

Anna Corazza, Simonetta Montemagni and Giovanni Semeraro (dir.)

Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016
5-6 December 2016, Napoli

Accademia University Press

Emojitalianobot and EmojiWorldBot

New online tools and digital environments for translation into emoji

Johanna Monti, Federico Sangati, Francesca Chiusaroli, Martin Benjamin and Sina Mansour

DOI: 10.4000/books.aaccademia.1811
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2016
Published on OpenEdition Books: July 26, 2017
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9788899982546

OpenEdition Books

http://books.openedition.org

Electronic reference
MONTI, Johanna ; et al. *Emojitalianobot and EmojiWorldBot: New online tools and digital environments for translation into emoji* In: *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016: 5-6 December 2016, Napoli* [online]. Torino: Accademia University Press, 2016 (generated 06 octobre 2023). Available on the Internet: <http://books.openedition.org/aaccademia/1811>. ISBN: 9788899982546. DOI: https://doi.org/10.4000/books.aaccademia.1811.

The text only may be used under licence CC BY-NC-ND 4.0. All other elements (illustrations, imported files) are “All rights reserved”, unless otherwise stated.

Emojitalianobot and EmojiWorldBot

New online tools and digital environments for translation into emoji

Johanna Monti
L'Orientale University
Naples, Italy
jmonti@unior.it

Federico Sangati
Independent Researcher
The Netherlands
federico.sangati@gmail.com

Francesca Chiusaroli
University of Macerata
Italy
f.chiusaroli@unimc.it

Martin Benjamin
EPFL
Lausanne, Switzerland
martin@kamusiproject.org

Sina Mansour
EPFL
Lausanne, Switzerland
mansour@ee.sharif.edu

Abstract

English. Emojitalianobot and EmojiWorldBot are two new online tools and digital environments for translation into emoji on Telegram, the popular instant messaging platform. Emojitalianobot is the first open and free Emoji-Italian and Emoji-English translation bot based on Unicode descriptions. The bot was designed to support the translation of Pinocchio into emoji carried out by the followers of the "Scritture brevi" blog on Twitter and contains a glossary with all the uses of emojis in the translation of the famous Italian novel. EmojiWorldBot, an off-spring project of Emojitalianobot, is a multilingual dictionary that uses Emoji as a pivot language from dozens of different languages. Currently the emoji-word and word-emoji functions are available for 72 languages imported from the Unicode tables and provide users with an easy search capability to map words in each of these languages to emojis, and vice versa. This paper presents the projects, the background and the main characteristics of these applications.

Italiano. *Emojitalianobot e EmojiWorldBot sono due applicazioni online per la traduzione in e da emoji su Telegram, la popolare piattaforma di messaggistica istantanea. Emojitalianobot è il primo bot aperto e gratuito di traduzione che contiene i dizionari Emoji-Italiano ed Emoji-Inglese basati sule descrizioni Unicode. Il bot è stato ideato per coadiuvare la traduzione di Pinocchio in emoji su Twitter da parte dei follower del blog Scritture brevi e contiene pertanto anche il glossario con tutti gli usi degli emoji nella traduzione del celebre romanzo per ragazzi. EmojiWorldBot, epigono di Emojitalianobot, è un dizionario multilingue che usa gli emoji come lingua pivot tra dozzine di lingue differenti. Attualmente le funzioni emoji-parola e parola-emoji sono disponibili per 72 lingue importate dalle tabelle Unicode e forniscono agli utenti delle semplici funzioni di ricerca per trovare le corrispondenze in emoji delle parole e viceversa per ciascuna di queste lingue. Questo contributo presenta i progetti, il background e le principali caratteristiche di queste applicazioni.*

1 Introduction

Emojitalianobot[1] and *EmojiWorldBot*[2] are two new translation bots[3] into and from emoji. These two bots were designed starting from the hypothesis of setting up an emoji multilingual dictionary and translator through a process of selection and assessment of conventional semantic values. Translation cases may show how images can convey common and universal meanings, beyond specific peculiarities, so as they can stand as models in the perspective of an interlanguage (Chiusaroli, 2015). The two

[1] https://telegram.me/emojitalianobot/
[2] https://telegram.me/emojiworldbot
[3] Computer programmes that carry out repetitive tasks and in their more sophisticated form can also simulate human behaviours.

bots ease the use of emojis but also collect, refine and make available valuable linguistic data by means of crowdsourcing and gamification approaches.

This contribution presents the state-of-the-art concerning the use of crowdsourcing and gamification approaches to linguistics in section 2, the *Emojitalianobot* and the *Pinocchio* project in section 3, the *EmojiWorldBot* in section 4 and finally conclusions and future work in section 5.

