Detection of Emerging Words in Portuguese Tweets

Afonso Pinto
ISCTE – Instituto Universitário de Lisboa, Portugal
http://www.iscte-iul.pt
adcmm@iscte.pt

Helena Moniz
CLUL/FLUL, Universidade de Lisboa, Portugal
INESC-ID, Lisboa, Portugal
UNBABEL, Lisboa, Portugal
http://www.inesc-id.pt
Helena.Moniz@inesc-id.pt

Fernando Batista
ISCTE – Instituto Universitário de Lisboa, Portugal
INESC-ID, Lisboa, Portugal
http://www.inesc-id.pt
fernando.batista@iscte-iul.pt

Abstract

This paper tackles the problem of detecting emerging words on a language, based on social networks content. It proposes an approach for detecting new words on Twitter, and reports the achieved results for a collection of 8 million Portuguese tweets. This study uses geolocated tweets, collected between January 2018 and June 2019, and written in the Portuguese territory. The first six months of the data were used to define an initial vocabulary on known words, and the following 12 months were used for identifying new words, thus testing our approach. The set of resulting words were manually analyzed, revealing a number of distinct events, and suggesting that Twitter may be a valuable resource for researching neology, and the dynamics of a language.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases Emerging words, Twitter, Portuguese language

Digital Object Identifier 10.4230/OASIcs.SLATE.2020.3

Funding This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

1 Introduction

Social networks are basically a way/facilitator for people to communicate and exchange ideas among themselves, so it is natural that they play an important role in the evolution of writing, reading, and take part in the introduction of new words and expressions. In recent years, social networks have become more and more part of the daily life of Portuguese society. Twitter started as a chat room with a limited number of people, and it is now a noisy place [16], where people, regardless of the age, gender, or social class, produce all possible content about the aspects of their daily lives, thus being the ideal place for a wide range of language studies.

According to [10], if the structure of language is necessary and is considered the learning basis for any language, then the role of vocabulary can also not be neglected since it provides the necessary means. Traditionally, two types of lexical variation have been identified [13]. The semasiological variation refers to the variation in the meaning of words, such as the

variation in the meaning of the word "cedo", which may denote various concepts depending on the context of the sentence in which it is found, may refer in temporal terms (e.g. *antes do tempo/before the time*), as it may be used in its verbal form, for example *ceder o lugar/give the place*, meaning to offer the seat to someone. The onomasiological variation refers to the variation of how the concepts are identified, such as, for example, the variation in the words to identify the waiting place for the bus, which in European Portuguese would be a "*paragem de autocarro/bus stop*" and in Brazilian Portuguese would be a "*ponto de ônibus/bus stop*". According to [9] semasiological variations involve changes in the meaning of words, while onomasiological variation involves changes in how words are identified, which includes the formation of new words.

Regarding the nature of words and distinctions of meaning, [17] presents four different ideas. The first one is that, except for technical words, two different words contain different meanings, so a literal translation may partially distort the true meaning that is intended to be conveyed. The second idea is that in most languages the number of words with only one meaning is very small. The third idea is that a word gets its meaning according to the context in which it is inserted. Finally, the fourth idea considers that in the same language there is no single word that can replace another, such as the word ocean and sea, although they are almost synonyms, they are used for a similar but different effect, therefore one cannot replace one by the other in certain contexts. So, two questions arise:

- What is the role that social networks play in the development of the vocabulary of a language?
- How social networks may be regarded as a linguistic resource for vocabulary development?

In addition to studying the words that appear for the first time on social networks, it is also important to understand whether people are aware of their importance. According to [2, 4, 11], there are blogs where users (students) can obtain knowledge by consulting them. Similarly, [14] highlighted the idea that online blogs and social networks may be didactic, since the students, by leaving comments and chatting with others, may improve their competence in a given subject. The importance of social networks for the development and evolution of the vocabulary of a language is now undeniable.

The area of the geographic spread of linguistic change is also of great importance, and has been the target of several theories proposed [5], being the "*wave model*" most notable, which has replaced the "*tree model*". The *wave model* predicts that new emerging words will be propagated radially with a central location limited only by physical distance, while the "hierarchical" model predicts that its propagation will also be conditioned by population density, thus propagating between urban cities and only later can it reach rural areas. Through empirical evidence, it was possible to understand that both models are related to each other [18, 3, 15], to that sense, the "counter-hierarchical model" was proposed in order to explain the spread between urban and rural areas. With this, the researchers demonstrated that with regard to propagation act the factors physical distance, population density and cultural patterns, however the intensity with which these factors intervene is not clear and seems to partially explain the propagation of new emerging words. [1] discovered demographic similarities between cities, which makes this an important factor in explaining the spread.

