A little bit of bella pianura: Detecting Code-Mixing in Historical English Travel Writing

Rachele Sprugnoli[1-2], Sara Tonelli[1], Giovanni Moretti[1], Stefano Menini[1-2]
[1]Fondazione Bruno Kessler, Trento
[2] Università di Trento
{sprugnoli,satonelli,moretti,menini}@fbk.eu

Abstract

English. Code-mixing is the alternation between two or more languages in the same text. This phenomenon is very relevant in the travel domain, since it can provide new insight in the way foreign cultures are perceived and described to the readers. In this paper, we analyse English-Italian code-mixing in historical English travel writings about Italy. We retrain and compare two existing systems for the automatic detection of code-mixing, and analyse the semantic categories mostly connected to Italian. Besides, we release the domain corpus used in our experiments and the output of the extraction.

Italiano. *Il code-mixing è l'alternanza di lingue diverse nello stesso testo. Questo fenomeno è particolarmente importante nel dominio dei viaggi, poiché aiuta a comprendere meglio il modo in cui vengono percepite e descritte culture diverse da quella dell'autore. In questo lavoro, analizziamo il code-mixing tra inglese ed italiano nei testi di viaggio scritti in inglese e aventi come soggetto l'Italia. A questo scopo confrontiamo due sistemi esistenti per il riconoscimento automatico del code-mixing dopo averli ri-addestrati e analizziamo le categorie semantiche connesse alle parole/espressioni italiane. Inoltre, rilasciamo il corpus e il risultato dell'estrazione.*

1 Introduction

Code-mixing is the alternation between two or more languages that can occur between sentences (inter-sentential), within the same utterance (intra-sentential), or even inside a single token (mixing of morphemes). This phenomenon has been widely studied from the linguistic, psycholinguistic, and sociolinguistic point of view (Gardner-Chloros, 1995; Grosjean, 1995; Ho, 2007) but there is no consensus on the terminology to be adopted. In this paper code-mixing is used as an umbrella term to indicate a manifestation of language contact subsuming other expressions such as code-switching, languaging, borrowing, language crossing (Muysken, 2000).

Code-mixing characterizes communication of post-colonial, migrant and multilingual communities (Papalexakis et al., 2014; Frey et al., 2016) and it emerges in different types of documents, for example parliamentary debates, interviews and social media posts (Carpuat, 2014; Das and Gambäck, 2015; Piergallini et al., 2016). Travel writings (e.g. guidebooks, travelogues, diaries, blogs, travel articles in magazines) are affected as well by this phenomenon that has been studied in particular by analyzing small corpora of contemporary tourism discourse through manual inspection (Dann, 1996). Even if code-mixing occurs in less than 1% of the cases (Cappelli, 2013), it has several important functions in the travel domain: it gives a "linguistic sense of place" (Cortese and Hymes, 2001), it adds authenticity to a narration, it provides translation of cultural-specific words and it is a mean to define social identity ("us" tourists *versus* "they" locals) (Jaworski et al., 2003).

In this work, we investigate the phenomenon of code-mixing in travel writings, but differently from previous works we shift the focus of analysis from contemporary to historical data and from manual to automatic information extraction. As for the first point, we present a corpus of more than 3.5 millions words of English travel writings published between the end of the XIX Century and the beginning of the XX Century, which we have retrieved from freely available sources and we release in a cleaned format. As for automatic information extraction, we retrain two state-of-the-art tools to identify English-Italian code-mixing and evaluate them on a sample of our dataset. We further launch the best system on the whole dataset and then we perform a semi-automatic refinement of the automatic annotation. The corpus, the training and test data and the outcome of the extraction are available online[1].

2 Related Work

Automatic language identification of monolingual documents has a long tradition in Natural Language Processing (Hughes et al., 2006; Lui and Baldwin, 2012). More recently a new hot topic of research has emerged, that is the detection of language at word level in code-mixing texts. Dedicated workshops and evaluation exercises have been organized on this task dealing with different pairs of languages and with social media data (Choudhury et al., 2014; Solorio et al., 2014; Molina et al., 2016). The most common approach of the proposed systems is based on Conditional Random Fields (CRFs) but there are also implementations of Logistic Regression and deep learning algorithms.

