

An application of linguistic methods and resources

NooJ for the Digital Humanities

Digital Humanities

- Researchers in the Social Sciences need tools to analyze their corpora.
- **Alceste, Lexico3, Hyperbase, Iramuteq, Sketch Engine, Trameur, TXM**
- To find what is interesting they detect abnormal frequencies...

Digital Humanities

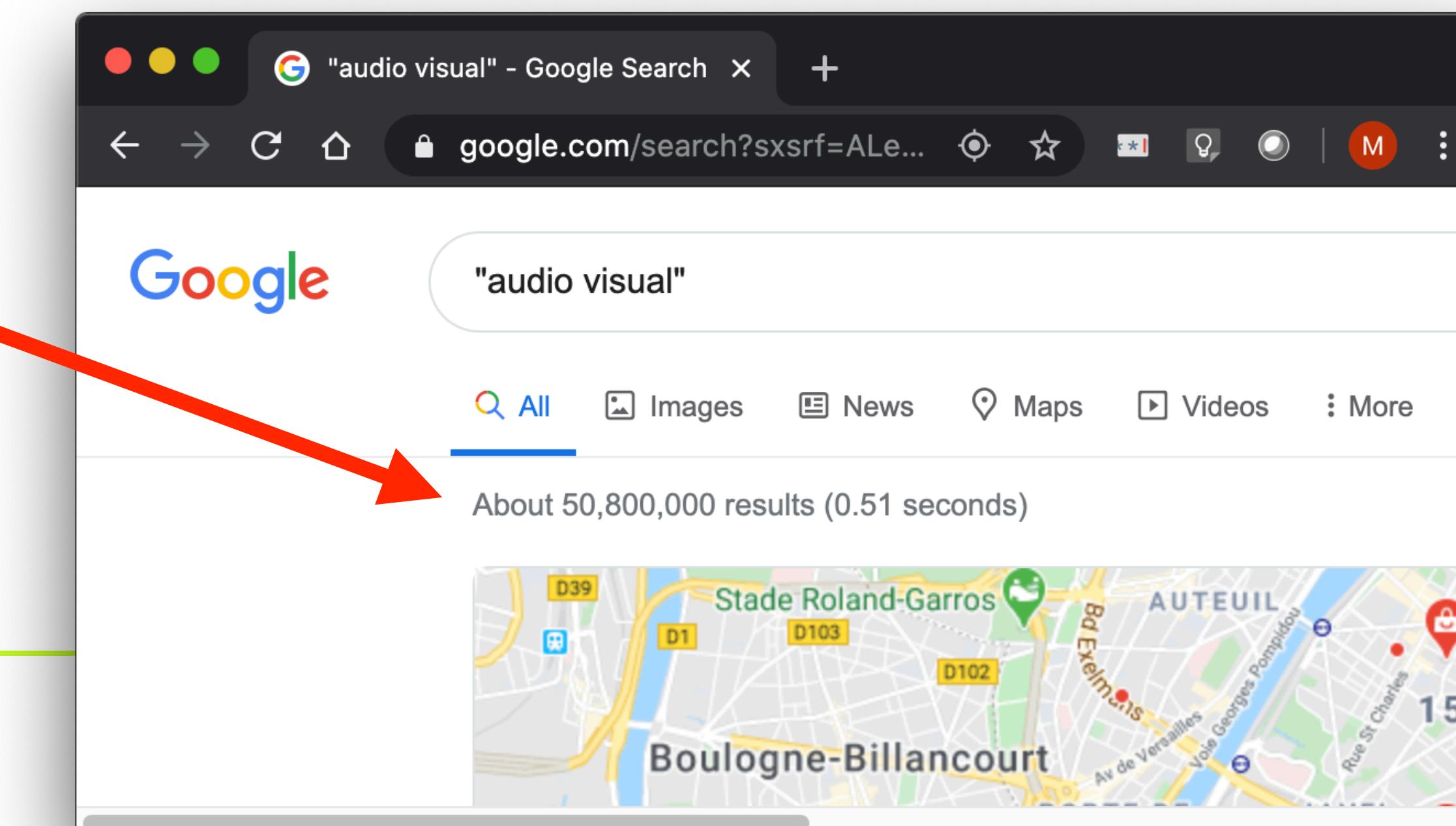
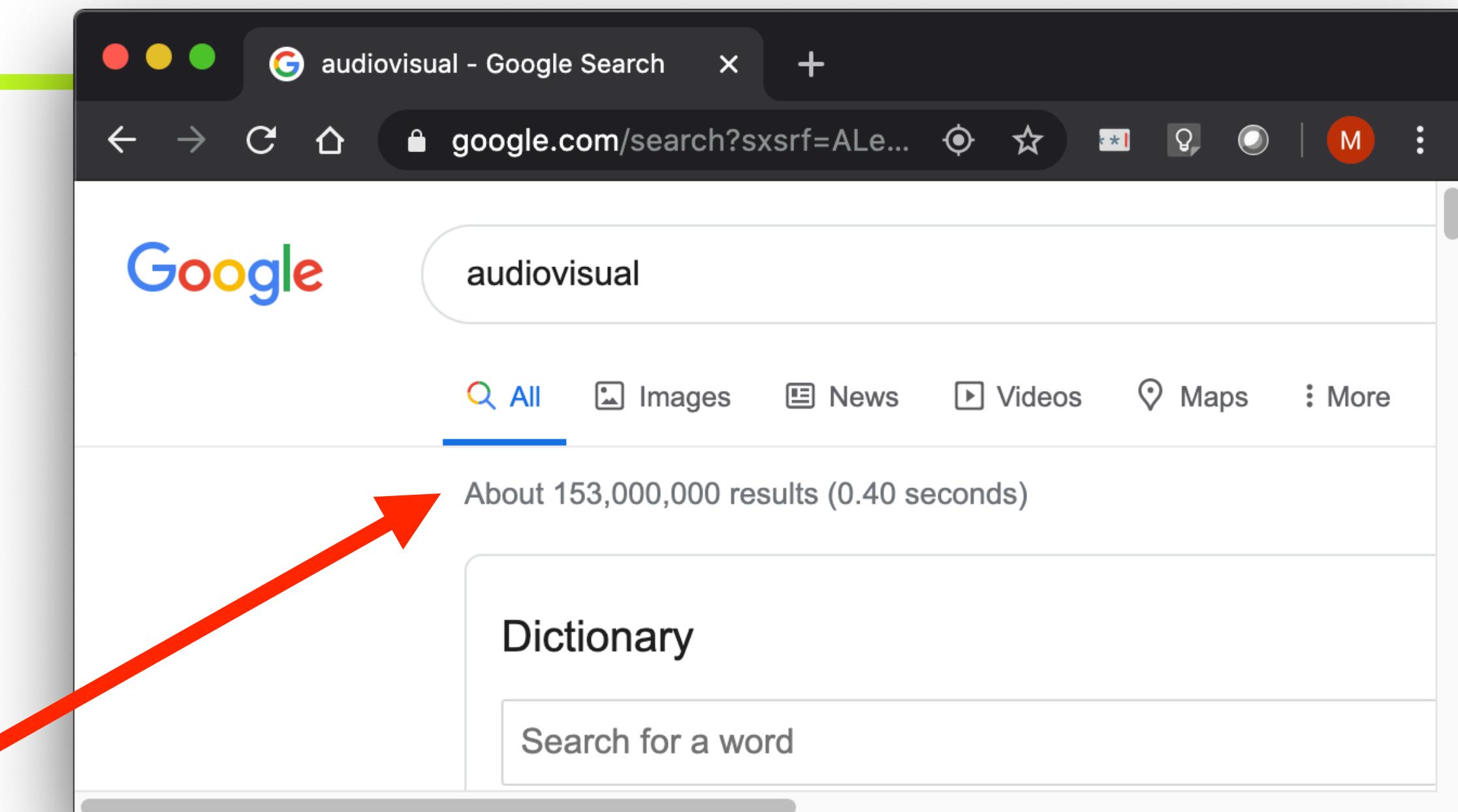
- Researchers in the Social Sciences need tools to analyze their corpora.
- **Alceste, Lexico3, Hyperbase, Iramuteq, Sketch Engine, Trameur, TXM**
- To find what is interesting they detect abnormal frequencies... **of word forms.**