2 Crowdsourcing and gamification

Crowdsourcing, i.e., the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call (Howe, 2006) is becoming a widespread practice on the Internet to develop linguistic resources (dictionaries, glossaries, translation memories, etc.) or services (translation, localisation, fansubbing, etc.) (Monti, 2012, 2014). It allows the large scale involvement of users who contribute with their knowledge, their ideas, and their skills, in this way performing an active role in the achievement of a common goal. Crowdsourcing can be used for the creation, maintenance and sharing of lexical/terminological data such as: i. lexical resources for online dictionaries, e.g., Wiki platforms such as Wiktionary[4] and Omegawiki[5], and recent forays by more traditional dictionary publishing companies like Collins, Oxford, and Macmillan; ii. terminological resources for online terminological databases, like TermWiki[6] , the terminological counterpart of Wiktionary or TaaS[7]; iii. lexical and semantic resources for Natural Language Processing (NLP) tasks, such as Word Sense Disambiguation (WSA), Sentiment Analysis, Computer Aided Translation, Machine Translation and so on, using platforms for distributing parts of large development projects to professional or occasional lexicographers such as Mechanical Turk[8]. To the best of our knowledge only very few projects so far have been tailored to mobile devices to gather linguistic data in the field, (i) to collect dialect data as in Dialectbot[9], (ii) to document endangered languages as in Aikuma[10] and Ma Iwaidja[11], or (ii) to gather grammaticality judgments (Madnani et al., 2011). The social dimension of these types of activities is sometimes connected and fed by social communities, where users discuss problems, give suggestions, and exchange ideas (Brabham, 2012; McGonigal, 2011). In order to loyalize social communities and improve their engagement, gamification is used very often. The use of games is a very effective tool for active participation since it provides a strong motivational framework which pushes people to act for good. Some effective uses of games are to create new habits or modify wrong actions. Wang et al. (2013) list Games with a purpose (GWAPs)[12] among the different types of crowdsourcing. Some good examples of games with a purpose in the lexicographic field are Phrase Detectives[13] and JeuxDeMots[14]. The main advantage of GWAPs is their high attractiveness, because people love playing games and it is easier to obtain their contribution in this way in comparison to other forms of crowdsourcing. The difficulty in designing such games is to match attractiveness with usefulness, i.e. an attractive game which produces valuable data.

3 Emojitalianobot and the Pinocchio project

Emojitalianobot is the first open and free Emoji-Italian translation bot on Telegram. It was developed to support the translation project of *Pinocchio* in emoji [15] launched on Twitter in February 2016 by F. Chiusaroli, J. Monti and F. Sangati. The translation of the famous children's novel was carried out by the followers of the *Scritture brevi* blog[16] (by F. Chiusaroli and F.M. Zanzotto) and the first fifteen chapters have been translated, which correspond to the original novel published by Collodi in 1881. Every day tweets with sentences taken from the novel were posted on Twitter and the followers suggested their translations

[4] https://en.wiktionary.org/
[5] http://www.omegawiki.org/Meta:Main_Page
[6] http://it.termwiki.com/
[7] https://term.tilde.com/
[8] https://www.mturk.com/mturk/welcome
[9] https://telegram.me/dialectbot/

[10] http://www.aikuma.org/aikuma-app.html
[11] https://itunes.apple.com/au/app/ma-iwaidja/id557824618?mt=8
[12] When a player without any special knowledge is put into a gaming environment and has to make decisions to win the game under the pressure of time or any game mechanics' constraints.
[13] https://anawiki.essex.ac.uk/phrasedetectives/
[14] http://www.jeuxdemots.org/jdm-accueil.php
[15] http://www.treccani.it/lingua_italiana/speciali/ludolinguistica/Chiusaroli.html
[16] https://www.scritturebrevi.it/

