



Sicilian Translator

Eryk Wdowiak

eryk@wdowiak.me

Project Napizia

October 1, 2020

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.



So why don't we have a Sicilian Translator?

- ▶ Because 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



<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 *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



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

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):

- 14,494 lines for training – 196,911 Sicilian words, 202,652 English words
- 121 hand selected for validation – 1836 Sicilian words, 1878 English words

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 splitting 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:
 - 22.5 on English-to-Sicilian translation
 - 25.2 on Sicilian-to-English translation

N

Comparing Models

- ▶ Translation quality improves with parallel text. As our dataset grew from 120,000 to 200,000 words, our BLEU scores increased:
 - from 11.4 to 22.5 on English-to-Sicilian translation
 - from 12.9 to 25.2 on Sicilian-to-English translation
- ▶ Within a given dataset, pushing the subword splitting 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

Next Steps

- ▶ We still have more issues of *Arba Sicula* to extract text from.
 - ▶ So we'll add more parallel text and further increase translation quality.
 - ▶ With English-Italian parallel text, we can enable “zero-shot translation” between Sicilian and Italian ([Johnson et al., 2016](#)).
-
- ▶ We may also use the Sicilian text to train word embeddings and create lists of context similar words for our dictionary, *Chìu dâ Palora*.
 - ▶ And we'll add more poetry, proverbs and prose to the dictionary too.



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 *Darrerì lu Sipariu*:
 - <https://translate.napizia.com/cgi-bin/darrerì.pl>
- ▶ Try our “Recipe for Low-Resource NMT”
 - <https://www.doviak.net/pages/ml-sicilian/ml-scn.p05.shtml>
- ▶ And check out our *Sicilian_Translator* repository at Github:
 - https://github.com/ewdowiak/Sicilian_Translator
- ▶ We hope you’ll join us. *Grazzi!*