Piötòst Ché Niènt, Mèi Piötòst - A Manually Revised Lombard-Italian Parallel Corpus

Edoardo Signoroni

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
e.signoroni@mail.muni.cz

Abstract. The Lombard language is a Gallo-Italic language spoken in the Northern Italian region of Lombardy and some surrounding areas by 3.5 million native speakers in varied spectrum of bilingual settings and fluency. However, it is currently listed as "definitely endangered" according to UNESCO. Despite some resurging interest in documenting, revitalizing, and using the language, no Natural Language Processing resource was specifically build for Lombard. The only existing Lombard-Italian parallel corpus was created as part of a bigger multilingual project by scraping aligned text from Wikipedia articles. However, we found the resulting corpus to be faulty, due to noise and erroneous alignments. Our work addresses these issues by providing a cleaner, human-revised version of this resource, which could be used as a stepping stone to build future NLP tools, such as a Machine Translation system.

Introduction

Lombard is a regional language[1] spoken in and around the Northern Italian region of Lombardy by about 3.1 million people,[2] where it exists alongside the official language, Italian, in varying degrees of bilinguality and fluency. It belongs to the Gallo-Romance-Cisalpine group of the Western Romance family of the Indo-European languages, and it is said to have between two and four varieties, the main ones being Western (in the provinces of Varese, Como, Lecco, Sondrio, Milan, Monza, Pavia and Lodi, in addition to Novara and Verbania in Piedmont and Canton Ticino in Switzerland) and Eastern Lombard (in the provinces of Bergamo, Brescia and Northern Cremona). These varieties, even with some phonetic, lexical, and grammatical differences, can

[1] This definition is preferred over the one commonly used today, even by some academics, of *dialetto* (en. "dialect", following Coseriu's (1981) [8] definition of so-called "primary dialects"), which is arguably both erroneous and derogatory. [5] As Chambers and Trudgill [4] state: "a dialect is a substandard, low status, often rustic form of language, generally associated with the peasantry, the working class, or other groups lacking in prestige".

[2] To these figures, which report numbers just from Lombardy, one must add speakers in neighboring regions and from Switzerland. Data according to Istituto Nazionale di Statistica (ISTAT) from 2015 https://www.istat.it/it/archivio/207961

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 105–112, 2022. © Tribun EU 2022

