Initiation au traitement automatique des corpus

Atelier pratique

Kirill Maslinsky & Aurelia Vasile

22/03/2023

CELIS / MSH
Université Clermont Auvergne

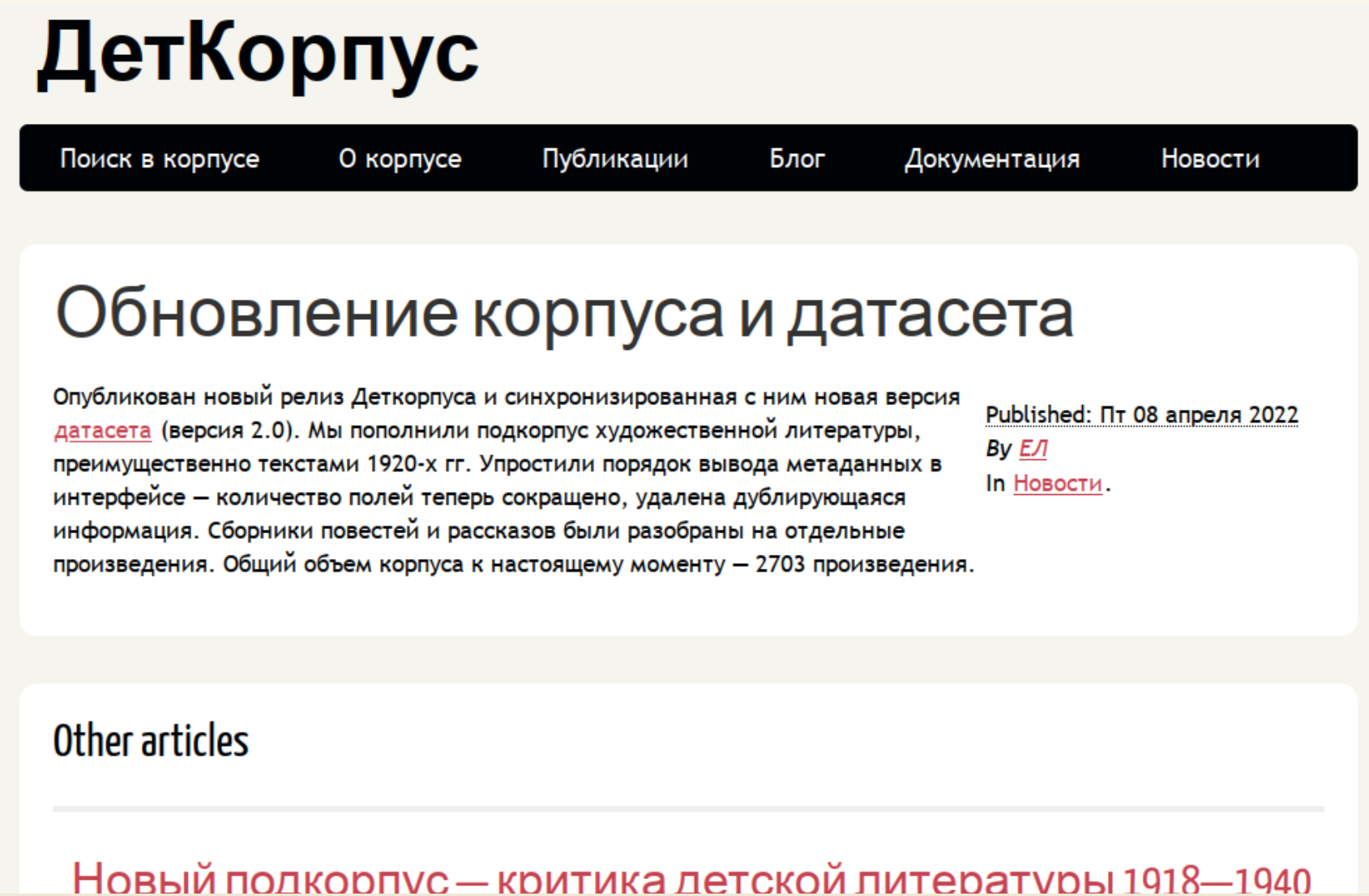
Introduction: my corpora

Corpus Bambara de Référence

http://cormand.huma-num.fr

Detcorpus: Littérature de jeunesse russe

http://detcorpus.ru

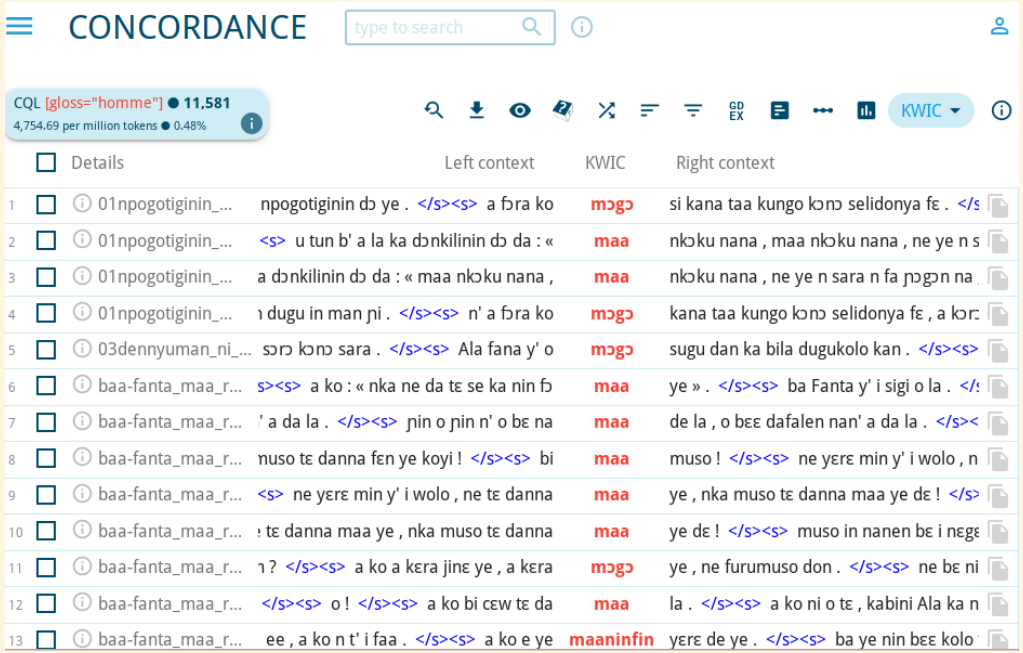


Rongorongo: écriture de l'île de Pâques

Workshop goals

The instrumental goal

- To be able to run a concordance on a text collection. This is our killer app. We will call it simply a corpus.



