Compass Aligned Distributional Embeddings with Language Shift

Artificial Intelligence - 2019/2020

Amrani Hamza 807386 - Carta Costantino 808417 - Vitali Simone 807792

Aim of the project

- Verify that the **vector representation of a word translated** in different languages is **similar**
- **Using TWEC to align** documents translated in different languages
- **TWEC AS-IS** vs **TWEC-IIS** (*Identity Injection Substitution*)
- Obtained results analysis

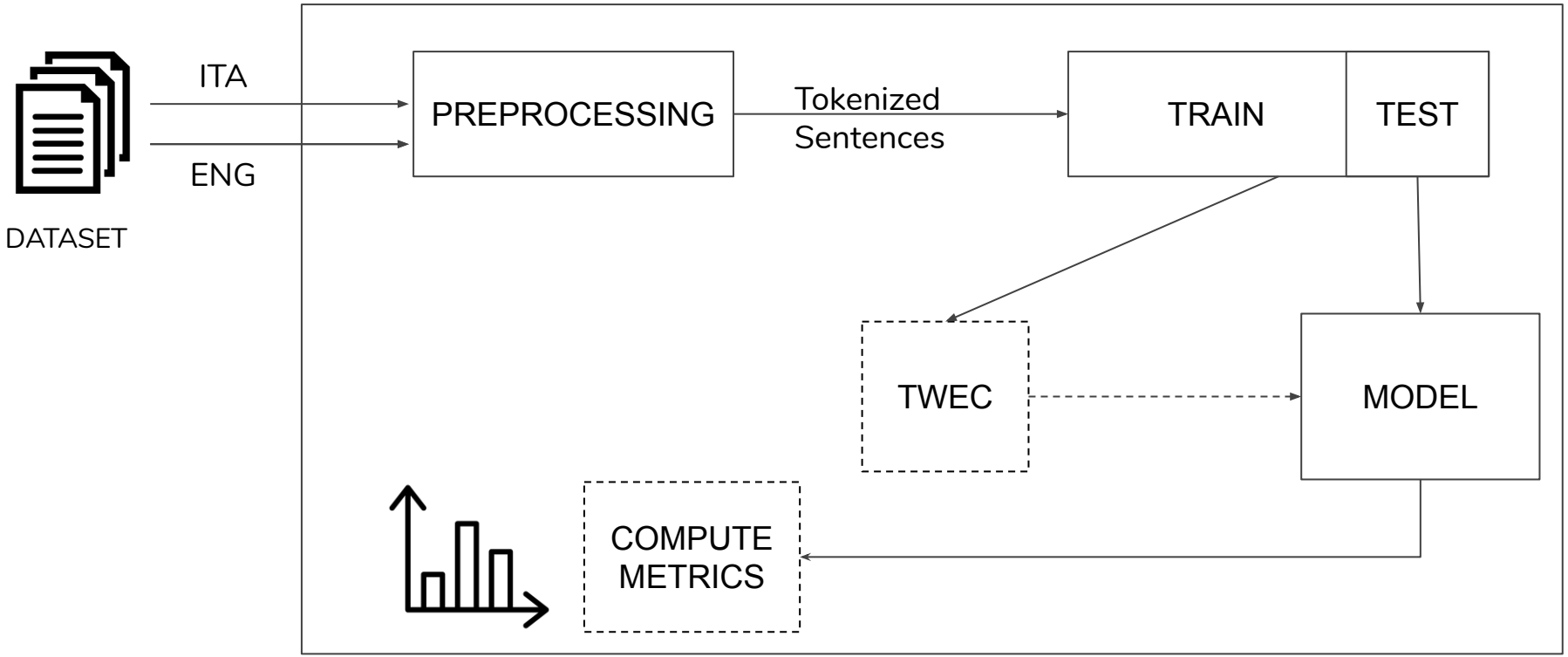
Dataset*

- Proceedings of the **EUROPEAN PARLIAMENT** (1996 - 2011)
- Extracted from the website of the European Parliament
- Composed by **1.946.253 non-structured sentences**
- Translated in **21 european** languages:
  - **italian**
  - **english**
  - ...