1. Orthography

- Lots of spelling hesitations

- “audiovisual” gets 153 M hits

- “audio visual” gets 51 M hits



1. Orthography: linguistic solution

audiovisual, audio visual, audio-visual → **audio_visual, ADJECTIVE**

Middle ages, Middle-Ages → **middle=ages, NOUN**
middle ages, middle-ages
middle Ages, middle-Ages

csar, czar, tsar, tzar → **csar, tsar, NOUN**
czar, tsar, NOUN
tzar, tsar, NOUN

2. No Lemmatization?

- e.g. **Lexico3**
- “liberté” is used in certain texts as a right-wing concept, e.g. “Entrepreneurial freedom”
- “libertés” is used in certain texts as a left-wing concept, e.g. “Human rights”

2. No Lemmatization?

- However, over 99% nouns do not carry different meanings in the Singular and in the Plural:

One demonstration → *Two demonstrations*

One idea → *Two ideas*

One table → *Two tables*

One cow → *Two cows*

One baker → *Two bakers*

One minister → *Two ministers*

...

3. Tagging

- Most NLP systems do lemmatize wordforms, e.g. **Sketch Engine**
- They use taggers, e.g. **TreeTagger** or **GATE**

The screenshot shows the Sketch Engine Concordance interface. At the top, there's a toolbar with various icons and a search bar containing "French Web 2017 (frT...)" with a magnifying glass icon. Below the toolbar, the word "CONCORDANCE" is displayed in large letters. A sidebar on the left contains several icons: a speech mark, a grid, a list, a circle, a dot, a list with a dot, a grid with dots, a downward arrow, and a "NE S" logo at the bottom.

The main area shows two search results for the lemma "aimer". The first result is labeled "doc#15462" and contains the following French text:

<s> C'est d'ailleurs au cours d'un voyage à Versailles qu'il rencontre Aimée. Arranet, dont il tombe éperdument amoureux dès le premier regard. </s>

This text is annotated with Part-of-Speech (POS) tags and lemmas. For example, "C'est" is tagged as PPSCNN0/ce, "Arranet" is highlighted in red, and "Aimée" is also highlighted in red. The second result is similar, showing another instance of the sentence structure.

3. Tagging

- *cours* not a plural feminine form of the noun *cour*
- *au* not a preposition
- *Aimée* not a past participle form of the verb *aimer*
- *rencontre, tombe, doute* and *pense* are not conjugated in the first person
- *Versailles* not a plural feminine noun
- *A* not a conjugated form of the verb *avoir*
- *qu'* not a conjunction.

The screenshot shows the Sketch Engine Concordance interface. The top bar displays the URL app.sketchengine.eu/#concordance?corpname=preloaded%2Ffrtenten17_f12&tab=advanced&queryselector.... The main title is "CONCORDANCE" and the search term is "French Web 2017 (frT...)".

Search parameters: lemme **aimer** 1871115 > filtre Arranet 1, 2 23 inférieur à 0.01

Results:

1 doc#15462 <s> C' est d' ailleurs au cours d' un voyage à Versailles qu' il rencontre Aimée
d' Arranet , dont il tombe éperdument amoureux dès le premier regard . </s>

2 doc#15462 <s> II doute aussi de son amour pour Aimée
d' Arranet et pense de plus en plus à la Satin : " A l' amour , feu couvant , qu' il continuait à lui porter , s' ajoutaient une estime , un respect et une

The interface includes a sidebar with various icons and a bottom navigation bar.

3. Tagging

- *cours* not a plural feminine form of the noun *cour*
- *au* not a preposition
- *Aimée* not a past participle form of the verb *aimer*
- *rencontre, tombe, doute* and *pense* are not conjugated in the first person
- *Versailles* not a plural feminine noun
- *A* not a conjugated form of the verb *avoir*
- *qu'* not a conjunction.
- multiword units not tagged → their component's tag irrelevant

The screenshot shows the Sketch Engine Concordance interface. The top bar displays the URL app.sketchengine.eu/#concordance?corpname=preloaded%2Ffrtenten17_f12&tab=advanced&queryselector.... The main title is "CONCORDANCE" and the search term is "French Web 2017 (frT...)".

Search filters applied are "lemme aimer 1871115" and "filtre Arranet 1, 2 23 inférieur à 0.01".

The interface includes various buttons and icons for navigating through the results, such as search, download, and zoom.

Two search results are shown:

- Result 1:** doc#15462. The sentence is: "C'est d'ailleurs au cours d'un voyage à Versailles qu'il rencontre Aimée." The word "Aimée" is highlighted in red. Part-of-speech tags and dependencies are shown below the tokens.
- Result 2:** doc#15462. The sentence is: "Il doute aussi de son amour pour Aimée." The word "Aimée" is highlighted in red. Part-of-speech tags and dependencies are shown below the tokens.

The left sidebar contains a vertical stack of icons representing different corpus types and features.

3. Tagging: Upenn TreeBank

A reference corpus

Battle-tested/SingularProperName Japanese/SingularProperName
industrial/Adjective managers/PluralNoun here/Adverb
always/Adverb buck/Verb up/Preposition nervous/Adjective
newcomers/PluralNoun with/Preposition the/Determiner
tale/SingularNoun of/Preposition the/Determiner first/Adjective
of/Preposition their/PossessivePronoun countrymen/NomPluriel to/to
visit/Verb Mexico/NomPropreSingulier/, a/Determiner
boatload/SingularNoun of/Preposition samurai/PluralNoun
warriors/PluralNoun blown/PastParticiple ashore/Adverb 375/Number
years/PluralNoun ago/Adverb ./.
From/Preposition the/Determiner beginning/SingularNoun ,/
it/PersonalPronoun took/PastParticiple a/Determiner
man/SingularNoun with/Preposition extraordinary/Adjective
qualities/PluralNoun to/TO-succeed/Verb in/Preposition
Mexico/SingularProperName, " /" says/VerbPresent3rdSingular
Kimihide/SingularProperName Takimura/ SingularProperName ,/
president/SingularNoun of/Preposition Mitsui/PluralNoun
group/SingularNoun 's/Possessive Kensetsu/SingularProperName
Engineering/SingularProperName-Inc./SingularProperName
unit/SingularNoun ./.