Conceptual learning

Building a corpus Corpus architecture, data and metadata, annotation, formats

Using a corpus Querying a corpus, corpus query languages

Running a corpus project management & data engineering

Acquaintance with tools

- Command line interfaces (bash)
- Version control (git)
- Linguistic annotation (spacy)
- Corpus exploration (TXM)
- A “programming glue” (python)

Hands-on experience

- follow the whole procedure on your own computer (and experience all the software caveats)

Corpus architecture

Data and metadata

```
LE MOT QUI TUE
L'agression de l'avenue de Villiers.'
Dans les assez longs séjours qu'il avait faits à Paris, Jim Jackson s'était créé d'amicales re
lations dans le monde de la police.
C'est ainsi que M. Manin, le chef de la Sûreté, avait la plus vive estime pour lui, se plaisai
t en sa société si instructive et que différents commissaires de police aimaient à recevoir à
leur table le célèbre détective américain.
Jim Jackson acceptait avec le plus vif plaisir les invitations de ces magistrats, mais c'est s
urtout chez M. Benoit, le commissaire de police du dix-septième arrondissement, que le fin lim
ier se rendait avec le plus de contentement.
M. Benoit était un homme remarquable qui aurait pu être un grand artiste s'il avait suivi enti
èrement ses goûts.
Il peignait, sculptait, écrivait des pièces de théâtre, et ses productions étaient bien supéri
eures à celles de la plupart des professionnels.
De plus, M. Benoit avait approché les plus grands artistes de son époque, et, à ce contact, so
n goût s'était affiné, son intelligence s'était développée et sa conversation s'était imprégné
e de cet esprit parisien unique au monde.
J im J ackson se plaisait énormément en la compagnie de cet esprit fin, disert, ingénieux qui
ne parlait pas seulement de crimes et de délits.
Il faut tout dire : le commissaire de police du dix-septicme arrondissement n'aimait qu'à demi
 son métier.
Il se reposait presque entièrement sur son secrétaire du soin d'élucider les affaires courante
s et n'intervenait que dans des cas exceptionnels.
Il avait donné le mot d'ordre à. son
subordonné et celui-ci savait qu'on pouvait trouver le magistrat dans son atelier de la rue Jo
uffroy où il préparait un tableau destiné au Salon, tableau auquel il travaillait avec acharne
ment et même passion.
Qu'importaient à M. Benoit les vols à l'américaine, les cambriolages, les suicides, les escroq
ueries et les affaires banales et journalières qu'Evariste Carbonel, son secrétaire, pouvait p
arfaitement instruire à sa place ! N'avait-il pas, lui, une œuvre d'art à accomplir?
Evariste Carbonel était tout jeune dans le métier et c'est pourquoi il y apportait un zèle inc
omparable.
Il était heureux que son patron lui laissât la bride sur le cou et, si l'expérience lui manqua
it, il la remplaçait par une ardente activité.
Tout était donc pour le mieux dans le commissariat et Jim Jackson riait en voyant que tout mar
chait aussi bien que si le magistrat eût tout instruit par lui-même.
—   Cependant, dit-il, un jour, à M. Benoit, si une affaire délicate se présentait, votre secr
étaire ne pourrait en venir à bout, car il n'a pas la ruse que
```