Aligned through *Gale-Church algorithm*

* Europarl Corpus

Workflow

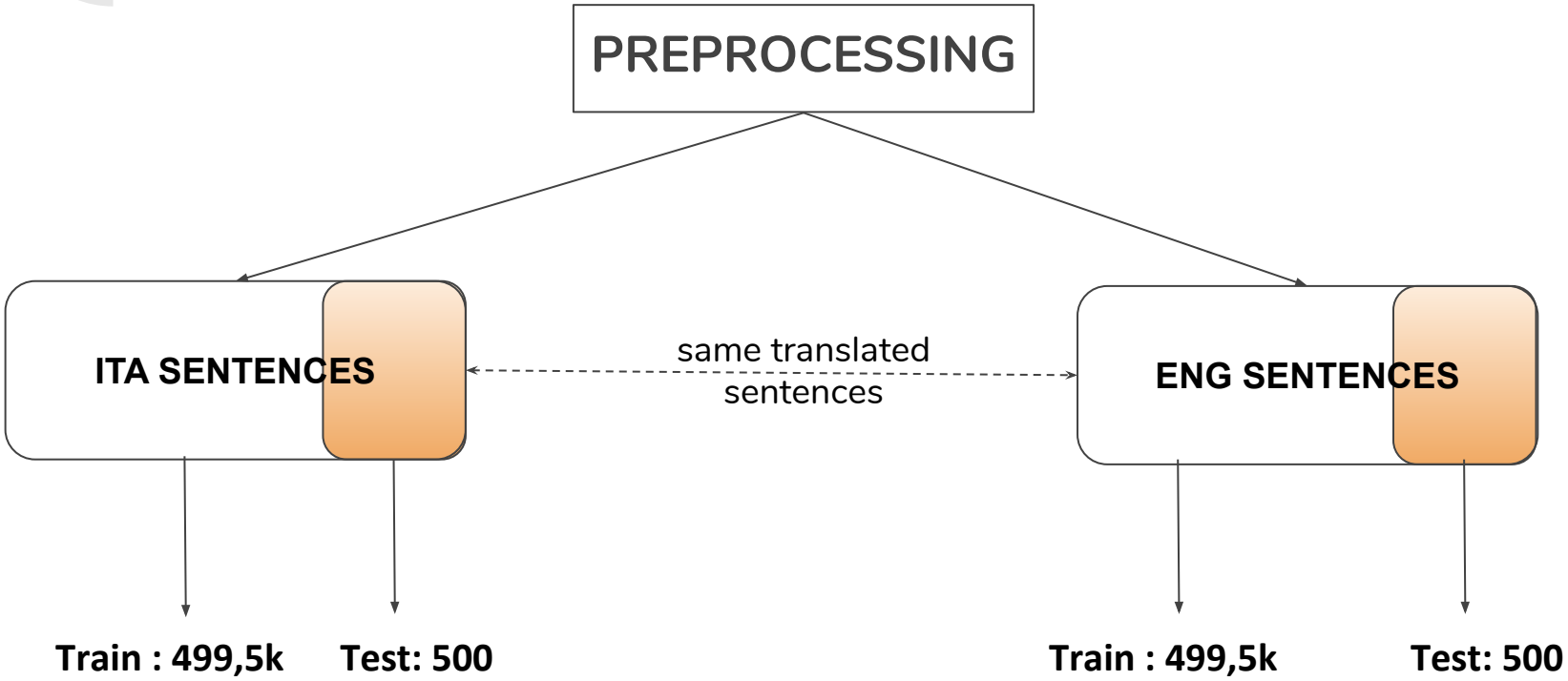

ITA
ENG
DATASET
PREPROCESSING
Tokenized
Sentences
TRAIN
TEST
TWEC
MODEL
COMPUTE
METRICS

