An application of linguistic methods and resources

NooJ for the Digital Humanities

MAX SILBERZTEIN, JUNE 5, 2020

Digital Humanities

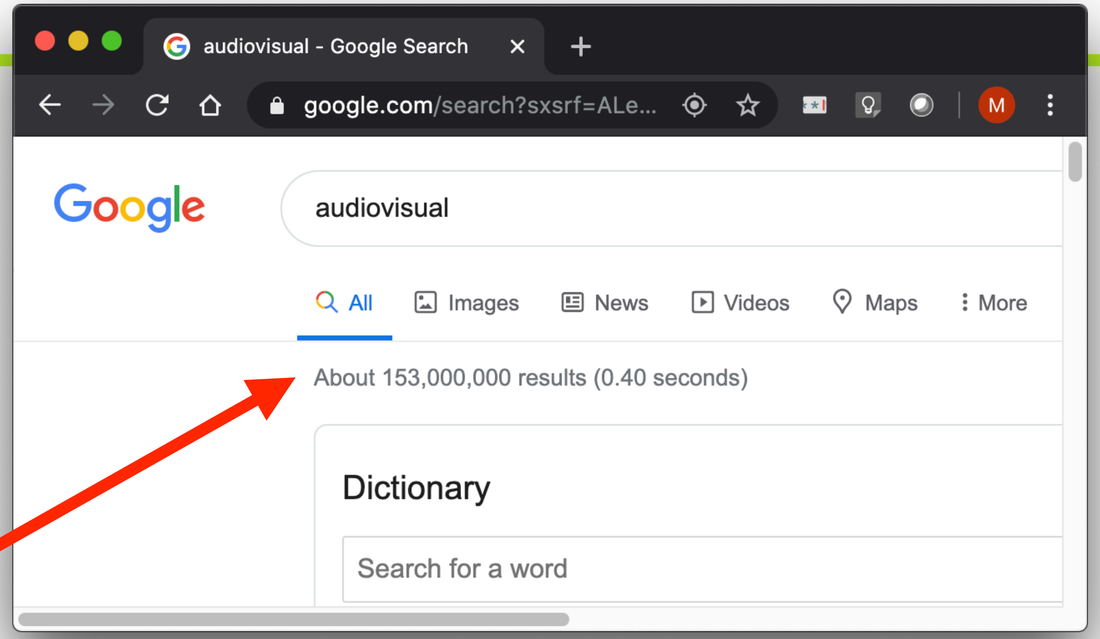
- Researchers in the Social Sciences need tools to analyze their corpora.
- **Alceste**, **Lexico3**, **Hyperbase**, **Iramuteq**, **Sketch Engine**, **Trameur**, **TXM**
- To find what is interesting they detect abnormal frequencies…

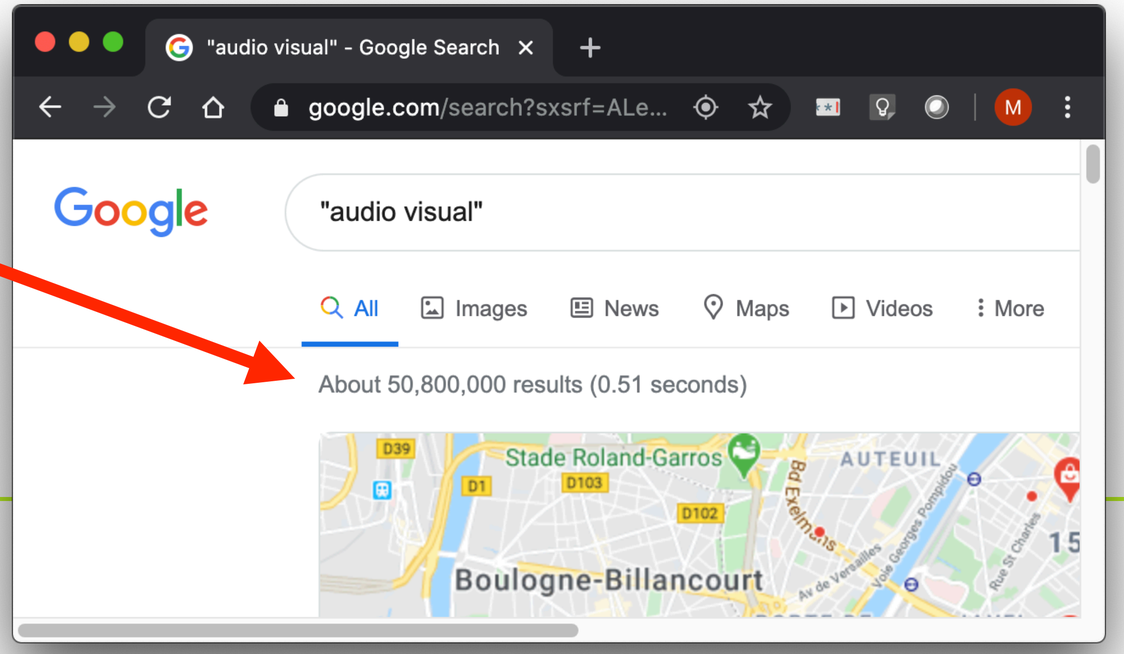
Digital Humanities

- 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, middle-ages
middle Ages, middle-Ages → `middle=ages,NOUN`

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?

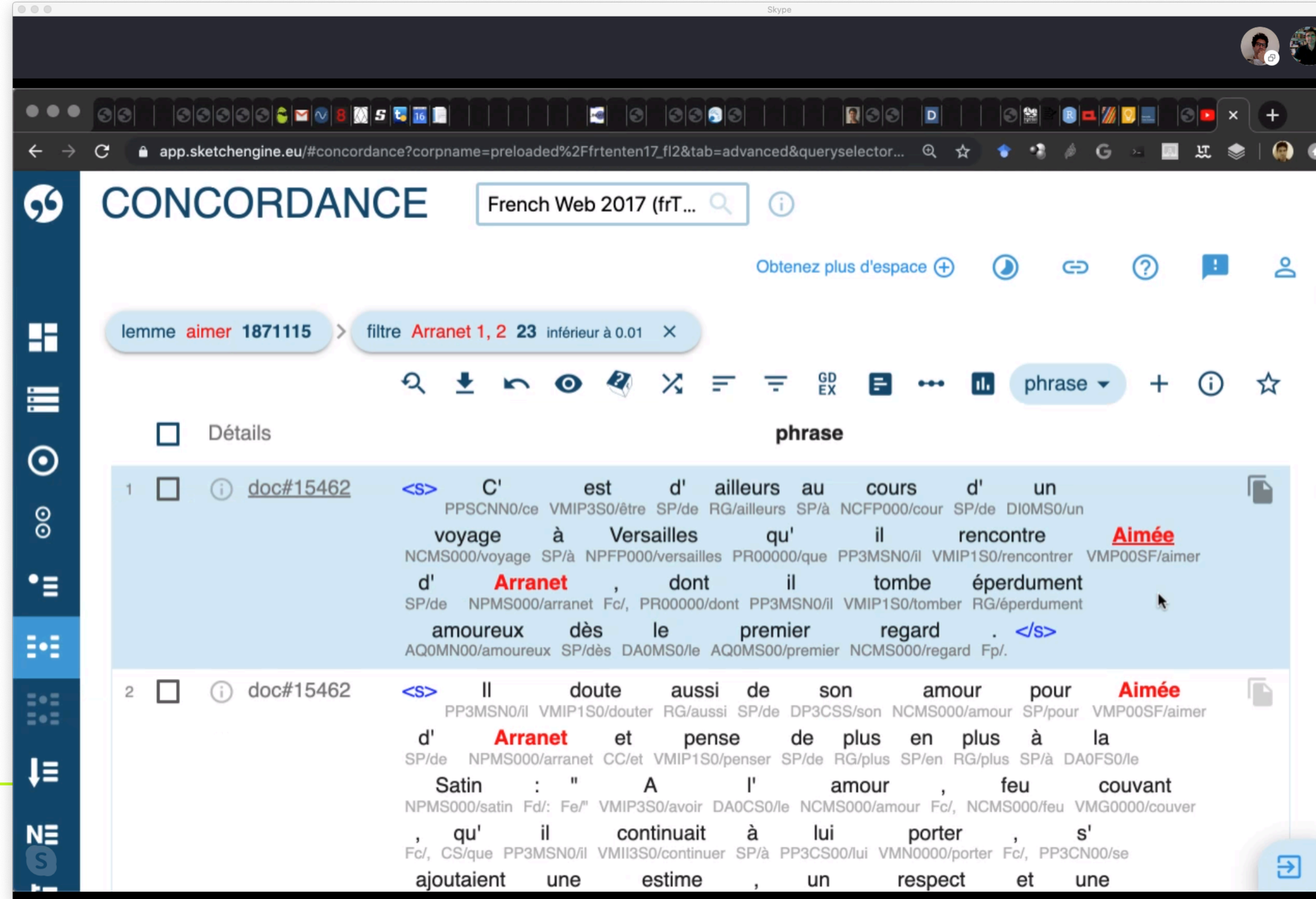
- However, over 99% nouns do not carry different meanings in the Singular and in the Plural:

