Word-length distribution	Ratio of words in length of 1–20
Syntactic	Function words ratio	Ratio of function words (e.g. ‘that’, ‘the’, ‘I’) to the total number of words
	Descriptive words to nominal words ratio	Adjectives and adverbs to the total number of nouns
	Personal pronouns ratio	Ratio of personal pronouns (e.g. ‘I’, ‘you’, ‘me’) to the total number of words
	Question words ratio	Proportion of wh-determiners, wh-pronouns, and wh-adverbs (e.g. ‘who’, ‘what’, ‘where’) to the total number of words
	Question mark ratio	Proportion of question marks to the total number of end of sentence punctuation
	Exclamation mark ratio	Proportion of exclamation marks to the total number of end of sentence punctuation
	Part-of-speech tag n-grams	Part-of-speech tag n-grams (e.g. ‘NP’, ‘VP’)
Content-specific	Word n-grams	Bag-of-word n-grams (e.g. ‘crossbar’, ‘very high’)

Syntactic features are indications of syntactical patterns present in sentences. To find these underlying syntactical patterns, the texts were parsed automatically using Frog (Bosch et al. 2007) for the Dutch soccer reports and the Stanford NLP parser (Klein and Manning 2003) for the English and German soccer reports. Besides the raw part-of-speech n-grams, syntactical feature groups such as function words, descriptive words and punctuation were captured as well.

The content-specific features used for this study were word uni-grams, bi-grams and tri-grams. These words or phrases could indicate certain topics in the text. Strategic data reduction was applied to the soccer reports for the content-specific features to reduce computational load and simultaneously reduce the chance that the classifier focuses on linguistically irrelevant features (the word *Manchester* might for instance be associated with a win, but this is not a good linguistic indicator). For this data reduction, words were lemmatized (e.g. *goals* to *goal*), stop words were removed (words like *the*, *who* and *are*), and named entities were again removed (e.g. *Arsenal*, *Luuk de Jong*). Furthermore, highly infrequent words (words appearing less than 10 times in the total corpus) were removed from the content-specific features.

Six machine learning classification algorithms were tested: Linear Support Vector Machines and Naïve Bayes, plus four tree-based algorithms: C4.5, AdaBoost, Random Forest, and XGBoost. Discriminating between wins, ties, and losses was done using either word statistics, syntactic or content-specific features as described above. Subsequently, the features from these three feature groups were combined using two different approaches: a supervector approach and a meta-classifier approach. The supervector approach pools all features together into a single vector to predict the type of report, regardless of the feature category. The meta-classifier approach takes the probabilistic outputs of each feature category and uses them as inputs for a higher-level classifier to predict the type of report, which has the potential to increase classification accuracy if the feature groups all contain some additional information that is not stored in a single feature group. The meta-classifier approach has been shown to increase performance in other classification tasks (Malmasi et al. 2015; van der Lee and van den Bosch 2017). Furthermore, a baseline was used that predicts the most frequent reports based on the training set.

The results show that all feature groups perform above baseline in all languages (Tables 16, 17 and 18). The lexical features (stylistic features such as word length and sentence length) classified the report types least well, with the syntactical features (e.g. POS n-grams and punctuation features) performing slightly better. The word-based content-specific features perform the best out of the feature groups, although the results can be improved by combining all three features in a meta-classifier. The best classifier was able to correctly label around 80% (compared to a 39% baseline) of the reports for each language, which confirms that there are clear linguistic differences between the descriptions of wins, ties, and losses for English, German and Dutch.

Interestingly, the most important word features (Tables 19, 20 and 21), as obtained using Gini importance scores for the best performing tree-based algorithm (Breiman et al. 2009), do not show salient differences in word use between win, tie or loss reports. These features are all expressions that occur rarely in the corpus,

Table 16 Classification performances of the best algorithm for the different feature groups and combination methods for the English subcorpus

Features	Algorithm	Precision (micro)	Recall (micro)	F-score (micro)	Accuracy
Lexical only	XGBoost	0.40	0.20	0.26	0.18
Syntactical only	AdaBoost	0.47	0.33	0.39	0.27
Content-specific only	XGBoost	0.84	0.70	0.76	0.68
Supervector	XGBoost	0.86	0.69	0.76	0.67
Meta-classifier	XGBoost	0.82	0.76	0.79	0.76
Baseline (stratified)	–	0.34	0.34	0.34	0.15