Dataset preprocessing

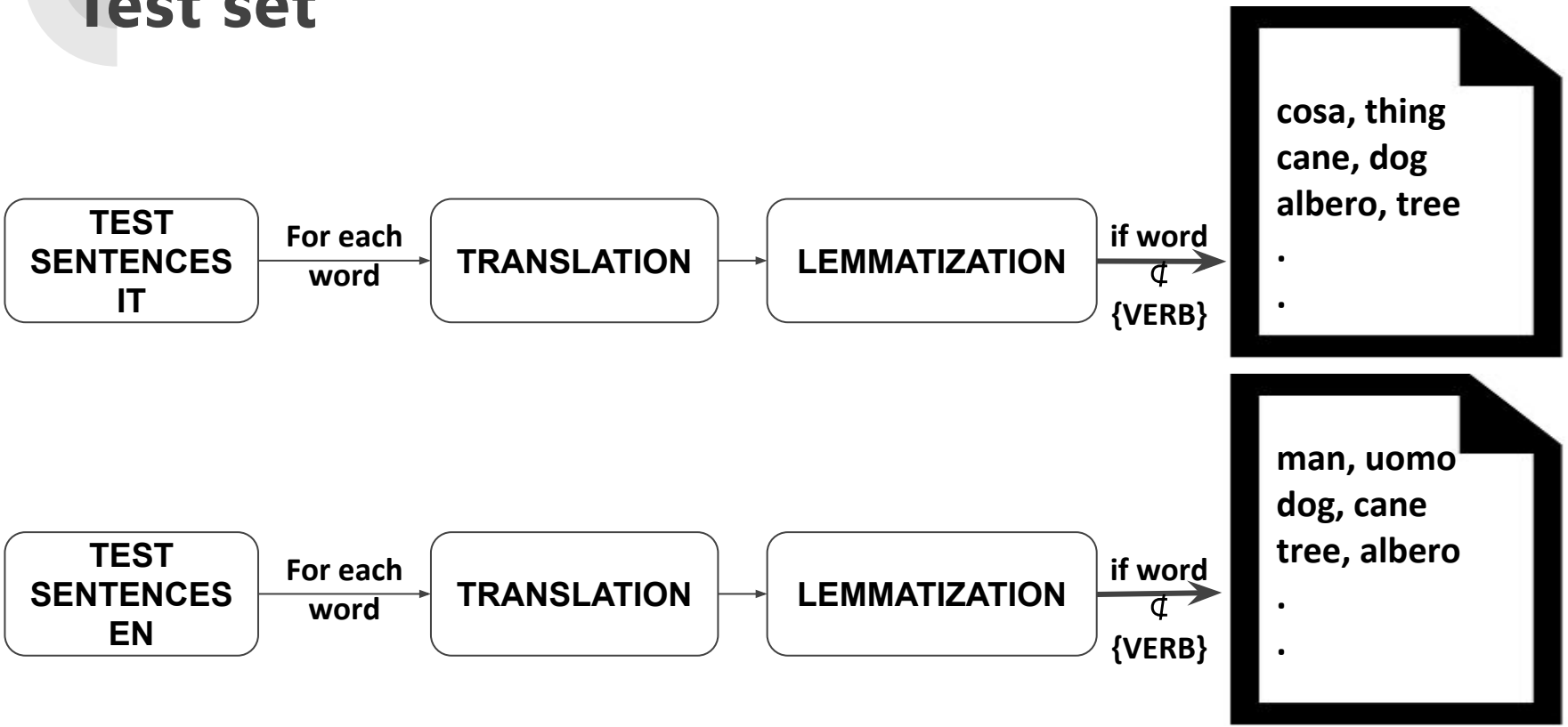
1. Subsampling **500k sentences**
2. **Tokenization**
3. **Data cleaning**:
   - to **lower** case
   - **deaccented** characters
   - removed words with length < 2 or > 15
   - removed digit numbers, date, time, punctuation and special characters
4. Removed **stopwords**
5. **Lemmatization** with "**spacy Lemmatizer"**
6. Removed **words** with **occurence equal or less than 3 in total**