3. Tagging: Upenn TreeBank

A reference corpus?

Battle-tested/SingularProperName Japanese/SingularProperName
industrial/Adjective managers/PluralNoun here/Adverb
always/Adverb buck/Verb up/Preposition nervous/Adjective
newcomers/PluralNoun with/Preposition the/Determiner
tale/SingularNoun of/Preposition the/Determiner first/Adjective
of/Preposition their/PossessivePronoun countrymen/NomPluriel
to/to visit/Verb Mexico/NomPropreSingulier,/ , a/Determiner
boatload/SingularNoun of/Preposition samurai/PluralNoun
warriors/PluralNoun blown/PastParticiple ashore/Adverb
375/Number years/PluralNoun ago/Adverb ./.
From/Preposition the/Determiner beginning/SingularNoun ,/
it/PersonalPronoun took/PastParticiple a/Determiner
man/SingularNoun with/Preposition extraordinary/Adjective
qualities/PluralNoun to/TO succeed/Verb in/Preposition
Mexico/SingularProperName ,/ , "/ says/VerbPresent3rdSingular
Kimihide/SingularProperName Takimura/ SingularProperName ,/
president/SingularNoun of/Preposition Mitsui/PluralNoun
group/SingularNoun 's/Possessive Kensetsu/SingularProperName
Engineering/SingularProperName Inc./SingularProperName
unit/SingularNoun ./.

3. Tagging: OANC (tagged with Annie/GATE)

- 20% of the tags are incorrect, e.g.:
abbreviate, abduct, abhor, abhors, etc. (~~Nouns~~),
about, agonized, bible, cactus, California, etc. (~~Adjectives~~)
expenditures, Japanese, many, initiatives, wimp, etc. (~~Verbs~~)
anomaly, back, because, by, of, out, particular, upon, etc. (~~Adverbs~~)...
- Systematic mistakes with agglutinated forms, e.g.:
audienceless, autodialed, barklike, etc. (~~Nouns~~)
- A large number of words in uppercase, e.g.:
Abacuses, ABATEMENT, Abnormal, Abolished, ALMOST, etc. (~~Proper names~~)
- incorrect disambiguation rules, etc. see (Silberztein 2018)

3. Tagging: the linguistic solution

(Silberztein 2016) shows how to use:

- dictionaries and morphological grammars to recognize all linguistic units
- local syntactic grammars to remove most false positives
- syntactic & distributional properties to remove most semantic ambiguities

(Silberztein 2018) shows how to correct the OANC automatically by using:

- dictionaries (no impossible tag would be inserted)
- morphological grammars (will recognize agglutinated words)
- local syntactic grammars (will avoid most systematic mistakes)

4. Derivation

- Processing only Inflection is not enough:

*... protesters **demonstrate** against the city's new drastic plans ...*

*... by **demonstrators** calling for the loosening of COVID-19 social-distancing ...*

*... podcasts related to **demonstrations** on France 24 ...*

4. Derivation

Word forms computed using purely orthographic rules.

“amour” [love]

→ *amours, amouracher, amoureusement...*

→ *amphigourique, amphithéâtre, ample, ampleur, amplifier, amusa, amusaient...*

→ No form of the verb *aimer* [to love]?

MAX SILBERZTEIN

C:\HYPERBAS\Rogon.tbk

Formes Lemmes Chercher un code Codes Chercher une séquence Structures

111 amour
1 amour»
1 amouracher
26 amoureuse
8 amoureusement
2 amoureuses
55 amoureux
40
1 amphigourique
2 amphithéâtre
1 ample
5 ampleur
1 amplifia
13 amusa
5 amusaient
25 amusait
16 amusant
2 amusante
1 amusantes

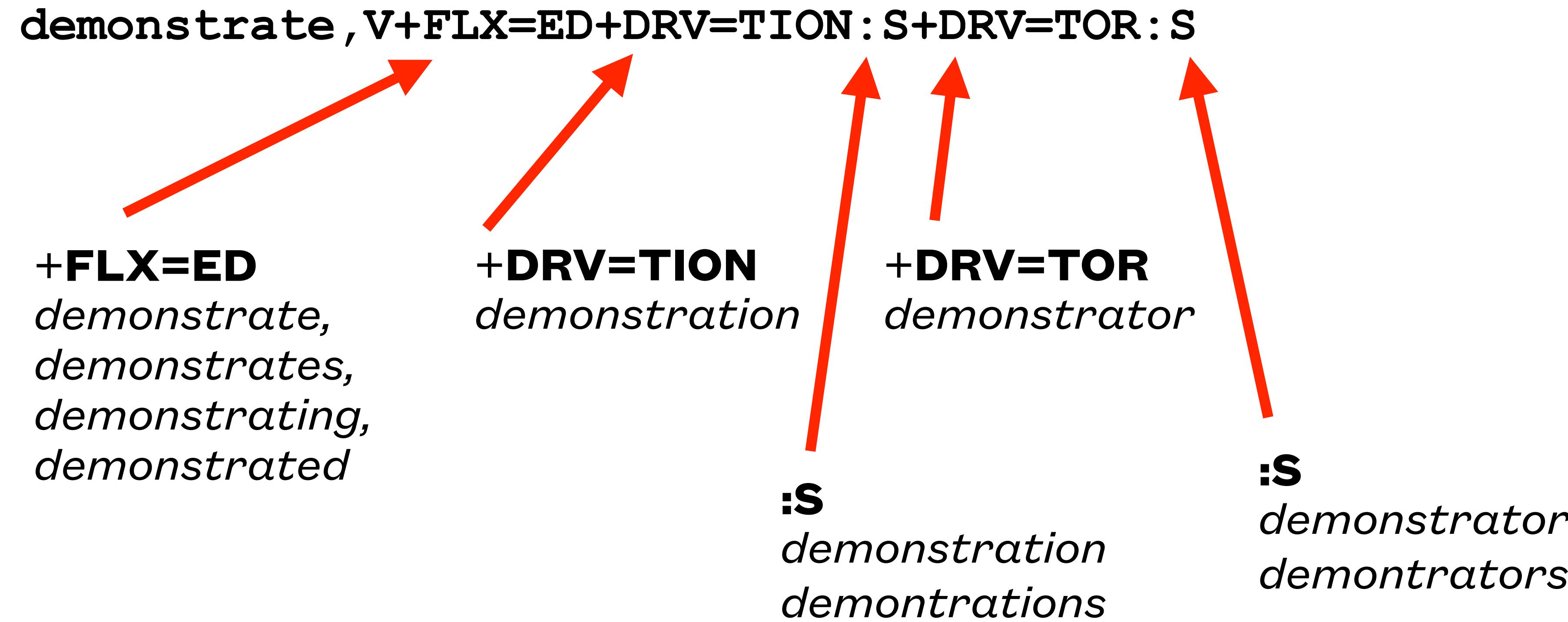
N° 1 TEX1 36
N° 2 TEX2 54
N° 3 TEX3 10
N° 4 TEX4 11
TOUS LES TEXTES

amour fréquence totale: 111
CLIQUEZ SUR UN TEXTE (ou sur TOUS) pour y repérer les contextes du mot "amour"
Cliquer AILLEURS dans cette fenêtre pour l'effacer