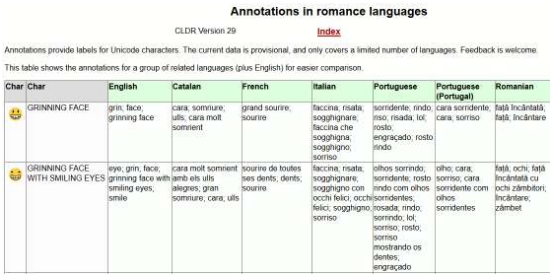
in emoji; at the end of each day, the official version of the translations was validated and published.[17] Translators used *Emojitalianobot* that contains (i) the Emoji-Italian dictionary, (ii) the Emoji-English descriptions based on Unicode and (iii) a glossary with all the uses of emoji in the translation of Pinocchio. The project was associated with the Emojitalia discussion group on Telegram, where users met to discuss problems, solutions, suggest improvements of the bot, in addition to the translation choices for *Pinocchio* and *communicate* in emoji. The *Pinocchio* translation project therefore allowed to crowdsource different linguistic data connected with the use of emojis as actual means of communication and not just simple graphics to express amusement or interest. In this respect the main findings of the project are twofold: the need to recur to compound multi-emoji expressions in order to express concepts which are not represented in the current set, as well as a related simple grammar to express syntactic relations among emojis, past and future tenses, etc. Unlike previous literary translation project in emojis, such as the translations of *Moby Dick* or *Alice in Wonderland*, this is the first attempt of a collective shared emoji code (vocabulary and grammar) based on a word for word translation totally in emojis. *Emojitalianobot* is an ideal test bench to experiment with new approaches like crowdsourcing and gamification in the field of Natural Language Processing (NLP). The *Pinocchio* project, games and features available in the bot to learn or guess the meaning of emoji are devised indeed both to enjoy the bot while using it and at the same time to give the opportunity to users to develop linguistic descriptions of emoji tailored on their actual perceptions. The most important reward for playing with the bot is the awareness of helping develop a linguistic resource for one's mother tongue, and the pride in contributing to it.

Since its release on Telegram, the project was an instant success, becoming a viral web phenomenon thanks to the *Scritture brevi* community and the *Pinocchio* translation in emojis, so that the bot has now almost 750 users. The Pinocchio translation project in emojis counts 611 tweets , 980 glossary entries which correspond to 2127 words, of which 185 are multi-emojis, i.e. compound emojis, such as 🔝👎 for the Italian word *peggio* (worst).

[17]The translation of *Pinocchio* in emoji can be followed on Twitter using #emojitaliano.

4 EmojiWorldBot

On the basis of (both linguistic and technological) experience with *Emojitalianobot*, the three Italian researchers together with Martin Benjamin and Sina Mansour of the Kamusi Project International[18] and EPFL (Switzerland) designed a new bot on Telegram in April 2016: *EmojiWorldBot*, a multilingual dictionary that uses Emoji as a pivot language from dozens of different languages. Currently the emoji-word and word-emoji functions are available for 70 languages imported from the Unicode tables [19] and provide users with an easy search capability to map words in each of these languages to emojis, and vice versa. Looking at the UNICODE descriptions (see Fig. 1) it is apparent that emojis are not annotated in a coherent way across languages, so some languages have more descriptions and some others, especially underrepresented languages, have less or in the most cases some languages are not represented at all.


Annotations in romance languages

CLDR Version 29 Index

Annotations provide labels for Unicode characters. The current data is provisional, and only covers a limited number of languages. Feedback is welcome.

This table shows the annotations for a group of related languages (plus English) for easier comparison.

Char	Char	English	Catalan	French	Italian	Portuguese	Portuguese (Portugal)	Romanian
😀	GRINNING FACE	grin; face; grinning face	cara; somriure; ulls; cara molt somrient	grand sourire; sourire	faccina; risata; sogghignare; faccina che sogghigna; sogghigno; sorriso	sorridente; rindo; riso; risada; lol; rosto; engraçado; rosto rindo	cara sorridente; cara; sorriso	față încântată; față; încântare
😄	GRINNING FACE WITH SMILING EYES	eye; grin; face; grinning face with smiling eyes; smile	cara molt somrient amb els ulls alegres; gran somriure; cara; ulls	sourire de toutes ses dents; dents; sourire	faccina; risata; sogghignare; sogghigno con occhi felici; occhi felici; sogghigno; sorriso	olhos sorrindo; sorridente; rosto rindo com olhos sorridentes; rosada; rindo; sorrindo; lol; sorriso; rosto; sorriso mostrando os dentes; engraçado	olho; cara; sorriso; cara sorridente com olhos sorridentes	față, ochi; față încântată cu ochi zâmbitori; încântare; zâmbet

Figure 1: Annotations in Romance languages

Our first goal with *EmojiWorldBot* is therefore to reach a uniform and comprehensive list of tags across multiple languages with a precise mapping between any language pair, which may serve to bootstrap a massive multilingual dictionary. The bot currently features:

- emoji-to-word and word-to-emoji translation for more than 70 languages
- *Eggs*, a tagging game for people to contribute to the expansion of these dictionaries or the creation of new ones for any additional language. Users can suggest additional tags for single emojis in any language (for example adding *egg* to the tag list for 🐣 in English).

