Grounding the Semantics of Part-of-Day Nouns Worldwide using Twitter

David Vilares
Universidade da Coruña
FASTPARSE Lab, LyS Group
Departamento de Computación
Campus de A Elviña s/n, 15071
A Coruña, Spain
david.vilares@udc.es

Carlos Gómez-Rodríguez
Universidade da Coruña
FASTPARSE Lab, LyS Group
Departamento de Computación
Campus de A Elviña s/n, 15071
A Coruña, Spain
carlos.gomez@udc.es

Abstract

The usage of part-of-day nouns, such as 'night', and their time-specific greetings ('good night'), varies across languages and cultures. We show the possibilities that Twitter offers for studying the semantics of these terms and its variability between countries. We mine a worldwide sample of multilingual tweets with temporal greetings, and study how their frequencies vary in relation with local time. The results provide insights into the semantics of these temporal expressions and the cultural and sociological factors influencing their usage.

1 Introduction

Human languages are intertwined with their cultures and societies, having evolved together, reflecting them and in turn shaping them (Ottenheimer, 2013; Dediu et al., 2013). Part-of-day nouns (e.g. 'morning' or 'night') are an example of this, as their meaning depends on how each language's speakers organize their daily schedule. For example, while the morning in English-speaking countries is assumed to end at noon, the Spanish term ('mañana') is understood to span until lunch time, which normally takes place between 13:00 and 15:00 in Spain. It is fair to relate this difference to cultural (lunch being the main meal of the day in Spain, as opposed to countries like the UK, and therefore being a milestone in the daily timetable) and sociopolitical factors (the late lunch time being influenced by work schedules and the displacement of the Spanish time zones with respect to solar time). Similar differences have been noted for different pairs of languages (Jäkel, 2003) and for cultures using the same language (Sekyi-Baidoo and Koranteng, 2008), based on manual study, field research and interviews with natives. Work on automatically extracting the semantics of part-of-day nouns is scarce, as classic corpora are not timestamped. Reiter and Sripada (2003); Sripada et al. (2003) overcome it by analyzing weather forecasts and aligning them to timestamped simulations, giving approximate groundings for time-of-day nouns and showing idiolectal variation on the term 'evening', but the work is limited to English.

The relation between language and sociocultural factors implies that the semantics of part-of-day nouns (e.g. 'end of the morning') cannot be studied in isolation from social habits (e.g. 'typical lunch time'). A relevant study of such habits is done by Walch et al. (2016), who develop an app to collect sleep habits from users worldwide. While they do not study the meaning of words, their insights are used for validation.

We propose a new approach to study the semantics of part-of-day nouns by exploiting Twitter and the time-specific greetings (e.g. 'good morning') used in different cultures. By mining tweets with these greetings, we obtain a large, worldwide sample of their usage. Since many tweets come with time and geolocation metadata, we can know the local time and country at which each one was emitted. The main contribution of the paper is to show how it is possible to learn the semantics of these terms in a much more extensive way than previous work, at a global scale, with less effort and allowing statistical testing of differences in usage between terms, countries and languages.

2 Materials and methods

To ground the semantics of greetings we used 5 terms as seeds: 'good morning', 'good afternoon', 'good evening', 'good night' and 'hello' (a time-unspecific greeting used for comparison). We translated them to 53 languages and variants using Bing translator.[1] We use *italics* to refer to greet-

[1] We used the mstranslator API for the Bing translator.

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 123–128
New Orleans, Louisiana, June 6, 2018.

ings irrespective of the language. 172,802,620 tweets were collected from Sept. 2 to Dec. 7 2016.

For some languages (e.g. Spanish), there is no differentiation between 'good evening' and 'good night', and they both are translated to the same expression. For some others, some expressions cannot be considered equivalent, e.g. 'good morning' is translated to 'bonjour' in French, which is however commonly used as 'hello', or simply as 'good day'.