Table 17 Classification performances of the best algorithm for the different feature groups and combination methods for the German subcorpus

Features	Algorithm	Precision (micro)	Recall (micro)	F-score (micro)	Accuracy
Lexical only	AdaBoost	0.38	0.30	0.33	0.24
Syntactical only	AdaBoost	0.51	0.40	0.44	0.32
Content-specific only	AdaBoost	0.91	0.82	0.86	0.78
Supervector	AdaBoost	0.91	0.78	0.84	0.76
Meta-classifier	XGBoost	0.89	0.87	0.88	0.86
Baseline (stratified)	–	0.33	0.34	0.33	0.15

Table 18 Classification performances of the best algorithm for the different feature groups and combination methods for the Dutch subcorpus

Features	Algorithm	Precision (micro)	Recall (micro)	F-score (micro)	Accuracy
Lexical only	AdaBoost	0.42	0.33	0.37	0.27
Syntactical only	AdaBoost	0.47	0.35	0.40	0.27
Content-specific only	XGBoost	0.85	0.76	0.80	0.72
Supervector	AdaBoost	0.87	0.75	0.80	0.71
Meta-classifier	Linear SVM	0.85	0.80	0.82	0.80
Baseline (stratified)	-	0.35	0.34	0.34	0.16

which suggests that the distinctiveness of wins, losses or ties is based on many features in combination rather than specific ones.

4 Conclusion and future work

This paper presented a new multilingual corpus, MEmoFC, consisting of pairs of reports for soccer matches, taken from the respective websites of the competing teams, combined with game statistics. The corpus can be used for linguistic emotion

Table 19 The ten most important word features for the English subcorpus with Normalized Gini Importance Scores

Feature name	Importance score
’s second goalkeeper	0.047244
2 rowe	0.033216
11 min game	0.032676
20 min lively	0.032019
’s square	0.027757
’s ORG game	0.019867
’s a8 yard	0.01563
0 PERSON healthy	0.015286
17 min team	0.013702
’s PERSON sure	0.01044

Table 20 The ten most important word features for the German subcorpus with Normalized Gini Importance Scores

Feature name	Importance score
’s mal	0.042092
17.9 19:00	0.038623
16 min Müller	0.037463
15 ohrenbetäubend	0.02865
+ 5 min	0.026977
12 min geben	0.026194
17.1	0.023036
17 nächst heimspiel	0.022955
+ 4 ORG	0.021822
15:30 ORG 30	0.019015

Table 21 The ten most important word features for the Dutch subcorpus with Normalized Gini Importance Scores

Feature name	Importance score
− 0 ijzersterk	0.062025
1 meteen vanaf	0.047749
− 0 vrijdag 27	0.045276
1 1 ruimte	0.041853
− 0 zetten tijdens	0.039825
1 treffer geel	0.031676
0 schoot verdwijnen	0.031082
1 60ste minuut	0.028526
− 1 doelpunt	0.026765
1 3 datum	0.0229

research and has been constructed to contribute to understanding how the production of written language is influenced by an author’s emotional state or the assumed state of the intended audience of a text (e.g., happy after a win and disappointed after a loss). After describing how the corpus was collected and preprocessed, we illustrated how the corpus can be used in three exploratory studies.

The three studies were each guided by a specific research question. In our first study, we investigated basking behavior on the use of first person plural pronouns. As expected, a trend appeared in the English and German subcorpora, indicating an increase of basking after won matches compared to lost or tied matches. However, this was not the case for the Dutch subcorpus. The second study was concerned with the use of specific words and phrases depending on game outcome. In three analyses, we first examined overall frequencies with TF-IDF, which, while already suggesting trends, proved less suitable for the task; we then moved on to keywords of the language and outcome subcorpora, which showed interesting, outcome-specific words that are used in the individual subcorpora, many of which were emotionally colored, although this seemed to be the case to different degrees in the different languages; finally, we zoomed in on VAD and emotion scores, which, while showing the expected patterns according to outcome, also differed in intensity in the three languages. The third study served as a demonstration of the possibility to classify the reports according to win, loss, and tie, which confirms that some linguistic features are more representative of the respective game outcomes and, hence, possibly also emotionality, than others.

