TRI: a tool for the diachronic analysis of large corpora and social media

Pierpaolo Basile
pierpaolo.basile@uniba.it

Incontro utenti ReCaS - Bari, 12 luglio 2019

Hello!

I am Pierpaolo Basile

Natural Language Processing
Distributional Semantics
Information Retrieval/Filtering

You can find me at pierpaolo.basile@uniba.it

Words change their meaning *(usage)*

Words change their meaning *(usage)*

Motivation

Detect meaning shift

When was this meaning introduced?

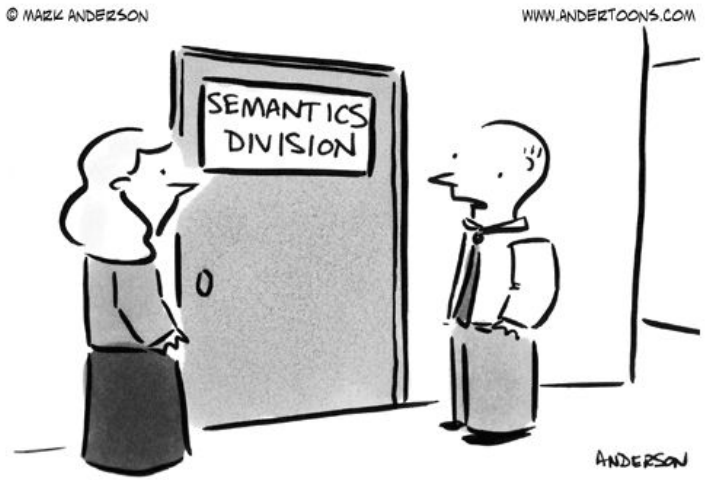
Surf!?!?! Or the web!?!?!?

surf the Net/Internet to use the Internet

Diachronic Linguistics

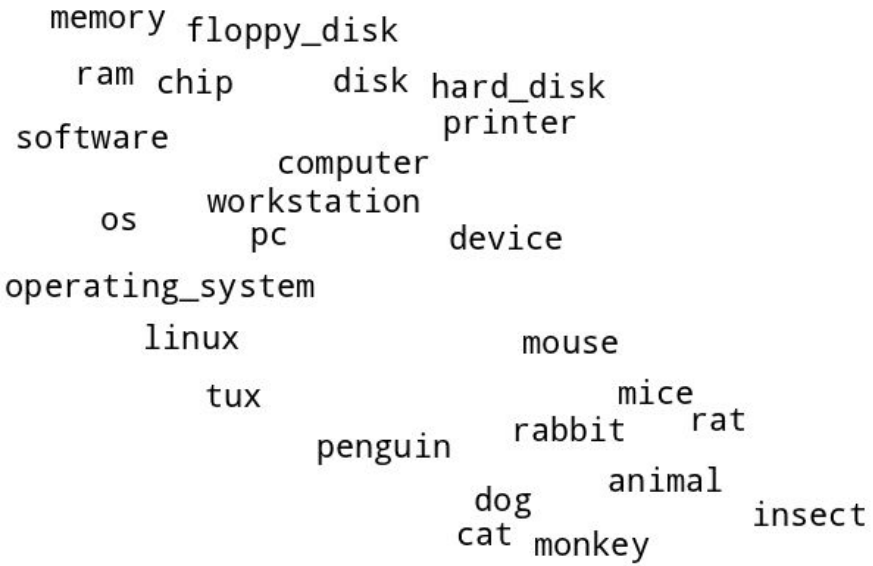
The scientific study of language change over time
(also called **Historical Linguistics**)

How to represent semantics?



Distributional Semantic Models

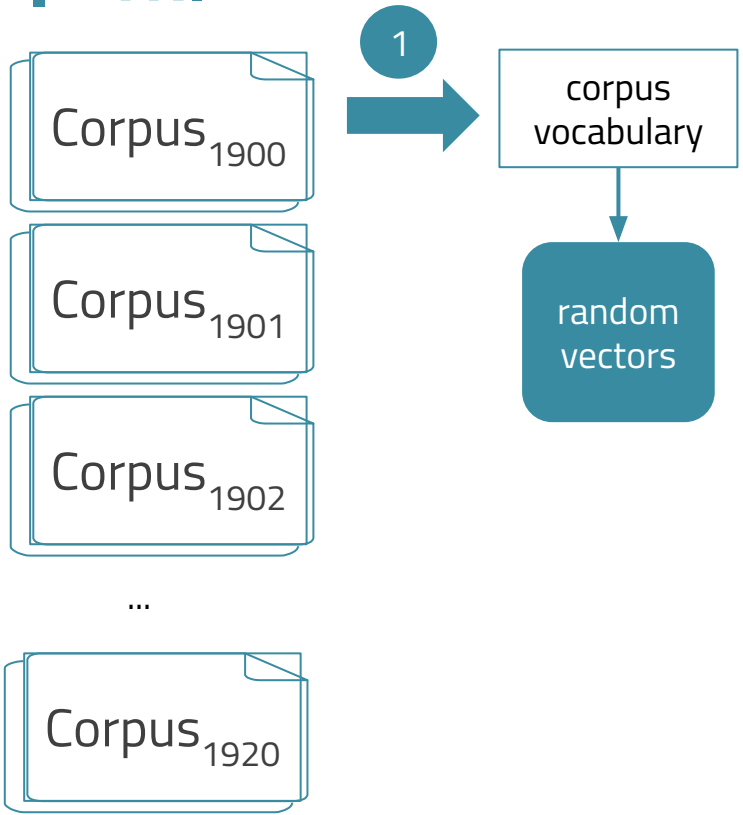
- Analysis of word-usage statistics over huge corpora
- Geometric space of concepts
- Similar words are represented close in the space

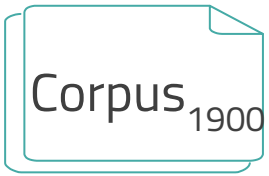


> *A **WordSpace** is a snapshot of a specific corpus, it does not take into account temporal information*

Temporal Random Indexing

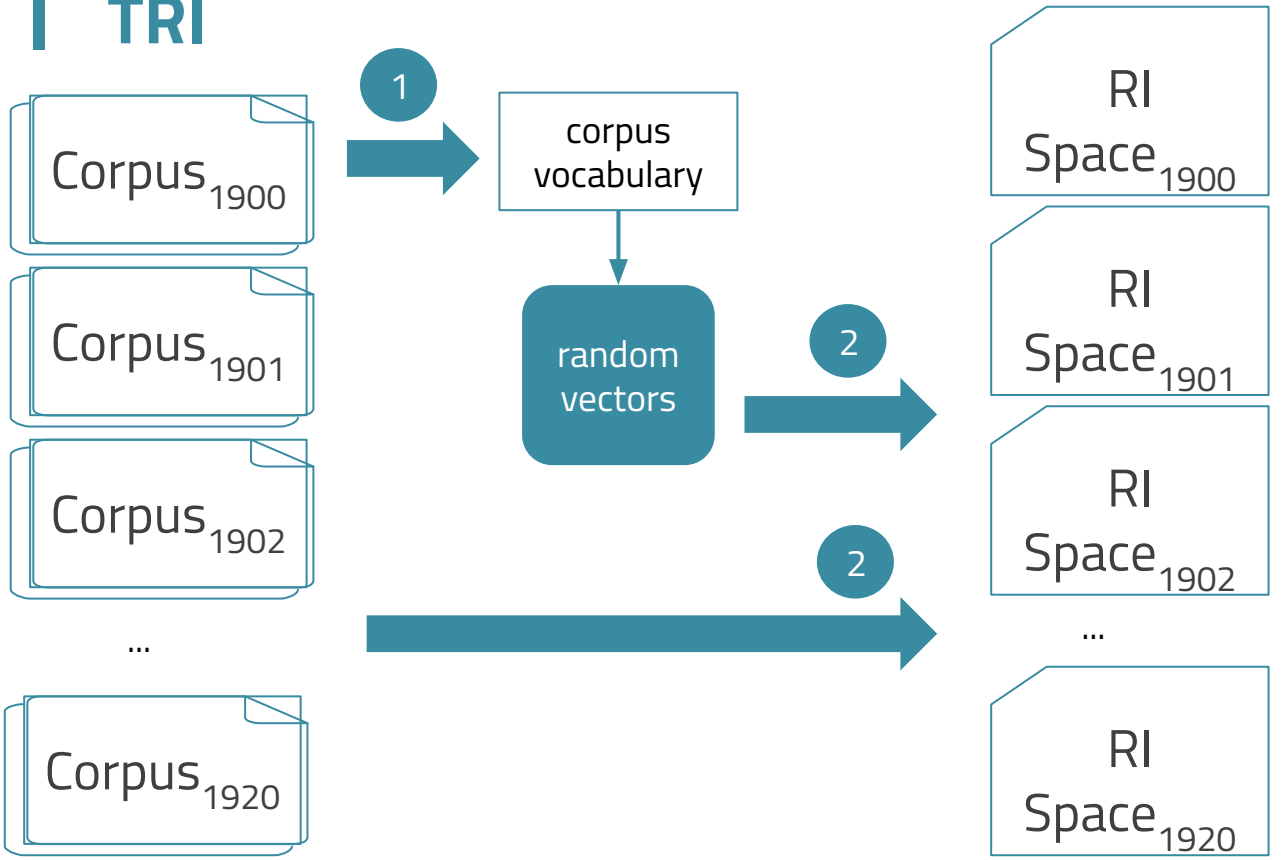
TRI





Temporal Random Indexing

TRI



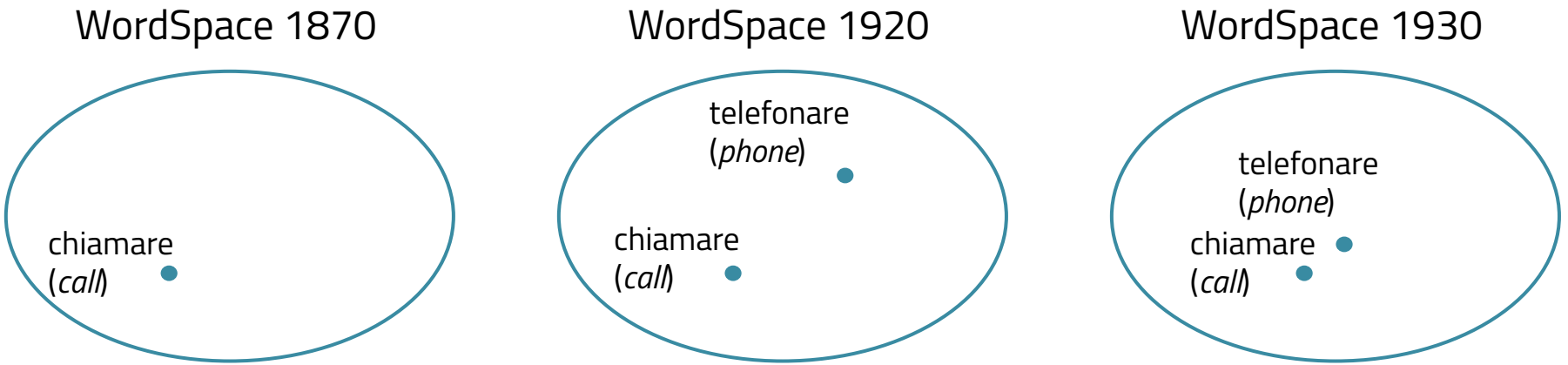
Semantic vector for a term is the sum of the context vectors co-occurring with the term in the same time period

Temporal Random Indexing
TRI

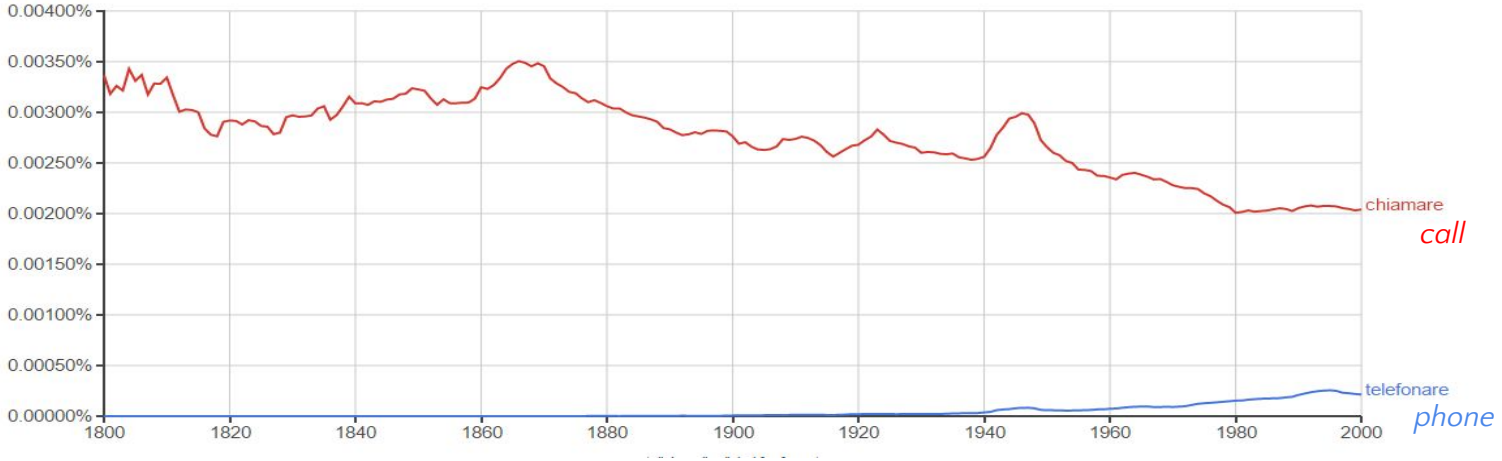
- Corpus with temporal information
  - split the corpus in several time periods
- Build a WordSpace for each time period using TRI
- Words in different WordSpaces are **comparable**!

P. Basile, A. Caputo, G. Semeraro. *Temporal random indexing: A system for analysing word meaning over time*. IJCoL vol. 1

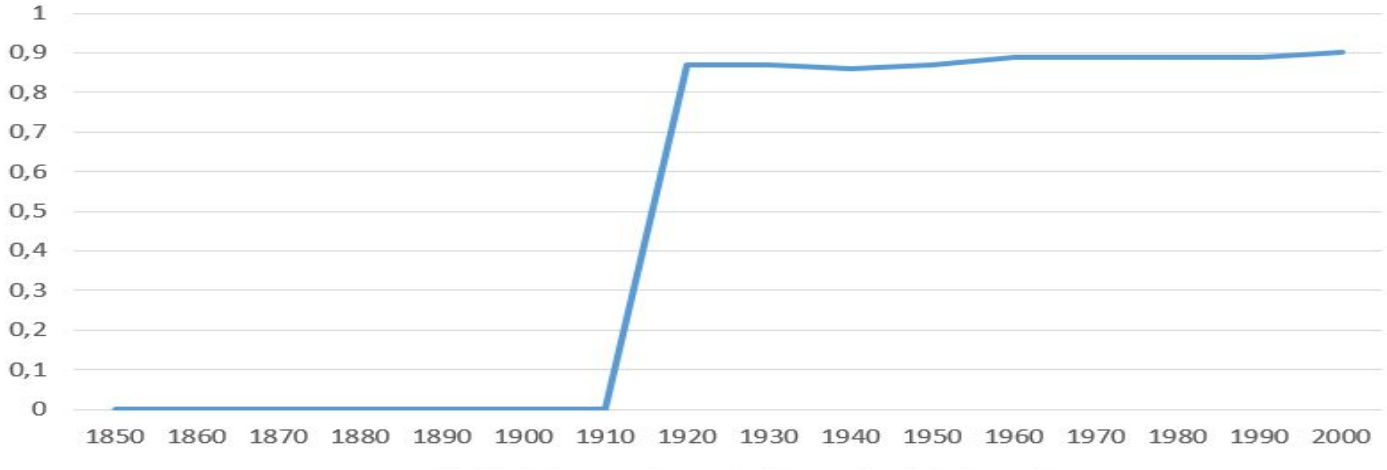
Similarity between words can change over time



Google N-gram



TRI



Methodology

TRI

Run TRI on a corpus split in time periods

Time Series

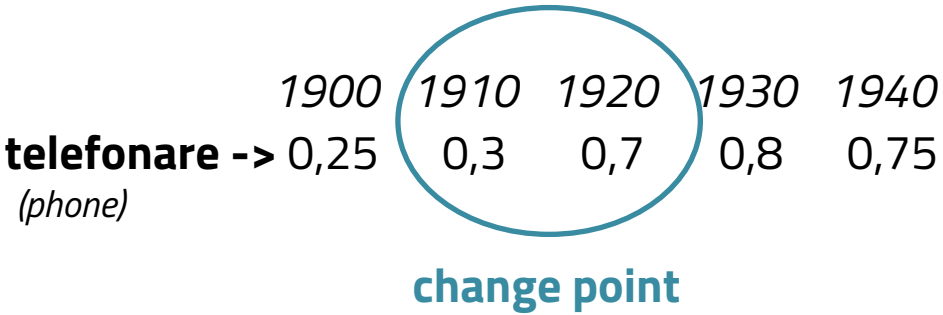
Provide a time series for each word

Change Point Detection

Detect significant changes in the time series

Change point
detection

- Track the word meaning change over time
- Build a time series by taking into account the semantic shift of each word
- Find significant change: **Mean shift model**



Social media

- Build TRI on **Twitter**
- About **500M tweets** (feb. 2012 – sep. 2015)
- Time interval = **1 month**
- Change point detection on the 1,000 most frequent hashtags

Social media
#bologna (august 2014)

CORRIERE DELLA SERA

UNA STRAGE CHE NON HA PRECEDENTI IN ITALIA. DUE IPOTESI: ATTENTATO O SCIAGURA

Scoppio apocalisse a Bologna 76 morti, 147 feriti in stazione

Pertini accorre - Nella notte prende corpo il sospetto di una bomba al plastico

L'OMBRA NERA SUL RIPOSO DEGLI ITALIANI

#bologna (august 2014)
vittime
bologna
memoria
strage
chiedere
festa
corsa
manifestazione

Social media
#euro (giugno 2015)

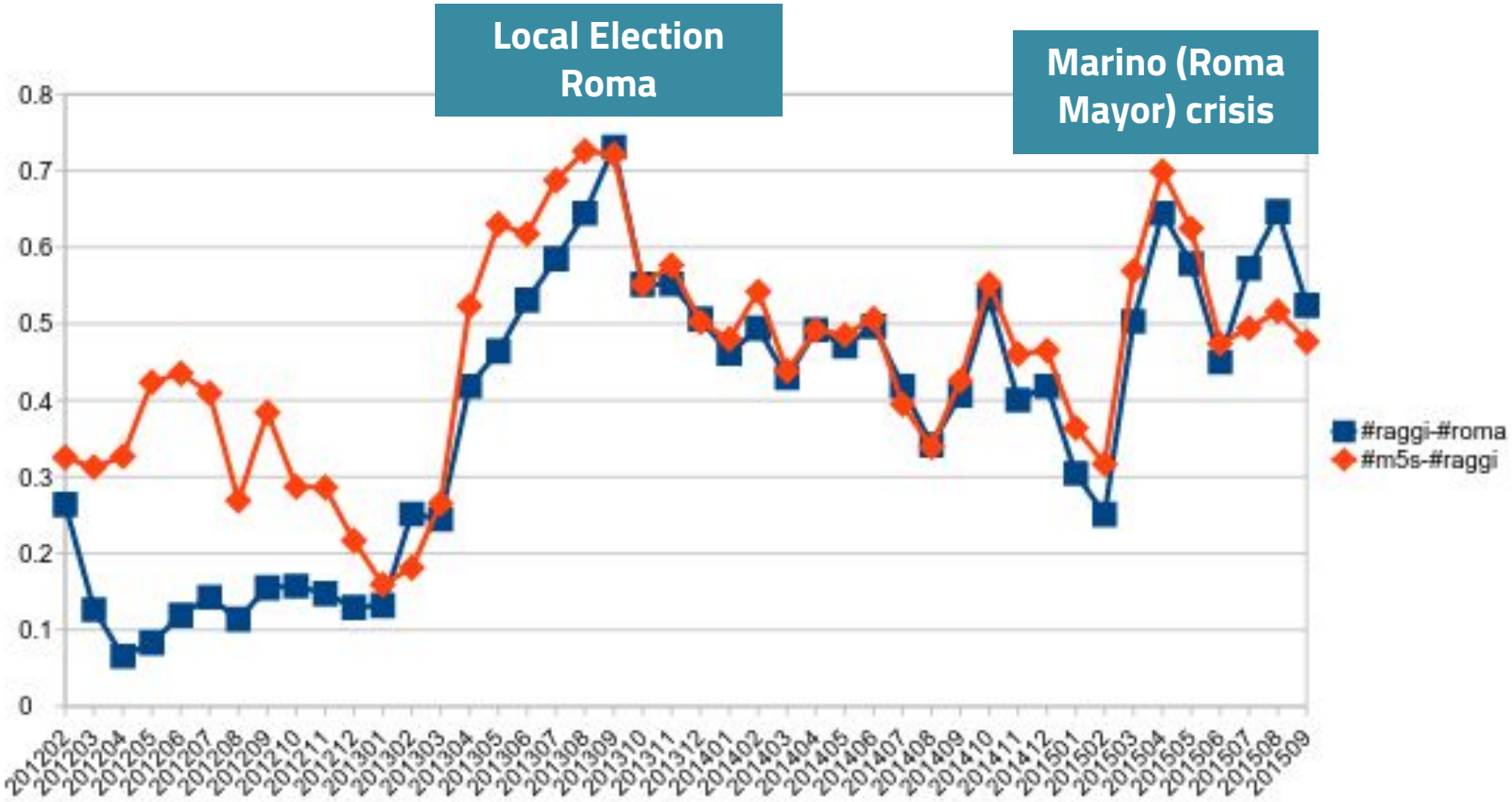
LA CRISI ELLENICA

Il parlamento greco approva il referendum. Tsipras chiede di votare «no»

–con un'article gallery di **Vittorio Da Rold** e un post di Econopoly | 28 giugno 2015

#euro (june 2015)
#europa
#grexit
ora
alternativa
acqua
invasione
uscita
economia

Social media



Ongoing work

Build a gold standard for the evaluation

http://dizionari.corriere.it/dizionario_italiano/

Evaluation

- Build word vectors using different approaches exploiting the Italian Google n-grams corpus
  - TRI, word embeddings alignment
- Evaluation using the gold standard
  - time period: 1900-2012

Thanks!!

Any questions?

You can find me at
pierpaolo.basile@uniba.it