Sicilian Translator

Eryk Wdowiak
eryk@wdowiak.me
Project Napizia

10 May 2021

Why don’t we have a Sicilian Translator?

- Google Translate doesn’t translate the Sicilian language.
- Nor does Bing Translator, Yandex Translate or DeepL Translator.
- Why not? What are we going to do about it?

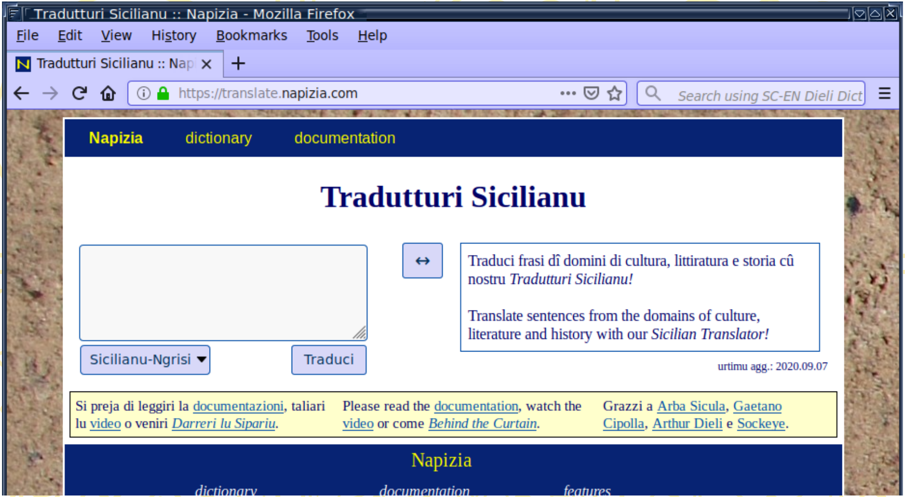
What is the Sicilian Language?

- The Sicilian School of Poets at the imperial court of Frederick II:
  - created the first literary standard in Italy (13th century)
  - inspired Dante, the "father of the Italian language"
- Sicilian emerged as a literary language before Italian.
- The people of Sicily, Calabria and Puglia speak it everyday.
  - They speak Italian at work.
  - But at home – with family and friends – they speak Sicilian.
  - More precisely, their own dialect of the language.
- And Sicilian is a language spoken here in Brooklyn, NY.

So why don’t we have a Sicilian Translator?

- No one had assembled the parallel text to train a translator.
- There’s plenty! For over 40 years, Arba Sicula has been:
  - organizing poetry recitals, concerts, cultural events and tours of Sicily
  - publishing books on Sicilian language, literature, history, cuisine, fiction, ...
  - translating Sicilian poetry and prose
  - publishing a bilingual journal, *Arba Sicula*.
- So we assembled the parallel text.

Sicilian Translator


Tradutturi Sicilianu :: Napizia - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
Tradutturi Sicilianu :: Napi
https://translate.napizia.com
Search using SC-EN Dieli Dict
Napizia
dictionary
documentation
↔
Sicilianu-Ngrisi
Traduci

Tradutturi Sicilianu

Traduci frasi dî domini di cultura, littiratura e storia cû nostru *Tradutturi Sicilianu!*

Translate sentences from the domains of culture, literature and history with our *Sicilian Translator!*

urtimu agg.: 2020.09.07

Si preja di leggiri la documentazioni, taliari lu video o veniri *Darreri lu Sipariu*.

Please read the documentation, watch the video or come *Behind the Curtain*.

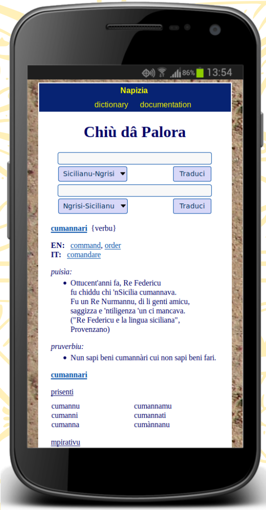
Grazzi a Arba Sicula, Gaetano Cipolla, Arthur Dieli e Sockeye.

https://translate.napizia.com

How did we do it?

- We did **NOT** start with data collection.
- We started by collecting the rules of Sicilian vocabulary and grammar.
  - Arthur Dieli's *Sicilian Vocabulary*
  - Kirk Bonner's *Introduction to Sicilian Grammar* (2001)
  - Gaetano Cipolla's *Mparamu lu sicilianu* (2013)
- And we created the *Chiù dâ Palora* (*More About the Word*) dictionary.
  - vocabulary annotated with grammar, proverbs, poetry, prose and examples
  - provides a reference for standardizing Sicilian language text

More About the Word



Napizia

dictionary documentation

Chiù dâ Palora

Sicilianu-Ngrisi ▾ Traduci

Ngrisi-Sicilianu ▾ Traduci

cumannari {verbu}

EN: command, order
IT: comandare

puisìa:

- Ottucent'anni fa, Re Federicu
  fu chiddu chi 'nSicilia cumannava.
  Fu un Re Nurmannu, di li genti amicu,
  saggizza e 'ntiligenza 'un ci mancava.
  ("Re Federicu e la lingua siciliana",
  Provenzano)

pruverbiu:

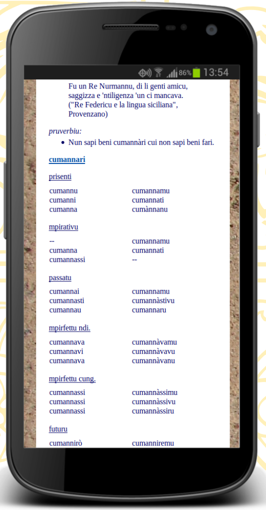
- Nun sapi beni cumannàri cui non sapi beni fari.

cumannari

prisenti

cumannu	cumannamu
cumanni	cumannati
cumanna	cumànnanu

mpirativu



Fu un Re Nurmannu, di li genti amicu,
saggizza e 'ntiligenza 'un ci mancava.
("Re Federicu e la lingua siciliana",
Provenzano)

pruverbiu:

- Nun sapi beni cumannàri cui non sapi beni fari.

cumannari

prisenti

cumannu	cumannamu
cumanni	cumannati
cumanna	cumànnanu

mpirativu

--	cumannamu
cumanna	cumannati
cumannassi	--

passatu

cumannai	cumannamu
cumannasti	cumannàstivu
cumannau	cumannaru

mpirfettu ndi.

cumannava	cumannàvamu
cumannavi	cumannàvavu
cumannava	cumannàvanu

mpirfettu cung.

cumannassi	cumannàssimu
cumannassi	cumannàssivu
cumannassi	cumannàssiru

futuru

cumannirò	cumanniremu

https://www.napizia.com/cgi-bin/cchiu-da-palora.pl

Then We Began Collecting Data

- Sources of parallel text:
  - the bilingual literary journal *Arba Sicula*
  - A. Dieli's translations of Sicilian poetry, proverbs and G. Pitrè's *Folk Tales*
  - examples from G. Cipolla's *Mparamu* and K. Bonner's *Introduction*
- Data Preparation
  - Selected Sicilian language text that could be edited to Standard Sicilian.
  - Used *hunalign* to identify translated sentence pairs.
  - Manually edited the text (both languages) for quality and standardization.
- Our parallel corpus (so far):
  - 16,891 lines for training – 262,582 Sicilian words, 266,740 English words
  - 121 hand selected for validation – 1836 Sicilian words, 1878 English words
  - the Italian-English subset of Farkas' *Books*

And We Began Modeling

- We trained our translation models with Sockeye.
- Adding parallel text always improves translation quality more than adjusting hyperparameters.
- But some ways of training a model are better than others.
- We avoid overfitting by training:
  - a self-attentional Transformer model (Vaswani et al., 2017)
  - a smaller network with fewer layers (Sennrich and Zhang, 2019)
  - with small subword vocabularies (Sennrich, Haddow and Birch, 2016)
  - with high-dropout parameters (Srivastava et al., 2014)
- Large empirical improvements when we added theoretical knowledge:
  - by pushing the subword distribution toward textbook desinences
  - by using textbook examples to give structure to the sequences

Evaluation Metrics

- BLEU score – measure of translation quality
  - Higher when sequences of words in the candidate translation match sequences in the reference translation.
  - Highly correlated with human judgements of translation quality.

- What's a good score?
  - In theory, BLEU score ranges from 0 to 100.
  - In practice, there are many ways to translate a sentence.
  - Scores reported by Vaswani et al. (2017):
    - 41.8 on English-to-French translation
    - 28.4 on English-to-German translation

- Our *Tradutturi Sicilianu* achieved BLEU scores of:
  - 35.1 on English-to-Sicilian translation
  - 34.6 on Sicilian-to-English translation

Comparing Models

- Translation quality improves with parallel text. As our dataset grew from 120,000 to 270,000 words, our BLEU scores increased:
  - from 11.4 to 25.1 on English-to-Sicilian translation
  - from 12.9 to 29.1 on Sicilian-to-English translation
- Within a given dataset, pushing the subword distribution toward textbook desinences increased BLEU scores:
  - from 20.3 to 22.4 on English-to-Sicilian translation
  - from 21.4 to 24.1 on Sicilian-to-English translation
- We also observed larger increases in BLEU scores when we added parallel text from textbook examples than from other sources.
  - We did not conduct any formal tests to confirm this observation.
  - With our eyes, we could see the structure that textbook examples added.

Multilingual Translation

- To further improve translation quality, we added the Italian-English subset of Farkas' *Books* (from the OPUS project) to our dataset.
- To enable multilingual translation, we added a directional token – ex. `<2it>` ("to Italian") – to the source sequence (Johnson et al., 2016).
- And we trained a larger model (Arivazhagan et al., 2019)
- This further improved translation quality. Our BLEU scores increased:
  - from 25.1 to 35.1 on English-to-Sicilian translation
  - from 29.1 to 34.6 on Sicilian-to-English translation
- And it enabled "zero-shot translation" between Sicilian and Italian.

Come to Napizia!

- So come to *Napizia* and try our *Tradutturi Sicilianu:*
  - https://translate.napizia.com
- To see how it works “behind the curtain,” come *Darreri lu Sipariu:*
  - https://translate.napizia.com/cgi-bin/darreri.pl
- Read our “Introduction to Sicilian NLP”
  - https://www.doviak.net/pages/ml-sicilian/
- And check out our `Sicilian_Translator` repository at Github:
  - https://github.com/ewdowiak/Sicilian_Translator
- We hope you’ll join us. *Grazzi!*