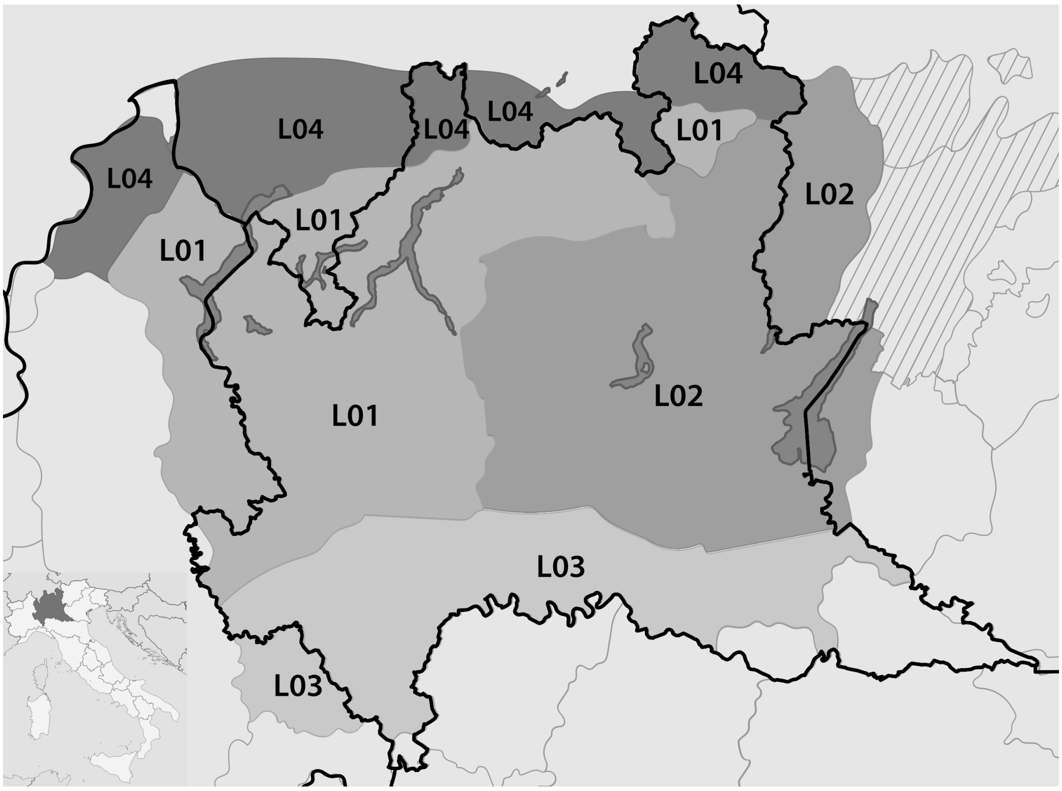

Fig. 1: Map showing the geographical distribution of the Lombard language and its varieties according to a fourfold subdivision. L01 denotes Western Lombard, L02 Eastern Lombard, L03 Southern Lombard, and L04 Alpine Lombard. Image retrieved from https://commons.wikimedia.org/wiki/File:Mappa_Dialetti_lombardi.svg

be loosely considered to be one language, since they are mutually intelligible. [7,2,11,6] At the present day, the language is mostly used in oral conversation and no unified orthography exists, with different approaches ranging from phonetic/phonemic, to historical or etymological ones. One of these is the proposal by Brasca (2011). [3] Figure 1 shows the location of Lombard and its variants in Northern Italy.

Despite the relatively large amount of speakers, and featuring literature and cultural activities in different forms, the current status of Lombard is of concern due to a plethora of reasons, being them historical, social, political, or legislative. Discussing these issues, most of which are complex and controversial (at least for an Italian audience), lies outside the scope of this paper.[3] UNESCO [14] lists Lombard as a ”Definitely endangered” language [4] According to other

[3] To more in-dept discussion on this topic, see the works referenced in the bibliography.

[4] A language that ”is no longer being learned as the mother tongue by children in the home. The youngest speakers are thus of the parental generation. At this stage, parents may still speak their language to their children, but their children do not typically respond in the language.”

metrics, such as EGIDS (Extended Graded Intergenerational Disruption Scale) [10], Lombard is between grades 6b ″threatened″ and ″moribund″. However, some interest in Lombard, and other regional languages of Italy, is resurging with some cultural and multimedia production, academic research, and even social network and Wikipedia[5] pages. Moreover, in 2016 a regional law[6] was passed for the protection and promotion of Lombard.

If a thorough effort towards this goal has to be made in the present day, Natural Language Processing (NLP) resources must to be developed. Among such technologies, a Machine Translation (MT) system and its foundational basis, a parallel corpus, can surely be beneficial to the preservation of the language. The only existing Lombard-Italian parallel corpus was created as part of a bigger multilingual project by scraping aligned text from Wikipedia articles. However, we found this corpus to be faulty, due to the widespread presence of noise and erroneous alignments. This work addresses this issue by providing a cleaner, human annotated version of this resource on top of which build NLP tools, such as a Machine Translation system.

This paper is structured as follows: Section 1 briefly surveys previous work on Lombard; Section 2 relates the methodology of this work and describes the resulting corpus; Section 3 discusses its limitations and outlines some future work to address them; Section 4 presents our conclusions.

1 Related Work

Regarding NLP work on Lombard, not much has been done. Glottolog[7] lists published research on Lombard variants from the 19th century to 2021. Most of the current research has focused on sociolinguistics and revitalization. Some systematic documentation of the language, or its variants, has been carried out in the form of lexical atlases, such as the one by the Fondazione Civiltà Bresciana.[8]

While books in Lombard (most likely one of its variants) can be found in physical circulation, the digitalization of textual sources is lacking, with not even a full text of the Bible[9] freely obtainable online, the only text available being dictionaries and parts of the Gospel.

As far as concrete NLP resources are concerned, Lombard monolingual corpora are available only as part of larger projects with Wikipedia dumps [15], such as W2C [12], and Deltacorpus [13]. To our knowledge, no monolingual corpus has been built specifically for Lombard.

[5] https://lmo.wikipedia.org/wiki/Pagina_principala

[6] Regional Law no. 130/2016

[7] https://glottolog.org/resource/languoid/id/lomb1257

[8] https://www.civiltabresciana.it/pubblicazioni/atlantelessicale.html

[9] The Bible is usually the go-to source for unresourced languages, since it is the most widely translated book in the world and comes with the advantage of having a built-in ″gold″ alignment in the form of verses.

```
Posso capire perché pensò a me.     A gh'è nissün che 'l pensa a mì.
```

Fig. 2: An example of a wrong alignment. The translation of the sentences are as follows: it. *"I can think why he/she thought about me."* lmo. *"There is no one who thinks about me."*

With regards to parallel corpora, the only readily available one is the parallel corpus in OPUS. [17][10] It consists of the Lombard-Italian section of the WikiMatrix corpus [16] automatically created by mining parallel sentences from Wikipedia articles trough multilingual sentence embedding similarity. [1] This resource was revealed to be very noisy and plagued by errors after our preliminary evaluation of a sample of the proposed sentence pairs.