This work studies the emergence of new words in the Portuguese territory, during one year time-span. The interest in this topic has been around the world for several years. In line with our goals, the study reported by [8] describes a quantitative method for identifying new emerging words over a long period of time and then describes the analysis of the lexical emergence on social networks in U.S. territory, based on a universe of millions of word

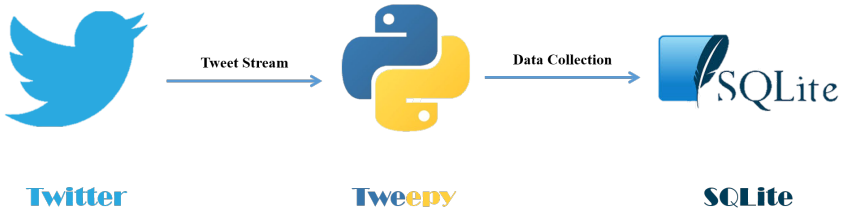

Figure 1 Data collection procedure.

creation, obtained over a period of one year through the internal Twitter API. The study identified 29 words and has examined them from various perspectives, in order to better understand the emergence process. To identify emerging words, two values were inserted into a data matrix, the relative frequency of each word at the beginning of the period and the degree of its frequency throughout that same period. After the list of potential emerging words was generated, several problems were identified: many of these words were names of people, products or company names, which did not meet the objective of the study, so they were manually excluded and 29 words were identified as emerging. The aim of this study was not to identify new words that only emerged during 2013 and 2014, but to identify words that emerged quickly on Twitter in 2014. Nevertheless, they used tools such as *Google Trends*[1] and *Urban Dictionary*[2] to identify the first appearance of the words in question. The analysis did not identify a large number of emerging words and there is a doubt as to whether only those identified will have emerged in Modern American English. A follow up of the above study was conducted by [7]. The same dataset of words was studied in order to *trace* the origin of words identified as *new words*. Through the mapping, five main regions where the appearance of new words is more productive were identified.

2 Corpus

Since tweets are an informal way of communicating, users feel more comfortable expressing themselves, which is a favorable factor for new words to emerge. This study uses a corpus of 7.721 million geolocated tweets, collected between January 2018 and June 2019 in the Portuguese territory, and corresponding to 18 months of data. Twitter allows, through the internal API, to obtain a large amount of data in a short period of time and provides a rich set of metadata, such as the username of the person who produced the content, the number of followers, the geolocation, date, among others. All data was obtained with geolocation information, date and time of publication. It is very common for the location to be obtained when publishing via a mobile phone (with the option of active geotracking), for example on the basis of geolocated data. Such information may be an important resource for finding patterns in the propagation path of a certain emerging word.

The data retrieval procedure is illustrated in Fig. 1. The data was collected directly from the Twitter platform API through python library *tweepy*, which allows access to the RESTful methods of the internal Twitter API. We have then stored the data into a lightweight SQLite database, which offers a reasonable performance, great portability, accessibility and may greatly reduce the complexity of the data analysis. Posts that have been *re-tweeted* have not been excluded from the database since they reflect an evidence of propagation of the emerging word.

[1] www.google.com/trends
[2] www.urbandictionary.com

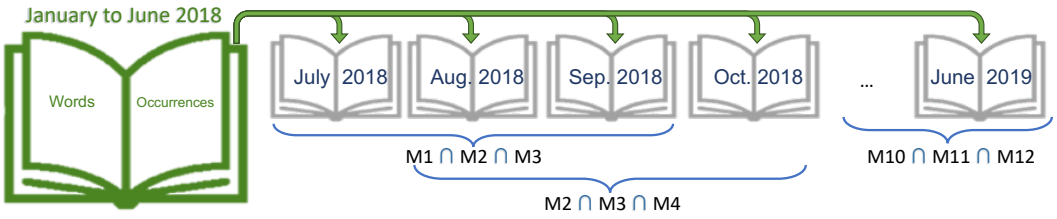

Figure 2 Approach for identifying new words.