To the best our knowledge, there is no previous work on the automatic identification of code-mixing in travel writing. Cappelli (2013) and Gandin (2014) have studied the phenomenon, but they have mainly used standard corpus linguistics tools, i.e. WordSmith (Scott, 2008), to analyse language contact in English guidebooks, travel blogs written by expatriates and travel articles from 2002-2012.

3 Corpus Description

Differently from the works cited in the previous Section, we focus on historical texts. To this end, we collect from Project Gutenberg[2] a corpus of travel writings about Italy written by English native authors and published between the country unification and the beginning of the 30's. We choose this period because in the second half of the XIX Century the tradition of the Grand Tour declined and leisure-oriented travels emerged. This radical transformation was enabled by technological, economic and sociological, factors, such as the development of steam-powered ships and of the railway network, the growth of Anglo-American economy and a greater emancipation of women with more female travelers (Schriber, 1995). Moreover, after unification, new routes to Southern Italy and the islands were opened, so that travelers' attention was no longer limited to the classic destinations in the North and Central Italy, such as Venice, Florence and Rome (Ouditt and Polezzi, 2012).

The corpus is made by 57 texts[3], divided into *travel narratives* (reports, diaries, collections of letters) and *guidebooks*, for a total of 3,630,781 tokens. We distinguish between these two types of text, following a standard classification of documents in the travel domain. However, the distinction was not so clear-cut in the period we take into account as it is now, since reports on personal travel experiences were often mixed with practical recommendations and long disquisitions on art and history. Therefore, we adopt as a rule of thumb the distinction suggested in (Santulli, 2007): travel narratives are those told in the first person, while guidebooks are written in impersonal form.

The authors of the selected texts belong to different nationalities (UK, US, Ireland, Australia) and are both male and female. Some books dwell on specific cities or regions, others cover different parts of Italy or even several countries: in the latter case we extracted only the chapters related to Italy. Although we made an effort to have a diverse, well-balanced corpus in terms of content, author's gender and nationality, this was only partially possible because of the limited availability of online travel books whose text is freely available and cleaned from OCR errors. The distribution of tokens according to the year of publication and type of text is shown in Fig. 1. Details about authors are given in a spreadsheet provided together with the corpus.

4 Code-Mixing Detection

In this Section we describe the experiments on code-mixing, comparing the performance of two available systems in different configurations. We also detail the post-processing step introduced to refine the output of the best performing system.

[1] https://dh.fbk.eu/technologies/code-mixing

[2] https://www.gutenberg.org/

[3] Thirty of these texts are also available in TEI-XML format on the website https://sites.google.com/view/travelwritingsonitaly.

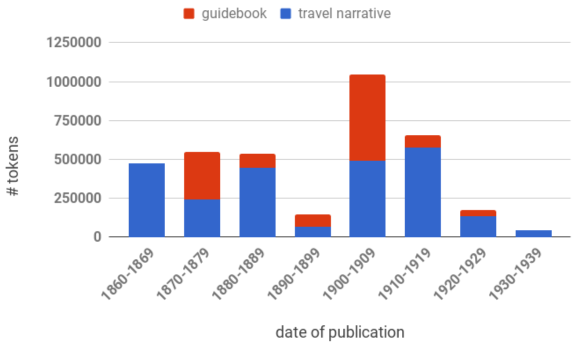

Figure 1: Distribution of tokens per year of publication and sub-genre.

4.1 Experimental Setting

In order to automatically extract Italian words, expressions and sentences from the corpus described in Section 3, we train and test two systems whose source code is available on the web. The first one (henceforth, *langid*) is based on character n-grams (n = 1 to 5) and adopts a weakly supervised approach, i.e. training data are monolingual texts of few thousand tokens (King and Abney, 2013). This system includes four classification algorithms: Conditional Random Field (CRF), Hidden Markov Model (HMM) and Maximum Entropy Model with and without generalized expectation criteria (MaxEnt-GE and MaxEnt). *langid* has been successfully evaluated on documents containing English texts mixed with 30 different minority languages such as Zulu and Chippewa[4].

For our experiments, we retrain *langid* using a collection of about 300,000 tokens taken from monolingual Italian and English books, of different genres, published in the same period of our corpus[5].