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

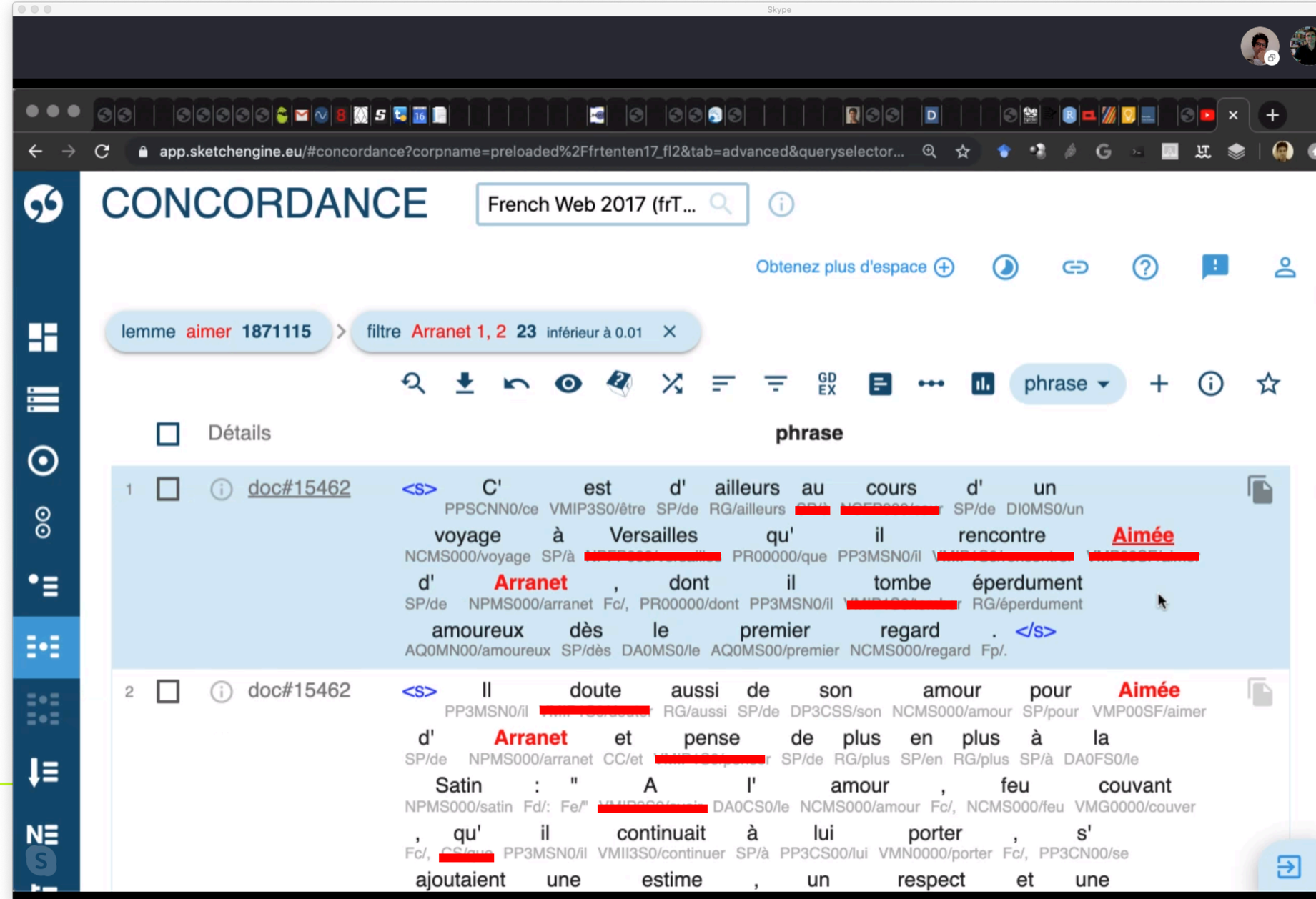
- Most NLP systems do lemmatize wordforms, e.g. **Sketch Engine**