3 Approach description

During a pre-processing stage, we have removed a number of tokens irrelevant for this study. Particularly, we have removed all the tokens starting with "*http*" or "@", which correspond to web addresses or user mentions, and all types of numbers. The word tokenization was performed using *TweetTokenizer* from the NLTK library, words with more than 3 repeating letters have also been normalized into a standard form keeping only 3 letters (e.g. the words "looooooove" is converted into "looove").

Our approach is illustrated in Fig. 2. We have started by creating an initial dictionary using the first six months of the data (between January and June 2018), corresponding to 2.9 million tweets. In a second stage we have identified all the possible new words for each of the following 12 months, considering only words appearing at least 50 times, by at least 15 different users. We have considered a minimum frequency of 50 in order to minimize the appearance of spelling errors. Finally, we have assumed that an emergent word would have to be used for at least a period of three consecutive months, so we have calculated the words appearing at least in three consecutive months. Table 1 shows the number of candidate words used at least in 3 consecutive months. It is interesting to notice that the number of words increases as the time period being considered is getting far away from the period used to create the initial dictionary, which suggests that the word usage changes over time.

In order to verify the existence of words in the Portuguese language, we have tested them using the DELAF[3] lexicon, which have all the words of the Portuguese language as well as their flexions, thus allowing to identify which of the emerging words are already in the dictionary of the Portuguese language.

FastText is a free open-source library that allows learning words and classifying text created by the AI (Artificial Intelligence) research laboratory. This library allows training sets of words supervised and unsupervised. The model allows the creation of algorithms to obtain vector word representations. The use of this library is intended to be a mechanism for explaining the origin motive of a given word. We have used two pre-trained word vectors for Portuguese[4], the first one trained on the Common Crawl[5], and the second trained on the Wikipedia[6]. The *fast*Text library has been used, since it is free and lightweight, when compared to other methods to achieve the same accuracy, (supervised) sentence vectors can be easily computed and *fast*Text works better in small datasets when compared with other libraries, such as *gensim*[7].

[3] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html
[4] https://fasttext.cc/docs/en/crawl-vectors.html
[5] https://commoncrawl.org
[6] https://www.wikipedia.org/
[7] https://pypi.org/project/gensim/

Table 1 Words shared by different periods of time.

set of months	shared words
2018-07, 2018-08, 2018-09	43
2018-08, 2018-09, 2018-10	70
2018-09, 2018-10, 2018-11	80
2018-10, 2018-11, 2018-12	83
2018-11, 2018-12, 2019-01	94
2018-12, 2019-01, 2019-02	104
2019-01, 2019-02, 2019-03	115
2019-02, 2019-03, 2019-04	129
2019-03, 2019-04, 2019-05	147
2019-04, 2019-05, 2019-06	146

4 Analysis of the results

We have combined all the words achieved in the previous stage, and reported in Fig. 1, into a single dictionary, totaling 496 words and containing all the candidate new words for one year time-span. We have also found that the 5 most prominent tokens are emojis/emoticons, possibly motivated by the appearance of a new application or by an update to an existent one. We have also found that some of the resulting tokens are, in fact, well-known words that were not present in the initial dictionary. That is due to several major factors: a) the initial dictionary, corresponding to the existing knowledge, was created from a very limited set of data; b) the typical vocabulary used in tweets is quite distinct from other sources, such as books or newspapers; c) the content produced in social networks and the corresponding word usage is highly influenced by ongoing events.

Finally, an important factor to take into consideration in order to identify a word as emerging is the number of different users using the word. In order to exclude words which had a reduced number of entries by different users it was required that they were mentioned by at least 15 different people. In fact, a few of the candidate words correspond to a typical situation were a person or a company starts using a given word in an exhausting way (e.g., *Saladumbras*, *Bonetto*), and they were discarded.

Table 2 presents the resulting top candidate tokens, after manually removing some of the entries mentioned above. Many of the identified tokens correspond to names of people, specially those related with soccer (e.g., *Keizer*, *Gudelj*, *Militao*, *Corchia*, *Phellype*, *Castaignos*, and *Manafá*). The remaining tokens were grouped as follows:

Emojis / Emoticons: emoticons, or textual portrayals of a writer's moods or facial expressions in the form of icons, usually used in conjunction with a sentence to express emotions.