1 cnappvspan
1 cnappvsppnsvdn
1 cnarpvr
1 cnasdpvv
1 cnasdvnsp
1 cnasdvnra
1 cnasn
1 cnasnvsrrvvdnssnsdn
1 cnasnvsdnn
1 cnavdn
1 cnavr
1 cnavrdan
1 cnavrvn
1 cnavsn
1 cnavsp
1 cnavspn
1 cnavsvnsnpan
2 cnav
1 cnavvdsndn
1 cnca
1 cncappvsna
1 cncav
1 cncdnv
1 cncdnvsnappv
1 cncnpvspan
1 cncnv
1 cncnvdn
1 cncnvdnasdnsa
1 cncnvnsppvdsndn
1 cncnvra
1 cndrv
1 cnm
20 cnn
1 cnncnvpvn
2 cnpv
1 cnpvspan
1 cnpvvdnsdn
1 cnnrvrsvpspn
1 cnnrvrdmn
8 cnnv
1 cnnvednvsn
2 cnnvdn
1 cnydnnnnnnnrsydcvcdna

Représentation graphique de la distribution du mot choisi

4. Derivation: the linguistic solution



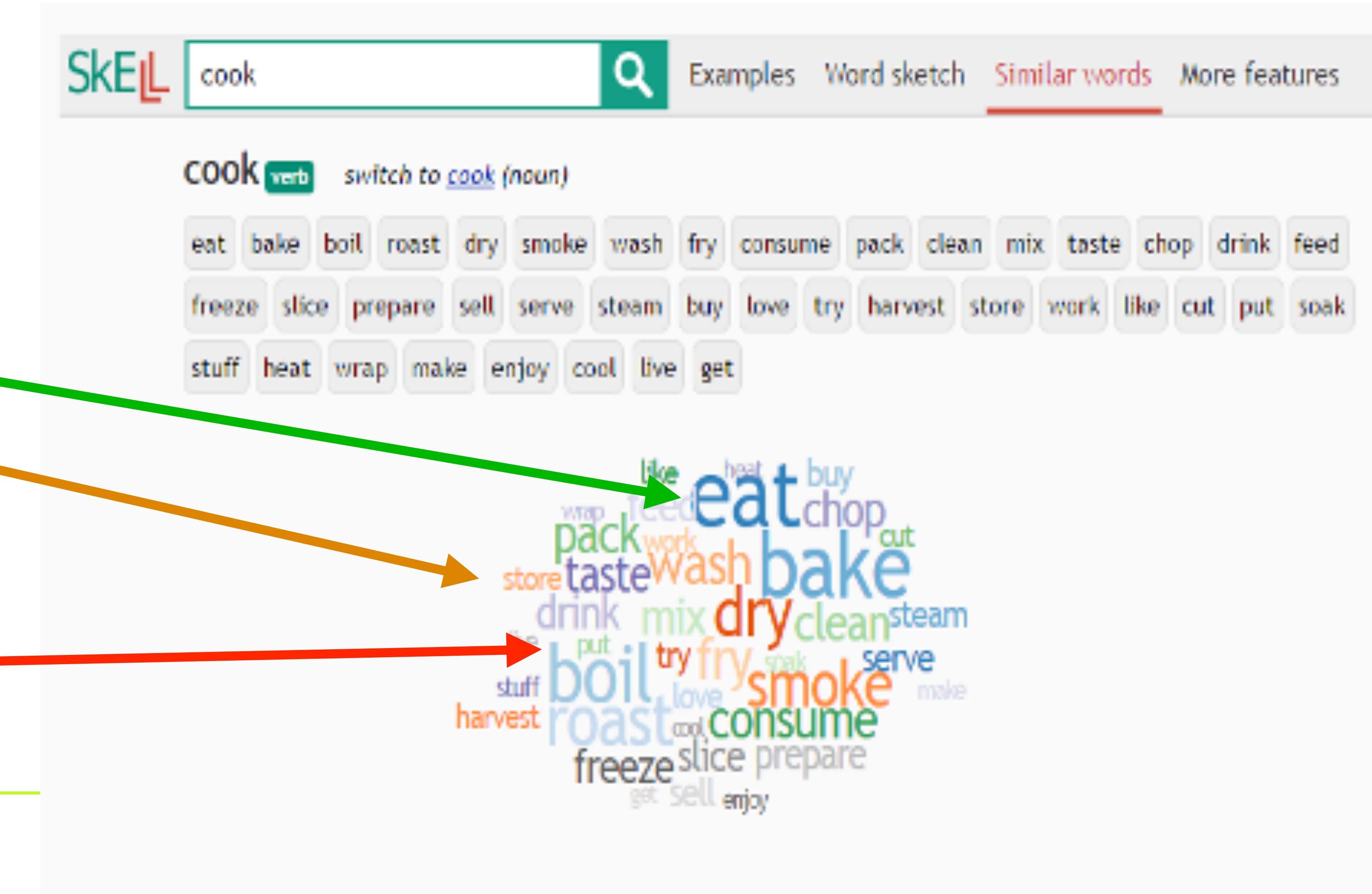
5. Lexical Fields

- These NLP applications retrieve related terms by processing word forms.

- “eat” and “bake”

- “buy”, “store”, “live”,
“clean”

- Grammatical words,
e.g. “put”, “try” and “get”



4. Lexical Fields: Linguistic solution

Verbs = <assassinate> | <decease> | <die> | <eradicate> |
<exterminate> | <kill> | <obliterate> | <perish> | <suicide> |
<strangle> ;

Nouns = <corpse> | <death> | <euthanasia> | <extermination> |
<fatality> | funerals | grim reaper | <guillotine> | <massacre>
| <mortality> | <murder> | <slaughter> | <tomb> | <widower> ;

Adjectives = fatal | deadly | morbid | mortal;

Expressions = <kick> the bucket | <buy> the farm | <put> to
death | <wipe> out ;

Exclude = death star | kill bill | <kill> time ;

Main = :**Verbs** | :**Nouns** | :**Adjectives** | :**Expressions** ;

NooJ for the Digital Humanities

- Processes any text (UTF8)

Text zones inside <tu> ... </tu>

- Import a web site
- Retrieve and import a page from Amazon, Facebook, Reddit or Twitter
- Use Google Search to find and import web pages

The screenshot shows two windows. The top window is titled "Construire un corpus" and displays options for importing text from various sources like Amazon France, Facebook France, Google France, Reddit France, and Twitter France. It also shows a preview of a corpus named "Corpus www.google.com: 19 pages aspirées dans le fichier 2020-06-01 tesla AND iphone.txt". The bottom window is a Google search results page for the query "tesla AND iphone". The search bar shows the query, and the results list includes links such as "Integration iPhone dans Model 3 - Tesla Model 3 - Forum ...", "Application Tesla via iPhone - Tesla Model 3 - Forum Automobile ...", "Lecture des sms (activation, bugs, ...) - Tesla Model 3 - Forum ...", "Ouverture avec iphone - Tesla Model 3 - Forum Automobile Propre", and "Son Iphone Bluetooth - Tesla Model 3 - Forum Automobile Propre".