Text preprocessing is not necessary: we rely on metadata, not on the tweet itself, and only the seed words are needed to categorize tweets within a part of day. To clean up the data, we removed retweets, as they last for hours, biasing the temporal analysis. Duplicate tweets were kept, as similar messages from different days and users (e.g. 'good night!') are needed for the task at hand. Tweets need to be associated with a timestamp and country-level geolocation. Tweets have a creation time, composed of a UTC time and a UTC offset that varies depending on the time zone. However, most tweets are not geolocated and we must rely on the data provided by the user. This may be fake or incomplete, e.g. specifying only a village. We used fine-grained databases[2] to do the mapping to the country level location and performed a sanity check, comparing the Twitter offset to the valid set of offsets for that country[3], to reduce the amount of wrongly geolocated tweets.[4] Comparing the solar and standard time could provide more insights, but this requires a fine-grained geolocation of the tweets. We obtained a dataset of 10,523,349 elements, available at https://github.com/aghie/peoples2018_grounding: 4,503,077 *good morning*'s, 599,586 *good afternoon*'s, 214,231 *good evening*'s, 880,003 *good night*'s and 4,359,797 *hello*'s.[5]

3 Results and validation

Given a country, some of the tweets are written in foreign languages for reasons like tourism or immigration. This paper refers to tweets written in official or *de facto* languages, unless otherwise specified. Also, analyzing differences according to criteria such as gender or solar time can be relevant. As determining the impact of all those is a challenge on its own, we focus on the primary research question: *can we learn semantics of the part-of-day nouns from simple analysis of tweets?* To verify data quality, *good morning* tweets were revised: out of 1 000 random tweets from the USA, 97.9% were legitimate greetings and among the rest, some reflected somehow that the user just started the day (e.g 'Didn't get any good morning SMS'). We did the same for Spain (98,1% legitimate), Brazil (97.8%) and India (99.6%).

Existing work and dated events are used to ratify the results presented below.

Country$_{lang}$	**morning**	**afternoon**	**night**	**hello**
Philippines$_{en}$	08:02:49	13:39:52	00:13:42	14:27:20
Japan$_{ja}$	08:07:28	15:46:50	01:04:19	*[6]
South Africa$_{en}$	08:10:07	14:50:52	22:51:48	13:40:19
Germany$_{de}$	08:16:41	13:15:18	23:29:38	14:35:06
Indonesia$_{in}$	08:17:18	16:25:11	19:02:09	13:55:00
Netherlands$_{nl}$	08:25:42	14:28:09	23:44:56	14:10:13
Ecuador$_{es}$	08:32:54	15:03:22	22:10:59	14:37:10
United States$_{en}$	08:33:23	13:26:25	21:06:00	13:33:13
Nigeria$_{en}$	08:34:37	14:11:49	17:19:19	13:40:23
Venezuela$_{es}$	08:37:03	15:04:00	21:18:05	14:11:07
Malaysia$_{en}$	08:39:17	13:31:41	01:02:33	13:56:49
Chile$_{es}$	08:39:38	15:06:52	00:10:43	14:11:56
Colombia$_{es}$	08:40:19	15:13:16	21:10:57	14:42:58
Canada$_{en}$	08:40:30	13:19:33	21:10:57	13:47:40
Mexico$_{es}$	08:51:04	15:26:35	21:58:24	14:25:37
India$_{en}$	08:51:24	13:40:00	00:03:12	14:12:54
United Kingdom$_{en}$	09:06:33	14:30:45	19:49:17	14:13:03
Turkey$_{tr}$	09:16:40	13:12:23	00:41:08	13:56:42
Australia$_{en}$	09:17:43	15:15:38	20:33:47	13:48:28
Brazil$_{pt}$	09:18:20	14:47:51	23:31:34	14:26:07
Pakistan$_{en}$	09:29:12	13:29:28	01:23:05	13:43:58
Russian Federation$_{ru}$	09:36:17	13:44:42	23:51:49	14:14:44
Spain$_{es}$	09:42:41	16:43:57	00:24:28	14:26:33
Argentina$_{es}$	09:43:47	16:20:05	00:26:55	14:02:03
Greece$_{el}$	09:46:11	17:12:35	23:28:56	15:01:05
Kenya$_{en}$	09:57:39	14:15:33	21:44:26	14:07:03
Portugal$_{pt}$	10:10:22	15:27:35	23:05:25	14:57:34
France$_{fr}$	12:37:09	*[7]	00:41:08	14:41:07

Table 1: Average local time for the greetings coming from the countries with most data, sorted by the average time for the greeting *good morning*. *Hello* was used as sanity check.