Benfiquistão, minguem, Vagandas: these emerging words correspond to derivations of existing ones. The later are used in a very ironic sense, playing around with the name "Varandas", Sporting's President.

kbk: correspond to the formation of new slang words [6], used to abbreviate commonly-used expressions.

Table 2 Emerging words, and their corresponding frequency over 12 months.

token	freq.	comment
phellype	125723	football player
corchia	59514	football player
castaignos	52735	football player
bozo	41613	Brazilian Portuguese word, and English word for stupid man
120M	11786	price paid for a football player
benfiquistão	10381	football coach
trotinetes	9836	the word exists in Portuguese
trotinetas	9785	the same as trotinetes
taki	9710	name of a music (Taki Taki)
militao	9665	Militão is the name of a football player
minguem	9635	Minguém the same as “ninguém” (nobody)
manafá	8594	football player
vagandas	7146	irony for Varandas, Sporting’s President
shallow	5654	name of a music; used an in the English word
sicko	5237	name of a music (sicko mode)
keizer	4147	football coach
kbk	4147	slang for kill or be killed
lomotif	3135	name of an app
legacies	3082	name of a series
gudelj	2824	football player
guaidó	2689	Venezuelan politician
🥺	525	pleading face emoji
🥵	271	hot face emoji
🥳	163	face with party horn and party hat emoji
🥴	125	woozy face emoji
🥶	66	cold face emoji

Table 3 fastText outputs, using word vectors trained using *Wikipedia*.

Phellype	*corchia*	*castaignos*	*bozo*
Wesllem	Corchia	Castaignos	fiuk
Rithelly	Forchia	direntes	buneco
Fellype	Porchia	cosmonômicos	yudi
Sueliton	Torchia	cosmonômico	raxo
Laionel	Acrorchis	castanos	vsf
Douglão	Archia	compotados	veei
Aélson	Torchiara	presssibutraminabepantoldieta	adogo
Ediglê	Erythrorchis	insetrônicos	veii
Neílson	Brasiliorchis	sementeiros	affz
Roniel	Xerorchis	escravinhos	zuei

bozo: despite corresponding to an existing word, it's current true meaning was changed during the time period being studied. It is often used to refer to Bolsonaro, the President of Brazil since 1 January 2019, in a very depreciative way and making use of the phonetic realization of his name as in its Italian origins Bol[z]onaro.

Lomotif: correspond to names of new stores, brands or applications that were introduced meanwhile.

taki, sicko, shallow, legacies: the appearance of these words is related to new music or TV series. *Taki*, *sicko* and shallow both refer to the name of a music (Taki Taki, Sicko Mode and Shallow), while Legacies is the name of a new TV series..

The first four groups correspond to emergent words, now being in use by the Twitter community. While the remaining groups represent new events, usually Named Entities, which may not be considered interesting cases of emerging words.

Using DELAF lexicons, in order to denote which words, represented in Fig. 2, belong in the Portuguese dictionary, it was noticed that only the word "trotinetes" already exists, all the others are new or derivations of existing ones.

The 4 words with the highest number of occurrences were selected to be tested in the *fast*Text library. Tables 3and 4 contain the outputs for the words *Phellype*, *corchia*, *castaignos*, and *bozo*. "Fellype" or "fellype" correspond to a variation of the word *Phellype*, closer to the widely used Portuguese proper noun *Filipe*. Also for the word *corchia*, derivations include the Italian words "Porchia" or "forchia", which may suggest that this word may be originated in Italy. Similarly, the word *castaignos* obtained results in another language, this time in Spanish, which may also suggest that the origin of this word derives from this language. Finally, the fourth word with the most occurrences, "bozo", suggests that it may be related to the word "zombozo". This is the name given to a cartoon portrayed by a clown and cruel with a dark sense of humor from the world of Ben 10, an American animated series, which suggests that the context where the word "bozo" is found has a clown connotation.

It is important to mention that Twitter does not facilitate tracing the entire path of the emerging words, especially because the provided API does not allow to retrieve all the required tweets in a easy way. Despite that, we have analyzed the appearance and most frequent usage of some of the emerging words, paying attention to the regions where the

Table 4 fastText outputs, using word vectors trained using *Common Crawl*.