While our exploratory analyses demonstrate how the corpus can be fruitfully used to investigate affective language production, there are also some limitations worth mentioning. For example, the fact that the authors of the texts in the corpus are mostly unknown means that possible effects of authorship cannot be studied well using MEmoFC. While the possibility that the individual authors' writing styles have an impact on the lexical choices and grammatical structures as well cannot be ruled out, we expect that the multitude of different reports coming from many different writers washes out the peculiarities of different writing styles. Although MEmoFC is smaller than other contemporary affective corpora that have been constructed, such as the Amazon corpus (McAuley and Leskovec 2013) or Twitter as a corpus (Pak and Paroubek 2010), its main strength is that it is controlled, combining pairs of descriptions for the same data. It was ensured that only certain leagues and time frames were collected, while also monitoring non-available texts, to keep the reports comparable. This made it difficult to scrape the texts automatically and called for a manual collection of reports, which also limited the scope. In general, we follow Borgman (2015, p. 4) in believing that "having the right data is usually better than having more data". However, the corpus is expandable to more seasons, other leagues, "neutral" reports from unrelated newspapers or other countries and in other languages. The latter might also include international matches and the respective reportage (e.g., the World Cup or the European Championship), where it would be possible to investigate cultural differences in (affective) language use by examining the perspectives of two countries instead of two clubs. An extension of the corpus with more texts could also make it possible for the classifier approach to find more robust individual features, meaning linguistic features that reliably reflect differences between reports describing a win, tie or loss in the three languages, which could not be detected now due to the scarcity of reoccurring bigrams/trigrams.

In its current form, we believe that MEmoFC will contribute to and improve the generation of sports narratives as a starting point for effectively training NLG systems. For example, based on the corpus, a data-to-text generation system that is able to generate multiple reports for a single match has already been developed (van der Lee et al. 2017). In the future, it could be interesting to look at how authors of match reports select which game events to report on based on the statistics collected for the leagues and seasons of MEmoFC since there might be a bias in the selection process due to the outcome of the match or due to cultural differences that are possibly traceable in the languages.

In a next step, we intend to conduct a laboratory study to directly investigate the effects of negative and positive emotions related to success and failure on language production. Similar to Baker-Ward et al.'s (2005) study that investigated the realization of negative and positive emotions in match narratives of children who played in two teams and participated in the same football match, it would be interesting to create a game setting experiment with participants producing the reports to the matches themselves. This setting will enable us to eliminate the issues of unknown authorship and uncertainty about the emotional involvement of the author.

MEmoFC is, to the best of our knowledge, the first corpus to include affective narratives about the same events from different perspectives, across different cultures and languages. The controlled selection process of the reports ensures the quality of the corpus, while still adding up to a respectable number of texts. In this paper, we demonstrated its usefulness both for linguists (e.g., to explore cultural differences in emotions and language production) and NLP/NLG researchers (for practical applications of such differences, see, e.g., PASS by Van der Lee et al. 2017). MEmoFC is available on request for research purposes.

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. A preliminary version of the corpus was presented at the International Conference on Natural Language Generation (INLG) in 2016 (Braun et al. 2016).

Appendix MEmoFC

See Tables 22, 23.

Table 22 Overview of participating football clubs, leagues, and abbreviations of club names for the UK, Germany, and the Netherlands of the season 2015/2016