3.1 Worldwide average greeting times

Table 1 shows the average greeting times for the countries from which we collected more data.

[2] http://download.geonames.org/export/dump/

[3] http://timezonedb.com

[4] Free geolocation API's have rate limits and their use is unfeasible with a large amount of tweets.

[5] The dataset does not contain tweets from the first two weeks of October due to logistic issues.

[6] Hello translated to 'Konnichiwa', as good afternoon.

[7] The French term for afternoon (après-midi) is not commonly used as part of a greeting.

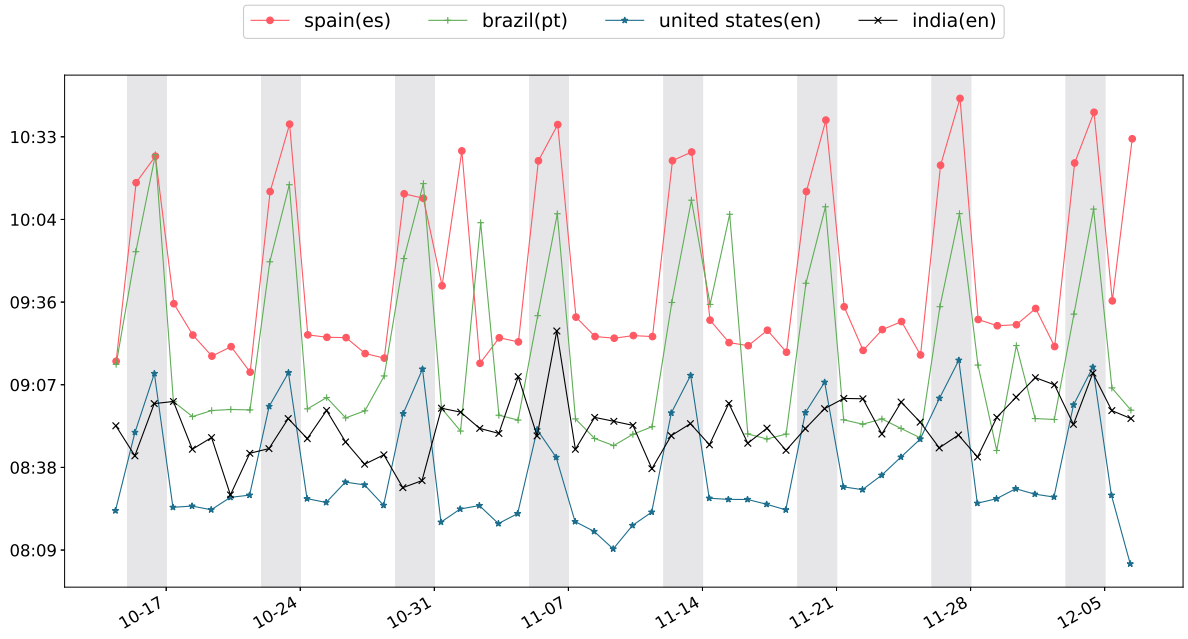

Figure 1: Average day time for the greeting *good morning* in different countries (USA, Brazil, Spain and India) for a period from mid October to early December, 2016. Weekends are shaded in gray.

Asian, African and American countries tend to begin the day earlier than Europe (with exceptions, e.g. Germany). The table reflects that countries in southern Europe (e.g. Spain, Portugal or Greece) start the day later than the northern ones (the Netherlands or UK). For some countries, e.g. France, this information is known to be biased, as *good morning* ('bonjour') is used all along the day. A validation at a fine-grained scale is unfeasible, but the results at the country level are in line with Figure 3 of Walch et al. (2016), e.g., they state that Japan, the USA or Germany have earlier wake up times than Spain, Brazil or Turkey.