Phellype	*corchia*	*castaignos*	*bozo*
fellype	forchia	pelignos	zombozo
ype	torchia	castrais	esbozo
mayke	corchiano	malignos	dozo
jaílton	exárchia	intersignos	calabozo
lenílton	vitorchiano	signos	jozo
josimar	torchiara	castrados	kozo
rogerinho	sinerchia	bretaigne	bozomal
raţ	archia	moos	yozo
clebinho	anarchia	montaigne	bozon
neílton	senerchia	châtaigneraie	logozo

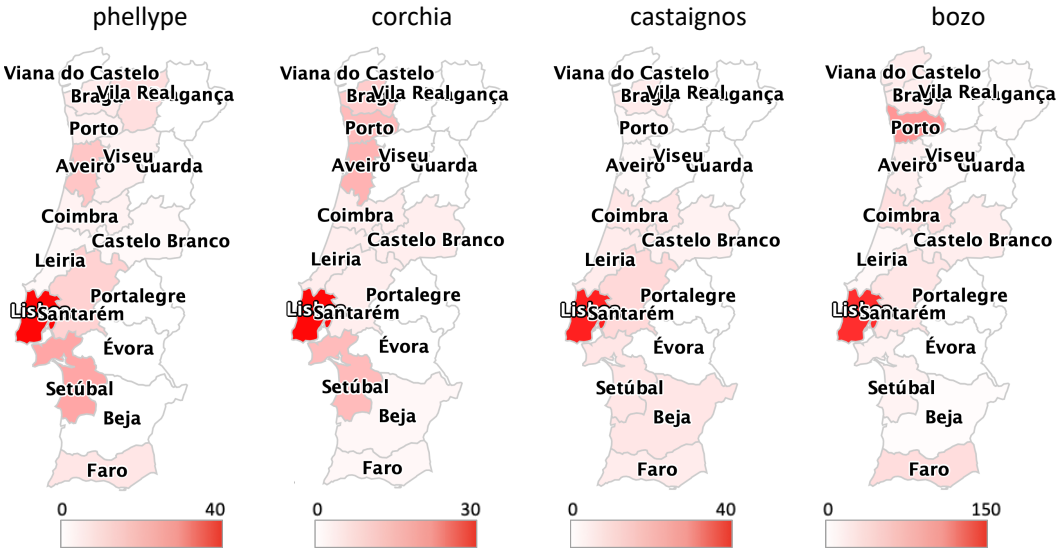

Figure 3 Frequency per district of words: *Phellype*, *corchia*, *castaignos*, and *bozo*.

word appeared first. Figure 3 shows the mapping of lexical innovation of the 4 words with the highest number of occurrences in the dataset that were identified as emerging words. Regions marked with a darker color correspond to the places where each word was more frequently used during the period in analysis. The figure reveals that they appear in urban areas where population density is higher and Internet is more frequently available [12], as expected.

We believe our study has made four general contributions:

- It is possible to observe regional patterns of word spread, even if it is not possible to affirm that the words occurred for the first time on social media.
- Words tend to follow a consistent path, as it is possible to see in Figure 3 that Lisbon will be the city with the center of propagation, registering a spread to the surrounding cities, with a natural decrease in the number of occurrences of the words.
- Population density has an important role in the spread of words and appears to be more fundamental than cultural or religious issues.
- Brazilian Portuguese will be one of the main sources of lexical innovation.

Twitter is only a source of virtual expression, which does not allow the generalization of results for the majority of the Portuguese population and presumably the words did not occur online for the first time. A corpus allows to partially detect patterns of lexical innovation in a language, especially in a modern world where virtual communications occur in the most varied platforms. However, since a large part of the words in twitter cover a common discourse we believe that the results achieved reflect a rather realistic scenario.

5 Conclusions and future work

We have presented an approach that uses a corpus containing a considerable number of tweets to detect the appearance of possible emerging words. We have applied our approach to tweets written in Portuguese, and produced in the Portuguese region over 18 months. Despite the limitations of our corpus, the proposed approach led to the discovery of a number of relevant linguistic phenomena in the achieved set of candidate words. The final set of possible emerging words was achieved by performing a manual analysis and selection of the retrieved words. This preliminary study has provided a methodological framework for future research in the field of neology. Our findings suggest that social networks, and Twitter in particular due to its nature, are a promising way of studying the dynamics of a language.