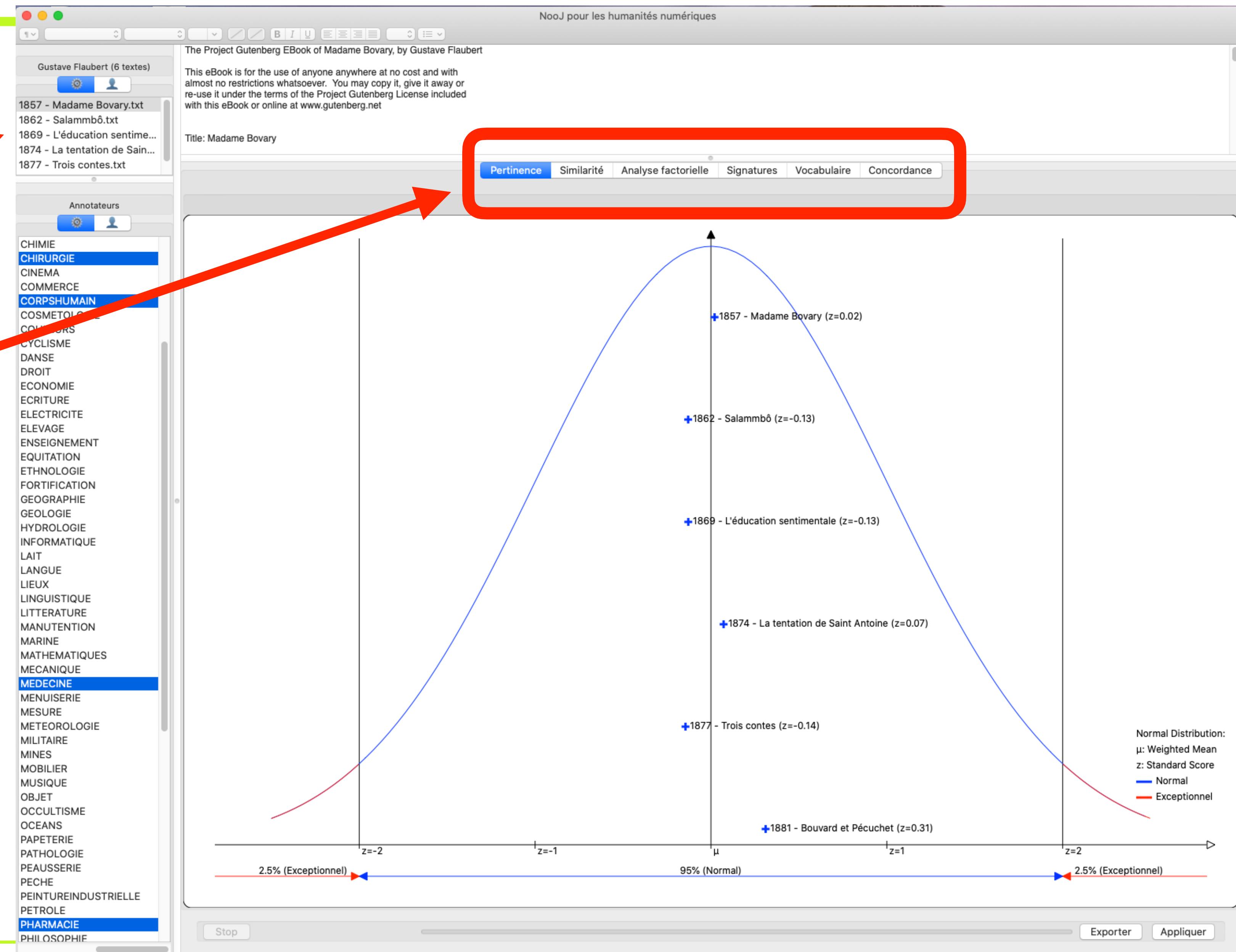
Relevance

Corpus

Statistical Analyses

Technical Terms:

*Surgery, Human Body,
Medicine, Pharmacy*



Similarity

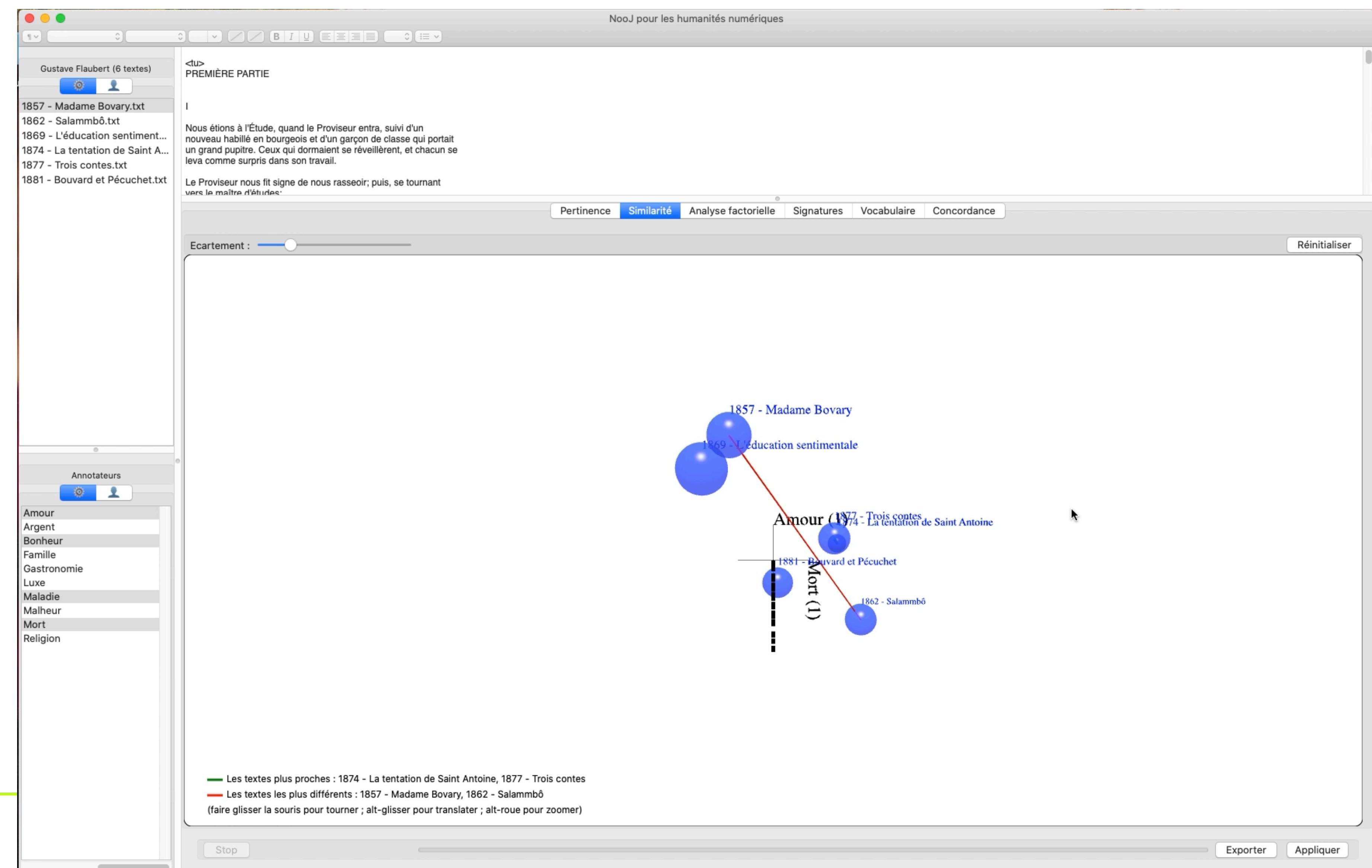
Amour [Love]