Club	League	Abbreviation
Arsenal	Premier League	A
AFC Bournemouth	Premier League	AFCB
Aston Villa	Premier League	AV
Chelsea	Premier League	CH
Crystal Palace	Premier League	CP
Everton	Premier League	EV
Leicester City	Premier League	LC
Liverpool	Premier League	LP
Manchester City	Premier League	MC
Manchester United	Premier League	MU
Norwich City	Premier League	NC
Newcastle United	Premier League	NU
Stoke City	Premier League	SC
Southampton	Premier League	SH
Sunderland	Premier League	SL
Swansea City	Premier League	SWA
Tottenham Hotspur	Premier League	TH
West Bromwich Albion	Premier League	WBA
Watford	Premier League	WF
West Ham United	Premier League	WHU
Barnsley	Sky Bet League 1	B
Burton Albion	Sky Bet League 1	BA
Bradford City	Sky Bet League 1	BF
Blackpool	Sky Bet League 1	BP
Bury	Sky Bet League 1	BU
Coventry City	Sky Bet League 1	C
Crewe Alexandra	Sky Bet League 1	CA
Chesterfield	Sky Bet League 1	CFC
Colchester United	Sky Bet League 1	CU
Doncaster Rovers	Sky Bet League 1	DFC
Fleetwood Town	Sky Bet League 1	FW
Gilingham	Sky Bet League 1	GFC
Millwall	Sky Bet League 1	MFC
Oldham Athletic	Sky Bet League 1	OA
Peterborough United	Sky Bet League 1	PB

Table 22 continued

Club	League	Abbreviation
Port Vale	Sky Bet League 1	PV
Rochdale	Sky Bet League 1	RD
Shrewsbury Town	Sky Bet League 1	SB
Southend United	Sky Bet League 1	SEU
Scunthorpe United	Sky Bet League 1	ST
Sheffield United	Sky Bet League 1	SU
Swindon Town	Sky Bet League 1	SW
Walsall	Sky Bet League 1	W
Wigan Athletic	Sky Bet League 1	WA
AFC Wimbledon	Sky Bet League 2	AFC
Accrington Stanley	Sky Bet League 2	AS
Barnet	Sky Bet League 2	BFC
Bristol Rovers	Sky Bet League 2	BR
Carlisle United	Sky Bet League 2	CAU
Cambridge United	Sky Bet League 2	CB
Crawley Town	Sky Bet League 2	CT
Dagenham and Redbrigde	Sky Bet League 2	DR
Exeter City	Sky Bet League 2	EC
Hartlepool United	Sky Bet League 2	HU
Leyton Orient	Sky Bet League 2	LO
Luton Town	Sky Bet League 2	LT
Morecambe	Sky Bet League 2	MC
Mansfield Town	Sky Bet League 2	MF
Notts County	Sky Bet League 2	NC
Northampton Town	Sky Bet League 2	NH
Newport County	Sky Bet League 2	NP
Oxford United	Sky Bet League 2	OU
Plymouth Argyle	Sky Bet League 2	PA
Portsmouth	Sky Bet League 2	PM
Stevenage	Sky Bet League 2	SA
Wycombe Wanderers	Sky Bet League 2	WW
York City	Sky Bet League 2	YC
Yeovil Town	Sky Bet League 2	YT
Bayer 04 Leverkusen	1. Bundesliga	BL
Borussia Mönchengladbach	1. Bundesliga	BMG
Hertha BSC	1. Bundesliga	BSC
Borussia Dortmund	1. Bundesliga	BVB
Eintracht Frankfurt	1. Bundesliga	EF
FC Augsburg	1. Bundesliga	FCA
FC Bayern München	1. Bundesliga	FCB
FC Ingolstadt 04	1. Bundesliga	FCI
1.FC Köln	1. Bundesliga	FCKO