The average greeting times for *good afternoon* reveal insights that may stem from cultural differences (e.g. lunch break time). Anglo-Saxon and South Asian countries have the earliest afternoon (with averages between 13:00 and 14:00), while in Mediterranean countries the morning lasts longer (average greeting times for *good afternoon* around 15:00 or 16:00). A number of countries under the influence of the United Kingdom, such as the United States, Pakistan or India show earlier *afternoon* times. The opposite happens in South America, historically influenced by Portuguese and Spanish colonialism during the Early modern period, which exhibits later *afternoon* times.

This poses interesting questions for future work, such as whether there is a particular reason that could justify this behavior, like having more similar cuisine practices. In this context, the adoption of food practices in colonialism has been already studied by anthropologists and historians (Earle, 2010). Trigg (2004) points out how in the early period of the Spanish colonialism in the Americas, they 'civilized' the Indigenous community by making them adopt manners, dress and customs. She points that the role of food was specially relevant due to its large social component, and was not limited to the way the food was eaten, but also prepared, served and consumed.

Twitter also reflects differences between countries regarding night life. On the one hand, Anglo-Saxon countries wish *good night* earlier (from 19:49 in the UK to 21:10 in Canada) than other societies. On the other hand, southern European countries go to bed later, and some of them even wish a *good night* after midnight (e.g. Spain). Comparing to Walch et al. (2016), we find similar tendencies. For example, in their study Spain, Turkey or Brazil use the smartphone until later than Canada, the USA or the UK, and therefore they go later to bed. Our Twitter approach also captures the particular case of Japanese mentioned by Walch et al.: they wake up very early, but use the smartphone until late in the night, suggesting a

later bed time.

A fine-grained analysis shows how Twitter captures other cultural and working differences. Figure 1 charts the average day time for *good morning* for the USA, Brazil, Spain and India during part of the polling period. The time peaks in the weekends for many of the countries, showing that Twitter captures how business and work are reduced during holidays, resulting in later wake up times.

However, this is not visible in some countries where working conditions are sometimes questioned (Mosse et al., 2002): for India the weekend peak is less pronounced, which can be considered as an indicator that a significant part of its population does not enjoy work-free weekends.

The usage of part-of-day expressions can be helpful to understand more complex issues, such as how foreigners integrate into a country and adapt to its daily schedule. We take the USA as example, as it has a large foreign community of Spanish speakers, mainly from Mexico (and in a smaller proportion from other Latin American countries). If we calculate the average day time for the Spanish form of 'good morning' ('buenos días') in the USA, we obtain that the result is 08:09, while the corresponding English greeting's average time is 08:33. This is reinforced by Figure 2, where 'buenos días' average day time is consistently lower than 'good morning'.[8] This would be in line to their presence in low-wage jobs that require to wake up earlier, e.g. waiter, cleaning or construction work (Flippen, 2012; Liu, 2013).

It is worth noting that, assuming that these 'buenos días' greetings come from latinos, those in the USA wake up even earlier than in their countries of origin (see Table 1).

Figure 1 also shows how national holidays influence societies. For example, Nov. 2 (Day of the Dead) and Nov. 15 (Proclamation of the Republic) are holidays in Brazil, producing a peak in that country's graph similar to the behavior in the weekends. Similarly, Nov. 1 (All Saints' Day) and Dec. 6 (Constitution Day) are holidays in Spain and similar peaks are observed too. From Figure 2 we can see how Thanksgiving (Nov. 24 in 2016) reflects a four-day weekend in the USA: many businesses allow employees to take this holiday from Thursday, resulting into a gradual and increasing peak that spans until Sunday. This is captured by the English *good morning*s, but not by the Spanish ones. The day after the USA 2016 elections (Nov. 9), a valley occurs on the *good morning* time for the States (Figure 1). The winner was not known until 03:00, suggesting that the distribution of greetings reflects social behaviors in other special events.

[8]The peak occurring on 29th October for the Spanish tweets is due to a case of spam that could not be avoided according to the procedure described in §2.