The **same preprocessing** has been executed for the **two corpus**

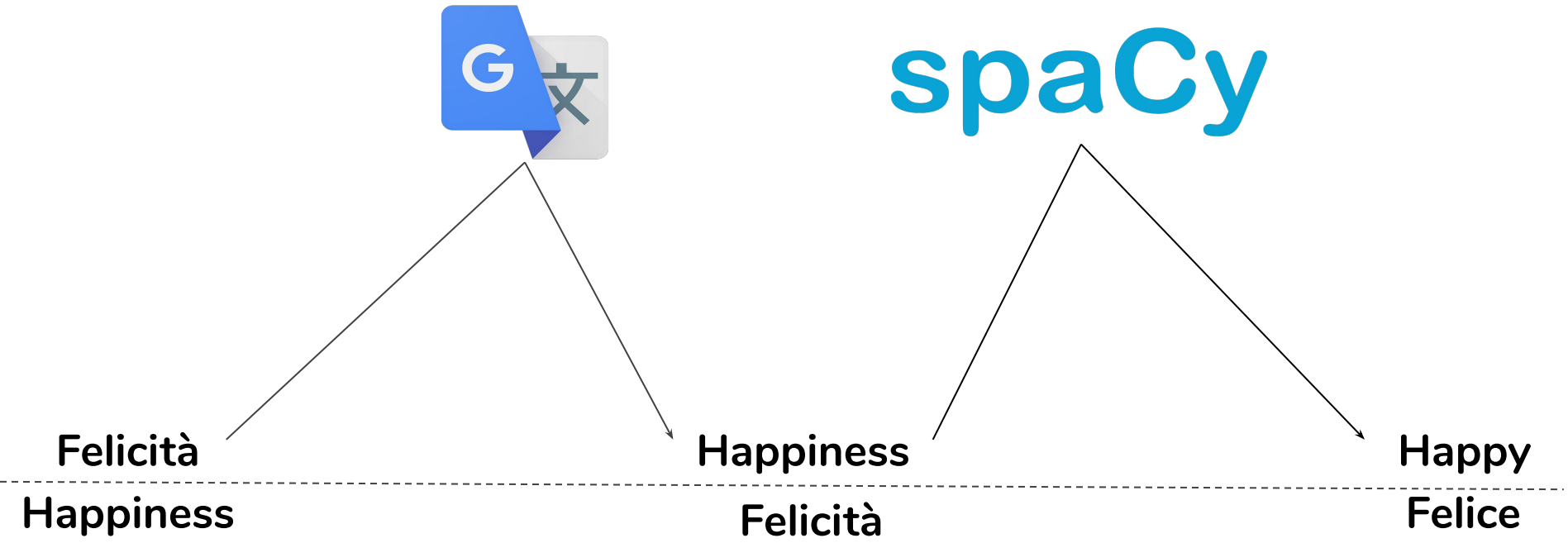
Train - Test Split



Test set


TEST SENTENCES IT
For each word
TRANSLATION
LEMMATIZATION
if word ¢ {VERB}
cosa, thing
cane, dog
albero, tree
.
.
TEST SENTENCES EN
For each word
TRANSLATION
LEMMATIZATION
if word ¢ {VERB}
man, uomo
dog, cane
tree, albero
.
.

