

GPT-based Language Models meet Emojitaliano: A Preliminary Assessment Test between Automation and Creativity

Francesca Chiusaroli^{1,†}, Tiberio Uricchio^{1,*}, Johanna Monti², Maria Laura Pierucci¹ and Federico Sangati³

¹Università degli Studi di Macerata

²Università di Napoli "L'Orientale"

³OIST Graduate University, Japan

Abstract

English. Starting from the crowdsourcing experience of Pinocchio in Emojitaliano [1], the present paper intends to test Chat-GPT's ability to take on the Emojitaliano grammar and dedicated glossary to verify and reapply the Emojitaliano rules in order to produce translations on its own. A test of re-translation of Pinocchio is presented here.

Italiano. A partire dall'esperienza in crowdsourcing di Pinocchio in Emojitaliano [1], il presente contributo intende testare la capacità di Chat-GPT di assumere la relativa grammatica e il glossario dedicato per verificare e riapplicare le regole della emojilingua allo scopo di svolgere traduzioni in proprio. Si presenta qui un test di ritraduzione di Pinocchio.

Keywords

Emojitaliano, LLM, Assessment, Evaluation

1. Introduction

Consisting today in over three thousand pictograms and symbols, and regularly updated by Unicode Consortium, the emoji international catalog contains signs for facial expressions (smileys) and for human gestures, portraits of people, plants and the animals, reproductions of food and objects for everyday activities and sports, symbols of travel and places. Whereas the visual content seems to provide an encyclopaedic catalog with a universal status, ideally able to signify language-independent meanings, the interpretation of emojis is, on the contrary, heavily arbitrary, subject to ambiguities and differences due to linguistic and cultural specificities [2].

Some efforts were made to develop an emoji based language that could be shared among different cultural peoples. The first notable project that made an effort of translating a classical novel ("Moby Dick" of Herman

Melville), was the Emoji Dick project¹ by Fred Benenson (2010). Starting from the English version of the novel, each sentence was translated into an emoji version via crowdsourcing. Each of Moby Dick's 6,438 sentences has been translated 3 times by different Amazon Mechanical Turk (MTurk) workers. The resulting emoji sentences were then chosen by voting by another set of workers, and the most popular version of each sentence was selected for inclusion in the book. The outcome is a wonderful but inconsistent translation of the same terms according to the wisdom of the crowd in good sense, but without any shared rules, structure or grammar, leading to the impossibility of recovering the original text or meaning. Another project was the translation of Lewis Carroll's "Alice's Adventures in Wonderland" by Joe Hale² (2014). In this case, each word was directly translated into a corresponding emoji. Consistency was thus guaranteed as the same word was translated with the same emoji, introducing a de-facto lexicon. Nonetheless, no grammar structure is developed as the translation follows verbatim the original text and its English-based word order.

In order to counteract the natural polysemy of emojis [3], Emojitaliano³ was created through a social community on Twitter (#scrittorebrevi #emojitaliano), devoted to the experimental crowdsourcing construction of an international emoji code 'emojilingua' [4, 5]. The aim of the project includes linguistic simplification and the possibility of reproducing a text in emoji that will be com-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 – Dec 02, 2023, Venice, Italy

*Corresponding author.

[†]These authors contributed equally.

✉ f.chiusaroli@unimc.it (F. Chiusaroli); tiberio.uricchio@unimc.it

(T. Uricchio); jmonti@unimc.it (J. Monti);

marialaura.pierucci@unimc.it (M. L. Pierucci);

federico.sangati@oist.jp (F. Sangati)

🌐 <https://docenti.unimc.it/f.chiusaroli> (F. Chiusaroli);

<https://docenti.unimc.it/tiberio.uricchio> (T. Uricchio)

📞 0000-0003-1923-3974 (F. Chiusaroli); 0000-0003-1025-4541

(T. Uricchio); 0000-0002-4563-5988 (J. Monti); 0000-0003-3637-2757

(M. L. Pierucci); 0000-0001-6088-415X (F. Sangati)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License

Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.emojidick.com>

²<https://www.joehale.info/visual-poetry/wonderland.html>

³<https://www.treccani.it/vocabolario/emojitaliano>

prehensible and readable in all languages of the world. For this reason, Emojitaliano consists in a unique project that provides a grammatical structure and a shared vocabulary.