[18]https://kamusi.org/

[19]http://www.unicode.org/cldr/charts/29/annotations/

- inline queries: type EmojiWorldBot and a word, and it will suggest a set of emojis for that word you can send in any Telegram conversation
- the possibility to add new languages.To date 56 new languages were added, such as Latin, Esperanto, Sardinian among others.

The basic idea of the *Eggs* game is to collect new tags to associate with emojis as shown in Fig. 2.



Figure 2: *Eggs* game

With fewer than 2000 official emojis, stretching the boundaries of their communicative potential makes them more useful. However, it also makes the dictionary more essential, so that someone who receives 🍆 in a chat in any language might look to see if it signifies something other than an *eggplant*. In the future, *Eggs* will experiment with multi-emoji terms (METs), building on the work of the *Pinocchio* translation project to Emoji, in an effort to build a larger pictorial vocabulary that is comprehensible across languages (Chiusaroli, 2015). A new version of the bot is already under development. It will feature *Ducks*, a second game where users are asked to map tags from a source language (e.g. English) to a target language (e.g. Swahili). In the example of Figure 1, several Romanian users would be shown the sense-specific definition of *grin* from Wordnet and all of the emojis that have been attached to that definition, and be asked which of the options among *față încântată față* and *încântare*, if any, is a good translation. The game would also be played for *face* and *grinning face*. In this way, all of the one-to-one relationships should be discovered, and all instances of a term that does not have a translation equivalent on the other side will be revealed. When it is known that no match from English exists, *Ducks* presents the definition, the emojis, and the English term, and asks the user to type in the best equivalent in their language. This is the method that will be most efficacious for new languages, bypassing the need to disentangle the many-to-many associations introduced through term clustering in the CLDR annotations. It should be noted that many terms will be removed from the game cycle through comparisons with Wordnets for available languages. For example, самолет appears in conjunction with English *airplane* in both the Emoji annotations and the Bulgarian Wordnet that are linked to the same English Princeton Wordnet (PWN) sense, which gives sufficient confirmation without needing a mass of human players. As of this writing, the project is in the process of importing and aligning Wordnet data for some 50 languages. In future work, terms from Wordnet synsets will be tested against the emojis with which they theoretically share a sense, e.g. asking crowd members whether 🚍 applies to other members of the PWN synset for *bus* (autobus, coach, jitney, motorbus, etc.), but the mechanism for doing so has not been finalized. In this way *EmojiWorldBot* employs crowd methods as part of an arsenal intended to conquer the walls of collecting data for numerous diverse languages. Data validation will be achieved via a consensus model through which answers are accepted as correct if the same result is provided by a threshold number of respondents. The new version of the bot will allow to:

- add new terms to the current languages (including the names of the countries for national flags)
- compare definitions across languages.

From the computational point of view, this project, as the *Emojitalianobot*, attempts to address the data chasm for natural language processing for most languages by distilling data collection to simple micro-tasks (Benjamin, 2015) using techniques adapted to least-common-denominator technology.

5 Conclusions

We described the *Emojitalianobot* and the *EmojiWorldBot* projects. Combining crowd-

sourcing, gamification and a smartphone app is a powerful strategy to collect, improve and refine valuable linguistic data easily and in a short time particularly for less-resourced languages (Benjamin and Radetzky, 2014).These may be the first crowdsourcing projects of this type to use bots for linguistic data collection and validation and are unique in their attempts at engaging participants for different languages.

References

Martin Benjamin. 2015. Crowdsourcing microdata for cost-effective and reliable lexicography. In *Proceedings of AsiaLex 2015 Hong Kong*, EPFL-CONF-215062, pages 213–221.

Martin Benjamin and Paula Radetzky. 2014. Multilingual lexicography with a focus on less-resourced languages: Data mining, expert input, crowdsourcing, and gamification. In *9th edition of the Language Resources and Evaluation Conference*, EPFL-CONF-200375.

Daren C Brabham. 2012. A model for leveraging online communities. *The participatory cultures handbook*, 120.

Francesca Chiusaroli. 2015. La scrittura in emoji tra dizionario e traduzione. *CLiC it*, page 88.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 508–513. Association for Computational Linguistics.

Jane McGonigal. 2011. *Reality is broken: Why games make us better and how they can change the world*. Penguin.

Johanna Monti. 2012. Translators' knowledge in the cloud: The new translation technologies. In *International Symposium on Language and Communication: Research Trends and Challenges(ISLC)*.

Johanna Monti. 2014. Dictionaries in the cloud: state of the art, trends and challenges. *Les Cahiers du dictionnaire*, (6):95–110.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.