Table 22 continued

Club	League	Abbreviation
FC Schalke 04	1. Bundesliga	FCS
1.FSV Mainz 05	1. Bundesliga	FSVM
Hannover 96	1. Bundesliga	HAN
Hamburger SV	1. Bundesliga	HSV
SV Darmstadt 98	1. Bundesliga	SVD
SV Werder Bremen	1. Bundesliga	SVW
TSG 1899 Hoffenheim	1. Bundesliga	TSG
VfB Stuttgart	1. Bundesliga	VFBS
VfL Wolfsburg	1. Bundesliga	VFLW
DSC Arminia Bielefeld	2. Bundesliga	DSC
Eintracht Braunschweig	2. Bundesliga	EB
1.FC Heidenheim 1846	2. Bundesliga	FCH
1. FC Kaiserslautern	2. Bundesliga	FCK
1. FC Nürnberg	2. Bundesliga	FCN
FC St. Pauli	2. Bundesliga	FCST
1.FC Union Berlin	2. Bundesliga	FCUB
Fortuna Düsseldorf	2. Bundesliga	FD
FSV Frankfurt 1899	2. Bundesliga	FSV
Karlsruher SC	2. Bundesliga	KSC
MSV Duisburg	2. Bundesliga	MSV
RB Leipzig	2. Bundesliga	RB
SC Freiburg	2. Bundesliga	SCF
SC Paderborn 07	2. Bundesliga	SCP
SpVgg Greuther Fürth	2. Bundesliga	SPVGG
SV Sandhausen 1916	2. Bundesliga	SVS
TSV 1860 München	2. Bundesliga	TSV
VfL Bochum 1848	2. Bundesliga	VFL
ADO Den Haag	Eredivisie	ADO
Ajax	Eredivisie	AX
AZ	Eredivisie	AZ
De Graapschap	Eredivisie	DG
Excelsior	Eredivisie	EX
FC Groningen	Eredivisie	FCG
FC Twente	Eredivisie	FCT
FC Utrecht	Eredivisie	FCU
Feyenoord	Eredivisie	FN
Heracles Almelo	Eredivisie	HA
NEC	Eredivisie	NEC
PEC Zwolle	Eredivisie	PEC
PSV	Eredivisie	PSV
Roda JC Kerkrade	Eredivisie	RJC
SC Cambuur	Eredivisie	SCC

Table 22 continued

Club	League	Abbreviation
SC Heerenveen	Eredivisie	SCH
Vitesse	Eredivisie	V
Willem II	Eredivisie	WII
Almere	Jupiler League	AC
Achilles ’29	Jupiler League	ACH
Den Bosch	Jupiler League	DB
Dordrecht	Jupiler League	FCD
Einhoven	Jupiler League	FCE
Emmen	Jupiler League	FCEM
Oss	Jupiler League	FCO
Telstar	Jupiler League	FCT
Volendam	Jupiler League	FCV
Fortuna Sittard	Jupiler League	FS
G.A.Eagles	Jupiler League	GAE
Helmond Sport	Jupiler League	HS
Jong Ajax	Jupiler League	JAX
Jong PSV	Jupiler League	JPSV
MVV	Jupiler League	MVV
NAC Breda	Jupiler League	NAC
RKC Waalwijk	Jupiler League	RKC
Sparta Rotterdam	Jupiler League	SR
VVV-Venlo	Jupiler League	VVV

Table 23 Most relevant bigrams in English, German, and Dutch for Win (W), Loss (L), and Tie (T) extracted with TF/I

	English			German			Dutch		
	W	L	T	W	L	T	W	L	T
1	back win	to condemn	fair result	wichtig siegen	unterliegen ORG	vom 1	LOC boeken	deze nederlaag	de remise
2	point safe	PERSON condemn	spoil be	schlagen ORG	Keine punkten	LOC trennen	deze zege	onderuit tegen	de puntendeling
3	this win	goal condemn	another point	ORG besiegen	hinnehmen müssen	Kein Sieger	ruim zege	nederlaag geleden	naar punt
4	another win	town slip	square in	Vorsprung an	kein punkten	ORG begnügen	overwinning boeken	nederlaag voor	pakken punt
5	impressive victory	down 3	GPE play_out	fahren_einen der	unterliegen beim	Punkt mit	ORG boeken	nederlaag lijden	delen de
6	record another	slump to	but krul	hochverdient mit	unterliegen am	erkämpfen Punkt	ORG langs	LOC lijden	gelijkspel voor
7	superb away	0 city	chance pass	klettern der	96 unterliegen	Fehler zu	winnen ORG	pijnlijk nederlaag	met gelijkspel
8	two superb	defeat leave	deserve point	drei wichtig	MISC unterliegen	LOC schaffen	rij boeken	vuist te	terecht uitslag
9	NORP see_out	down 1	deserved share	flachen rechts	ORG unterlagen	Nummer 21	de ontlading	groot nederlaag	voelen als
10	PERSON get_off	pools at	equaliser earn	LOC besiegen	unterliegen in	Punkteteilung mit	zege boeken	met nederlaag	ORG delen

The following tables are the output of LIWC and the concordance tool (AntConc) and are, due to their length, better fit as Excel tables. They can be freely accessed through: https://surfdrive.surf.nl/files/index.php/s/PrGfvhVNDEeNZrZ

References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL* (pp. 7–13).