The second system (henceforth, *CodeSwitching*), has been developed to detect languages in texts mixing Latin and Middle English (Schulz and Keller, 2016). It implements a CRF classifier with features generated from TreeTagger models and word lists of both languages[6]. Differently from *langid* that classifies words as belonging to one language rather than the other, this latter system performs a fine-grained annotation by distinguishing five classes (see below). Since this system is fully supervised, we create a training set by manually annotating 3,900 tokens from 4 samples extracted from our corpus, a size in line with the training data used in the original paper. The training data were annotated with 5 different classes: Italian tokens (*i*), English tokens (*e*), punctuation (*p*), named entities (NEs) (*n*), and ambiguous tokens that belong to the dictionary of both languages (*a*).

Both *langid* and *CodeSwitching* were evaluated on the same test set, i.e. two samples of texts (one from a travel narrative and one from a guidebook) of 1,623 tokens. The test set was annotated by assigning to each token a label for English or Italian, as required by *langid*, and also marking punctuation, NEs and ambiguous tokens, following *CodeSwitching* scheme. Since the performance of *CodeSwitching* is sensitive to the length of the input file, we split the test set in batches of 40 sentences, replicating the experimental setting presented in (Schulz and Keller, 2016).

4.2 Evaluation

Table 1 presents the performances of *langid* on the test set: contrary to the results achieved by King and Abney (2013), HMM – not CRF – proved to be the best approach. This is likely due to the greater sparseness of the code-mixing phenomenon in our dataset with respect to what was registered in the original corpus, where languages different from English cover the 56% of the overall number of tokens.

Table 2 reports Precision, Recall and F-measure of the retrained *CodeSwitching* system. Even if the overall performance is slightly better than the one obtained with HMM in *langid*, the scores for the detection of Italian tokens (*i*) are lower (0.82 versus 0.90 in terms of F-measure). Punctuation (*i*) and ambiguous tokens (*a*) are generally detected with a good performance, while NEs (*e*) represent the most challenging class. Given that we are mainly interested in recognising English and Ital-

[4] http://www-personal.umich.edu/~benking/resources/langid_release.tar.gz

[5] For Italian: "Le Avventure di Pinocchio" by C. Collodi, "Una donna" by S. Aleramo, "Il Valdarno da Firenze al mare" by G. Carocci, "La vita operosa" by M. Bontempelli, "Dopo il divorzio" by G. Deledda, "Novelle umoristiche" by A. Albertazzi, "Lezioni e Racconti per i bambini" by I. Baccini. For English: "The Adventures of Tom Sawyer" by M. Twain, "Pioneers of the Old Southwest" by C. L. Skinner, "The Happy Prince, and Other Tales" by O. Wilde, "Vanished Arizona" by M. Summerhayes, "The Tale of Peter Rabbit" by B. Potter, "The Strange Case of Dr. Jekyll and Mr. Hyde" by R. L. Stevenson.

[6] https://github.com/sarschu/CodeSwitching

	CRF	HMM	MaxEnt	MaxEnt-GE
P	1	**0.89**	0.59	0.82
R	0.51	**0.92**	0.90	0.47
F	0.67	**0.90**	0.71	0.60

Table 1: Results of the evaluation on the retrained *langid* system in terms of precision (P), recall (R), and F-Measure (F).

	i	*e*	*a*	*n*	*p*	ALL
P	**0.83**	0.98	0.98	0.85	0.98	0.92
R	**0.80**	0.99	0.90	0.85	0.96	0.90
F	**0.82**	0.99	0.94	0.85	0.97	0.91

Table 2: Results of the evaluation on the retrained *CodeSwitching* system in terms of precision (P), recall (R), and F-Measure (F) for each class and the macro-average of all classes.

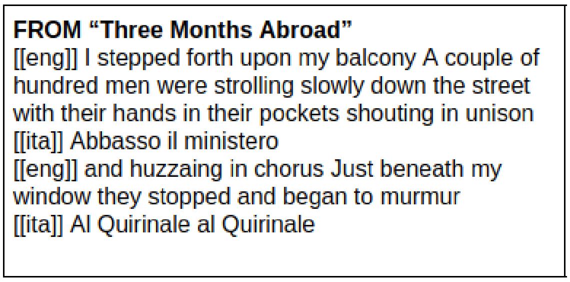
FROM "Three Months Abroad"
[[eng]] I stepped forth upon my balcony A couple of hundred men were strolling slowly down the street with their hands in their pockets shouting in unison
[[ita]] Abbasso il ministero
[[eng]] and huzzaing in chorus Just beneath my window they stopped and began to murmur
[[ita]] Al Quirinale al Quirinale

Figure 2: Examples of *langid* output.

ian terms, and that on this task *langid* performs better, we run this tool on the whole corpus.