- They use taggers, e.g. **TreeTagger** or **GATE**



3. Tagging

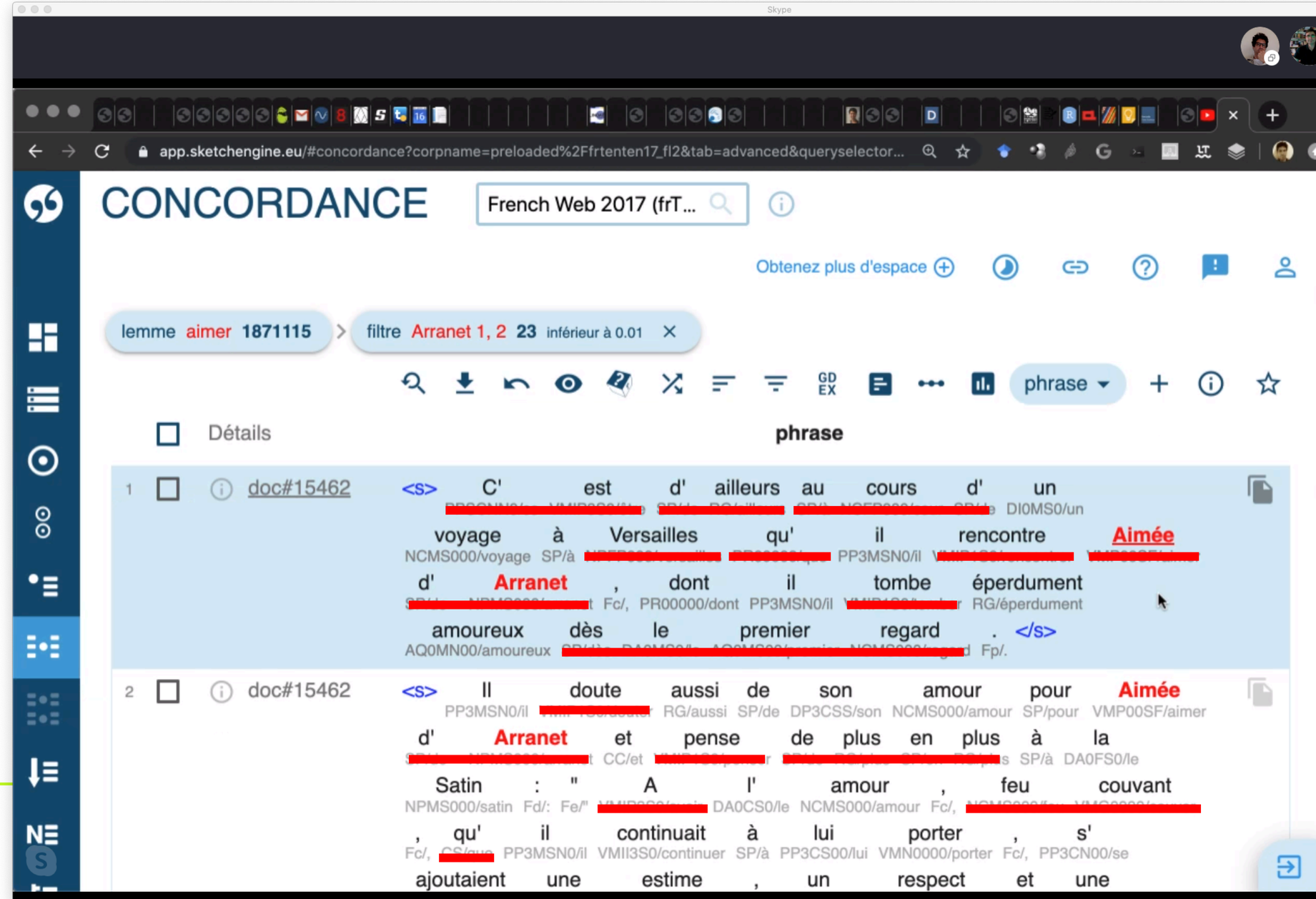
- *cours* not a plural feminine form of the noun *cour*
- *au* not a preposition
- *Aimée* not a past participle form of the verb *aimer*
- *rencontre*, *tombe*, *doute* and *pense* are not conjugated in the first person
- *Versailles* not a plural feminine noun
- *A* not a conjugated form of the verb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



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



MAX SILBERZTEIN

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

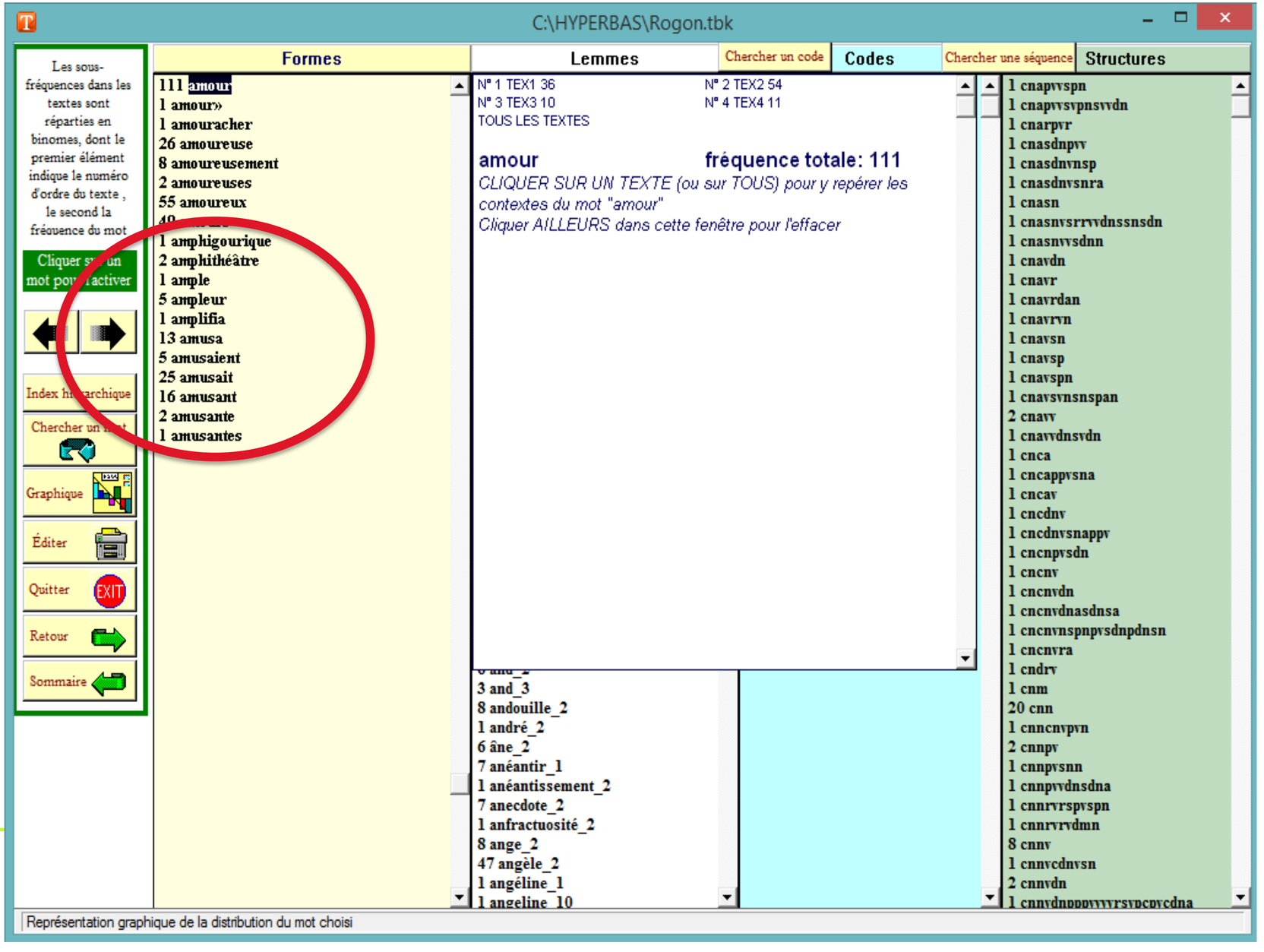
Word forms computed using purely orthographic rules.

“amour” [love]