Baker-Ward, L. E., Eaton, K. L., & Banks, J. B. (2005). Young soccer players' reports of a tournament win or loss: Different emotions, different narratives. *Journal of Cognition and Development, 6*(4), 507–527.

Basile, V. (2013). *Sentiment analysis on Italian tweets.* Paper presented at the Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

Bateman, J. A., & Paris, C. (1989). *Phrasing a text in terms the user can understand.* Paper presented at the IJCAI.

Bautin, M., Vijayarenu, L., & Skiena, S. (2008). *International sentiment analysis for news and blogs.* Paper presented at the ICWSM.

Beukeboom, C. J., & Semin, G. R. (2006). How mood turns on language. *Journal of Experimental Social Psychology, 42*(5), 553–566.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world.* Cambridge: MIT Press.

Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occasional Series, 7*, 191–206.

Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems, 28*(2), 55–63.

Braun, N., Goudbeek, M., & Krahmer, E. (2016). *The Multilingual Affective Soccer Corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch.* Paper presented at the INLG.

Breiman, L., Friedman, J., & Olshen, R. (2009). *Stone, cj (1984) classification and regression trees.* Belmont: Wadsworth.

Buechel, S., & Hahn, U. (2018). Representation mapping: A novel approach to generate high-quality multi-lingual emotion lexicons. arXiv preprint arXiv:1807.00775.

Chen, X., & Lawrence Zitnick, C. (2015). *Mind's eye: A recurrent visual representation for image caption generation.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34*(3), 366.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*(10), 687–693.

dos Santos, C., & Gatti, M. (2014). *Deep convolutional neural networks for sentiment analysis of short texts.* Paper presented at the Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.

Downs, E., & Sundar, S. S. (2011). "We won" vs. "They lost": Exploring ego-enhancement and self-preservation tendencies in the context of video game play. *Entertainment Computing, 2*(1), 23–28.

Ekman, P. (1992). Are there basic emotions?

Feng, Y., & Lapata, M. (2010). *Topic models for image annotation and text illustration.* Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychological Bulletin, 117*(1), 39.

Forgas, J. P. (1999). On feeling good and being rude: Affective influences on language use and request formulations. *Journal of Personality and Social Psychology, 76*(6), 928.

Forgas, J. P. (2013). Don't worry, be sad! On the cognitive, motivational, and interpersonal benefits of negative mood. *Current Directions in Psychological Science, 22*(3), 225–232.

Forgas, J. P., & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology, 44*(5), 1362–1367.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research, 61*, 65–170.

Gatti, L., Guerini, M., & Turchi, M. (2016). SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing, 7*(4), 409–421.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications, 40*(16), 6266–6282.

Gilbert, E., & Hutto, C. J. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text.* Paper presented at the Eighth International Conference on Weblogs and Social Media (ICWSM-14). http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf. Accessed 20 Apr 2016.

Glorot, X., Bordes, A., & Bengio, Y. (2011). *Domain adaptation for large-scale sentiment classification: A deep learning approach.* Paper presented at the Proceedings of the 28th International Conference on Machine Learning (ICML-11).

Hancock, J. T., Landrigan, C., & Silver, C. (2007). *Expressing emotion in text-based communication.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *The Journal of Abnormal and Social Psychology, 49*(1), 129.

Honnibal, M., & Johnson, M. (2015). *An improved non-monotonic transition system for dependency parsing.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

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

Isah, H., Trundle, P., & Neagu, D. (2014). *Social media analysis for product safety using text mining and sentiment analysis.* Paper presented at the Computational Intelligence (UKCI), 2014 14th UK Workshop on.