Emojitaliano is thus based on the assessment of conventional meanings, capable of guaranteeing the sharing of sense by means of intersemiotic translation [6]. The iconic property of emoji, together with the coded glossary and grammar, makes Emojitaliano a unique tool for communicative accessibility and for multilingual and language (L1 and L2) teaching. Born with the translation of Collodi's *Pinocchio, The Story of a Puppet*⁴ [1], the structure and glossary of Emojitaliano have been later usefully reapplied for the translation of texts of different genres such as the technical declaratory prose of the Italian *Constitution*, the narrative prose of moral tales (i.e. *The Wolf and the Lamb*), Dante's allegorical poetry of *The Comedy*, Giacomo Leopardi's lyrical poem *The infinite*⁵.

The process of such translation relied heavily on manual labor and human expertise, often time-consuming and subject to human limitations.

With the rapid advancements in artificial intelligence, a new era has dawned upon the world of translation. Large Language Models (LLMs), such as BLOOM [7], GPT-3 [8], LLaMa-1 [9], have revolutionized the way we approach language translation. Recent chat interfaces enable an easy human interaction with them, that has led to a rapid and widespread adoption among the public, also due to the recent high performance closed models Chat-GPT and GPT-4, which remarkable are reported to be able to pass several human admission exams⁶. This preliminary assessment, thus, is about evaluating and establishing the utility of such models, even in the challenging application that is Emojitaliano, where figurative and idiomatic writing is needed, beyond the basic rules of the language.

2. Preliminaries on Emojitaliano: lexicon and grammar

Emojitaliano consists of an iconic-based shared conventional code, first of all a simplified grammar of an isolating and analytical type, constructed on the model of 'interlanguage'; it is anyway not a truly formalized algebraic language, since it is adopted as a conventional code in a social media environment, as a 'living human language'⁷. As for the vocabulary, Emojitaliano is a semantic-based code. In fact, word-emoji pairings are fixed by exploiting the semiotic value of the icons, such as 'scarpa'

CAPITOLO 1

Come andò che Maestro Ciliegia, falegname trovò un pezzo di legno che piangeva e rideva come un bambino.

— C'era una volta...

— Un re! — diranno subito i miei piccoli lettori.

— No, ragazzi, avete sbagliato. C'era una volta un pezzo di legno.

Non era un legno di lusso, ma un semplice pezzo da catasta, di quelli che d'inverno si mettono nelle stufe e nei caminetti per accendere il fuoco e per riscaldare le stanze.

Non so come andasse, ma il fatto gli è che un bel giorno questo pezzo di legno capitò nella bottega di un vecchio falegname, il quale aveva nome maestr'Antonio, se non che tutti lo chiamavano maestro Ciliegia, per via della punta del suo naso, che era sempre lustra e paonazza, come una ciliegia matura.

Appena maestro Ciliegia ebbe visto quel pezzo di legno, si rallegrò tutto; e dandosi una fregatina di mani per la contentezza, borbottò a mezza voce:

— Questo legno è capitato a tempo; voglio servirmene per fare una gamba di tavolino.

Detto fatto, prese subito l'ascia arrotata per cominciare a levargli la scorza e a digrossarlo; ma quando fu lì per lasciare andare la prima ascia, rimase col braccio sospeso in aria, perchè sentì una vocina sottile sottile, che disse raccomandandosi:

Figure 1: A few sentences from the first chapter of *Pinocchio*, given to the model to derive rules. Best seen zoomed in the PDF version.