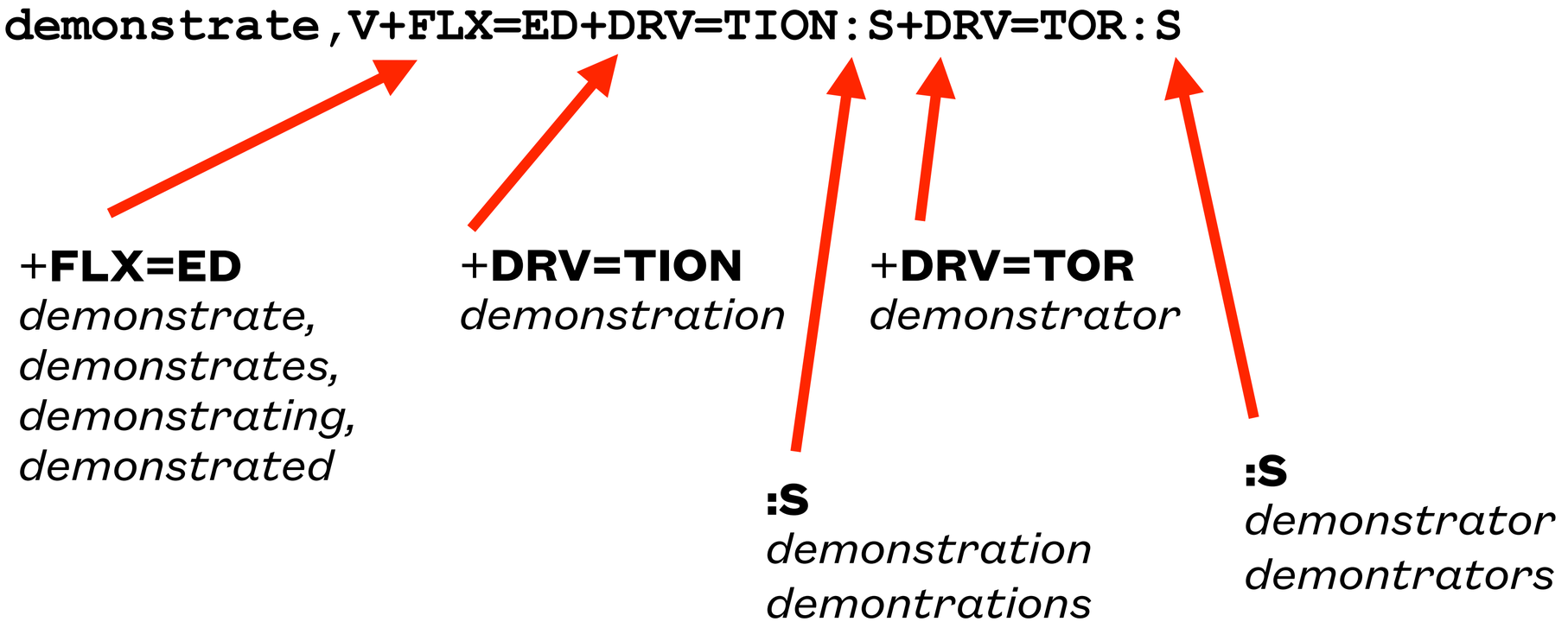
⟶ *amours*, *amouracher*, *amoureusement*…

⟶ ~~*amphigourique, amphithéâtre, ample, ampleur, amplifier, amusa, amusaient*…~~

