



# Sicilian Translator

Eryk Wdowiak  
[eryk@wdowiak.me](mailto:eryk@wdowiak.me)

Arba Sicula

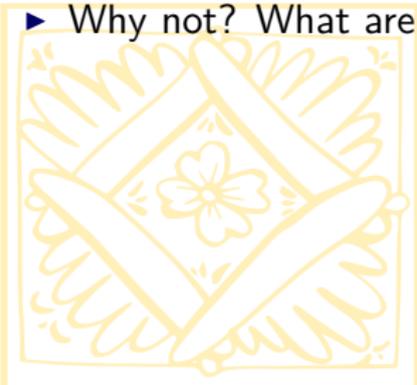
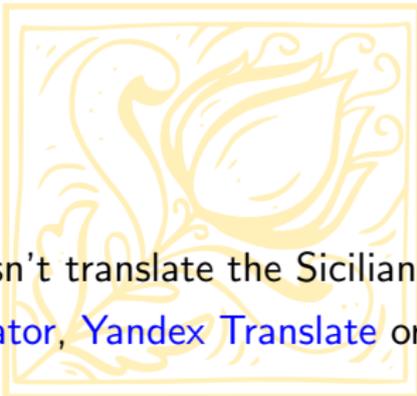
27 Sep 2021

# N

## 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?



# N

## 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.

# N

## 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

The screenshot shows a web browser window with the title "Traduttori Sicilianu :: Napizia - Mozilla Firefox". The address bar displays "https://translate.napizia.com". The website header includes "Napizia", "dictionary", and "documentation". The main heading is "Traduttori Sicilianu". Below this is a text input area, a "Traduci" button, and a dropdown menu set to "Sicilianu-Ngrisi". A right-hand box contains the text: "Traduci frasi di cultura, littiratura e storia cù nostru Traduttori Sicilianu!" and "Translate sentences about culture, literature and history with our Sicilian Translator!". The footer includes the text "urtimu agg.: 2021.06.05" and a yellow banner with three columns of text: "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](#)."; and "Grazzi a [Arba Sicula](#), [Gaetano Cipolla](#) and [Arthur Dieli](#)."

<https://translate.napizia.com>

Napizia

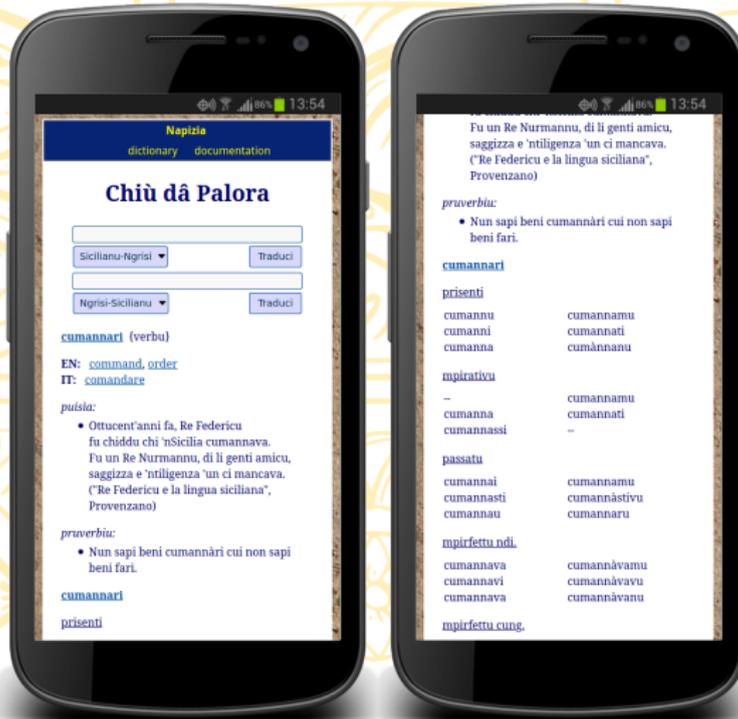
# N

## 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 *Chiu' 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



# N

## 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):
  - 12,357 lines of bilingual text – 237,456 Sicilian words, 236,568 English words
  - 4,660 lines of trilingual textbook exercises – Sicilian, English and Italian
  - 121 hand-selected lines for validation (in all three languages)
  - the Italian-English subset of *Farkas' Books*

# N

## 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

# N

## 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 *Traduttori Sicilianu* achieved BLEU scores of:
  - 35.0 on English-to-Sicilian, 36.8 on Sicilian-to-English
  - 36.5 on Italian-to-Sicilian, 30.9 on Sicilian-to-Italian

# N

## 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.

# N

## 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](#)) that “bridges” ([Fan et al., 2020](#)) with trilingual textbook exercises.
- 
- ▶ This further improved translation quality. Our BLEU scores increased:
    - from 25.1 to 35.0 on English-to-Sicilian
    - from 29.1 to 36.8 on Sicilian-to-English
  - ▶ And it yielded good translation quality between Sicilian and Italian:
    - 36.5 on Italian-to-Sicilian
    - 30.9 on Sicilian-to-Italian

# N

## Come to Napizia!

- ▶ So come to *Napizia* and try our *Traduttori Sicilianu*:
  - <https://translate.napizia.com>
- ▶ To see how it works “behind the curtain,” come *Darrerri lu Sipariu*:
  - <https://translate.napizia.com/cgi-bin/darrerri.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](https://github.com/ewdowiak/Sicilian_Translator)
- ▶ We hope you’ll join us.

# N

## Acknowledgements

- 
- ▶ Thank you to [Arba Sicula](#), [Gaetano Cipolla](#) and [Arthur Dieli](#):
    - for developing the resources that made this project possible
    - for their support and encouragement
    - for sponsoring and developing Sicilian language and culture

▶ *Grazzi!*