(shoe). When lacking direct matches, compound emojis are created ('bottega', *atelier* = [casa + attrezza, *house* + *tools*]), also adopting the techniques of symbolic and figurative transposition, such as metaphors and similes ('volare', *to fly* = ["aereo", *airplane*]). Linguistic simplification is achieved through synonymy and semantic generalization, based on the fact that the very same emoji may have more than one meaning (i.e. the 'monkey' translates 'birba', 'monello', 'capriccio'; the compound emoji 'man+heart' stays for 'Geppetto' and for 'babbo'). Each new translation experiment subsequent to Pinocchio has been conducted by the community and also by new groups, university and high school students in particular, and sometimes by single translators: everyone was required to use the fixed grammar and, if already present, the fixed vocabulary, and to reapply the simple rules for the creation of new vocabulary, starting from the semiotic value of the emoji. In this way, Emojitaliano has been able to benefit from an ever-growing community, capable of using creativity within a codified scheme of rules. Hosted in a specific bot on Telegram (@emojitalianobot) [10]⁸, the Emojitaliano language consists today in 3.522 recorded matches⁹.

⁴ https://it.wikisource.org/wiki/Le_avventure_di_Pinocchio

⁵ www.scritturaebrevi.org

⁶ <https://openai.com/research/gpt-4>

⁷ https://www.treccani.it/magazine/lingua_italiana/speciali/ludolinguistica/Chiusaroli.html

3. GPT-4 meets Emojitaliano

Given the project's goal of establishing an international emoji code, we assumed that LLMs can be a useful tool to speed up translation, as well as to spread the language beyond the currently active community. For this purpose, we chose GPT-4 as our reference model, reported as currently the best general performing model in language tasks [11] also for translation [12], even compared by several users in benchmarks with other open source models¹⁰. Since we do not have access to the training data of such model, we preliminarily verified that it does not know about Emojitaliano making multiple empirical questions (e.g. "What is Emojitaliano?")¹¹. Then, we proceeded to gradually instruct the model using the prompt method [13], by providing pairs of Italian and Emojitaliano translated text or a description of the grammar and lexicon of Emojitaliano.

3.1. Automatic grammar derivation

Our first experiment aimed at observing the capability of the model in acting as a human translator of Emojitaliano, its capability to apply grammar rules, and its ability to use the already known tags by Unicode Consortium to creatively associate Italian words, to memorize and enlarge the lexicon. We developed a prompt where we provided a passage from the first chapter of *Pinocchio in Emojitaliano*, initially without providing grammar and the lexicon, asking it to derive them, instead. Then we asked the model:

1. to derive the syntactic and semantic rules;
2. to translate on its own some sentences by applying the derived rules;
3. to correct and refine the rules by comparing the personal version and the official one.

⁸Developed and updated by Federico Sangati: <https://fede.sangati.me/project/emojitaliano/>

⁹On July 30, 2023

¹⁰<https://chat.lmsys.org/?arena>

¹¹The first answer to "What is Emojitaliano?" was generic and incorrect, that is: "Emojitaliano is a term used to describe the use of Italian words and phrases in combination with emojis to create a unique form of communication. It is a type of digital language that blends written words with visual symbols, allowing people to express emotions, thoughts, and ideas in a more dynamic and creative way. Emojitaliano is particularly popular among young people in Italy and has become a common feature of social media and messaging platforms. It has also led to the development of new emojis that reflect Italian culture and traditions, such as the pizza, the espresso, and the gondola. The use of Emojitaliano is not limited to Italy, however, and it has spread to other countries where Italian is spoken or appreciated. It is an interesting example of how digital communication is transforming language and culture, and how people are adapting to new forms of expression in the digital age."