As a follow-up for the present study, we plan to use other sources of information, such as newspapers, blogs, and other social networks as means to trace a complete path of the emerging words, thus contributing to the better understanding of the evolution of the vocabulary in a given language. We plan to better characterize each word by providing an extended analysis over time, by mapping its usage, adoption by communities, and mapping its propagation path. We realize that there are no standard steps to address this issue, but most data sources are now providing geolocation and time information, which constitute a relevant advantage for tracing reliable geo-temporal propagation paths of a word in the near future.

References

1 David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014. doi:10.1111/josl.12080.

2 Rebecca Blood. Weblogs: A history and perspective, September 2000. URL: http://www.rebeccablood.net/essays/weblog_history.html.

3 Charles Boberg. Geolinguistic diffusion and the U.S. - Canada border. *Language Variation and Change*, 12:1–24, March 2000. doi:10.1017/S0954394500121015.

4 Marilyn Dyrud, Rebecca Worley, and Marie Flatley. Blogging for enhanced teaching and learning. *Business Communication Quarterly*, 68, March 2005. doi:10.1177/108056990506800111.

5 Alexandre François. Trees, Waves and Linkages: Models of Language Diversification. In Claire Bowern and Bethwyn Evans, editors, *The Routledge Handbook of Historical Linguistics*, chapter Trees, Waves and Linkages: Models of Language Diversification, pages 161–189. Routledge, London, June 2014. doi:10.4324/9781315794013.ch6.

6 Jonathon Green and David Kendal. Writing and publishing green's dictionary of slang. *Dictionaries: Journal of the Dictionary Society of North America*, 38:82–95, January 2017. doi:10.1353/dic.2017.0003.

7 Jack Grieve. Dialect variation. In Douglas Biber and RandiEditors Reppen, editors, *The Cambridge Handbook of English Corpus Linguistics*, Cambridge Handbooks in Language and Linguistics, pages 362–380. Cambridge University Press, Cambridge (UK), 2015. doi:10.1017/CBO9781139764377.021.

8 Jack Grieve, Andrea Nini, and Diansheng Guo. Analyzing lexical emergence in modern american english online. *English Language and Linguistics*, 21(1):99–127, 2017. doi:10.1017/S1360674316000113.

9 Stefan Grondelaers, Dirk Geeraerts, and Dirk Speelman. Lexical variation and change. In Dirk Geeraerts and Hubert Cuyckens, editors, *The Oxford Handbook of Cognitive Linguistics*, pages 988–1011. Oxford University Press, 2012. doi:10.1093/oxfordhb/9780199738632.013.0037.

10 Jeremy Harmer. The Practice of English Language Teaching. *SERBIULA (Sistema Librum 2.0)*, January 2001.

11 Sara Kajder, Glen Bull, and Emily Van Noy. A space for "writing without writing.". *Learning and Leading with Technology*, 31:32–35, 2004. URL: https://files.eric.ed.gov/fulltext/EJ695756.pdf.

12 T. Lapa, Jorge Vieira, J. Azevedo, and G. Cardoso. As desigualdades digitais e a sociedade portuguesa: divisão, continuidades e mudanças. In *Desigualdades Sociais: Portugal e a Europa*, pages 257–257. Mundos Sociais, Lisboa, 2018. URL: http://www.mundossociais.com/livro/desigualdades-sociais/112.

13 M Lynne Murphy. Theories of lexical semantics by Dirk Geeraerts. *Journal of Linguistics*, 47:231–236, January 2011. doi:10.2307/41261748.

14 Dilip Mutum and Qing Wang. Consumer generated advertising in blogs. In M.S. Eastin, T. Daugherty, and N. Burns, editors, *Handbook of Research on Digital Media and Advertising: User Generated Content Consumption*, chapter 13, pages 248–261. IGI Global, 2010. doi:10.4018/978-1-60566-792-8.ch013.

15 John Nerbonne. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3821–3828, 2010. doi:10.1098/rstb.2010.0048.

16 João Pedro Pereira. Era uma vez o Twitter em Portugal. *Público*, 77(3):95–106, 2016.

17 S.M. Shahid. Teaching of English an Introduction. *Majeed Book Depot Urdu Bazar Lahore*, 2002.

18 Peter Trudgill. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2):215–246, 1974. URL: http://www.jstor.org/stable/4166764.