Translation and Lemmatization

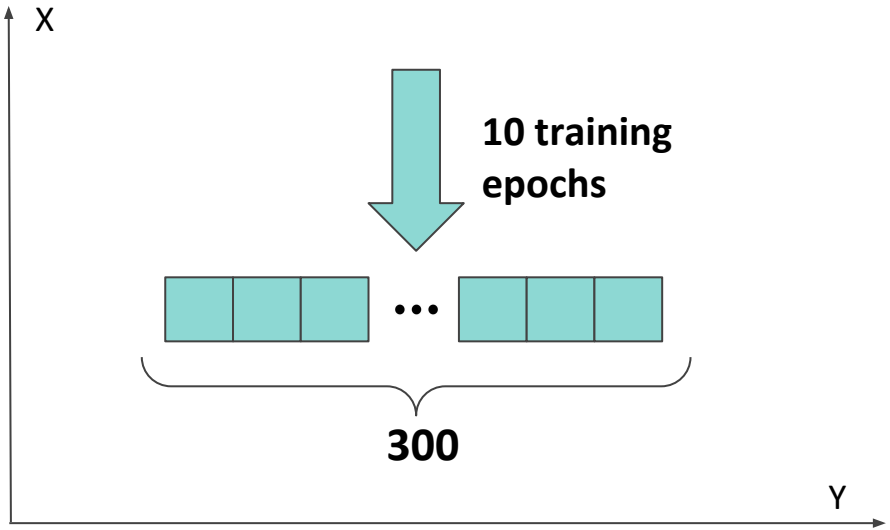


https://pypi.org/project/googletrans/

https://spacy.io/api/lemmatizer

TWEC* with CBOW on train set

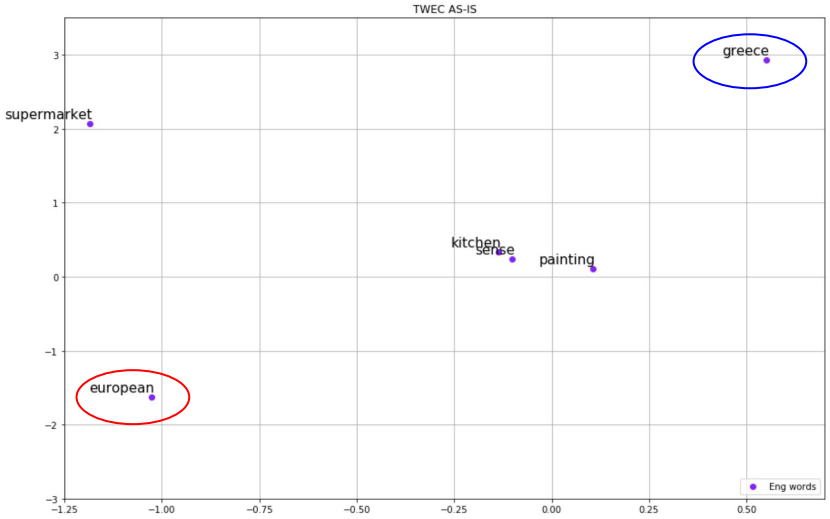
- CBOW with negative sampling:
  - **Concatenation of the two corpus tokens (Save T)**
  - **Italian tokens (Target fixed = T)**
  - **English tokens (Target fixed = T)**
- Parameters:
  - Dimension: **300**
  - Epochs: **10**
  - # of negative sampling: **5**
  - Initial LR: **0.025**
  - Init Mode: **hidden**

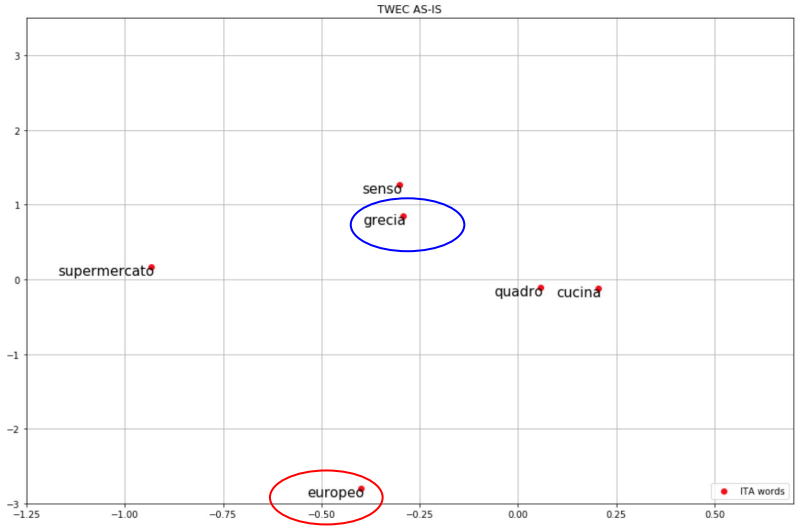


* Carlo, Valerio & Bianchi, Federico & Palmonari, Matteo (2019) Training Temporal Word Embeddings with a Compass

TWEC AS-IS

- Word embeddings from **two different languages** are **difficult** to compare

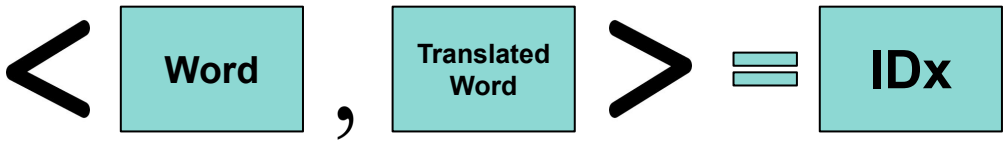




* Word vectors are visualized with PCA

TWEC IIS (Identity Injection Substitution) - Mapping

- Creation of a **mapping** between **words translated** in the **two languages**
- **Aim:** represent the same word in different languages with the same representation



Example:

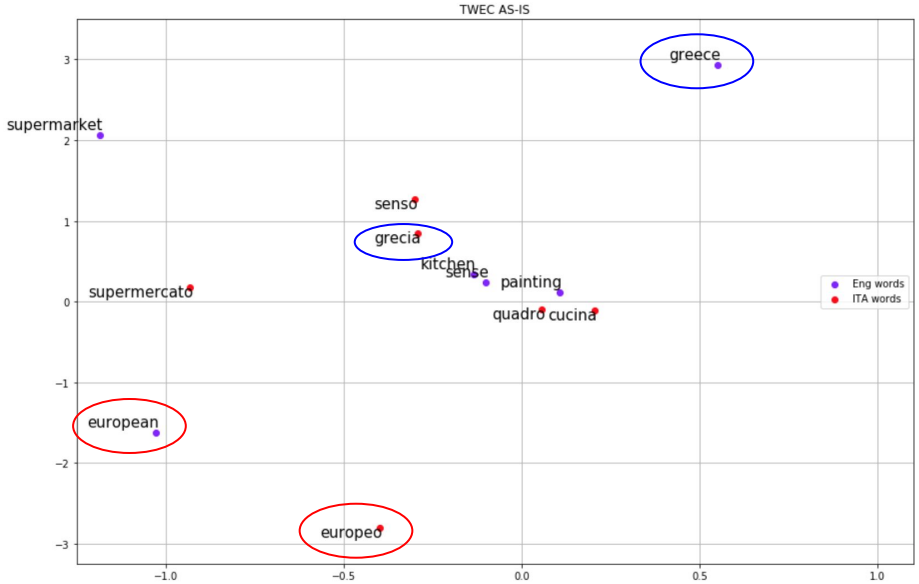
```
cosa, thing, id0
anno, year, id1
uomo, man, id2
giorno, day, id3
volta, time, id4
casa, home, id5
```

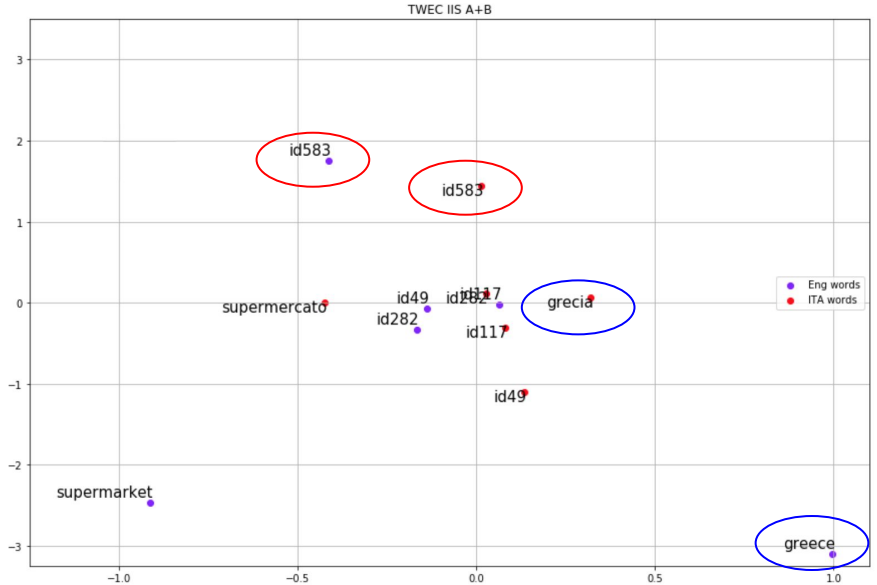
TWEC IIS - Mapping taxonomy

A. Most **frequently** used words in the two languages:
   - **Italian**
     - https://it.wiktionary.org/wiki/Utente:.mau./Le_mille_parole_pi%C3%B9_comuni
   - **English**
     - https://en.wikipedia.org/wiki/Dolch_word_list
B. Most **frequent** words that appear within the two corpus:
   - 350 most frequent words for each corpus

For each type of mapping a **dictionary** has been created composed up by ~**500 mappings.**

TWEC AS-IS vs TWEC IIS

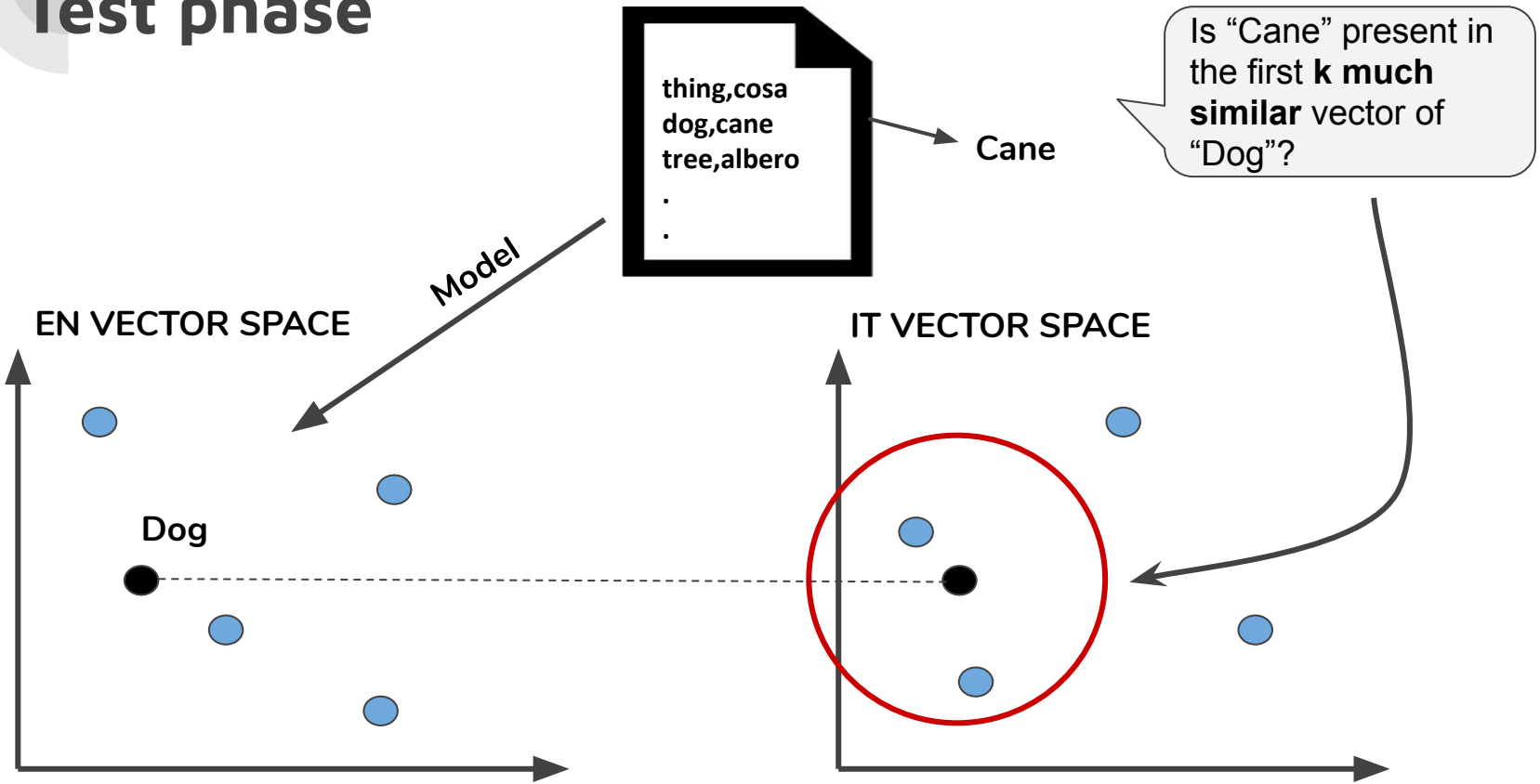




* Word vectors are visualized with PCA

< european , europeo > = id583
< kitchen , cucina > = id282
< painting , quadro > = id117
< sense , senso > = id49

Test phase


thing,cosa
dog,cane
tree,albero
.
.
Cane
Is “Cane” present in the first **k much similar** vector of “Dog”?
Model
EN VECTOR SPACE
IT VECTOR SPACE
Dog

Evaluation metrics

- **MRR** (Mean Reciprocal Rank):

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{Relevance\_Label\_Value}{\text{rank}_i}$$

- **HITS@k:** if correct translation appear within the top-k elements of measure results list

- **Cosine similarity**: $$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Results on Test set

	MODELS	MRR	HITS@1	HITS@3	HITS@5	HITS@10
Ita ➔ Eng	TWEC1	0.1403	0.1095	0.1807	0.2102	0.2418
	TWEC2-A	0.5422	0.4909	0.6066	0.6350	0.6881
	TWEC2-B	0.6036	0.5496	0.6699	0.6941	0.7212
	TWEC2-A+B	**0.6439**	**0.6009**	**0.6974**	**0.7304**	**0.7629**
Eng ➔ Ita	TWEC1	0.1141	0.9500	0.1386	0.1776	0.2165
	TWEC2-A	0.4676	0.4092	0.5442	0.5782	0.6248
	TWEC2-B	0.5566	0.5056	0.6153	0.6521	0.6898
	TWEC2-A+B	**0.5943**	**0.5397**	**0.6616**	**0.6928**	**0.7278**

Conclusions

- Good results
- **TWEC AS-IS** doesn't work well in Language shift as in Temporal shift
- The mapping between words **improve** the accuracy of the model in terms of **cosine similarity** between the original and the translated word representation

Future work...

- Compare corpus in other languages
  - *other european languages (e.g: fr, de, etc..)*
- Increase Train and Test data quantity
- Improve preprocessing
  - *N-grams handling, ...*
- Try other algorithms of embedding and alignment
- Further analysis on obtained results

Thank you!