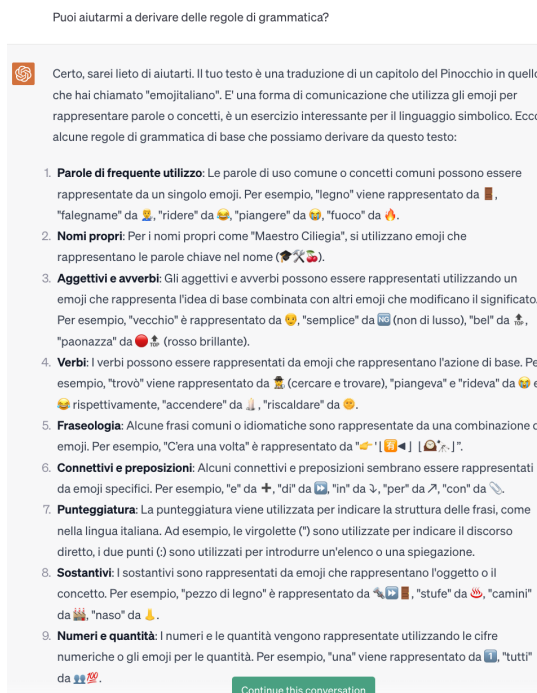


Figure 2: Grammar rules automatically derived by the model. Best seen zoomed in the PDF version.

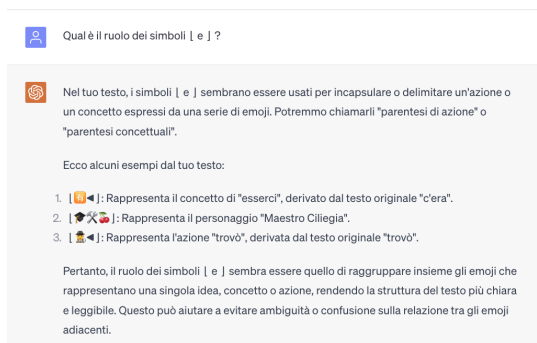


Figure 3: The role of the delimiters, automatically derived by GPT-4. Best seen zoomed in the PDF version.

In Fig. 1, we report the sentences from the first chapter that were given to the model, with the subsequent request to derive the language rules. The output is reported in Fig. 2. The interpretation of conceptual bracket signs in Emojitaliano was then requested, as considered a relevant information missing in the first answer. The result is particularly interesting since a metalinguistic label is assigned by the model, as can be seen in Fig. 3



Figure 7: Further sentences and subsequent derivations of new rules. Best seen zoomed in the PDF version.

player' emoji for 'tirare un colpo'. Although the model is familiar with the rule for the semantic plain translation of figurative language, it does not succeed in applying it, as in 'occhi fuori dal capo', which is translated literally 'eyes going out of the head' instead of rendering the meaning of 'to be shocked' or 'to be upset'.

A relevant issue emerged when Chat-GPT was asked to translate some specific words: for example, 'legno' ('pezzo di legno', 'piece of wood') is translated with the 'fir' ('abete') emoji and, somewhere later, with the 'wooden door' emoji. This is against one of the main Emojitaliano lexical rules which aims at reducing semantic ambiguity. In fact, each word within the same text, should always be translated the same way. Chat-GPT is to be trained accordingly.

We also noticed that grammar and rules mistakes can be corrected by the model upon casually reminding rules in long interactions. The model leaned to progressively forget the rules and, thus, a restart of the session was required after a few sentences. We believe that this is due to the limited window of attention of LLMs and the encoding of emoji that require several tokens for each of them.

