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

Edoardo Signoroni
e.signoroni@mail.muni.cz

NLP Centre, Masaryk University

Dic 10, 2022

Outline

1. Linguistic background
2. Related work
3. Methodology
4. Corpus
5. Limitations and future work

Some linguistic background...

Many think that the Italian linguistic landscape looks like this:

Some linguistic background...

Or like this:

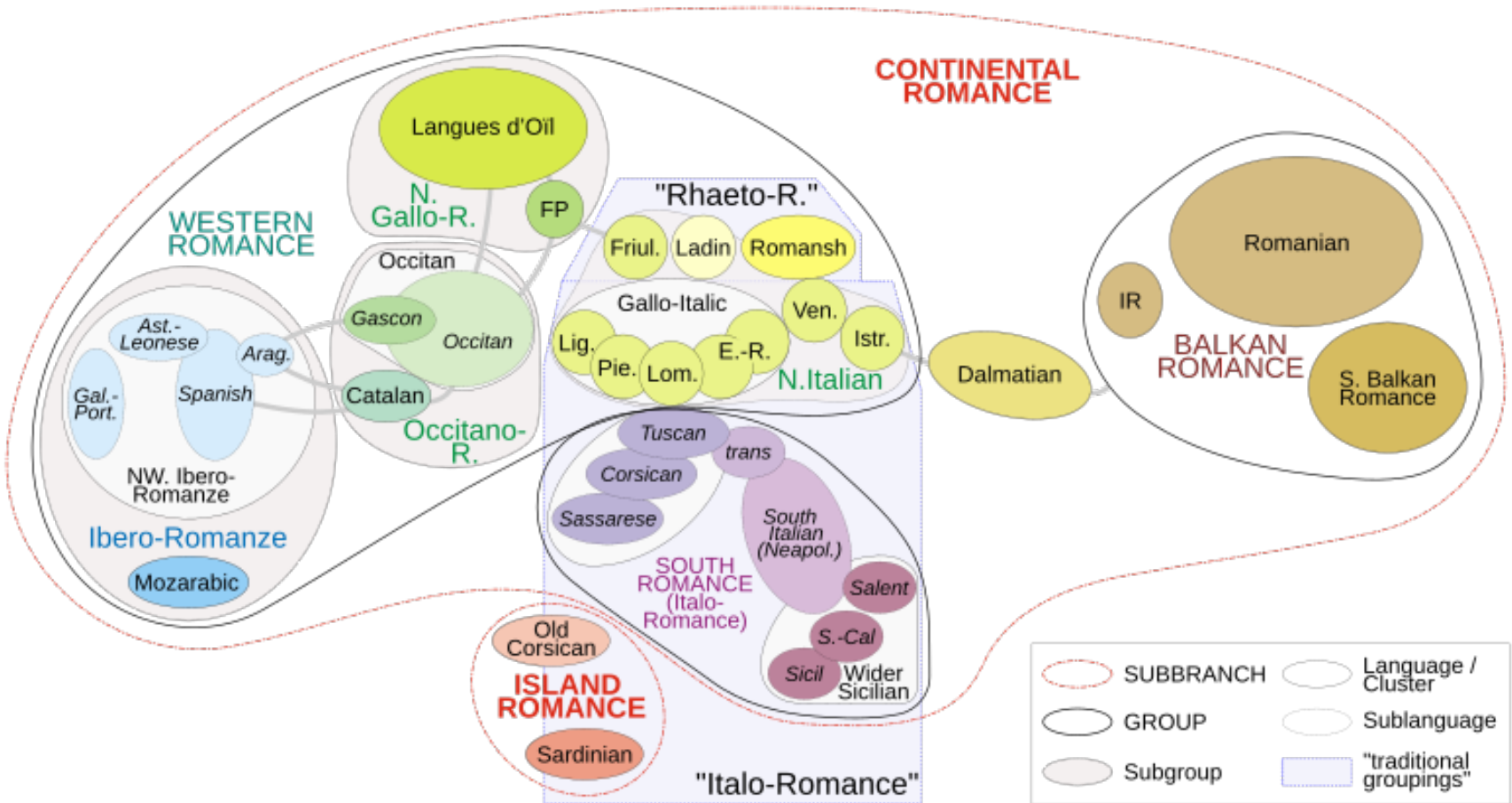
Some linguistic background...