Data and metadata

C	D	E	G	H	I
Dublin Core:Title	Dublin Core:Creator	Dublin Core:Date	xml_link	collection	
Le mot qui tue	Saintillac, Hector (1876-1962)	1916.0	BUCA_Bastaire_Collection_Aventures_C50578.xml	Collection d'Aventures	
La Tringle et Paolo	Kerlecq, Jean de (1882-1969)	[19..]	BUCA_Bastaire_Collection_Aventures_C91953.xml	Collection d'Aventures	
Le martyre de Paolo, le petit acrobate	Kerlecq, Jean de (1882-1969)	[19..]	BUCA_Bastaire_Collection_Aventures_C91954.xml	Collection d'Aventures	
Dick, le petit télégraphiste. [I], Les fugitifs	Hales, A. G. (1870-1936)	[191.]	BUCA_Bastaire_Collection_Aventures_C91955.xml	Collection d'Aventures	
Dick le petit télégraphiste. [III], Les Nubiens sans nez	Hales, A. G. (1870-1936)	[191.]	BUCA_Bastaire_Collection_Aventures_C91956.xml	Collection d'Aventures	
Le maître de la banquise. [I]	Moselli, José (1882-1941)	[19..]	BUCA_Bastaire_Collection_Aventures_C91959.xml	Collection d'Aventures	
Le secret du chiffonnier. [I]	Valle, Jo (1865-1949)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91968.xml	Collection d'Aventures	
Maurice Gillar, détective. [I]	Idiers, Marcel (1886-1960)	[191.?]	BUCA_Bastaire_Collection_Aventures_C50583.xml	Collection d'Aventures	
Le roi des boxeurs. Les terroristes du Kurdistan	Moselli, José (1882-1941)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91957.xml	Collection d'Aventures	
En avant, Fanfan-la-Tulipe	C..., J. de (18..-19..)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91958.xml	Collection d'Aventures	
Le maître de la banquise. [III], La cité du pôle|L'espion de	Moselli, José (1882-1941)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91960.xml	Collection d'Aventures	
L'étrange voyage. [I], Les hommes-singes	Laurian, Marcel (18..-19..)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91961.xml	Collection d'Aventures	
L'héritage de Doublezède. [I]	Adam, Pierre (1883-1943)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91962.xml	Collection d'Aventures	
L'héritage de doublezède. [III], La dette du crime	Adam, Pierre (1883-1943)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91963.xml	Collection d'Aventures	
Trois coeurs de France|Combats tragiques dans les mont	Adam, Pierre (1883-1943)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91964.xml	Collection d'Aventures	
La nièce du geolier	Salmon, Paul (1884-1965)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91965.xml	Collection d'Aventures	
Les bandits de Paris-New-York	Aleyrac, Jean (18..-19..)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91966.xml	Collection d'Aventures	
La brousse aux loups	Véran, Réginald (1873-1962)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91967.xml	Collection d'Aventures	
Le secret du chiffonnier. [II], Le compagnon de chaîne	Valle, Jo (1865-1949)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91969.xml	Collection d'Aventures	
Les pillards mexicains. [II], La machine infernale	Idiers, Marcel (1886-1960)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91970.xml	Collection d'Aventures	
L'explorateur fantôme. [II], Le cratère du Diable	Choquet, Gaston (1875-1917)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91971.xml	Collection d'Aventures	
Les chevaliers de la forêt. [II], Le spectre vivant|La gorge d	Aleyrac, Jean (18..-19..)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91972.xml	Collection d'Aventures	
Justus Wise, détective. [II], La chasse à l'homme|Le pli rou	Romagny, A. (18..-19..)	[191.?]	BUCA_Bastaire_Collection_Aventures_C91973.xml	Collection d'Aventures	

Why metadata

Metadata is what links textual data to interesting research questions.

Structural metadata

- date (year)
- genre/type/collection
- author (if contrasted to others)
- author metadata
- location
- source/media/etc.

Descriptive metadata

- Title
- Bibliographic info
- Author (if not contrasted to others)

Text structure: tokens and annotation

- Token is a primitive unit of corpus structure. An atom for all further processing.
- Attribute is what makes a token interesting. Often used to assign token to some category.
- Format can technically be anything that keeps the order of tokens and their attributes (and is supported by your corpus software).

```
<?xml version='1.0' encoding='UTF-8'?>
<text>
  <w pos="ET" lefffLemma="<feff>" spacyLemma="<feff>"><feff></w>
  <w pos="DET" lefffLemma="le" spacyLemma="la">LA</w>
  <w pos="NC" lefffLemma="guerre" spacyLemma="guerre">GUERRE</w>
  <w pos="P" lefffLemma="dans" spacyLemma="dans">DANS</w>
  <w pos="DET" lefffLemma="le" spacyLemma="le">LA</w>
  <w pos="NC" lefffLemma="prairie" spacyLemma="prairie">PRAIRIE</w>
  <w pos="NPP" lefffLemma="chapitre" spacyLemma="chapitre">CHAPITRE</w>
  <w pos="NPP" lefffLemma="i" spacyLemma="i">I</w>
  <w pos="NPP" lefffLemma="coucou" spacyLemma="coucou">Coucou</w>
  <w pos="PONCT" lefffLemma="," spacyLemma=",">,</w>
  <w pos="P+D" lefffLemma="au" spacyLemma="au">au</w>
  <w pos="NC" lefffLemma="prix" spacyLemma="prix">prix</w>
  <w pos="P" lefffLemma="de" spacyLemma="de">de</w>
  <w pos="DET" lefffLemma="mille" spacyLemma="mille">mille</w>
  <w pos="NC" lefffLemma="danger" spacyLemma="danger">dangers</w>
  <w pos="PONCT" lefffLemma="," spacyLemma=",">,</w>
  <w pos="V" lefffLemma="avoir" spacyLemma="avoir">avait</w>
  <w pos="VPP" lefffLemma="pu" spacyLemma="pouvoir">pu</w>
  <w pos="VINF" lefffLemma="échapper" spacyLemma="échapper">échapper</w>
  <w pos="P+D" lefffLemma="aux" spacyLemma="à">aux</w>
  <w pos="NPP" lefffLemma="kioways" spacyLemma="kioway">Kioways</w>
  <w pos="PROREL" lefffLemma="qui" spacyLemma="qui">qui</w>
  <w pos="CLR" lefffLemma="clr" spacyLemma="se">s'</w>
  <w pos="V" lefffLemma="apprêter" spacyLemma="apprêter">apprêtaient</w>
  <w pos="P" lefffLemma="à" spacyLemma="à">à</w>
  <w pos="CLO" lefffLemma="l'" spacyLemma="le">l'</w>
  <w pos="VINF" lefffLemma="attacher" spacyLemma="attacher">attacher</w>
  <w pos="P+D" lefffLemma="au" spacyLemma="au">au</w>
  <w pos="NC" lefffLemma="poteau" spacyLemma="poteau">poteau</w>
  <w pos="P" lefffLemma="de" spacyLemma="de">de</w>
  <w pos="NC" lefffLemma="torture" spacyLemma="torture">tortures</w>
  <w pos="PONCT" lefffLemma="." spacyLemma=".">.</w>
  <w pos="NPP" lefffLemma="harrassé" spacyLemma="harrassé">Harrassé</w>
  <w pos="P" lefffLemma="par" spacyLemma="par">par</w>
  <w pos="DET" lefffLemma="un" spacyLemma="un">une</w>
  <w pos="NC" lefffLemma="fuite" spacyLemma="fuite">fuite</w>
  <w pos="P" lefffLemma="de" spacyLemma="de">de</w>
  <w pos="DET" lefffLemma="plusieurs" spacyLemma="plusieurs">plusieurs</w>
  <w pos="NC" lefffLemma="jours" spacyLemma="jour">jours</w>
  <w pos="PONCT" lefffLemma="," spacyLemma=",">,</w>
  <w pos="CLS" lefffLemma="il" spacyLemma="il">il</w>
```