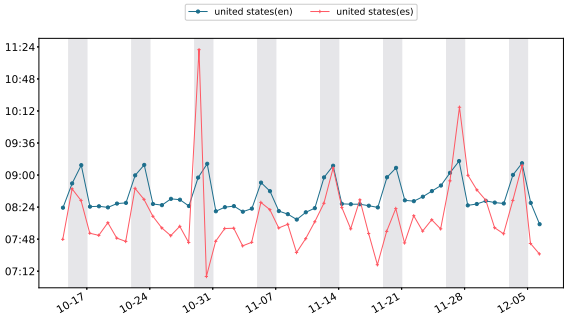

Figure 2: Average day time for the greeting 'good morning' and its Spanish form in the USA.

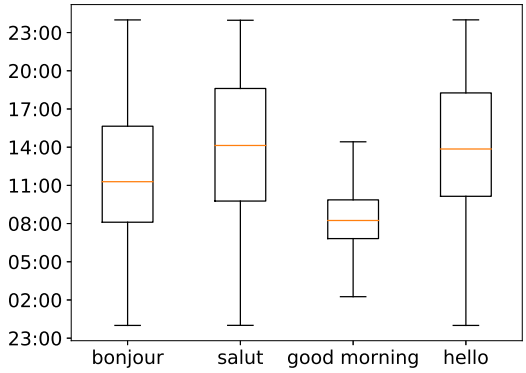

Figure 3: Box & whisker plot for the French and English *good morning*'s and *hello*'s in Canada.

3.2 Daily analysis

Twitter can be used to do a time-of-day analysis, e.g., as said in §3.1, 'bonjour' is assumed to be used all along the day. To test this, we take Canada, where French and English are official languages. Figure 3 shows how 'bonjour' and 'salut' ('hello') are used all along the day, while 'good morning' is used in the morning hours. English and French *hello*'s share a similar distribution.

Figure 4 shows a greeting area chart for the USA, showing how 'good evening' and 'good afternoon' are well differentiated, with the transition happening over 16:30. This contrasts to countries such as Spain (Figure 5), where the language has

a single word ('tarde') for 'evening' and 'afternoon', whose greeting spans from over 14:00, as the morning ends late (see §1), to 21:00.

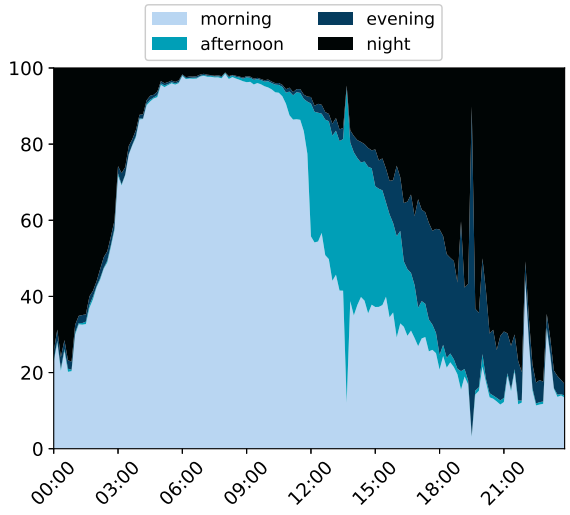

Figure 4: Stacked area chart for the greetings in the USA: % (y axis) vs time (x axis).

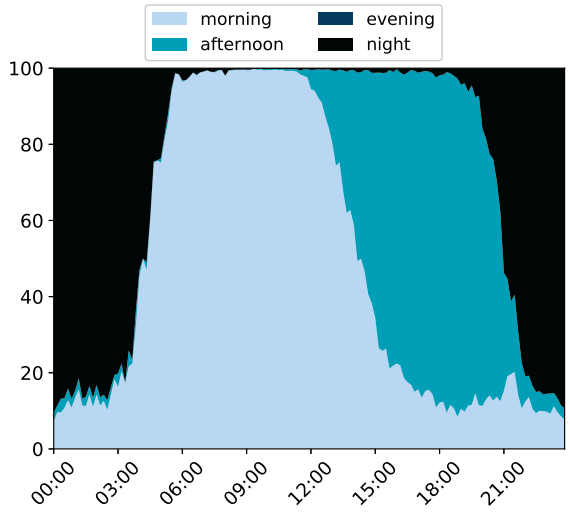

Figure 5: Same as Figure 4, but for Spain.