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

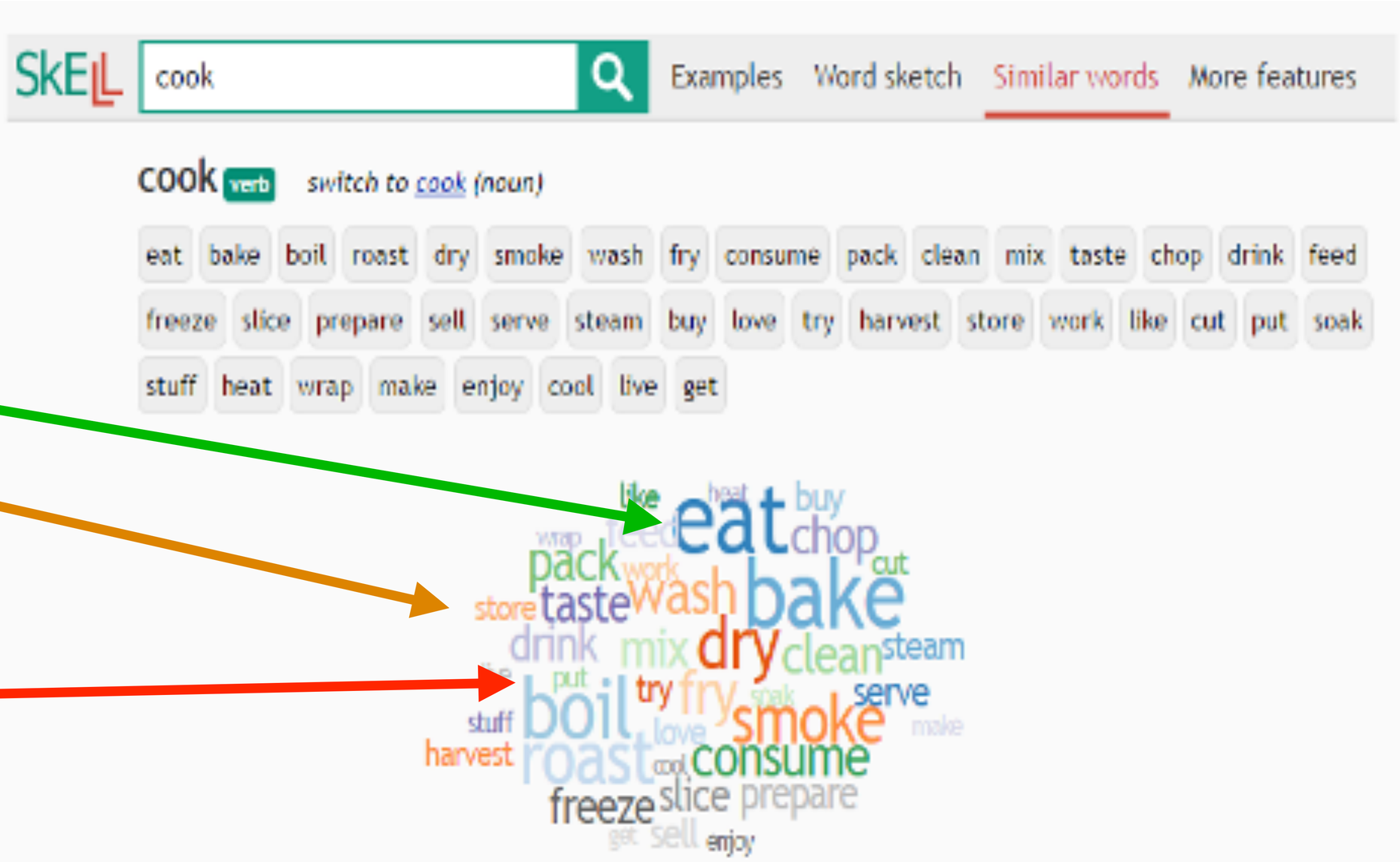


4. Derivation: the linguistic solution



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

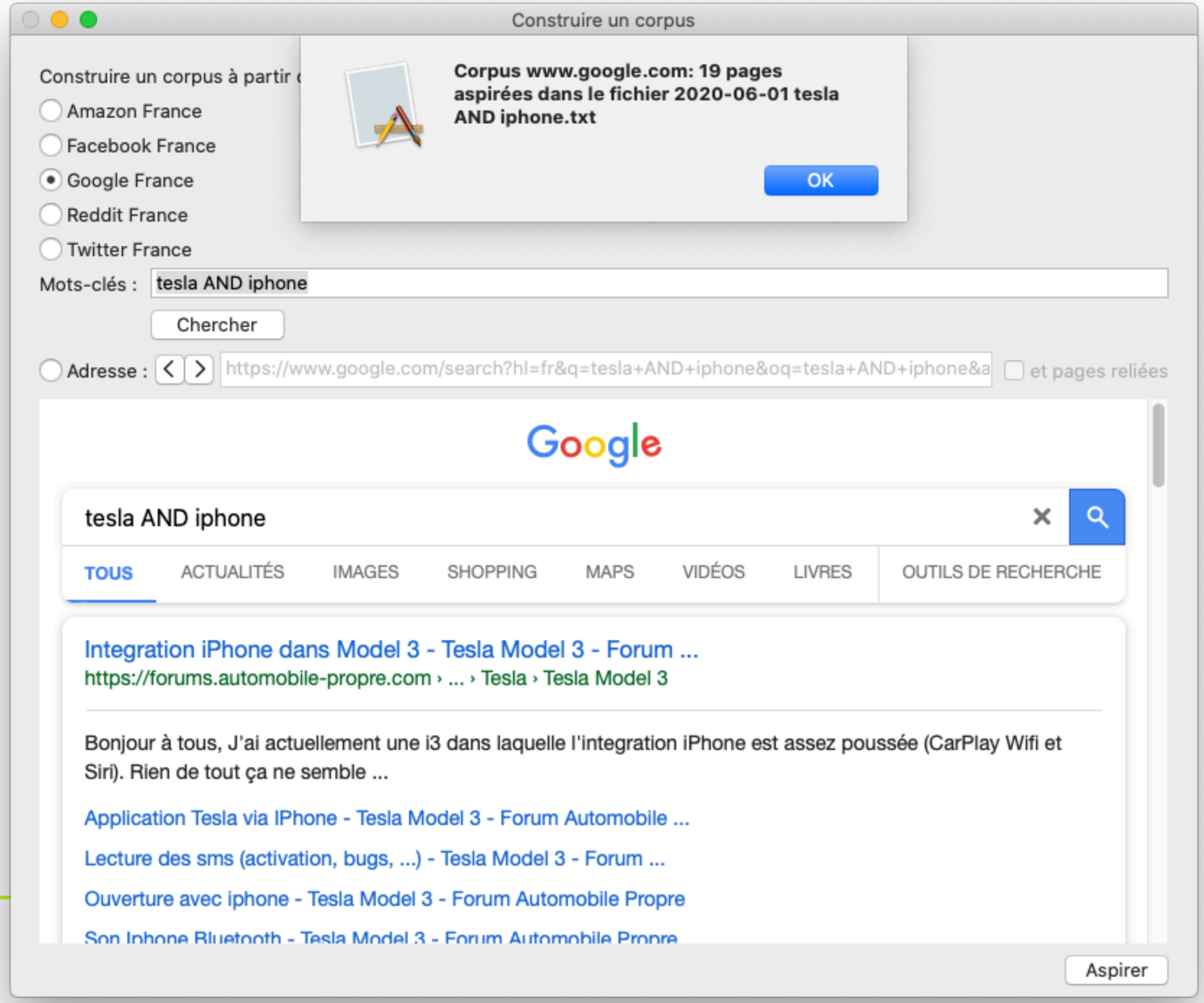
```
Verbs = <assassinate> | <decease> | <die> | <eradicate> |
<exterminate> | <kill> | <obliterate> | <perish> | <suicide> |