4.3 Post-processing

In order to refine the output of *langid* (see Figure 2), we perform three post-processing steps. First of all, we check whether tokens tagged as Italian are included in Morph-it, an Italian lexicon of inflected forms (Zanchetta and Baroni, 2005): in this way we are able to detect false positives. Then, we run the Polyglot Python module on the corpus to find out if the processed documents contain other languages beside English and Italian[7]. Indeed 27 books result to have a high probability of including text written also in Latin, French, Germany or Greek. These books are likely to be problematic given that *langid* recognizes only English and Italian. Information obtained in these two steps are then used to manually check the outcome of *langid* extraction and correct it semi-automatically. Furthermore, we employ the USAS Italian semantic tagger (Piao et al., 2015) to obtain a categorization of the terms tagged as Italian. Based on the 21 semantic classes recognised by USAS, we are able to understand in which cases and why writers used to switch their narration from English to Italian.

5 Discussion

The classification performed with the USAS tagger shows that Italian is adopted to express concepts covered by 20 semantic classes, both in guidebooks and in travel narratives. Only one USAS class, the one related to "Science and technology", is not found in the corpus. Table 5 shows frequency and examples for each detected class. As in contemporary travel writings (Francesconi, 2007), food is well represented: traditional dishes, drinks and products (e.g. *polenta*, *Chianti*, *mortadella*) appear together with fruits, vegetables (e.g. *mandarini*, *finocchio*) and also eating establishments (e.g. *osteria*, *trattoria*, *locanda*). The attention for Italian art and architecture manifests itself through the use of many specialized terms (*cassettoni*, *gotico*, *giallo antico*). The semantic areas of emotions and psychological processes are not recorded in previous work on contemporary texts but are frequent especially in travel reports (e.g. *addolorata*, *trionfo*, *simpatico*). As for NEs, city names reveal an increasing interest for towns in Central regions (for example, *Perugia* has a high frequency of occurrence in both genres). Moreover, following Italy unification, travellers discovered several locations in the South (e.g. *Ragusa*, *Catanzaro*). Among the most mentioned people, there are representatives of past Italian politics (e.g. *Lorenzo and Cosimo de Medici*), artists (e.g. *Giotto*, *Dante*) and religious figures (e.g. *Madonna*, *San Michele*).

In many cases, the use of Italian is not limited to single words or multi-token expressions (e.g. *appartamento signorile*) but longer utterances are reported. Texts of both genres contain proverbs (e.g. *chi tardi arriva mal alloggia*) and citations, not only from the canon of Italian literature, such as Leopardi's poems, but also from the popular tradition, such as Tuscan songs (*O rosa O rosa O rosa gentillina*). The main difference between travel narratives and guidebooks is the greater presence in the former of dialogues or expressions heard by the author during his/her stay in Italy (*voi siete un*

[7] http://polyglot.readthedocs.io/en/latest/Installation.html

GUIDEBOOKS			TRAVEL NARRATIVES		
SEMANTIC CLASS	**#**	**EXAMPLES**	**SEMANTIC CLASS**	**#**	**EXAMPLES**
names & grammar	29,927	*Pisa*	names & grammar	28,694	*Donatello*
architecture	3,070	*villa*	social elements	3,134	*popolo*
movement	2,294	*automobile*	architecture	3,065	*palazzo*
social elements	1,590	*trinità*	environment	1,311	*lago*
materials & objects	717	*fontana*	movement	1,207	*vetturino*
environment	713	*campagna*	materials & objects	965	*rosso*
general/abstract terms	580	*essere*	general/abstract terms	943	*fare*
measurement	340	*alto*	food & farming	665	*trattoria*
arts & crafts	231	*stucco*	life	479	*fiore*
time	225	*nuovo*	measurement	464	*grande*
life	222	*agnello*	time	379	*primavera*
body	211	*cintola*	body	350	*braccio*
public domain	205	*podestà*	psyche	330	*vedere*
psyche	198	*volere*	entertainment	319	*marionetta*
food & farming	162	*maccaroni*	money & commerce	269	*dazio*
entertainment	141	*giuoco*	communication	268	*dire*
emotion	137	*amore*	public domain	260	*carabiniere*
communication	131	*motto*	arts & crafts	206	*arte*
money & commerce	127	*soldo*	emotion	176	*evviva*
education	22	*università*	education	135	*maestro*

Table 3: Italian word frequency for each semantic class

cattivo; e voi siete bella).