Area plots like these give a clear picture of the semantics of part-of-day nouns, as they depict the exact times when they are used. The precise semantics can be grounded more rigorously using statistical testing to know the exact time intervals at which people significantly use a specific greeting.

For example, to know when to switch from *good morning* to *good afternoon* in Spanish, we can: (1) group the number of 'buenos días' ('good morning') and 'buenas tardes' ('good afternoon') by intervals of 10 minutes, and (2) apply a binomial test to each interval, to determine if one of the greetings is significantly more likely to occur than the other (assuming equal probability of occurrence). For example, for Spain, we obtain that the morning ends at 14:00 (p-value=2×10^{-8} at 14:00, 0.09 at 14:10) and the afternoon starts at 14:40 (p-value becomes statistically significant again with 4×10^{-7}, showing a significant majority of *good afternoon*).

4 Conclusion

We crawled Twitter to study the semantics of part-of-day nouns in different countries and societies, showed examples from the polled period and ratified them against existing research and dated events. For space reasons we cannot show insights for all scenarios, but full results are at https://github.com/aghie/peoples2018_grounding.

Acknowledgments

DV and CGR receive funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), from the TELEPARES-UDC project (FFI2014-51978-C2-2-R) and the ANSWER-ASAP project (TIN2017-85160-C2-1-R) from MINECO, and from Xunta de Galicia (ED431B 2017/01).

References

Dan Dediu, Michael Cysouw, Stephen C. Levinson, Andrea Baronchelli, Morten H. Christiansen, William Croft, Nicholas Evans, Simon Garrod, Russell D. Gray, Anne Kandler, and Elena Lieven. 2013. Cultural evolution of language. In Peter J. Richerson and Morten H. Christiansen, editors, *Cultural Evolution: Society, Technology, Language, and Religion*, volume 12 of *Strüngmann Forum Reports*, pages 303–332. MIT Press, Cambridge, MA.

Rebecca Earle. 2010. "if you eat their food . . . ": Diets and bodies in early colonial spanish americarebecca earle"if you eat their food . . . ". *The American Historical Review*, 115(3):688–713.

Chenoa A Flippen. 2012. Laboring underground: The employment patterns of hispanic immigrant men in Durham, NC. *Social Problems*, 59(1):21–42.

Olaf Jäkel. 2003. 'Morning, noon and night': Denotational incongruencies between English and German.

In Cornelia Zelinsky-Wibbelt, editor, *Text, Context, Concepts*, pages 159–178. Mouton de Gruyter, Berlin/New York.

Cathy Yang Liu. 2013. Latino immigration and the low-skill urban labor market: The case of Atlanta. *Social Science Quarterly*, 94(1):131–157.

David Mosse, Sanjeev Gupta, Mona Mehta, Vidya Shah, Julia fnms Rees, and KRIBP Project Team. 2002. Brokered livelihoods: Debt, Labour Migration and Development in Tribal Western India. *The Journal of Development Studies*, 38(5):59–88.

Harriet Joseph Ottenheimer. 2013. *The Anthropology of Language: An Introduction to Linguistic Anthropology*, 3rd edition. Wadsworth, Cengage Learning, Englewood Cliffs, NJ.

Ehud Reiter and Somayajulu G. Sripada. 2003. Learning the meaning and usage of time phrases from a parallel text-data corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, pages 78–85.

Yaw Sekyi-Baidoo and Louisa A. Koranteng. 2008. English general greetings in the Ghanaian sociolinguistic context. *The International Journal of Language, Society and Culture*, pages 113–126.

Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003. Exploiting a parallel text-data corpus. In *Proceedings of Corpus Linguistics 2003*, pages 734–743.

Heather Trigg. 2004. Food choice and social identity in early colonial new mexico. *Journal of the Southwest*, pages 223–252.

Olivia J Walch, Amy Cochran, and Daniel B Forger. 2016. A global quantification of "normal" sleep schedules using smartphone data. *Science advances*, 2(5):e1501705.