2 Methodology

2.1 Preliminary evaluation

Our work started with evaluating a sample of the corpus available on OPUS. We manually analyzed 500 sentence pairs and determined that 157 were incorrect. This amounts to 31.4% of the sample being judged either as errors or noise. The most common instances of these were duplication of the sentence on both sides, a fully or partially incorrect alignment, or similar sentences or context that were nonetheless incorrect translations. In some cases, these where loose paraphrases or summarizations of the Italian text. Where these could be easily fixed, that is if the extent of the error was roughly under half of the overall length of sentence, we modified the Italian sentence to match the Lombard one. We did not modify the Lombard side of the alignments to avoid the injection of further noise in the data, e.g. through subjective spellings or orthographical choices.

It is relevant to note that some of the removed examples contained well formed sentences on the Lombard side. Recovering and complementing these phrases is left to future work, but it signals that a bigger amount of data may be available to be exploited. Figure 2 gives an example of an incorrect alignment to be removed.

2.2 Manual annotation

We then moved on to manually revise the whole parallel corpus, which amounts to 10.533 sentence pairs. These were divided among five different annotators, all native bilingual speakers of Italian and Lombard, more precisely the Brescian variety of Eastern Lombard.[11] While it can be argued that this annotator group may bias the results, we maintain that this risk, while present, is very low for the task we carried out. Our reasoning is the following.

[10] https://opus.nlpl.eu/

[11] The author of this paper is also among the annotators.

Total	Correct	Removed	Modified
10.533	4915 *46.67%*	5227 *49.62%*	391 *3.71%*

Table 1: Number of correct, removed, and modified alignments against the starting total.

First, the annotators did not provide or choose any kind of data for the corpus; their task was to judge the correctness of the alignments, which were independently generated by an automated method. Moreover, as stated in Section 2.1, even in the instances in which the alignments were manually corrected, instead of being removed all-together, only the Italian side was modified in order to avoid the insertion of subjective forms and orthography in the text.

Second, similar work [9] found that relatively simple annotation tasks such as evaluating the correctness of a sentence alignment can be carried out effectively even by annotators with little or no proficiency of the languages under scrutiny. The annotators were all native bilingual speakers of Italian and a Lombard variety. Recall from the Introduction that the varieties of Lombard are to a great extent mutually intelligible, thus being proficient in one of them should suffice for this annotation task.

In our manual revision, we removed 5227 pairs, or 49.62% of all the alignments, and modified a further 391, the 3.71% of the total. The pairs deemed to be already correct were 4915, amounting to 46.67% of the total. Thus, the final corpus has a total size of 5306 sentence pairs. Table 1 gives the numbers of correct, removed, and modified pairs against the original size of the corpus.

2.3 Corpus

After the revision the corpus has 5306 sentence pairs, the 50.37% of the initial 10.533. The Lombard side has 122.550 tokens,[12] the Italian one has 113.385, for a total of 236.264 tokens. The average sentence length in tokens is 23.10 for Lombard and 21.37 for Italian. Table 2 summarizes these statistics.

N. of pairs	**N. of words**		**Avg. sentence length**	
	LMO	IT	LMO	IT
5306	122.550	113.385	23.10	21.37

Table 2: Some figures about the revised corpus: the total number of sentence pairs, the number of whitespace-separated tokens, and the average sentence length for each side.

[12] Here a token is intended a string separated by whitespace.

3 Limitations and Future work

Despite being cleaner, the corpus is definitely small, both in scale and scope. It will have to be expanded with data from other domains and sources to be more impactful for the training of an MT system which can generalize well across different domains.

Another limitation of the corpus, which is however inherent in Lombard text, is the lack of standardisation in orthography. As you may recall from the Introduction, at the present moment, there is no generally accepted orthography for Lombard and its varieties. This is reflected in the Lombard Wikipedia, where pages are written in one of the proposed orthographies and varieties, which is signalled by a disclaimer on the top of the page. This is an issue if this corpus is used in the training of a NLP system, since words with the same meaning and contexts of use will be present in different forms, with lower frequencies and thus, with worse representations.

Future work will aim to solve these issues from a NLP perspective. Applying Optical Character Recognition tools to existing text may be worthy of investigation as a way to augment the size of the corpus. A tool to convert text to a uniform orthography could be devised leveraging existing dictionaries and standardisation proposals.