<strangle> ;

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



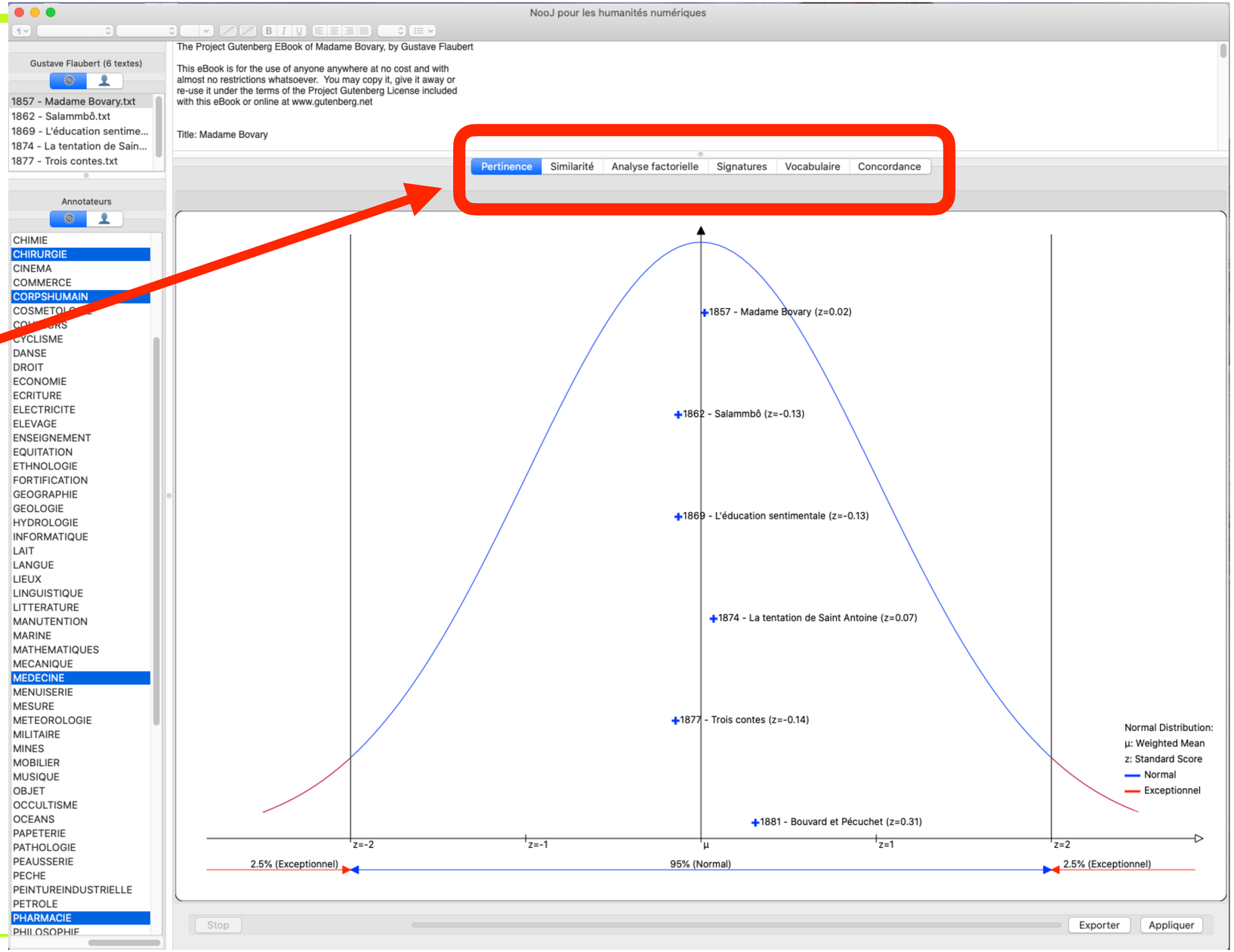
Relevance

Corpus

Statistical Analyses

Technical Terms:

Surgery, Human Body, Medicine, Pharmacy



MAX SILBERZTEIN

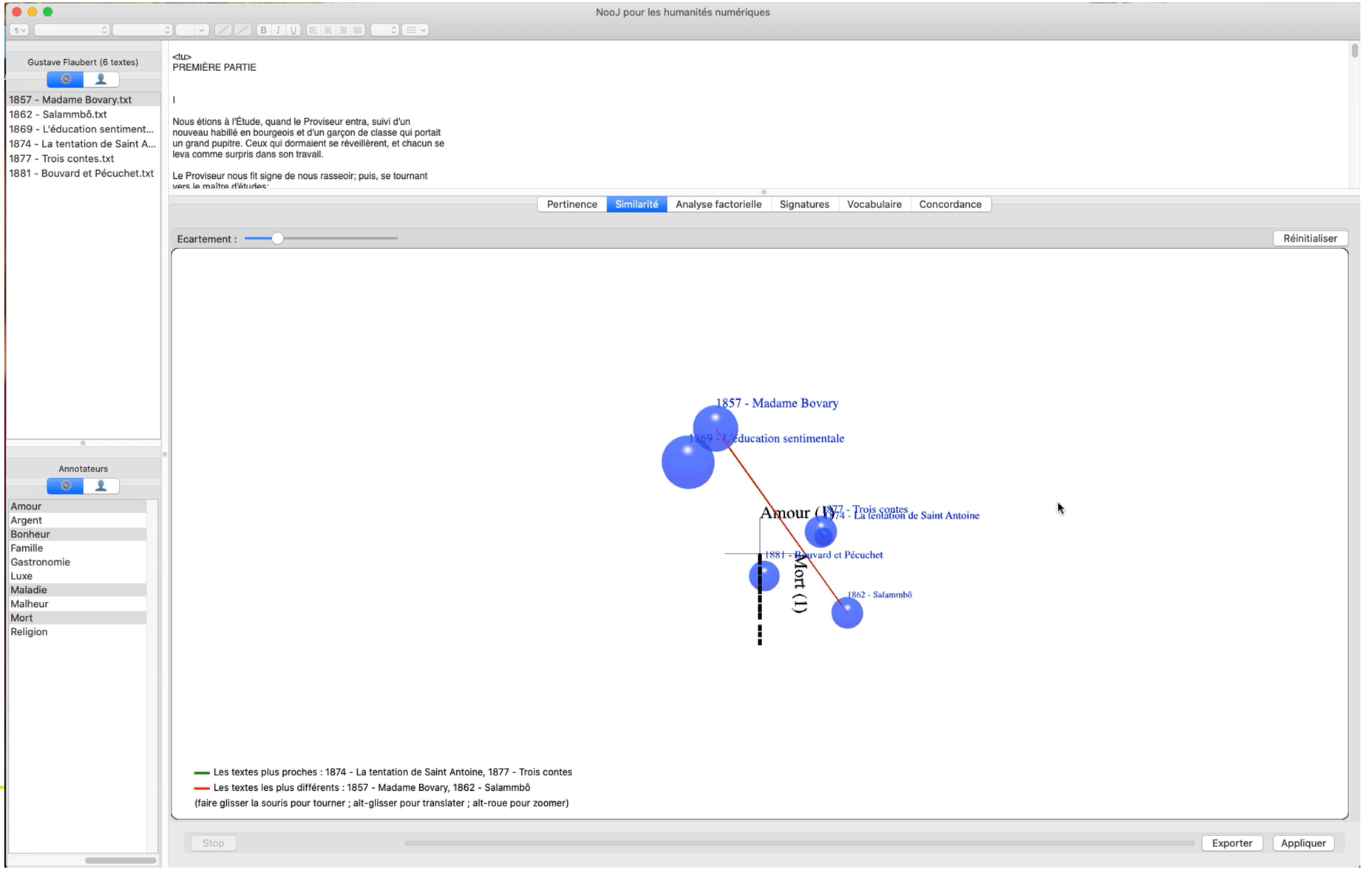
Similarity

Amour [Love]

Bonheur [Happiness]

Maladie [Decease]

Mort [Death]