When, in fact, it is more like this:

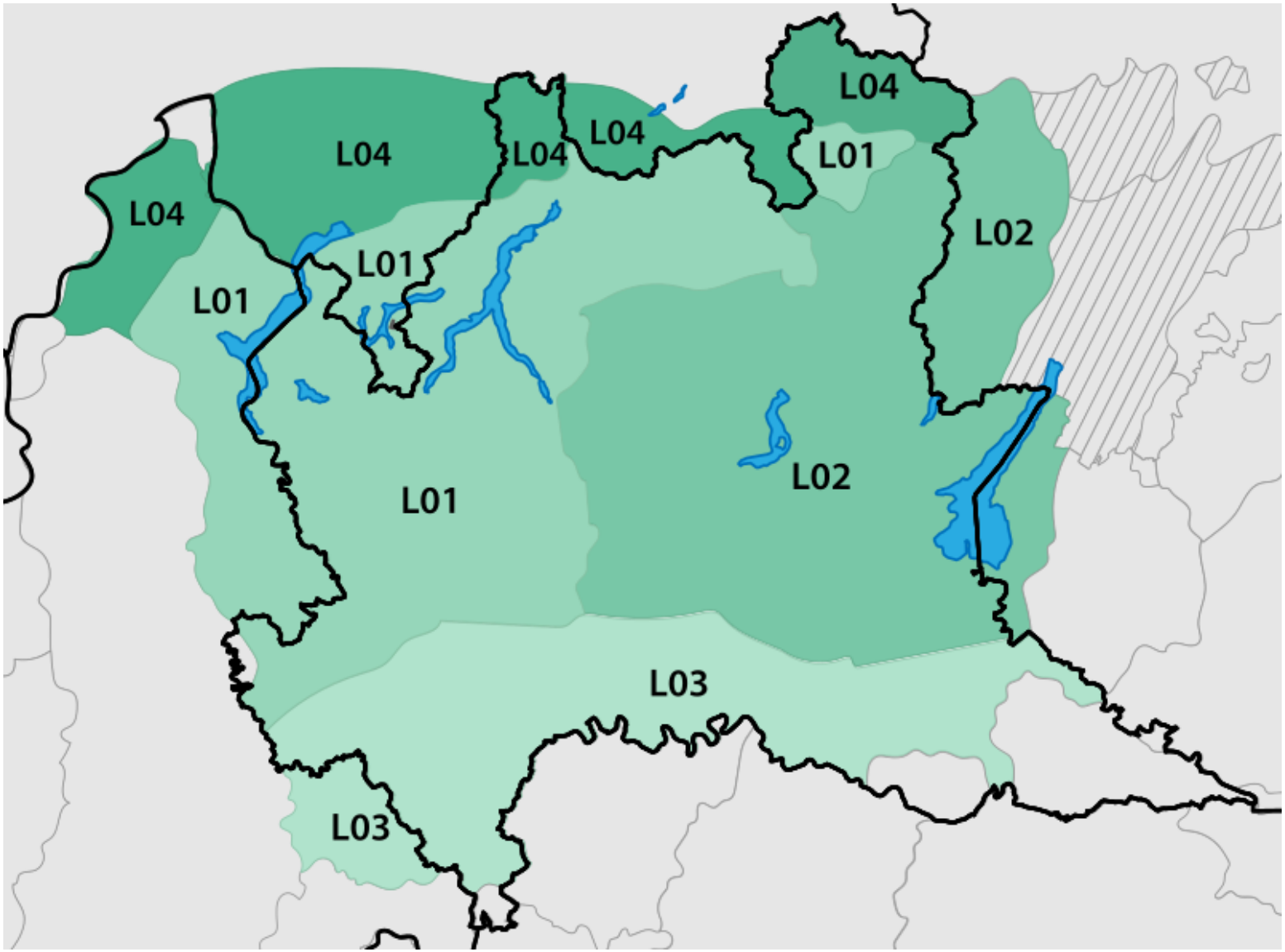
Some linguistic background...