4. Performance evaluation

According to our preliminary exploration, we established that GPT-4 is able to derive the semantic rules and translate text to Emojitaliano. To evaluate the latter, we per-

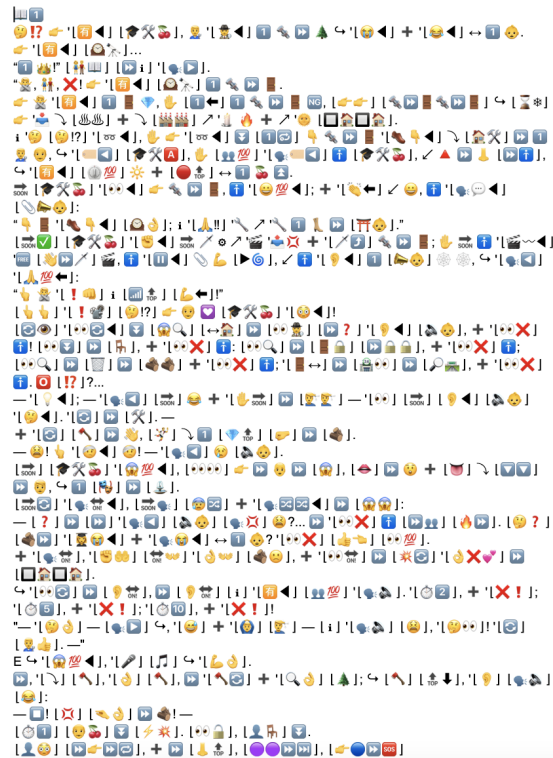


Figure 8: The 1st chapter of Pinocchio translated by the model. Best seen zoomed in the PDF version.

formed a more extensive evaluation by building a dataset of text pairs Italian-Emojitaliano and asking the model for the translation.

We constructed the dataset by considering the first 3 chapters of Pinocchio [14], previously translated in Emojitaliano [1]. The chapters are split respectively in 27, 50, and 45 sentences for a total of 122, ranging from 5 to ~80 Italian words and from 2 to ~70 emojis each. For each sentence, we constructed a pair made of the original Italian text and the relative human translation. Each sentence is given to the model for translation independently from the others.

To perform the evaluation, we constructed a textual prompt where the grammar and the basic rules are carefully explained in Italian, where we include as the training set, the first chapter as given examples of translation. The remaining two chapters are used as the test set. Measuring quantitatively the quality of the translation is more challenging than the typical translation tasks metrics, given the creative use of emojis and their combinations in expressing a meaning. Given the low number of samples, we resorted to human evaluation and the use of GPT-4 as a judge following [15]. For both human and GPT-4 evaluators, we provided the rules of grammar,

| Translation by | GPT-4 Evaluation | | | Human Evaluation | | |
|----------------|------------------|-------|-------|------------------|-------|-------|
| | Human | GPT-4 | Equal | Human | GPT-4 | Equal |
| Preferences | 27 | 51 | 17 | 46 | 36 | 13 |
| Average Score | 7.23 | 7.80 | / | 7.34 | 7.21 | / |

Table 1
Results of the performance evaluation.

the original text and blindly the translated Emojitaliano from the ground truth and the output of the model. The evaluators were asked to vote for the best translation (i.e. choose the preferred translation) according to relevance, accuracy, creativity, correct use of grammar. In addition to choosing the preferred translation, we also asked the evaluators to provide a quality score from 1 to 10 for each sentence.

Results are reported in the Table 1.

GPT-4 and Human evaluators disagree on their preference of translations. The Human evaluators, generally, tend to prefer the Human translations while GPT-4 the opposite. From the evaluators and GPT-4 feedbacks, we noted that the Human evaluators put more emphasis on the correct structure of the sentences (e.g. the subject verb object rule), while GPT-4 generally reported better scores for creativity and direct matching of the emojis (e.g. emojis that match the words). This is consistent, since the translation in the ground truth was realized in 2017, when most of the modern emojis were still not defined at the time. Due to the absence of a proper matching, many emojis were chosen even if they were distant from the corresponding words. Moreover, GPT-4 has consistently not fully caught the rules of Emojitaliano, leading to less awareness of errors in the sentences structure.