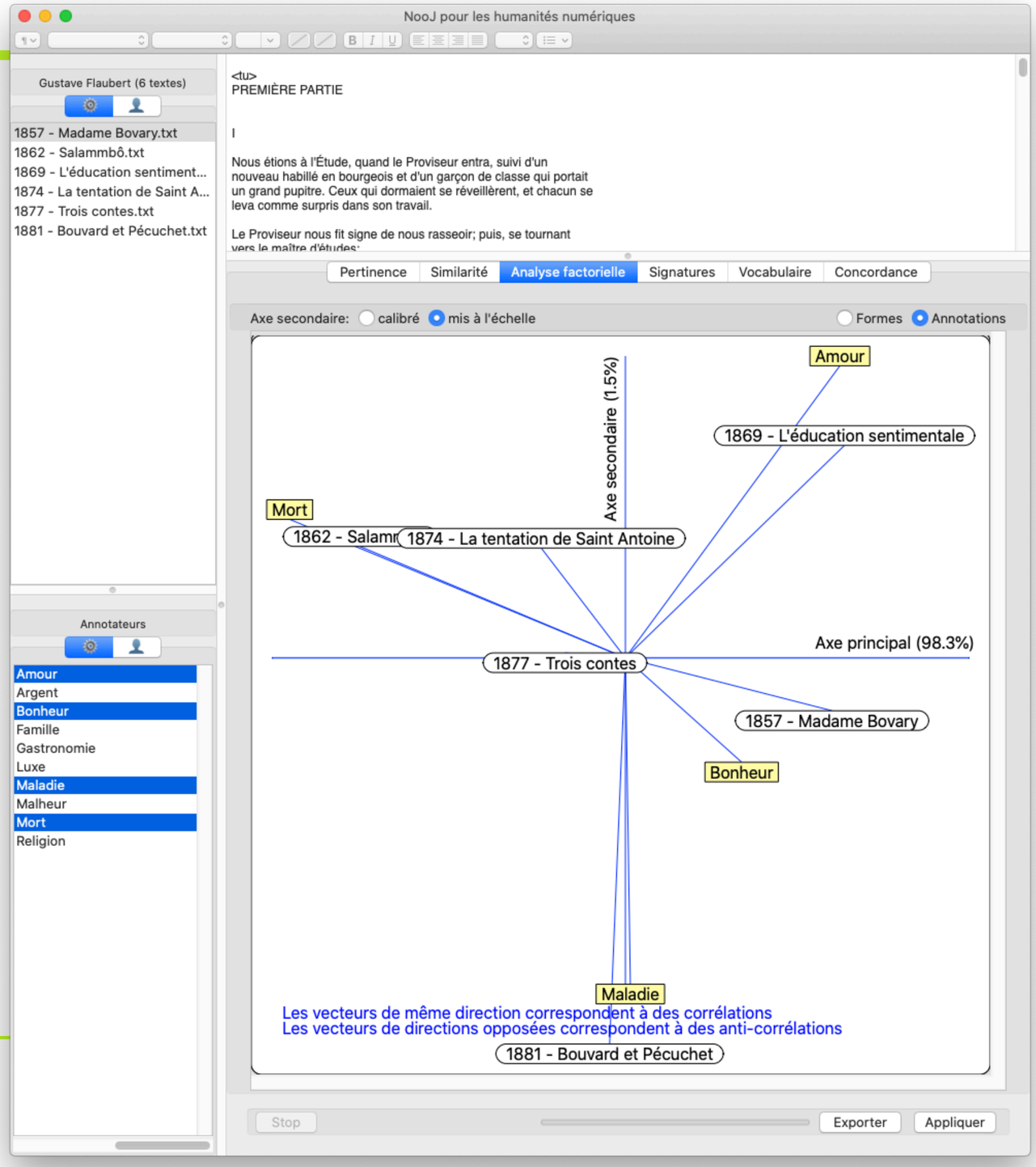
MAX SILBERZTEIN

Correspondence Factor Analysis

Mort [Death] ↔ *Salammbô*

Amour [Love] ↔ *L'éducation sentimentale*

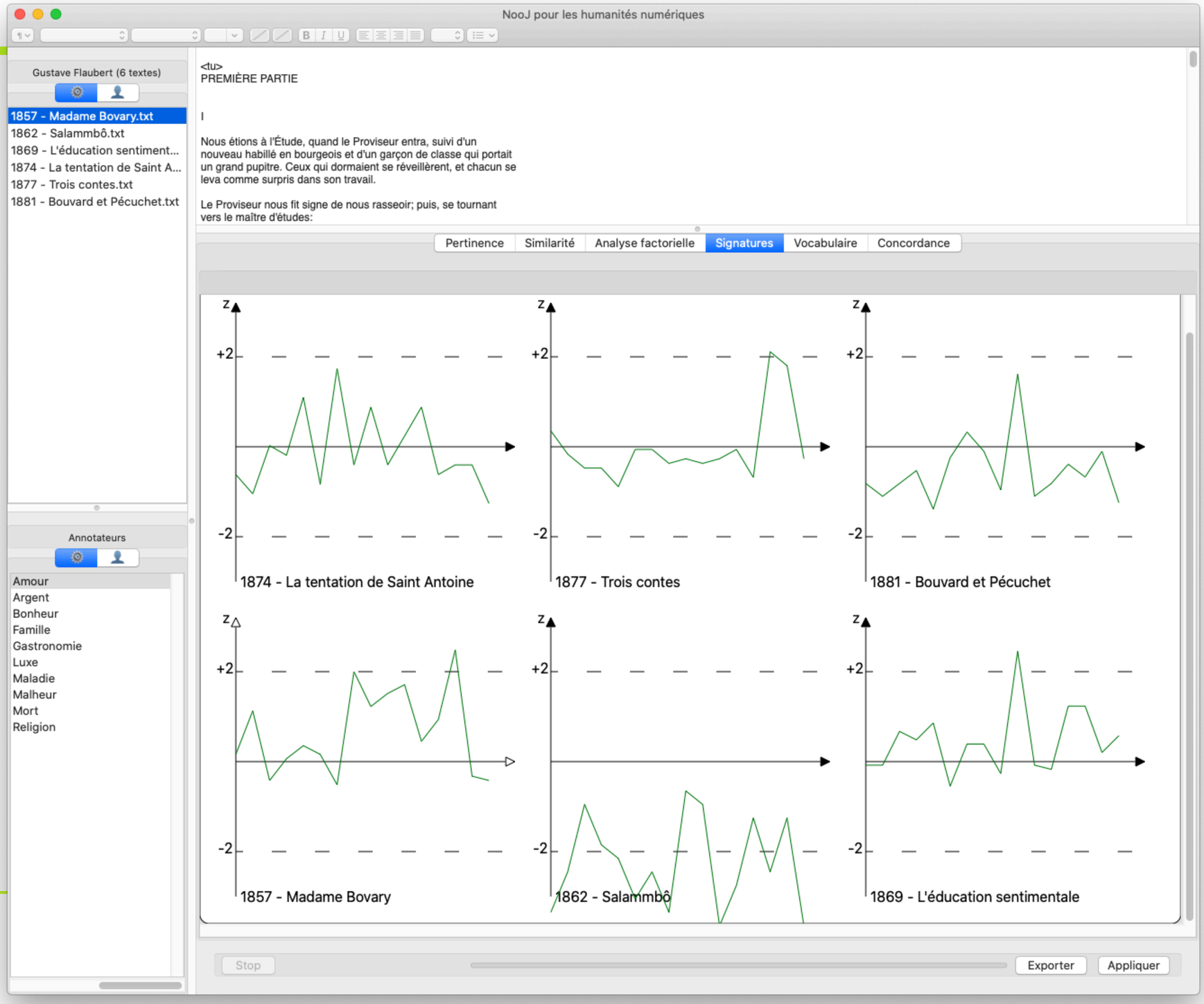
Maladie [Decease] ↔ *Bouvard et Pécuchet*



MAX SILBERZTEIN

Narrative Profiles

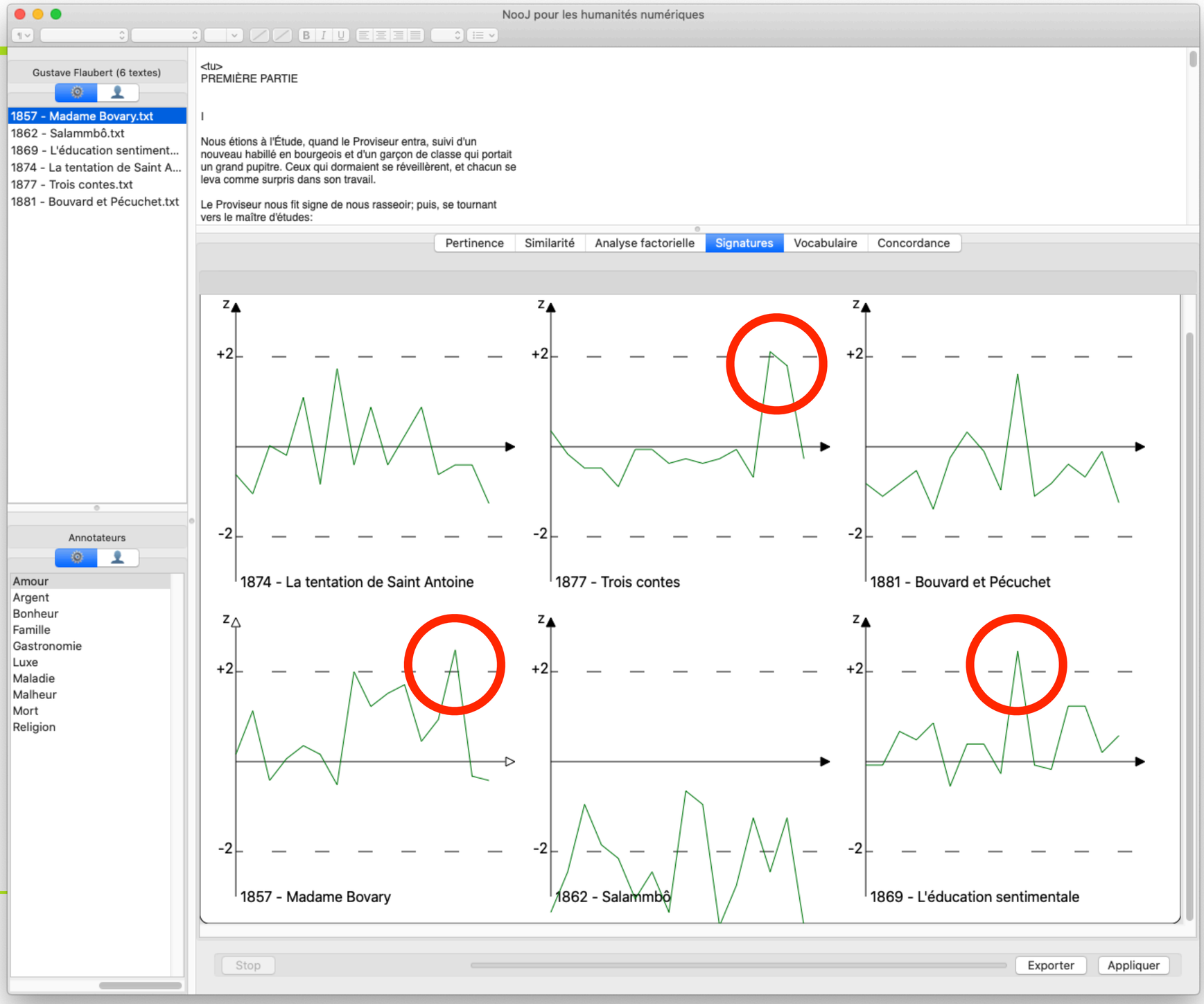
Amour [Love]



MAX SILBERZTEIN

Narrative Profiles