Bonheur [Happiness]

Maladie [Decease]

Mort [Death]

MAX SILBERSTEIN



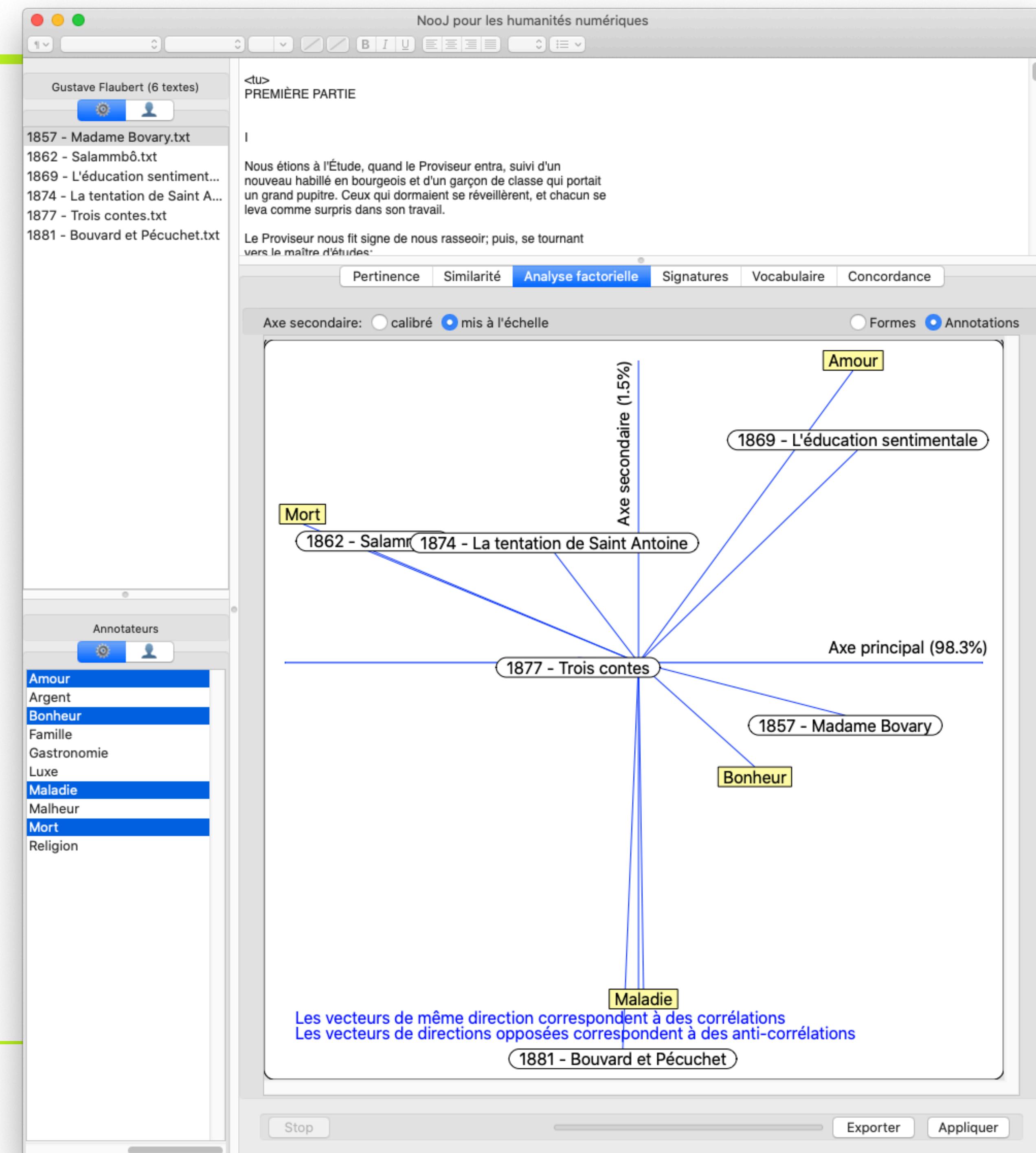
Correspondence Factor Analysis

Mort [Death] ↔ *Salammbô*

Amour [Love] ↔ *L'éducation sentimentale*