5. Conclusions and work in progress

Emojitaliano was born thanks to the free dedication and commitment of an enthusiast devoted Twitter social community, then also of student groups, willing to share the goal of building an emoji-based artificial language model, to be used as a communicative code across language barriers [5, 1]. The effort to adapt to the rules and to join the common glossary, as well as to expand it according to the common rules, was challenging as well as a hard task, but it was the only way to ensure an essential linguistic basis, by giving rise to a language, validated and practiced by a community of ‘speakers’. The intensive crowdsourcing experience made Emojitaliano a unique case among the (actually not many) examples of integral translations in emoji, which are mostly represented by intentionally non-systematic or solipsistic works. The regular expansion of the international emoji set by the Unicode Consortium

constantly extends the range of choice by enriching the emoji-language with ‘emoji-neologisms’, as happens in every living natural language, but the core of the Emojitaliano glossary and grammar provides a settled authoritative translation method. Translating Pinocchio into Emojitaliano today would certainly involve new, and sometimes more relevant, pairings, synonymic pairs that do not exclude the previous ones; but the method remains fixed, because the syntax alone guarantees, through the instrument of translation, mutual understanding. Teaching Emojitaliano to GPT-4 (and the like) does not mean replacing a human translator with a machine, but rather is like having a tool to enhance human work to the maximum: automation ensures the speed, the iconic base of the emoji embeds and guides creativity, therefore setting limits against the arbitrary drift of individual subjective interpretation. Following our design, the year of work spent in the ‘human’ translation of the original 15 chapters of Pinocchio will be matched by a few minutes’ work in the translation of the entire work (35 chapters) by Chat-GPT, and in the translation of other works from any world’s language. Extreme speed is comfortable and convenient, but the results cannot be achieved without training: that is, by learning a “language” and its rules.

Acknowledgments

This work has been funded by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY - CUP D83C22000710005.

References

- [1] F. Chiusaroli, J. Monti, F. Sangati, Pinocchio in Emojitaliano, Apice libri, Sesto Fiorentino, 2017.
- [2] V. Evans, The emoji code: How smiley faces, love hearts and thumbs up are changing the way we communicate, Michael O’Mara Books, 2017.
- [3] M. Danesi, The semiotics of emoji: The rise of visual language in the age of the internet, Bloomsbury Publishing, 2017.
- [4] F. Chiusaroli, Da emojipedia a pinocchio in emojitaliano: l’“emojilingua” tra scritte e riscritte, in: Homo Scribens 2.0. Scritture ibride della modernità, Franco Cesati, 2019, pp. 45–87.
- [5] F. Chiusaroli, Emoji e semplificazione linguistica, in: Comunicare il patrimonio culturale. Accessibilità comunicativa, tecnologie e sostenibilità, FrancoAngeli, 2021, pp. 164–193.
- [6] F. Chiusaroli, La scrittura in emoji tra dizionario e traduzione, in: Proceedings of the Second Italian Conference on Computational Lin-

- guistics CLiC-it 2015, Accademia University Press, Torino, 2015. URL: <http://books.openedition.org/aaccademia/1437>. doi:<https://doi.org/10.4000/books.aaccademia.1437>.
- [7] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
- [10] J. Monti, F. Sangati, F. Chiusaroli, B. Martin, M. Sina, et al., Emojitalianobot and emojiworldbot-new online tools and digital environments for translation into emoji, in: *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, 2016.
- [11] J. A. Baktash, M. Dawodi, Gpt-4: A review on advancements and opportunities in natural language processing, arXiv preprint arXiv:2305.03195 (2023).
- [12] W. Jiao, W. Wang, J. Huang, X. Wang, Z. Tu, Is chatgpt a good translator? yes with gpt-4 as the engine, arXiv preprint arXiv:2301.08745 (2023).
- [13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM Computing Surveys* 55 (2023) 1–35.
- [14] C. Collodi, *Le avventure di Pinocchio. Storia di un burattino*, illustrata da Carlo Chiostri., [etc.] Bemporad & figlio., 1907.
- [15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.