Text structure: tokens and annotation

Some popular formats include :

- XML
- TSV
- CSV
- JSON

```xml
<?xml version='1.0' encoding='UTF-8'?>
<text>
  <w pos="ET" lefffLemma="<feff>" spacyLemma="<feff>"><feff></w>
  <w pos="DET" lefffLemma="le" spacyLemma="la">LA</w>
  <w pos="NC" lefffLemma="guerre" spacyLemma="guerre">GUERRE</w>
  <w pos="P" lefffLemma="dans" spacyLemma="dans">DANS</w>
  <w pos="DET" lefffLemma="le" spacyLemma="le">LA</w>
  <w pos="NC" lefffLemma="prairie" spacyLemma="prairie">PRAIRIE</w>
  <w pos="NPP" lefffLemma="chapitre" spacyLemma="chapitre">CHAPITRE</w>
  <w pos="NPP" lefffLemma="i" spacyLemma="i">I</w>
  <w pos="NPP" lefffLemma="coucou" spacyLemma="coucou">Coucou</w>
  <w pos="PONCT" lefffLemma="," spacyLemma=",">,</w>
  <w pos="P+D" lefffLemma="au" spacyLemma="au">au</w>
  <w pos="NC" lefffLemma="prix" spacyLemma="prix">prix</w>
  <w pos="P" lefffLemma="de" spacyLemma="de">de</w>
  <w pos="DET" lefffLemma="mille" spacyLemma="mille">mille</w>
  <w pos="NC" lefffLemma="danger" spacyLemma="danger">dangers</w>
  <w pos="PONCT" lefffLemma="," spacyLemma=",">,</w>
  <w pos="V" lefffLemma="avoir" spacyLemma="avoir">avait</w>
  <w pos="VPP" lefffLemma="pu" spacyLemma="pouvoir">pu</w>
  <w pos="VINF" lefffLemma="échapper" spacyLemma="échapper">échapper</w>
  <w pos="P+D" lefffLemma="aux" spacyLemma="à">aux</w>
  <w pos="NPP" lefffLemma="kioways" spacyLemma="kioway">Kioways</w>
  <w pos="PROREL" lefffLemma="qui" spacyLemma="qui">qui</w>
  <w pos="CLR" lefffLemma="clr" spacyLemma="se">s'</w>
  <w pos="V" lefffLemma="apprêter" spacyLemma="apprêter">apprêtaient</w>
  <w pos="P" lefffLemma="à" spacyLemma="à">à</w>
  <w pos="CLO" lefffLemma="l'" spacyLemma="le">l'</w>
  <w pos="VINF" lefffLemma="attacher" spacyLemma="attacher">attacher</w>
  <w pos="P+D" lefffLemma="au" spacyLemma="au">au</w>
  <w pos="NC" lefffLemma="poteau" spacyLemma="poteau">poteau</w>
  <w pos="P" lefffLemma="de" spacyLemma="de">de</w>
  <w pos="NC" lefffLemma="torture" spacyLemma="torture">tortures</w>
  <w pos="PONCT" lefffLemma="." spacyLemma=".">.</w>
  <w pos="NPP" lefffLemma="harrassé" spacyLemma="harrassé">Harrassé</w>
  <w pos="P" lefffLemma="par" spacyLemma="par">par</w>
  <w pos="DET" lefffLemma="un" spacyLemma="un">une</w>
  <w pos="NC" lefffLemma="fuite" spacyLemma="fuite">fuite</w>
  <w pos="P" lefffLemma="de" spacyLemma="de">de</w>
  <w pos="DET" lefffLemma="plusieurs" spacyLemma="plusieurs">plusieurs</w>
  <w pos="NC" lefffLemma="jours" spacyLemma="jour">jours</w>
  <w pos="PONCT" lefffLemma="," spacyLemma=",">,</w>
  <w pos="CLS" lefffLemma="il" spacyLemma="il">il</w>
```

Text structure: tokens and annotation

Some popular formats include :

- XML
- TSV
- CSV
- JSON

```
<s>
Да  да  CONJ
,   ,   c
мне я   SPRO    ед|1-л  дат|пр
все все SPRO    мн  вин|им
еще еще ADV
кажется кажется ADV вводн       1_знать_говорить_становиться    3_знать
_знать_жить_говорить
,   ,   c
что что CONJ
это это SPRO    ед|сред|неод    вин|им
только  только  PART
сон сон S   муж|неод    вин|ед|им   16_спать_ночь_сон   31_спать_ночь_у
паться
,   ,   c
но  но  CONJ
—   —   c
увы увы INTJ            12_глаз_лицо_девушка    20_гость_рука_очень 25_
!   !   c
—   —   c
это это SPRO    ед|сред|неод    вин|им
не  не  PART
сон сон S   муж|неод    вин|ед|им   16_спать_ночь_сон   31_спать_ночь_у
паться
!.. !.. c
—   —   c
Петербург   петербург   S   гео|муж|неод    вин|ед|им   69_поезд_вагон_
н_станция 211_господин_гимназия_гимназист
!   !   c
—   —   c
раздался    раздаваться V   нп  ед|изъяв|муж|прош|сов   44_голос_петь_п
рона  22_голос_слышать_звук
за  за  PR
моей    мой APRO        дат|ед|жен|пр|род|твор
спиной  спина   S   жен|неод    ед|твор 12_глаз_лицо_девушка    10_бежа
ать_рука_кричать
голос   голос   S   муж|неод    вин|ед|им   44_голос_петь_песня 10_бежа
ос_слышать_звук
кондуктора  кондуктор   S   муж|од  вин|ед|им|мн|род    69_поезд_вагон_
н_станция 207_вагон_поезд_паровоз
```

Sources and derivatives

Source files those files that serve as your primary data.