Maladie [Decease] ↔ *Bouvard et Pécuchet*

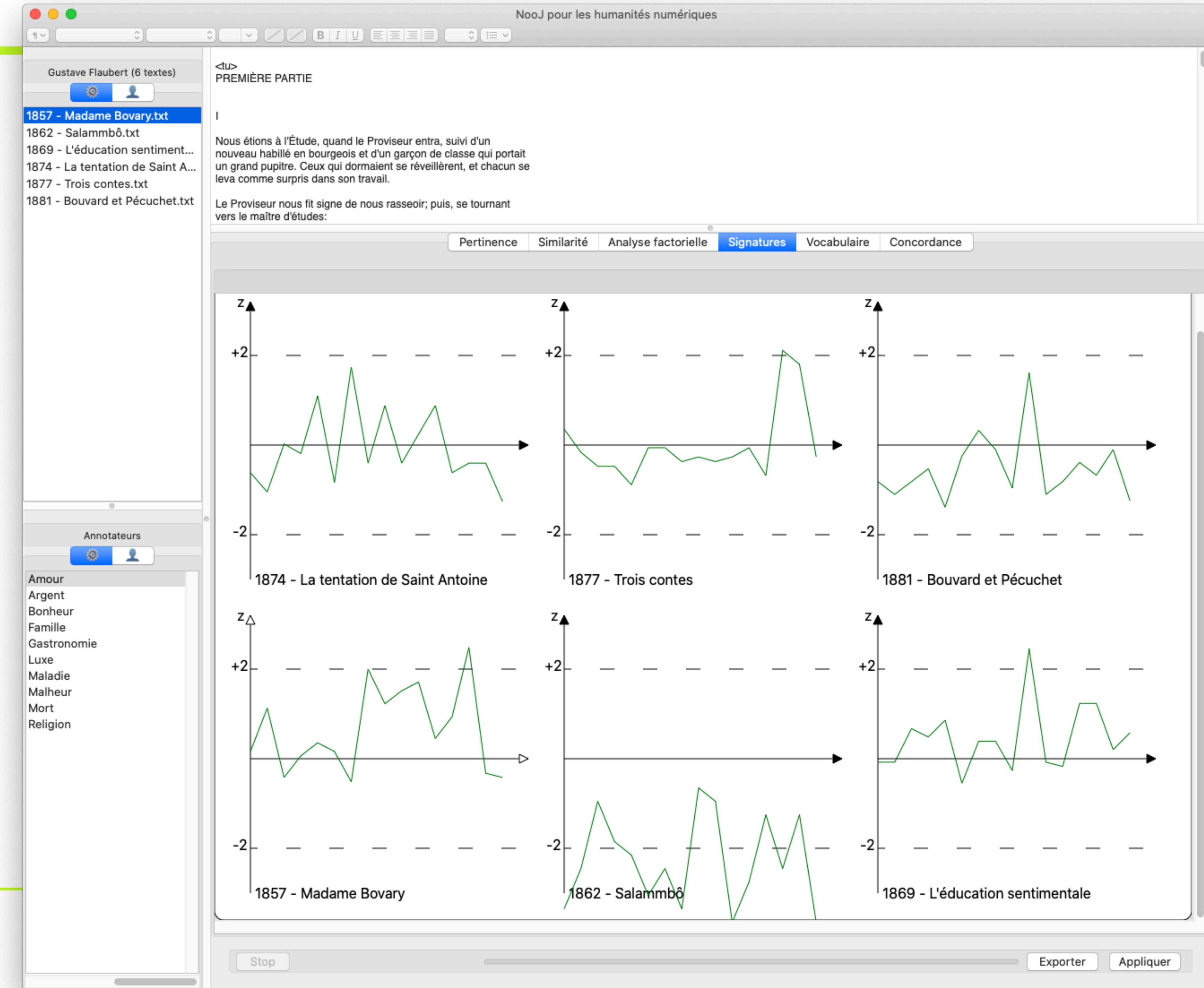
MAX SILBERZTEIN



Narrative Profiles

Amour [Love]

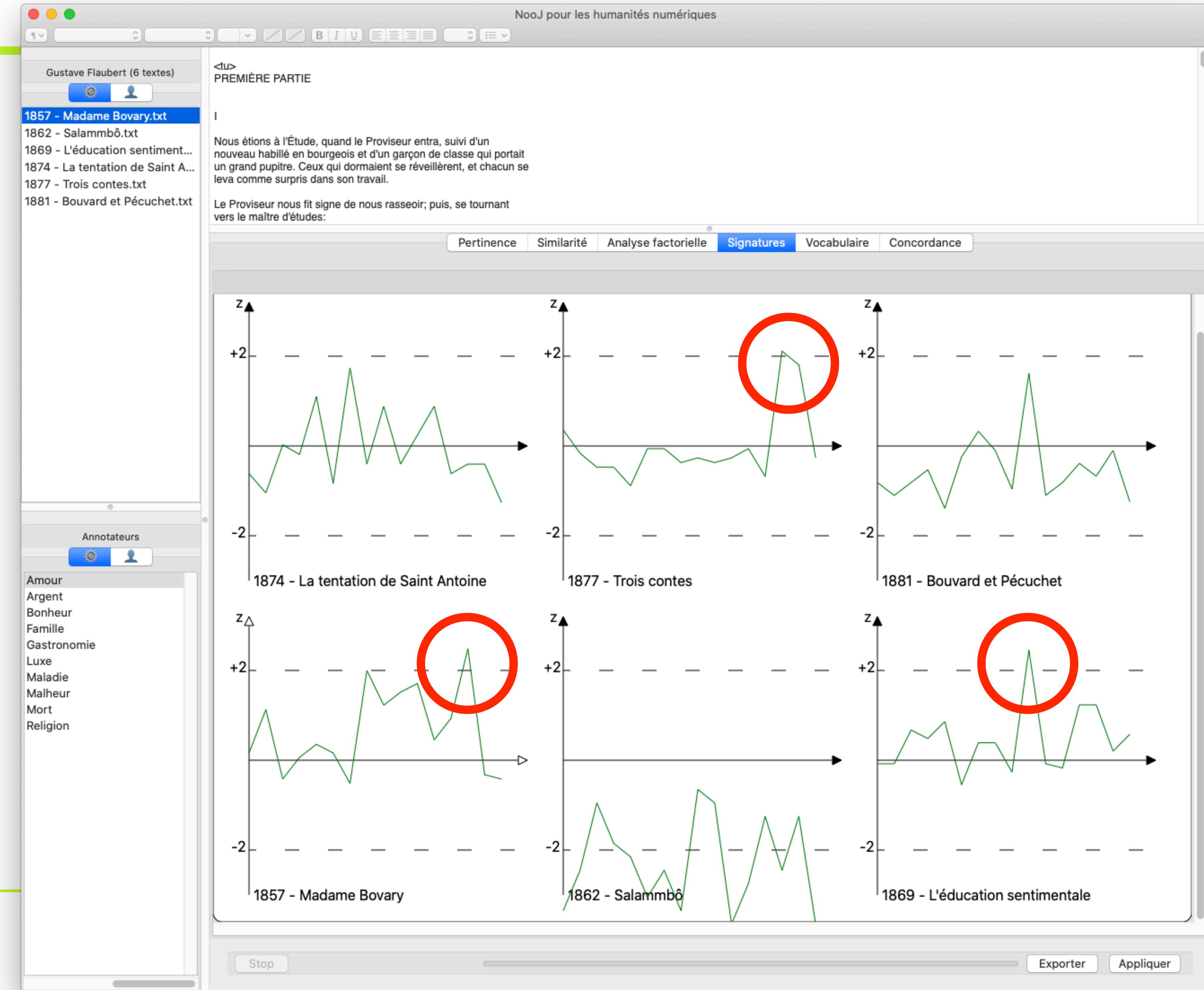
MAX SILBERZTEIN



Narrative Profiles

Amour [Love]