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. In (pp. 1–1024). Prentice Hall, Pearson Education International.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Klein, D., & Manning, C. D. (2003). *Accurate unlexicalized parsing.* Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Koch, A. S., Forgas, J. P., & Matovic, D. (2013). Can negative mood improve your conversation? Affective influences on conforming to Grice's communication norms. *European Journal of Social Psychology, 43*(5), 326–334.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012). *Collective generation of natural image descriptions.* Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.

Lee, G., Bulitko, V., & Ludvig, E. A. (2014). Automated story selection for color commentary in sports. *IEEE Transactions on Computational Intelligence and AI in Games, 6*(2), 144–155.

Lin, C.-Y., & Hovy, E. (2000). *The automated acquisition of topic signatures for text summarization.* Paper presented at the Proceedings of the 18th Conference on Computational Linguistics-Volume 1.

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review, 48*(4), 499–527.

Losada, D. E., & Gamallo, P. (2020). Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, *54*(1), 1–24. https://doi.org/10.1007/s10579-018-9423-1.

Mahamood, S., & Reiter, E. (2011). *Generating affective natural language for parents of neonatal infants.* Paper presented at the Proceedings of the 13th European Workshop on Natural Language Generation.

Malmasi, S., Refaee, E., & Dras, M. (2015). *Arabic dialect identification using a parallel multidialectal corpus.* Paper presented at the International Conference of the Pacific Association for Computational Linguistics.

McAuley, J., & Leskovec, J. (2013). *Hidden factors and hidden topics: Understanding rating dimensions with review text.* Paper presented at the Proceedings of the 7th ACM Conference on Recommender Systems.

Mihalcea, R., & Strapparava, C. (2009). *The lie detector: Explorations in the automatic recognition of deceptive language.* Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., et al. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45*(1), 169–177.

Morales, M., Scherer, S., & Levitan, R. (2017). *A cross-modal review of indicators for depression detection systems.* Paper presented at the Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality.

Nguyen, D., Smith, N. A., & Rosé, C. P. (2011). *Author age prediction from text using linear regression.* Paper presented at the Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.

Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining.* Paper presented at the LREc.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1–135.

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

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90–93.

Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012). *Learning sentiment lexicons in Spanish.* Paper presented at the LREC.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., et al. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence, 173*(7–8), 789–816.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.

Rayson, P., & Garside, R. (2000). *Comparing corpora using frequency profiling.* Paper presented at the Proceedings of the Workshop on Comparing corpora-Volume 9.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems.* Cambridge: Cambridge University Press.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121–1133.

Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases, applications and range. *European Review of Social Psychology, 2*(1), 1–30.

Smith, M. K., & Montgomery, M. B. (1989). The semantics of winning and losing. *Language in Society, 18*(1), 31–57.

Snyder, C. R., Lassegard, M., & Ford, C. E. (1986). Distancing after group success and failure: Basking in reflected glory and cutting off reflected failure. *Journal of Personality and Social Psychology, 51*(2), 382.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank.* Paper presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

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

Strapparava, C., & Mihalcea, R. (2017). *A computational analysis of the language of drug addiction.* Paper presented at the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.

Taboada, M., Anthony, C., & Voll, K. D. (2006). *Methods for creating semantic orientation dictionaries.* Paper presented at the LREC.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Tsakalidis, A., Papadopoulos, S., Voskaki, R., Ioannidou, K., Boididou, C., Cristea, A. I., et al. (2018). Building and evaluating resources for sentiment analysis in the Greek language. *Language Resources and Evaluation, 52*, 1021–1044.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review, 29*(4), 402–418.

van der Lee, C., Krahmer, E., & Wubben, S. (2017). *PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences.* Paper presented at the Proceedings of the 10th International Conference on Natural Language Generation.

van der Lee, C., & van den Bosch, A. (2017). *Exploring lexical and syntactic features for language variety identification.* Paper presented at the Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial).

Vo, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods, 41*(2), 534–538.

Wann, D. L., & Branscombe, N. R. (1990). Die-hard and fair-weather fans: Effects of identification on BIRGing and CORFing tendencies. *Journal of Sport and Social Issues, 14*(2), 103–117.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica, 54*(2), 85–98.

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond, 32*, 271–281.