- "Language" of Italy (it. *lingua*) vs "Dialect" of Italian (it. *dialetto*)
- Widespread diglossic bilingualism

The Lombard language



The Lombard language



The Lombard language

- 3.5 million speakers in Lombardy and the neighboring regions
- "Definitely Endangered" according to the UNESCO
- No unified orthography, but several standardization attempts

La mèrla

I mèrli ’na ólta i ghìa le pène biànche, ma chèl envéren lé l’éra stàt en bèl envéren e lé, la mèrla, la gà dìt: "Zenér de la màla gràpa, per tò despèt gó i uzilì ’ndela gnàta." A lü, ’l Zenér, gh’è nìt adòs ’n pó de ràbia, e ’l gà dìt: "Spèta, mèrla, che te la faró mé adès a té, e se te sét biànca mé te faró ègner négra." E pò dòpo ’l gà dit amò: "Dù ghe i ó e giü ’n prèstet el töaró e se te sét biànca, mé te faró ní négra." E alùra ’l gà fàt nì fò ’n frèt che se n’ìa mài vést giü compàgn.
Lé la mèrla la saìa piö che fà cói sò uzilì ndèla gnàta, e isé l’è nàda a rifügiàs endèla càpa del camì; dré al camì va sö ’l föm e lùr i uzilì i è déentàcc töcc négher, e quànche i è nicc fò de là, la mèrla la gh’ìa mìa piö le pène biànche, ma la ghe i éra négre. Alùra Zenér, töt sudisfàt, el gà dìt: "Tò mèrla, che te l’ó fàda mé staólta: se te se stàda biànca mé t’ó fàt ní négra e isé te làset lé de seghetà a tiràm en gìr."

The She-Blackbird

Once upon a time blackbirds had white feathers, but in that time winter had been mild and a she-blackbird scorned January saying: "Bad-headed January, in spite of you I have got a brood in my nest." Hearing this, January got angry and he said: "Just wait a bit, you she-blackbird, I will fool you and I will turn you from white into black." Then he said: "I have got two, and I will borrow one, and I will turn you from white to black." And he brought forth a cold as there had never been before.
The she-blackbird did not know how to cope with her brood in the nest, so she sheltered in the hood of a chimney, and the smoke turned all the birds black; so when they came out the blackbirds did not have white feathers anymore, but black ones. And January, very happy, said: "This time it was me that fooled you, blackbird: you were white and I turned you black, this will teach you to stop teasing me."

Linguistic research

- Linguistic research and dictionaries from the 19th century up to 2021
- Sociolinguistics and revitalization
- Lexical atlases

NLP resources

- lmo resources are part of bigger multilingual projects:
  - lmo sections of Wikipedia [1, 2, 3] (monolingual)
  - **lmo-xx parallel corpora on OPUS** [4]
- these are automatically aligned and quite noisy
- Other are cleaner:
  - FLORES-200 [1]
- Lombard fits the definition of "(very) low-resource language"

[1] https://github.com/facebookresearch/flores/tree/main/flores200

Revision process

- Manual revision of lmo-it alignments on OPUS
- Preliminary evaluation: 157/500 (31.4%) alignments were wrong
- Five Italian-Eastern Lombard bilingual annotators
- Alignments judged as wrong were removed
- Easy fixes (e.g. partial alignments) were manually corrected

Revised corpus

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.

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

Figure 1: Example of a removed sentence pair.

Limitations

- No standard orthography for Lombard, the language is mostly spoken.
- Thus, the annotators had different levels of proficiency with written Lombard.
- Accordingly, corrections were limited to the Italian side, to avoid injection of subjective spelling.
- The corpus was small, and now it is tiny.

Future work

- **expand the corpus** by scraping, crowdsourcing, and digitalization of texts (literary works, social media, developing L2 community, OCR)
- partial (and sometimes wrong) alignments hint to the possibility of extracting more data from Wikipedia
- explore the feasibility of a **MT system**

Conclusions

- we focused on **Lombard**, **a regional language** of Northern Italy
- and **revised** the only readily available **Lombard-Italian parallel corpus**.
- More than **half** of the (already small set of) alignments **was** judged as **wrong**
- thus, we need **better automatic alignment methods**,
- but also **data collection, documentation, NLP resources and tools**, ...

And to close on a happy tone:

Serenada de ’l burtulì

NLP Centre

e.signoroni@mail.muni.cz

GitHub: edoardosignoroni

Twitter: @edo_signoroni

References I

[1] Martin Majliš. *W2C – Web to Corpus – Corpora*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2011. URL: http://hdl.handle.net/11858/00-097C-0000-0022-6133-9.

[2] Rudolf Rosa. *Plaintext Wikipedia dump 2018*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2018. URL: http://hdl.handle.net/11234/1-2735.

[3] Holger Schwenk et al. "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia". In: *CoRR* abs/1907.05791 (2019). arXiv: 1907.05791. URL: http://arxiv.org/abs/1907.05791.

References II

[4] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

MUNI
FACULTY
OF INFORMATICS