6 Conclusions and Future Work

In this work, we presented the first automated analysis of code-mixing in historical travel writings. In particular, we focus on English documents about Italy, and we compare guidebooks and travel narratives, analysing the semantic categories mostly related to code-mixing.

In the future, we plan to investigate how code-mixing phenomena relate to content types in travel writings (Sprugnoli et al., 2017). Besides, we are planning to implement an algorithm to automatically link code-mixing quotations to their original source text. Finally, we would like to extend our experiments to recognise code-mixing in multiple languages, and compare the semantic domains specific to each language.

References

Gloria Cappelli. 2013. Travelling words: Languaging in english tourism discourse. *Travels and translations*, pages 353–374.

Marine Carpuat. 2014. Mixed-language and code-switching in the canadian hansard. In *Proceedings of EMNLP 2014*, page 107.

Monojit Choudhury, Gokul Chittaranjan, Parth Gupta, and Amitava Das. 2014. Overview of fire 2014 track on transliterated search. In *Proceedings of FIRE*.

Giuseppina Cortese and Dell Hymes. 2001. Languaging in and across human groups. *Perspectives on difference and asymmetry. Textus. English Studies in Italy*, 14(2).

Graham MS Dann. 1996. *The language of tourism: a sociolinguistic perspective*. Cab International.

Amitava Das and Björn Gambäck. 2015. Code-mixing in social media text: the last language identification frontier? *Revue TAL*, pages 41–64.

Sabrina Francesconi. 2007. Italian borrowings from the semantic fields of food and drink in English tourism texts. *The Languages of Tourism: turismo e mediazione, Milano: Unicopli*, page 129.

Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W Stemle. 2016. The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. In *Proceedings of CLIC-it*.

Stefania Gandin. 2014. Investigating loan words and expressions in tourism discourse: A corpus driven analysis on the bbctravel corpus. *European Scientific Journal*, 10(2).

Penelope Gardner-Chloros. 1995. Code-switching in community, regional and national repertoires: the myth of the discreteness of linguistic systems. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 68–89.

François Grosjean. 1995. A psycholinguistic approach to code-switching: The recognition of guest words by bilinguals. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 259–275.

Judy Woon Yee Ho. 2007. Code-mixing: Linguistic form and socio-cultural meaning. *The International Journal of Language Society and Culture*, 21.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proc. International Conference on Language Resources and Evaluation*, pages 485–488.

Adam Jaworski, Crispin Thurlow, Sarah Lawson, and Virpi Ylänne-McEwen. 2003. The uses and representations of local languages in tourist destinations: A view from British TV holiday programmes. *Language Awareness*, 12(1):5–29.

Ben King and Steven P Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of HLT-NAACL*, pages 1110–1119.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Giovanni Molina, Nicolas Rey-Villamizar, Thamar Solorio, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, and Mona Diab. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of EMNLP 2016*, pages 40–49.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Sharon Ouditt and Loredana Polezzi. 2012. Introduction: Italy as place and space. *Studies in Travel Writing*, 16(2):97–105.

Evangelos Papalexakis, Dong-Phuong Nguyen, and A Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics.

Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. Association for Computational Linguistics.

Mario Piergallini, Rouzbeh Shirvani, Gauri S Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in swahili-english language data. In *Proceedings of EMNLP 2016*.

Francesca Santulli. 2007. Le parole ei luoghi: descrizione e racconto. *Antelmi, Donelli/Held, Gudrun/Santulli, Francesca*, pages 81–153.

Mary Suzanne Schriber. 1995. Women's place in travel texts. *Prospects*, 20:161179.

Sarah Schulz and Mareike Keller. 2016. Code-switching ubique est – Language identification and part-of-speech tagging for historical mixed text. In *Proceedings of LaTeCH Workshop*.

Mike Scott. 2008. WordSmith tools version 5. *Liverpool: Lexical Analysis Software*, 122.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Rachele Sprugnoli, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. 2017. The Content Types Dataset: a new Resource to Explore semantic and functional Characteristics of Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 260–266, Valencia, Spain, April. Association for Computational Linguistics.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).