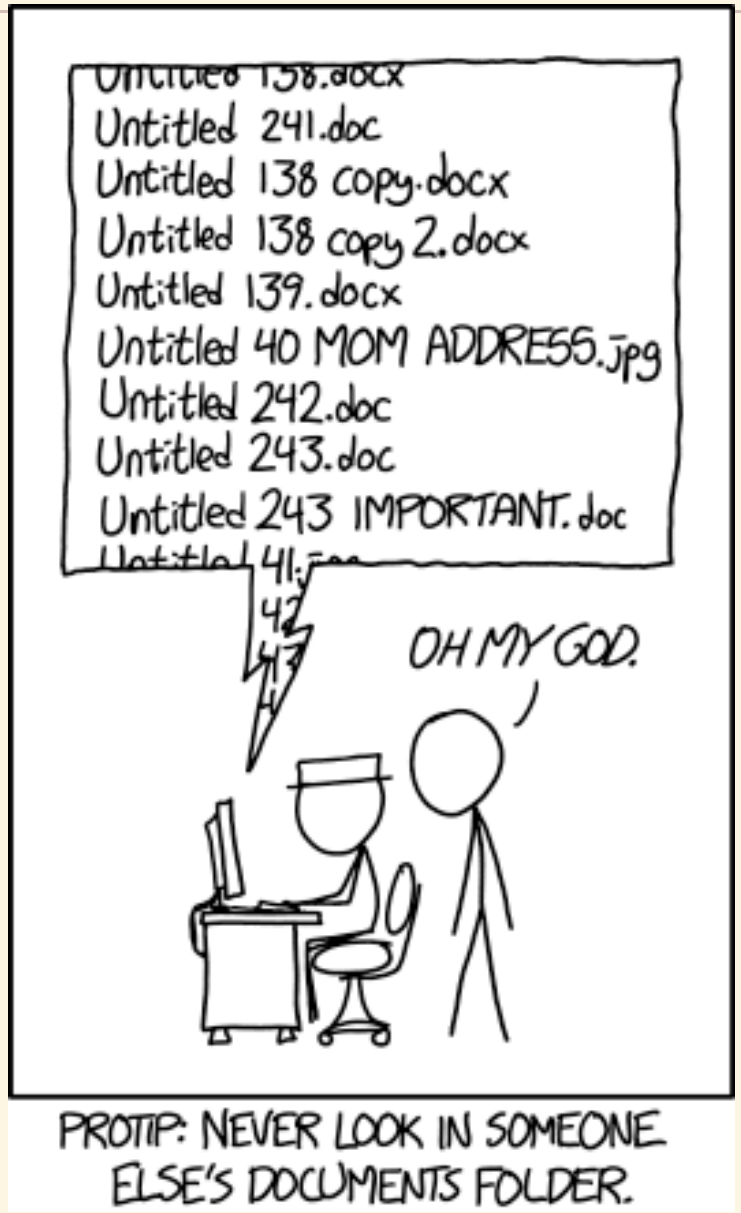
Derivative files those files that result from running some tools/algorithms on text, and are required to build a corpus.

Rules of thumb

1. You keep your source files under version control (git) in your corpus repository. They're precious. You fix all OCR issues and other problems with the source representation here.
2. You don't keep derivative files, unless the computation needed to produce them is too expensive. They're dispensible. You re-generate them when something either in your sources or in your tools change.

Files in order

- You need a file naming scheme and probably some system of folders.
- And version control, even if you don't know what it is. Enter git.



Annotating text

A classic NLP pipeline

Tokenization Raw text → list of tokens.

PoS tagging Token (in context) → part of speech tag

Lemmatization Token (in context) → lemma, a normalized (dictionary) form.

Parsing Sentence (list of tokens) → syntactic tree, implying a syntactic role for each word.

Higher level processing named entities extraction, coreference resolution, sentiment analysis etc.

Tokenization is not trivial

- `qu'il` — one token or two?
- `Y a-t-il` — how many tokens
- `Allons-nous` — is dash a part of any token? which one?
- `?..` — one or three?
- `))` — one or two?
- `A+` — what words do we have here? How to tokenize?

Lemmatization is useful

Subjunctive

Present

que je **sache**
que tu **saches**
qu'il **sache**
que nous **sachions**
que vous **sachiez**
qu'ils **sachent**

Past

que j'aie **su**
que tu aies **su**
qu'il ait **su**
que nous ayons **su**
que vous ayez **su**
qu'ils aient **su**

Imperfect

que je **susse**
que tu **susses**
qu'il **sût**
que nous **sussions**
que vous **sussiez**
qu'ils **sussent**

Pluperfect

que j'eusse **su**
que tu eusses **su**
qu'il eût **su**
que nous eussions **su**
que vous eussiez **su**
qu'ils eussent **su**

Conditional

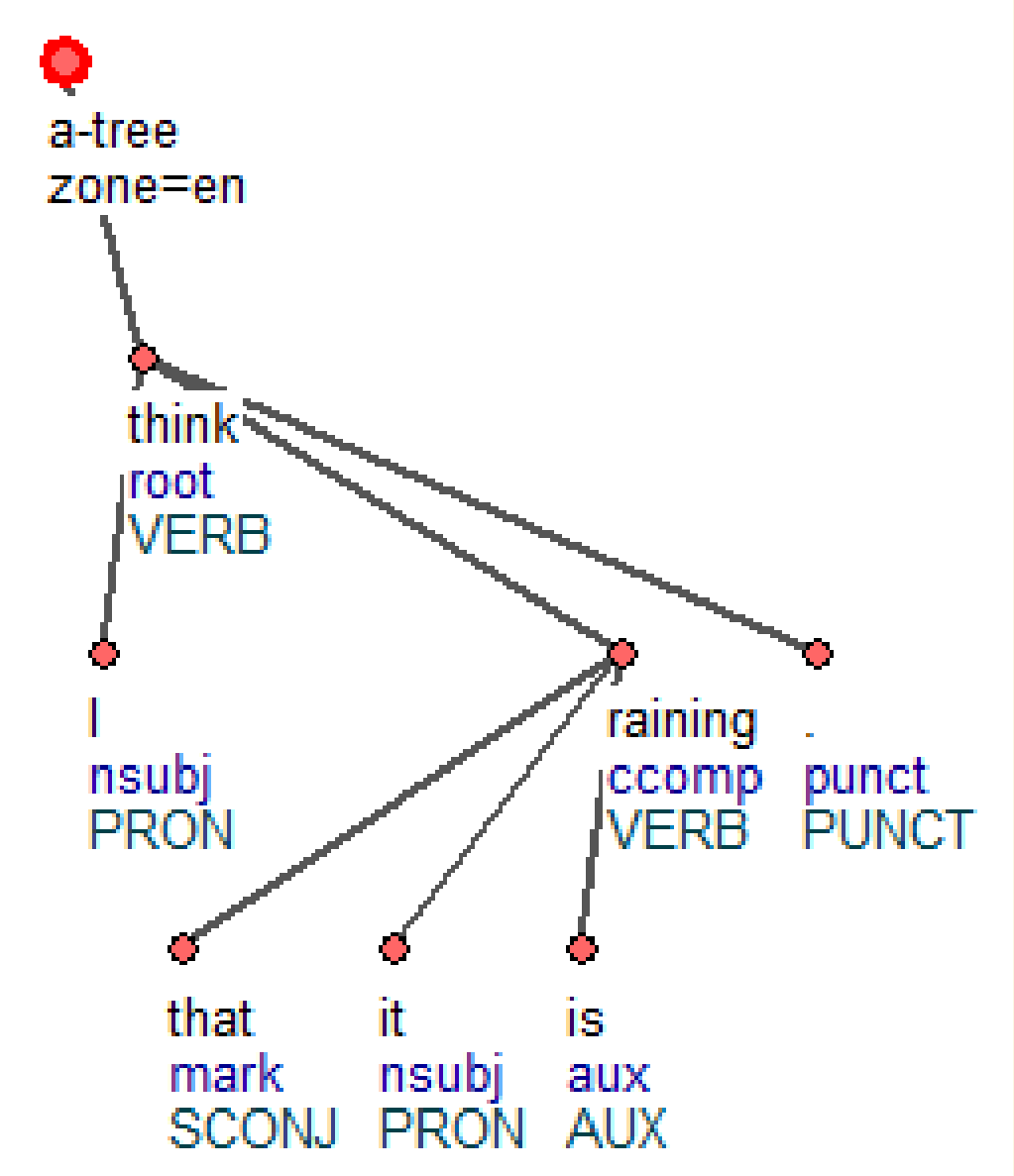
Present

je **saurais**
tu **saurais**
il **saurait**
nous **saurions**
vous **sauriez**
ils **sauraient**

Past

j'aurais **su**
tu aurais **su**
il aurait **su**
nous aurions **su**
vous auriez **su**
ils auraient **su**

Syntax serves those in the know



Syntax serves those in the know

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC
# sent_id = ParisStories_2019_cuisineApproximative_1
# text = donc comment je fais les gougères ?
# sound_url = https://api.nakala.fr/data/10.34847/nkl.d6eb01m0/161acd44d78cab860706b76ff16515e5918b84a3
1	donc	donc	ADV	_	_	4	advmod	_	_
2	comment	comment	ADV	_	_	4	advmod	_	_
3	je	il	PRON	_	Number=Sing|Person=1|PronType=Prs	4	nsubj	_	_
4	fais	faire	VERB	_	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin	0	root	_	_
5	les	le	DET	_	Definite=Def|Number=Plur|PronType=Art	6	det	_	_
6	gougères		gougère	NOUN	_	Gender=Fem|Number=Plur	4	obj	_	_
7	?	?	PUNCT	_	_	4	punct	_	_

# sent_id = ParisStories_2019_cuisineApproximative_2
# text = bah d'abord, enfin je suis pas quelqu'un qui fait euh, des recettes très très carrées etc.
# sound_url = https://api.nakala.fr/data/10.34847/nkl.d6eb01m0/161acd44d78cab860706b76ff16515e5918b84a3
1	bah	bah	INTJ	_	_	9	discourse	_	_
2	d'	de	ADP	_	_	9	advmod	_	ExtPos=ADV|Idiom=Yes|SpaceAfter=No
3	abord	abord	NOUN	_	Gender=Masc|Number=Sing	2	fixed	_	InIdiom=Yes|SpaceAfter=No
4	,	,	PUNCT	_	_	2	punct	_	_
5	enfin	enfin	ADV	_	_	9	discourse	_	ExtPos=INTJ
6	je	il	PRON	_	Number=Sing|Person=1|PronType=Prs	9	nsubj	_	_
7	suis	être	AUX	_	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin	9	cop	_	_
8	pas	pas	ADV	_	Polarity=Neg	9	advmod	_	_
9	quelqu'un		quelqueun	PRON	_	_	0	root	_	_
10	qui	qui	PRON	_	_	11	nsubj	_	_
11	fait	faire	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	9	acl:relcl	_	_
12	euh	euh	INTJ	_	_	15	discourse	_	SpaceAfter=No
```

A selection of NLP tools

- TreeTagger (C++?) — ye olde tool, PoS tagger, lemmatizer; fast
- SpaCy (python) — many languages, PoS tagger, lemmatizer; slow
- Stanza (python) — many languages, PoS tagger, lemmatizer, syntactical parsing; slow
- StanfordNLP (java) — several languages, PoS tagger, lemmatizer, syntax, named entities; fast

NLP models wax and wane

MA

In the following paragraph from a novel in French, find all the mentioned named entities: people (fictional characters), places, and organizations. Return the passage with inserted markup tags spanning the named entities. Use xml attribute "type" on a tag to indicate the type of a named entity (PERS, LOC, ORG). Make sure the returned result is a valid XML. The passage follows:

NLP models wax and wane

<?xml version="1.0" encoding="UTF-8"?>
Dambara découvrit dans un large sourire sa dentition éclatante et saine, puis chargea sur ses propres épaules le bagage de son maître en plus» du sien. La marche reprit à travers l'enchevêtrement des lianes et des plantes grimpantes sous lequel la piste disparaissait complètement.
Raynal prospectait la partie orientale de l'<namedEntity type="LOC">Ouban-gui</namedEntity> pour le compte d'une exploitation forestière de <namedEntity type="LOC">Brazzaville</namedEntity>. Depuis plusieurs années, il parcourait l'Afrique Equatoriale, reconnaissant la richesse des terrains boisés acquis par sa <namedEntity type="ORG">Compagnie</namedEntity>, étudiant les possibilités de leur mise en valeur. Il connaissait donc la brousse, ses pièges, ses dangers naturels.
Pour vaincre ces derniers, <namedEntity type="PERS">Dambara</namedEntity> lui était d'un grand secours. Le nègre, originaire des forêts du Congo, comprenait à merveille les idiomes et les mœurs des tribus de l'intérieur. Maintes fois, il avait sauvé son maître de situations périlleuses, faisant preuve d'une intelligence et d'un sang-froid rares chez les hommes de sa race.

Still, there's always TEI

```
<div xml:id="n26" type="chapitre" n="23">
 <fw place="top-left" type="pageNum">164</fw>
 <head>
  <lb/>Comment <persName>Panurge</persName> faict discours
 <lb/>pour retourner a <persName>Raminagro-
  <lb rend="hyphen"/>bis</persName>.<space unit="cm" quantity="1.5"/>Chap. 23.
 </head>
 <p>
  <lb/>
  <hi rend="larger">R</hi>Etournons (dist <persName>Panurge</persName> continuant)
 <lb/>l'admonester de son salut. Allons on
 <lb/>nom, allons en la vertus Dieu. Ce
 <lb/>sera oeuvre charitable a nous faicte. Au
 <lb/>moins s'il perd le corps &amp; la vie, qu'il
 <lb/>ne damne son <sic>asne</sic>. Nous le induirons
 <lb/>a contrition de son peché: a requerir par-
 <lb rend="hyphen"/>don es dictz tant beatz peres absens com-
 <lb rend="hyphen"/>me praesens. Et en prendrons acte, affin
 <lb/>qu'apres son trespas ilz ne le declairent
 <lb/>hereticque &amp; damné comme les Farfa-
 <lb rend="hyphen"/>detz feirent de la praevoste d'<placeName type="ville">Orleans</placeName>: &amp;
 <lb/>leurs satisfaire de l'oultraige, ordon-
 <lb rend="hyphen"/>nant par tous les couvens de ceste pro-
 <lb rend="hyphen"/>vince aux bons peres religieulx force bri-
 <lb rend="hyphen"/>bes, force messes, force obitz &amp; anni-
 <lb rend="hyphen"/>versaires. Et que au jour de son trespas
 <lb/>sempiternellement ilz ayent tous quin-
 <lb rend="hyphen"/>tuple pitance: &amp; que le grand bourra-

 <pb n="173" xml:id="B372616101_3537_0173"/>
  <fw place="top-right" type="pageNum">165</fw>
  <lb rend="hyphen"/>baquin plein du meilleur trote de ranco
 <lb/>par leurs tables, tant des Burgotz, Layz,
 <lb/>&amp; Briffaulx, que des presbtres &amp; des
 <lb/>clercs: tant des novices, que des profes.
 <lb/>Ainsi pourra il de Dieu pardon avoir.
 </p>
```

CQL : Corpus query language

Single-word queries

Word

```
"je"
```

Token. Any token

```
[]
```

Token with a given attribute value

```
[word="je"]
```

Single-word queries

Word

```
"je"
```

Token. Any token

```
[]
```

Token with a given attribute value

```
[word="je"]
```

Single-word queries

Word

```
"je"
```

Token. Any token

```
[]
```

Token with a given attribute value

```
[word="je"]
```

Combining attribute conditions

Token matching both lemma AND PoS

```
[lefflemma="je" & pos="CLS"]
```

Token that is either a common noun (NC) OR a proper noun (NPP)

`[pos="NC" | pos="NPP"]`, or similarly `[pos="NC|NPP"]`

Combining attribute conditions

Token matching both lemma AND PoS

```
[lefflemma="je" & pos="CLS"]
```

Token that is either a common noun (NC) OR a proper noun (NPP)

`[pos="NC" | pos="NPP"]`, or similarly `[pos="NC|NPP"]`

Regex patterns

All words starting with j

```
[word="j.*"]
```

All verb forms (PoS starts with V)

```
[pos="V.*"]
```

Regex patterns

All words starting with j

```
[word="j.*"]
```

All verb forms (PoS starts with V)

```
[pos="V.*"]
```

Multiword queries

”je” followed by a verb

`"je" [pos="V"]` or, similarly `[lefflemma="je"] [pos="V"]`

All adjectives in preposition

`[pos="ADJ"] [pos="N.*"]`

Multiword queries

”je” followed by a verb

`"je" [pos="V"]` or, similarly `[lefflemma="je"] [pos="V"]`

All adjectives in preposition

`[pos="ADJ"] [pos="N.*"]`

Multiword queries with gaps (co-occurences)

je followed by savoir with a gap of 0—3 words