4 Conclusions

This work focused on Lombard, a Gallo-Romance regional language spoken in and around the Northern Italian region of Lombardy. Despite having more than 3.5 million speakers, no NLP resource has ever been created specifically for this language, with most of the research concentrating on documentation and sociolinguistics issues.

This work thus focused on providing a first foundational NLP resource for Lombard, a manually revised parallel corpus starting from the only Lombard-Italian resource available on-line. This corpus was created automatically mining parallel text from Wikipedia, and was found to be noisy. Thanks to the manual revision of five annotators, all bilingual native speakers of both Italian and Lombard, we obtained a cleaner corpus, which is available on GitHub.[13]

While being small[14], this is a first step towards providing NLP tools to users of the Lombard language, hopefully securing its precarious position in the diverse and complex linguistic landscape of Italy.

Acknowledgments The author wishes to thank the volunteer annotators, whose help was vital. We also want to thank the reviewers for their useful comments. The author's work is partially supported by the Internal Grant Agency of Masaryk University, project MUNI/IGA/1334/2021, and by Lexical Computing.

[13] https://github.com/edoardosignoroni/piotost

[14] That is

References

1. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics **7**, 597–610 (nov 2019). https://doi.org/10.1162/tacl_a_00288, `https://doi.org/10.1162%2Ftacl_a_00288`
2. Bonfadini, G.: lombardi, dialetti. In: Eds., T. (ed.) Enciclopedia dell'italiano. Treccani, online at https://www.treccani.it/enciclopedia/dialetti-lombardi_(Enciclopedia-dell%27Italiano)/#Studi (2010)
3. Brasca, L.: Scriver Lombard. Menaresta, Monza, 1st adj. edn. (2011)
4. Chambers, J.K., Trudgill, P.: Dialectology. Cambridge Textbooks in Linguistics, Cambridge University Press, 2 edn. (1998). https://doi.org/10.1017/CBO9780511805103
5. Coluzzi, P.: The new speakers of lombard. Multilingua **38**(2), 187–211 (2019). https://doi.org/doi:10.1515/multi-2018-0017
6. Coluzzi, P., Brasca, L., Scuri, S.: Revitalizing contested languages: The case of lombard. In: Tamburelli, M., Tosco, M. (eds.) Contested Languages: The hidden multilingualism of Europe, chap. 9, pp. 163–182. John Benjamins, Amsterdam (2021)
7. Coluzzi, P., Brasca, L., Trizzino, M., Scuri, S.: Language planning for italian regional languages: the case of lombard and sicilian. In: Stern, D., Nomachi, M., Belić, B. (eds.) Linguistic Regionalism in Eastern Europe and Beyond: Minority, Regional and Literary Microlanguages, pp. 274–298. Peter Lang, Frankfurt am Main (2018)
8. Coseriu, E.: Los conceptos de dialecto, nivel y estilo de lengua y el sentido propio de la dialectología (1981)
9. Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P.O., Orife, I., Ogueji, K., Rubungo, A.N., Nguyen, T.Q., Müller, M., Müller, A., Muhammad, S.H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B.F.P., Dlamini, S., de Silva, N., Ballı, S.Ç., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I.A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., Adeyemi, M.: Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics **10**, 50–72 (2022). https://doi.org/10.1162/tacl_a_00447, `https://doi.org/10.1162%2Ftacl_a_00447`
10. Lewis, M.P., Simons, G.F.: Assessing endangerment: Expanding fishmans's gids (2010), `https://www.lingv.ro/RRL%202%202010%20art01Lewis.pdf`
11. Loporcaro, M.: Profilo Linguistico dei Dialetti Italiani. Manuali Laterza, Editori Laterza, Bari, 1st edn. (2009)
12. Majliš, M.: W2C – web to corpus – corpora (2011), `http://hdl.handle.net/11858/00-097C-0000-0022-6133-9`, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
13. Mareček, D., Yu, Z., Zeman, D., Žabokrtský, Z.: Deltacorpus 1.1 (2016), `http://hdl.handle.net/11234/1-1743`, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
14. Moseley, C., Nicholas, A.: Atlas of the World's Languages in Danger, Memory of Peoples, vol. 19. UNESCO, Paris, 3rd edn. (2010)
15. Rosa, R.: Plaintext wikipedia dump 2018 (2018), `http://hdl.handle.net/11234/1-2735`, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

16. Schwenk, H., Chaudhary, V., Sun, S., Gong, H., Guzmán, F.: Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. CoRR **abs/1907.05791** (2019), `http://arxiv.org/abs/1907.05791`
17. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), `http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf`