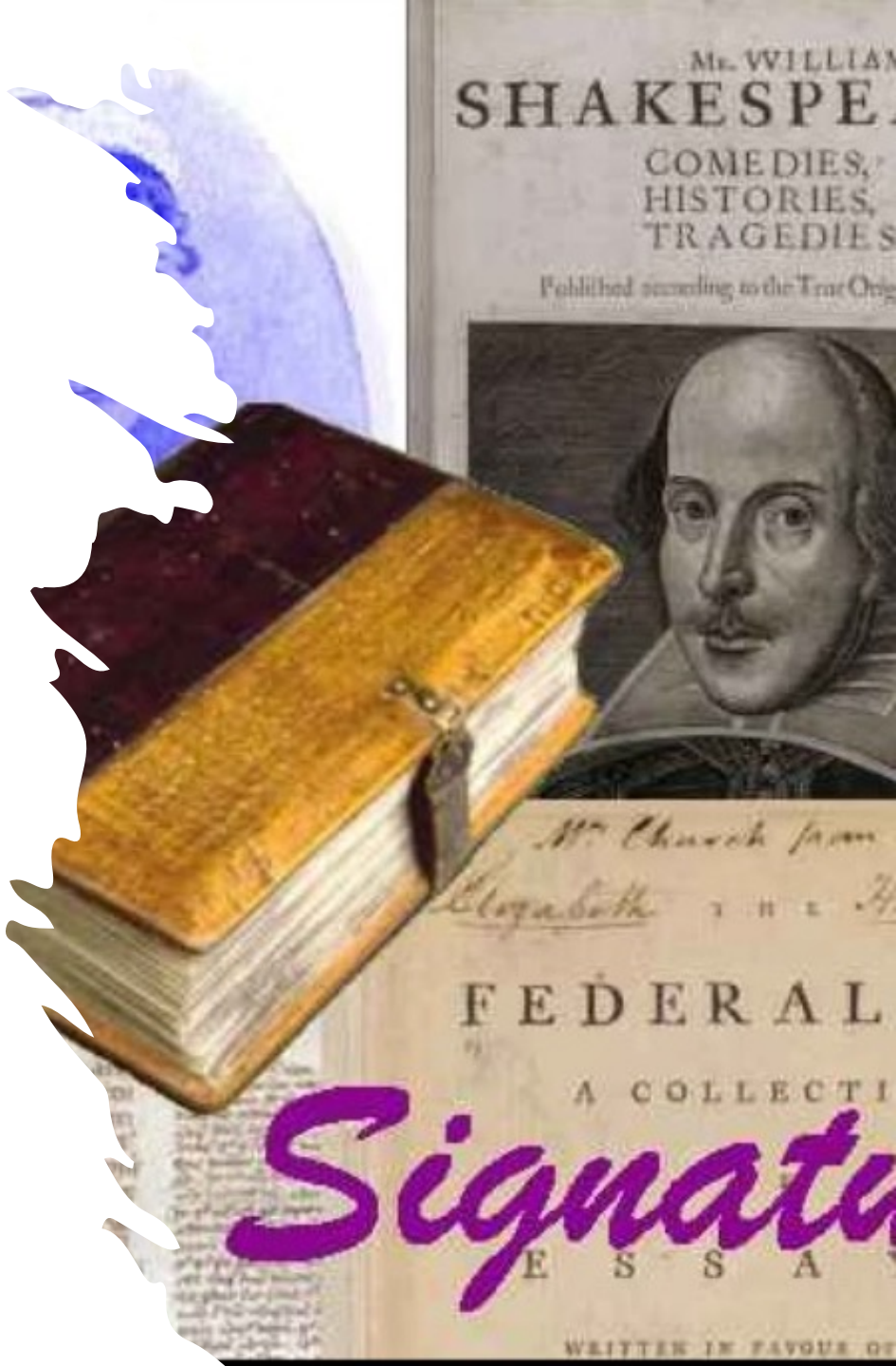
ESCUELA NACIONAL DE CIENCIAS FORENSES
UNAM

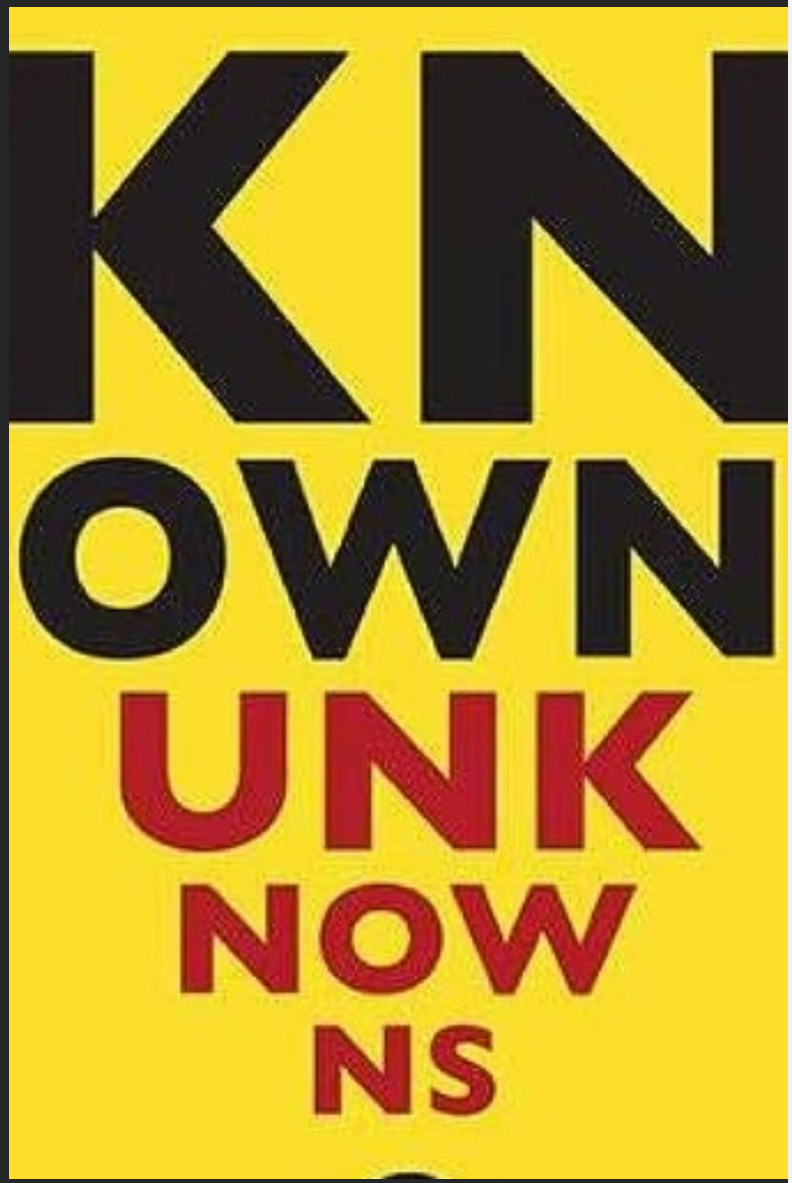
SAUTEE: Sistema Automático de Estudios Estilométricos

Fernanda López Escobedo

Authorship Attribution

Authorship attribution is a branch of Forensic Linguistics that employs linguistic knowledge in analyzing texts that constitute linguistic evidence. This is to determine whether or not their authorship can be attributed to a particular suspect in the investigation of a crime.





Known texts

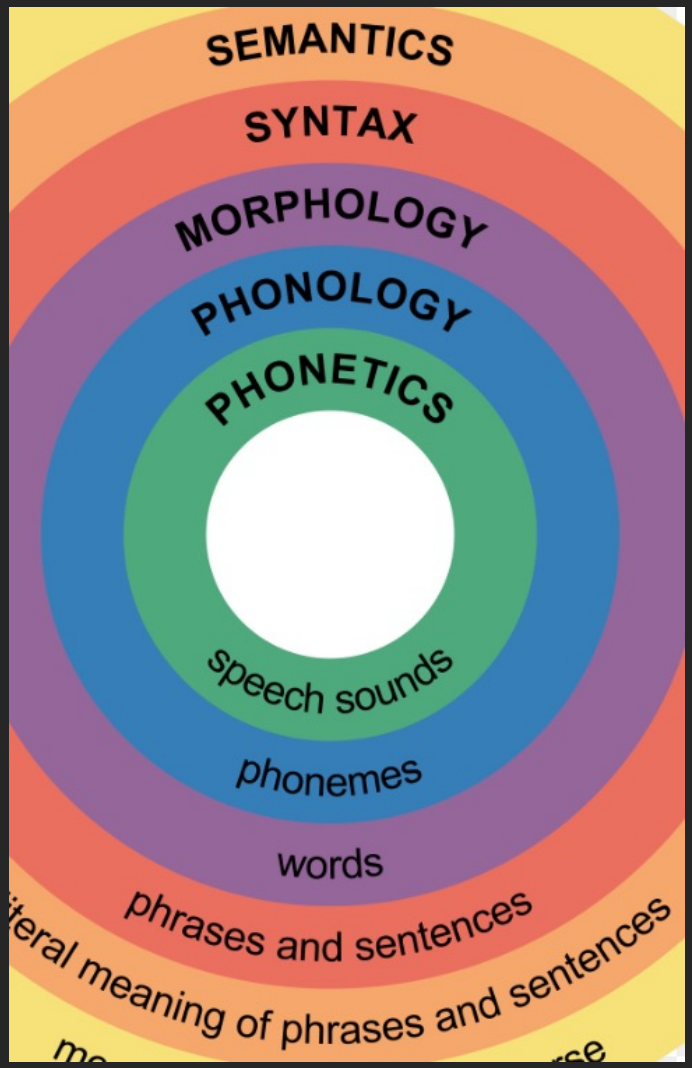
The list of possible suspects is established by the defendant or by the prosecution (Ministerio Público). The work of a forensic linguist involves guiding the defendant or the prosecution to find known samples by providing advice on various aspects such a:

- selecting the same textual genre or similar genres,
- ensuring the length and number of samples are appropriate,
- and other relevant considerations

Methodology

A comparison of texts whose authorship is questioned (questioned texts) with others whose authorship is known (known texts).

After analyzing the texts through a long process of linguistic analysis, the final step is comparison. The units of text are measured, calculated, and classified to identify the distinctive stylistic features of the author, also known as identifying marks.

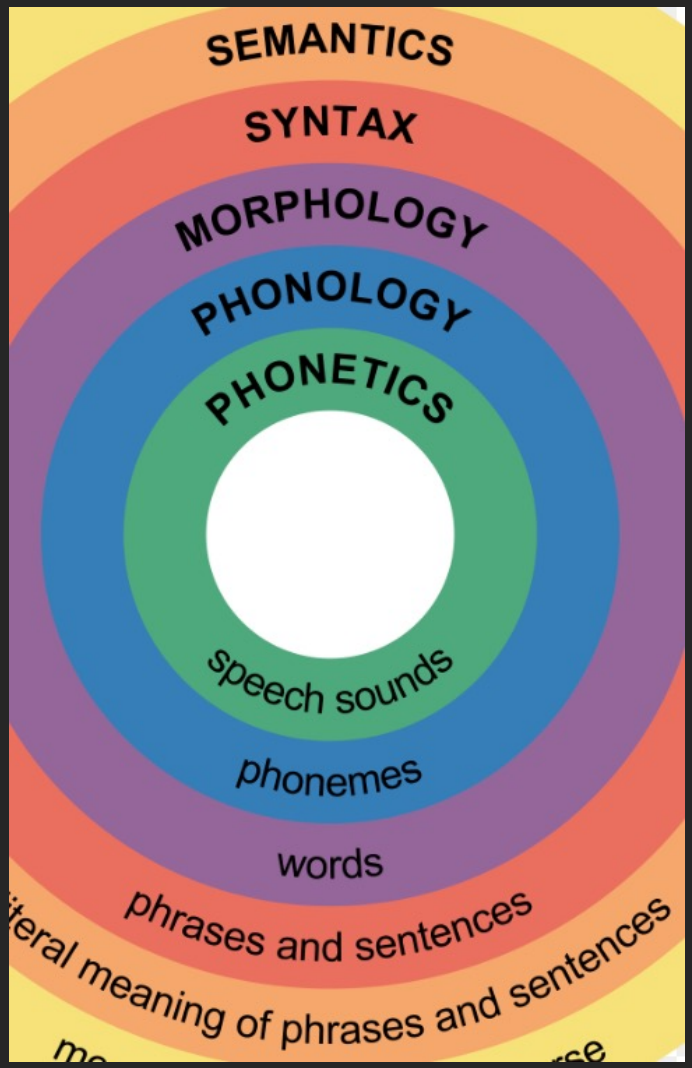


Identifying marks/features

It is difficult to define what is an identifying mark/feature since any linguistic element of an individual's writing could be a mark when it presents certain characteristics.

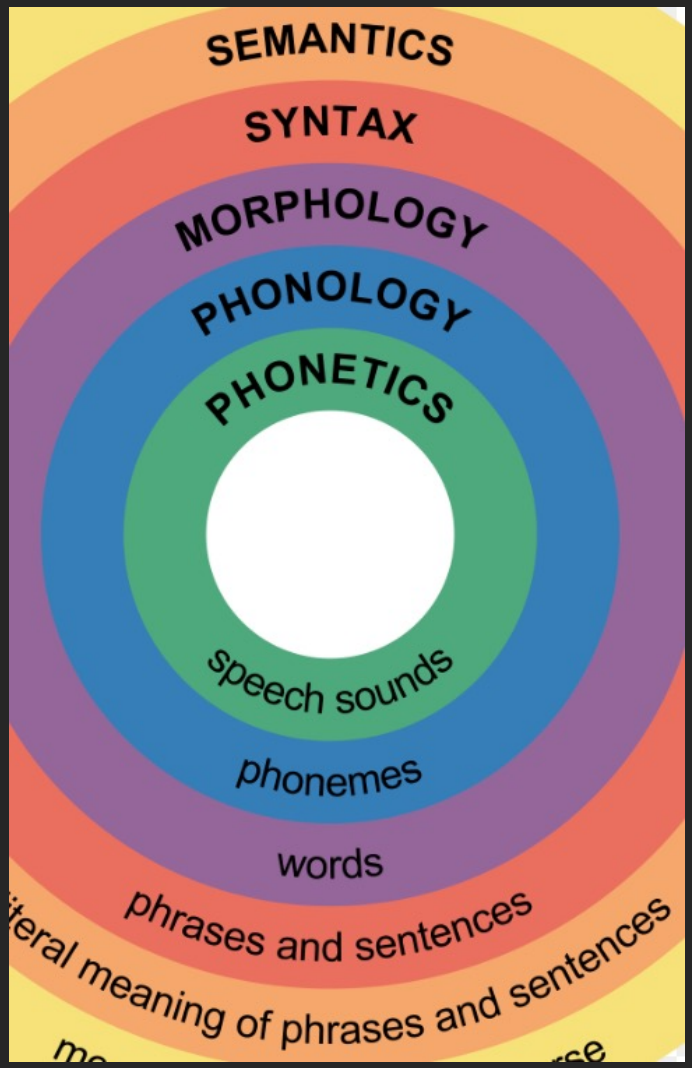
Bailey (1979) suggests general criteria for selecting identifying marks:

Prominence: refers to words that statistically stand out when comparing one subcorpus to another or a subcorpus to a complete corpus.



Identifying marks/features

Relative independence of conscious control: refers to the extent to which a language user is aware of and able to control their choices during the process of linguistic production. When a linguistic unit appears to be less manipulable by the individual's conscious control, it is more likely to be considered an identifying mark.



Identifying marks/features

Distribution: the number of occurrences of the mark in each text. In a forensic context, if the mark is present in all available documents, it is more likely to be an identifying mark. The mark is found in all the available documents, including both questioned and known texts, and its frequency allows for quantification.

Stylometric variables

- Character: is an indivisible text unit that includes graphemes (letters), digits, punctuation marks and blank spaces.

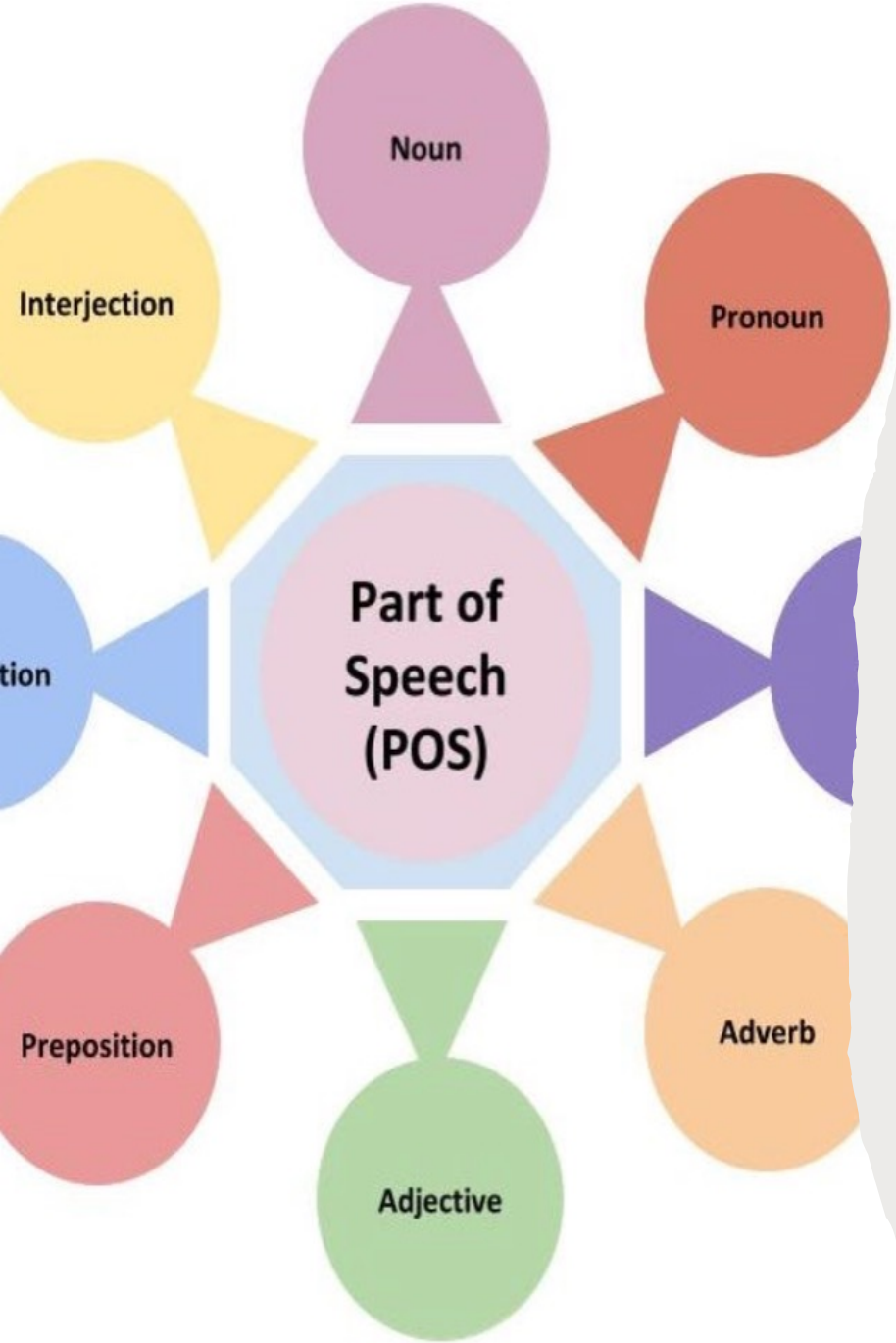
- Grapheme: a letter of the alphabet.

- Word: a string of graphemes and/or digits.

- Sentence: a continuous string of characters excluding question marks and exclamation marks

- n-word collocation: is a sequence of n words

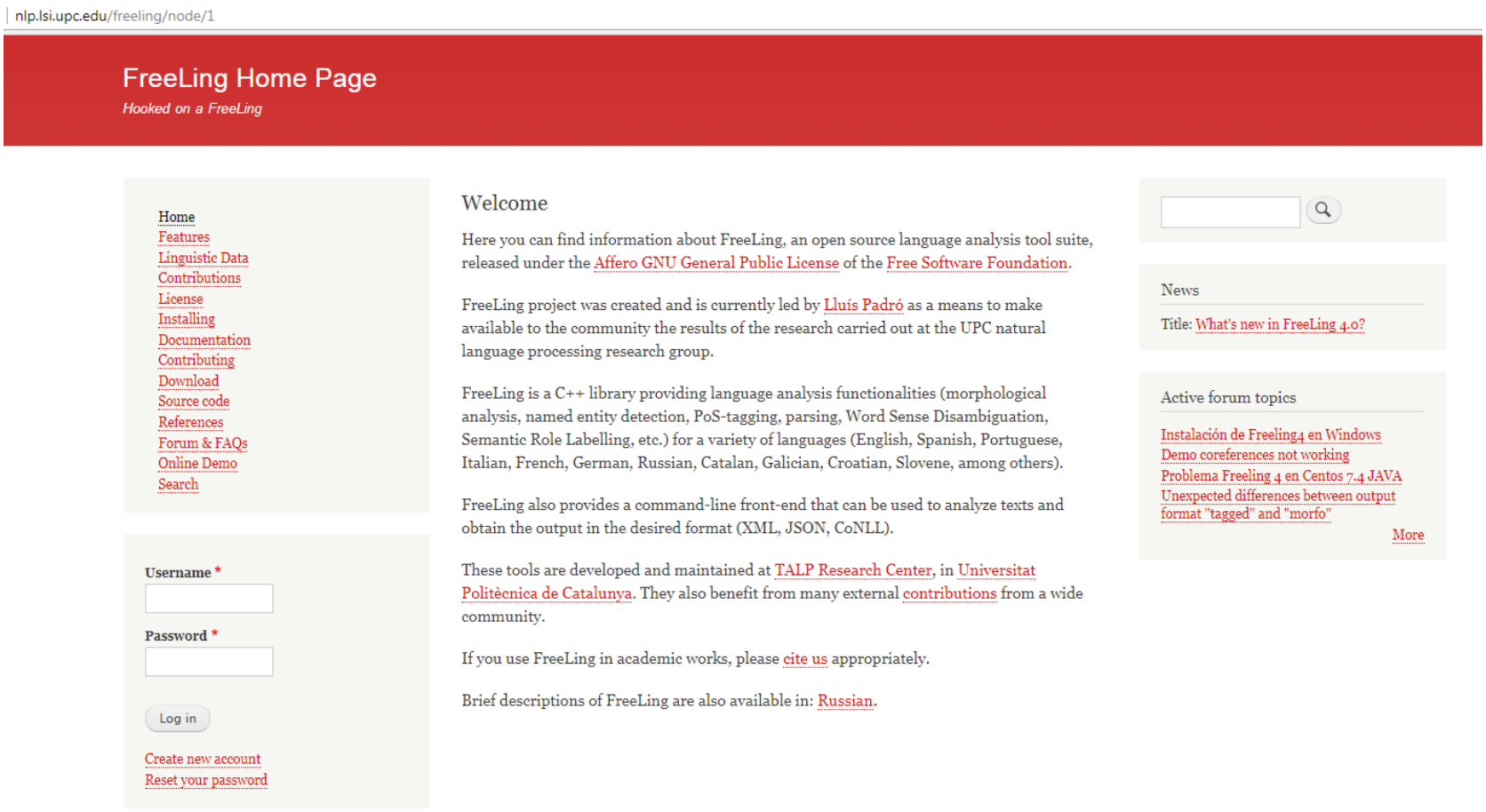
- n-gram: a sequence of n characters.



POS Tagging

Part-of-speech (POS) tagging is one of the most important addressed areas and main building block and application in the natural language processing discipline. POS tagging, also called grammatical tagging, is the automatic assignment of part-of-speech tags to words in a sentence. A POS is a grammatical classification that commonly includes verbs, adjectives, adverbs, nouns, etc. (Chiche and Yitagesu, 2022)

Freeling


nlp.lsi.upc.edu/freeling/node/1
FreeLing Home Page
Hooked on a FreeLing
Home
Features
Linguistic Data
Contributions
License
Installing
Documentation
Contributing
Download
Source code
References
Forum & FAQs
Online Demo
Search
Username *
Password *
Log in
Create new account
Reset your password
Welcome
Here you can find information about FreeLing, an open source language analysis tool suite, released under the Affero GNU General Public License of the Free Software Foundation.
FreeLing project was created and is currently led by Lluís Padró as a means to make available to the community the results of the research carried out at the UPC natural language processing research group.
FreeLing is a C++ library providing language analysis functionalities (morphological analysis, named entity detection, PoS-tagging, parsing, Word Sense Disambiguation, Semantic Role Labelling, etc.) for a variety of languages (English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian, Slovene, among others).
FreeLing also provides a command-line front-end that can be used to analyze texts and obtain the output in the desired format (XML, JSON, CoNLL).
These tools are developed and maintained at TALP Research Center, in Universitat Politècnica de Catalunya. They also benefit from many external contributions from a wide community.
If you use FreeLing in academic works, please cite us appropriately.
Brief descriptions of FreeLing are also available in: Russian.
News
Title: What's new in FreeLing 4.0?
Active forum topics
Instalación de Freeling4 en Windows
Demo coreferences not working
Problema Freeling 4 en Centos 7.4 JAVA
Unexpected differences between output format "tagged" and "morfo"
More

Example

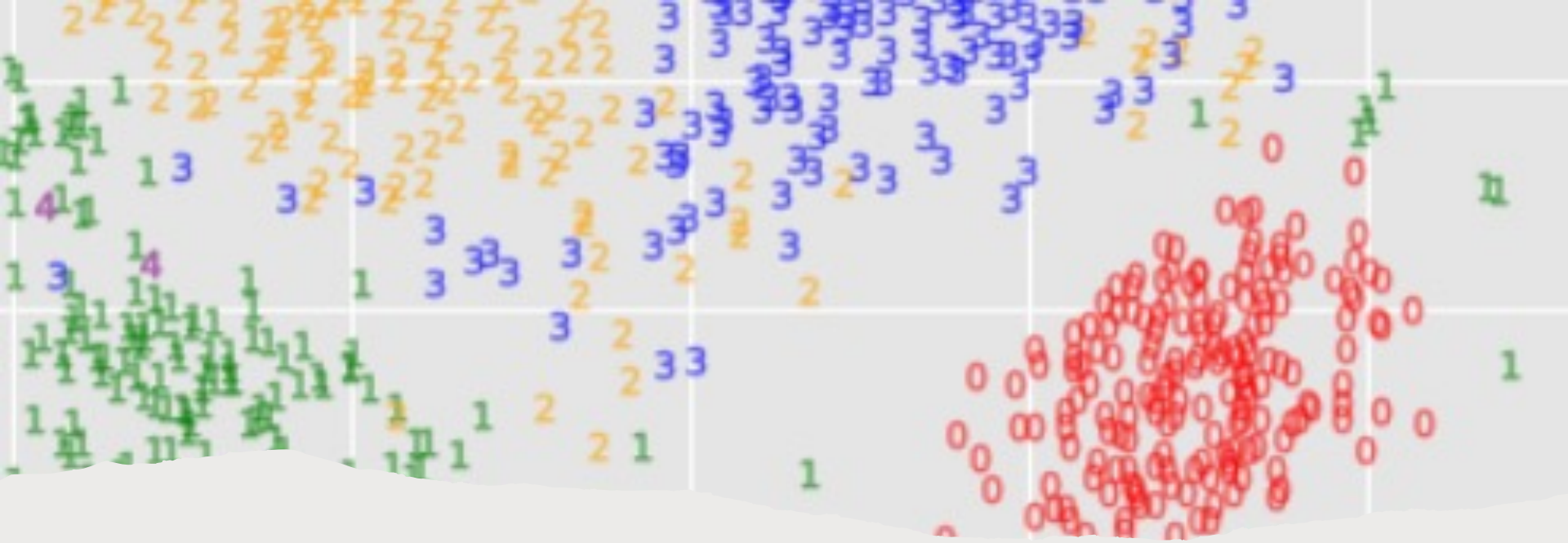
▼ **Sentences**

Sentence 1

The	house	is	red	and	the	car	is	green
the	house	be	red	and	the	car	be	green
DT	*NN*	*VBZ*	*JJ*	*CC*	*DT*	*NN*	*VBZ*	*JJ*

▶ CoNLL format

Stylometric variables in SAUTEE



Multidimensional Scaling

Multidimensional scaling refers to a set of techniques that aim to represent data through a configuration of points when certain information about proximities between objects is known.

A visual representation of distances or dissimilarities between sets of objects is created. In the context of authorship attribution, texts are the objects being analyzed, and the distances represent similarities between certain stylometric variables found in the texts. The closer two texts are in proximity, the more likely they are to have similar writing styles or stylometric features.

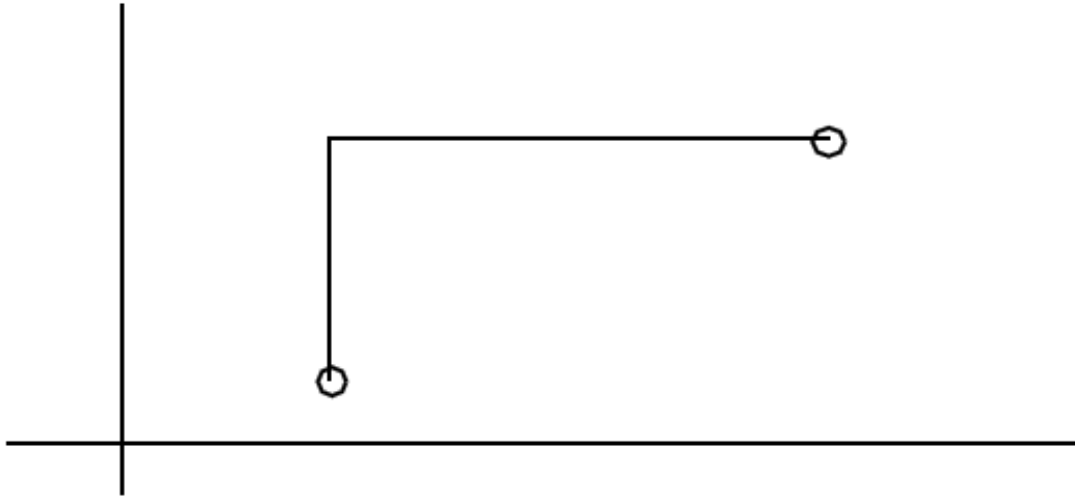
Distance measure

Euclidean distance: measures the shortest straight line between two points in a multi-dimensional space.

For a two-dimensional space, the Euclidean distance between two points X, Y is:

$$d(X,Y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Other distance measures



Manhattan distance: It is also called "City" distance or Average distance since the distance is measured as if we were walking the streets of a city as shown in the figure. Is the distance between two points measured along axes at right angles

SAUTEE

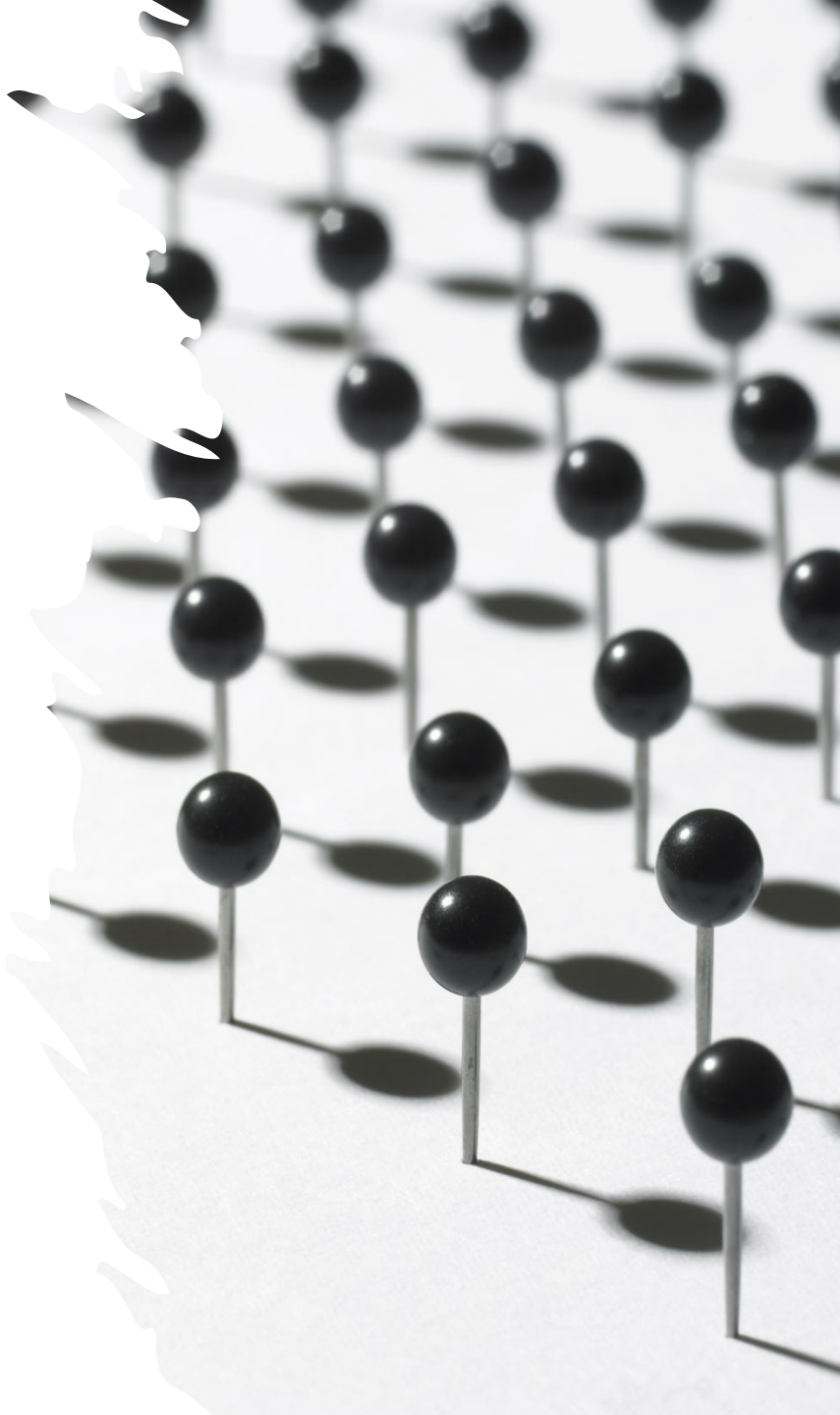

SAUTEE
Conectado como Fernanda López Escobec
Sistema Automático para Estudios Estilométricos
Inicio
Acerca de
Aplicación
Participantes
Publicaciones
1. Seleccionar documentos
2. Seleccionar marcadores estilométricos
3. Seleccionar método
4. Resultados
Método
Distancia euclideana
Delta de Burrows
Distancia Manhattan
Distancia Canberra
Descrip
La distancia euclidiana es igual a la raíz cuadrada de la suma del cuadrado de las diferencias de cada dimensión:
$\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$
¿Listo? Presiona el botón para obtener los resultados con las opciones seleccionadas
Ten en cuenta que esto puede tardar unos minutos dependiendo de la cantidad de documentos y de la cantidad de marcadores estilométricos seleccionados

Burrows's Delta

Burrows's Delta is a widely known distance measure in authorship attribution studies based on the Manhattan distance-assumes a Laplace distribution of data.

Is based, like many other measures and techniques in authorship attribution, on differences in the frequencies of the most frequent words in a group of texts.

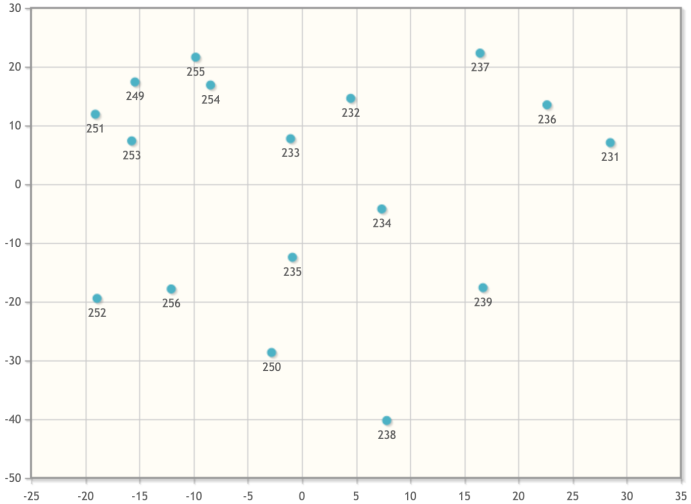
Burrows, John. 2002. Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing 17(3). 267–287



Results SAUTEE

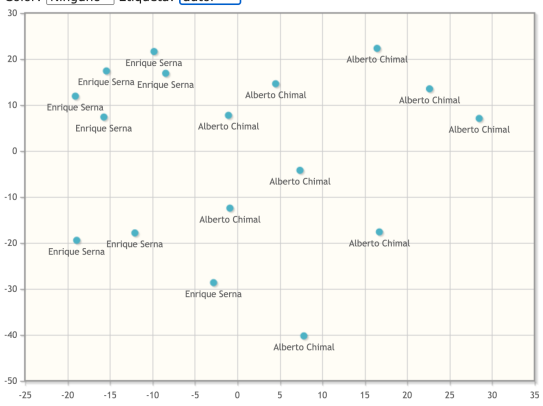
1. Seleccionar documentos | 2. Seleccionar marcadores estilométricos | 3. Seleccionar método | 4. Resultados

Color: Ninguno Etiqueta: id



231	D171.txt	Alberto Chimal	artículo	El carnaval de Ray Bradbury
232	D172.txt	Alberto Chimal	ensayo	La ciudad invisible
233	D173.txt	Alberto Chimal	ensayo	Manifiesto del cuento mutante
234	D174.txt	Alberto Chimal	cuento	El juego más antiguo
235	D175.txt	Alberto Chimal	ensayo	La idea de México
236	D176.txt	Alberto Chimal	artículo	JLB y la CF
237	D177.txt	Alberto Chimal	artículo	Lo fantástico en México: la vida en el margen
238	D178.txt	Alberto Chimal	cuento	Mogo

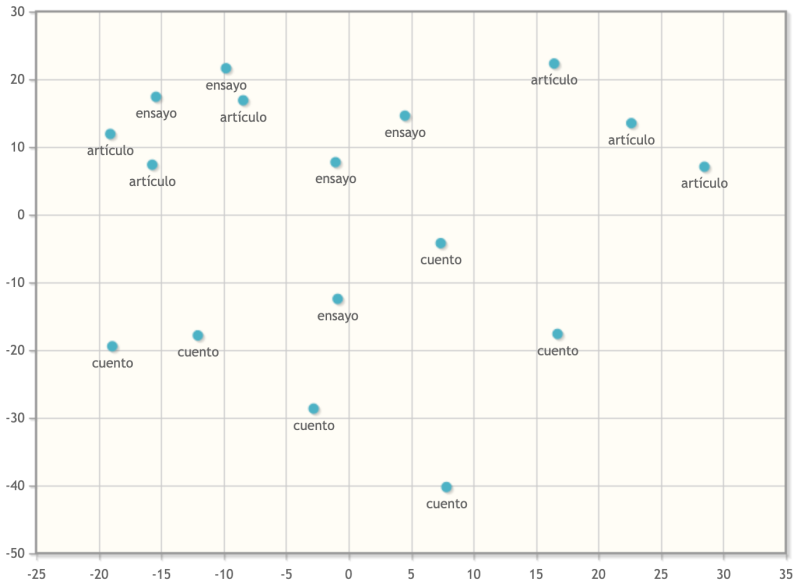
Color: Ninguno Etiqueta: autor



231	D171.txt	Alberto Chimal	artículo	El carnaval de Ray Bradbury
232	D172.txt	Alberto Chimal	ensayo	La ciudad invisible
233	D173.txt	Alberto Chimal	ensayo	Manifiesto del cuento mutante
234	D174.txt	Alberto Chimal	cuento	El juego más antiguo
235	D175.txt	Alberto Chimal	ensayo	La idea de México
236	D176.txt	Alberto Chimal	artículo	JLB y la CF
237	D177.txt	Alberto Chimal	artículo	Lo fantástico en México: la vida en el margen
238	D178.txt	Alberto Chimal	cuento	Mogo

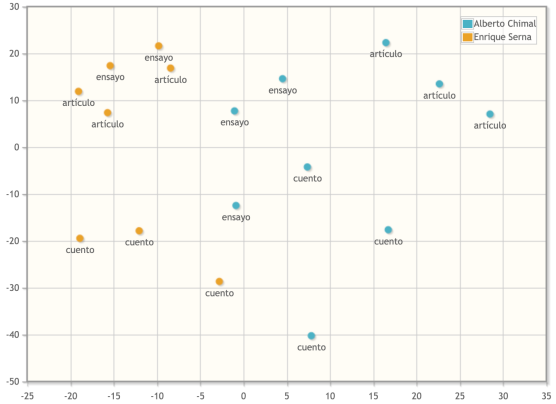
Results SAUTEE

Color: Ninguno Etiqueta: género



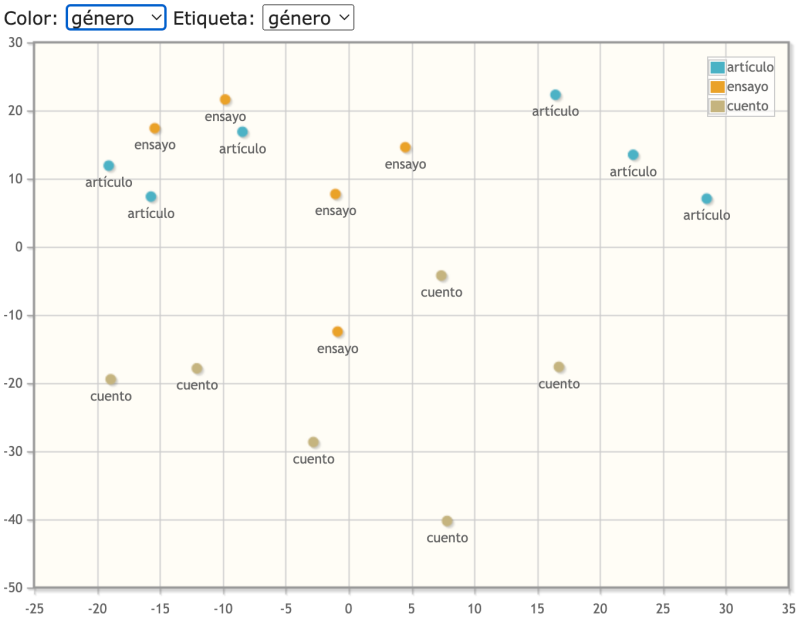
231	D171.txt	Alberto Chimal	artículo	El carnaval de Ray Bradbury
232	D172.txt	Alberto Chimal	ensayo	La ciudad invisible
233	D173.txt	Alberto Chimal	ensayo	Manifiesto del cuento mutante
234	D174.txt	Alberto Chimal	cuento	El juego más antiguo
235	D175.txt	Alberto Chimal	ensayo	La idea de México
236	D176.txt	Alberto Chimal	artículo	JLB y la CF
237	D177.txt	Alberto Chimal	artículo	Lo fantástico en México: la vida en el margen
238	D178.txt	Alberto Chimal	cuento	Mogo

Color: autor Etiqueta: género



231	D171.txt	Alberto Chimal	artículo	El carnaval de Ray Bradbury
232	D172.txt	Alberto Chimal	ensayo	La ciudad invisible
233	D173.txt	Alberto Chimal	ensayo	Manifiesto del cuento mutante
234	D174.txt	Alberto Chimal	cuento	El juego más antiguo
235	D175.txt	Alberto Chimal	ensayo	La idea de México
236	D176.txt	Alberto Chimal	artículo	JLB y la CF
237	D177.txt	Alberto Chimal	artículo	Lo fantástico en México: la vida en el margen
238	D178.txt	Alberto Chimal	cuento	Mogo

Results SAUTEE



231	D171.txt	Alberto Chimal	artículo	El carnaval de Ray Bradbury
232	D172.txt	Alberto Chimal	ensayo	La ciudad invisible
233	D173.txt	Alberto Chimal	ensayo	Manifiesto del cuento mutante
234	D174.txt	Alberto Chimal	cuento	El juego más antiguo
235	D175.txt	Alberto Chimal	ensayo	La idea de México
236	D176.txt	Alberto Chimal	artículo	JLB y la CF
237	D177.txt	Alberto Chimal	artículo	Lo fantástico en México: la vida en el margen
238	D178.txt	Alberto Chimal	cuento	Mogo

Catálogo de aplicaciones

SAUTEE

Sistema Automático para Estudios Estilométricos

Let's try SAUTEE!

http://www.corpus.unam.mx/geco/