```
[lefffflemma="je"] []{0,3} [lefffflemma="savoir"]
```

je followed by savoir within a context window of 5 tokens

```
[lefffflemma="je"] []* [lefffflemma="savoir"] within
5
```

and the same, maybe in a reversed order

```
([lefffflemma="je"] []* [lefffflemma="savoir"]) |
([lefffflemma="savoir"] []* [lefffflemma="je"])
within 5
```

Multiword queries with gaps (co-occurences)

je followed by savoir with a gap of 0—3 words

```
[leffflemma="je"] []{0,3} [leffflemma="savoir"]
```

je followed by savoir within a context window of 5 tokens

```
[leffflemma="je"] []* [leffflemma="savoir"] within
5
```

and the same, maybe in a reversed order

```
([leffflemma="je"] []* [leffflemma="savoir"]) |
([leffflemma="savoir"] []* [leffflemma="je"])
within 5
```

Multiword queries with gaps (co-occurences)

je followed by savoir with a gap of 0—3 words

```
[lefflemma="je"] []{0,3} [lefflemma="savoir"]
```

je followed by savoir within a context window of 5 tokens

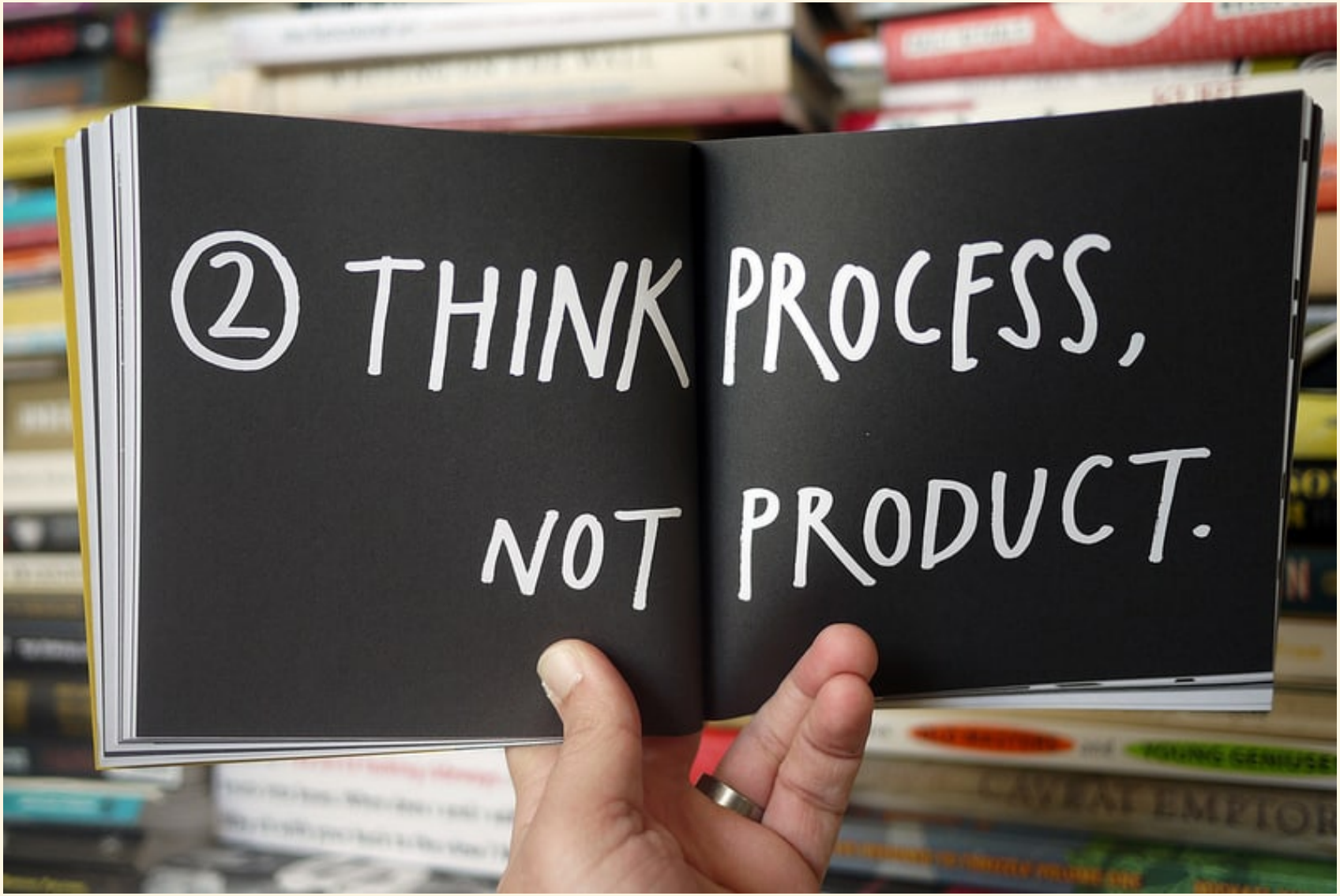
```
[lefflemma="je"] []* [lefflemma="savoir"] within
5
```

and the same, maybe in a reversed order

```
([lefflemma="je"] []* [lefflemma="savoir"]) |
([lefflemma="savoir"] []* [lefflemma="je"])
within 5
```

Running a corpus

What it takes to run a corpus



Roles in a corpus team

- Data contributor
- Data curator
- Data engineer

Reproducibility is key

- All source texts+metadata are kept in git.
- Every repetitive operation is done with code.
- All code is kept in git.
- Documenting procedures is crucial (what is built from what and how). This could be done formally, as code (Makefiles).

What we learned

- Errors are inevitable. We need a sustainable infrastructure and procedures to make sure that correcting them is simple and reliable.
- Formalized testing (python unit-tests) save a lot of time and effort for the QA.