MAX SILBERZTEIN

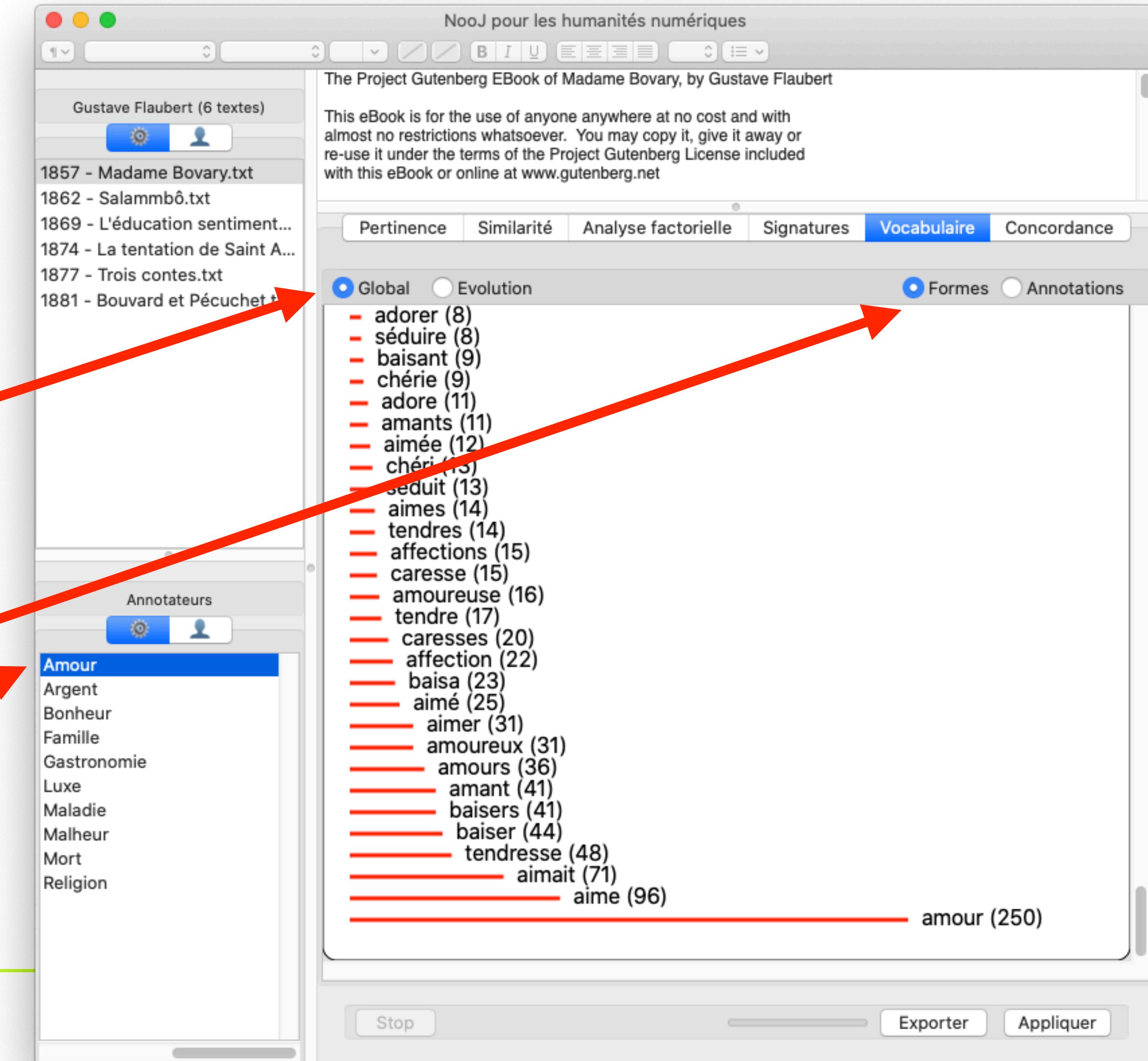


Vocabulary

Global analysis, or Evolution

Matching sequences, or Annotations

Amour [Love]



Concordance

Collocations

(depend on the length of left and right contexts)

Amour [Love]

MAX SILBERZTEIN

NooJ pour les humanités numériques

<tu>
PREMIÈRE PARTIE

1
Nous étions à l'Étude, quand le Proviseur entra, suivi d'un nouveau habillé en bourgeois et d'un garçon de classe qui portait un grand pupitre. Ceux qui dormaient se réveillèrent, et chacun se

Pertinence Similarité Analyse factorielle Signatures Vocabulaire Concordance

Gauche : Droit : 1128 occurrences. Collocations : Frédéric (43 occ.), aime (33 occ.), homme (32 occ.)

Texte	Gauche	Occurrence	Droit
1862 - Salammbô.txt	remercia encore une fois les Mercenaires; il leur baisait Sicca, et on leur avait dit avec toutes sortes de caresses ras en remuant ses doigts pour mieux sentir cette caresse rrêtant. Déjà Narr'Havas s'avancait vers lui. Il baissa et des plumes d'autruche. Le Libyen, ébahie de ces caresses s s'accoutumaient à ses services; il s'en faisait aimer récipitaient vers les gardes de la Légion et leur baissaient es monts comme la roue d'un char. «O Tanit! tu m' aimes squ'à mes pieds, passe dans ma chair... c'est une caresse oignaient à la fois du même principe, et Salammbô adorait endre! «--Un génie--reprit-elle--me pousse à cet amour à moitié, reprit: «Elle inspire et gouverne les amours nspire et gouverne les amours des hommes. «--Les amours avec du sang, et puisque tu ne peux assouvir ton amour rises dans le sac de villes, que l'on fatiguait d' amour es. On leur offrit des étalons d'Hécatompyle; ils aimèrent osité orientales ils se firent des excuses et des caresses , puisque tu vas avoir leur force dans les mains? Aimes à Carthage, entre les collèges des pontifes, qui baiseront , un cône de pierre; Mâtho, en passant à côté, se baissa ainsi sur le bas de son gros ventre,--poli par les baisers le se confondait avec la Déesse elle-même; et son amour uffle de ton haleine! Que mes lèvres s'écrasent à baiser n défaillant contre les coussins du lit. «--Je t' aime les mains; enfin, il les félicita du banquet, tout : «--Vous êtes les sauveurs de Carthage! Mais vous qui lui coulait sur le corps. Des espoirs de vengeance deux pouces en signe d'alliance, rejetant la c , hésitait à y répondre ou à s'en exaspérer. Mais . Cependant ils attendaient un ambassadeur de Car les pieds. La litière s'avancait sur les épaules , n'est-ce pas? Je t'ai tant regardée! Mais non! t qui m'enveloppe, et je me sens écrasée comme si un la Déesse en sa figuration sidérale. Une influence . J'ai gravi les marches d'Eschmoûn, dieu des plan des hommes. «--Les amours des hommes!» répéta Sal des hommes!» répéta Salammbô, rêvant. «--Elle est , gorge ta haine; elle te soutiendra!» Mâtho rappelant qu'elles étaient jeunes, qu'on accablait de ce mieux de l'argent. Puis ils demandèrent qu'on leur . Puis les soldats réclamèrent, comme une preuve d' tu mieux périr le soir d'une défaite, misérablement tes sandales; et si le voile de Tanit te pèse encore la main droite. La première chambre était très haute de la foule. Puis ils se retrouvèrent à l'air libre s'en dégageait plus fort, comme les grands lotus que tes mains! «--Laisse-moi voir!--disait-elle.--Plus !» criait Mâtho. Elle balbutia:--«Donne-le!» Et il <p>Stop Exporter Appliquer</p>		

1. Linguistics can enhance current NLP tools

2. NooJ's methods and resources (for over 30 languages!) should have more impact in the Digital Humanities.

3. Easy-to-use functionalities are key

4. Feedback from users will help add other functionalities...

Conclusion