Amour [Love]



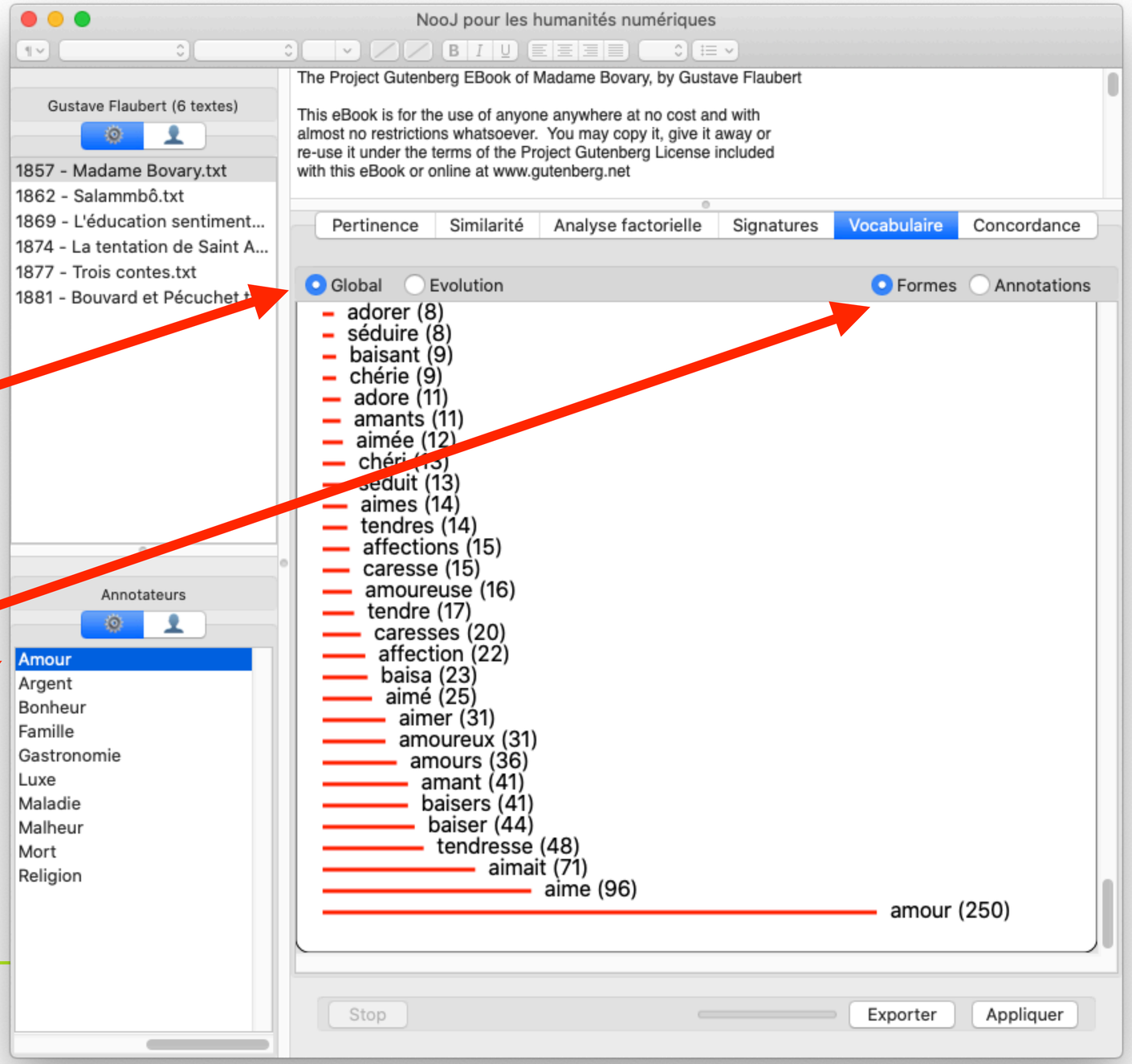
MAX SILBERZTEIN

Vocabulary

Global analysis, or Evolution

Matching sequences, or Annotations

Amour [Love]

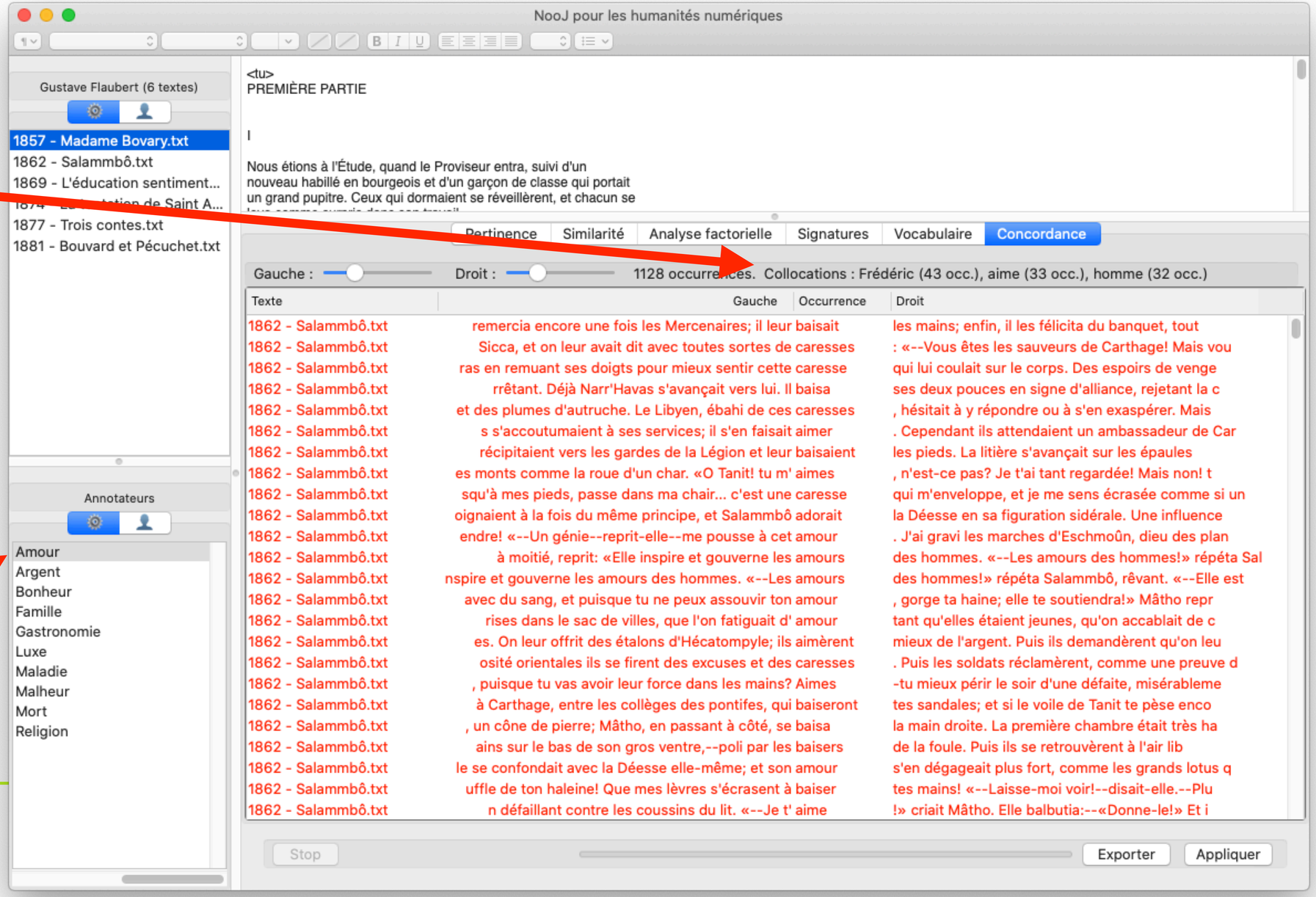


Concordance

Collocations

(depend on the length of left and right contexts)

Amour [Love]



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