Corpus GRAFE : Workflow

michael.nauge@univ-poitiers.fr

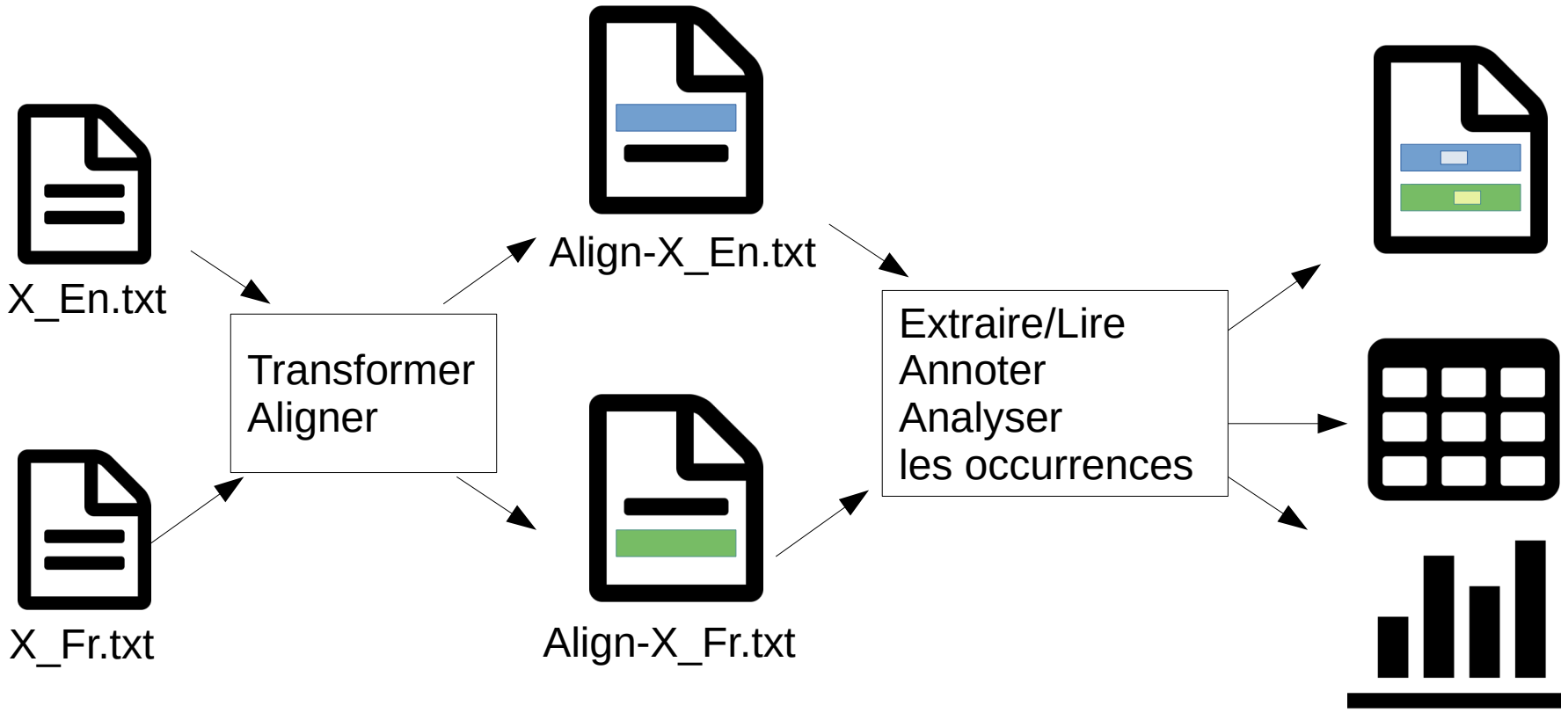
FoReLLIS
Unité de Recherche 15076
Université de Poitiers

Objectifs des corpus parallèles

Permettre la lecture et l'annotation de relevés d'occurrences de *motifs d'interêts (en contexte)*

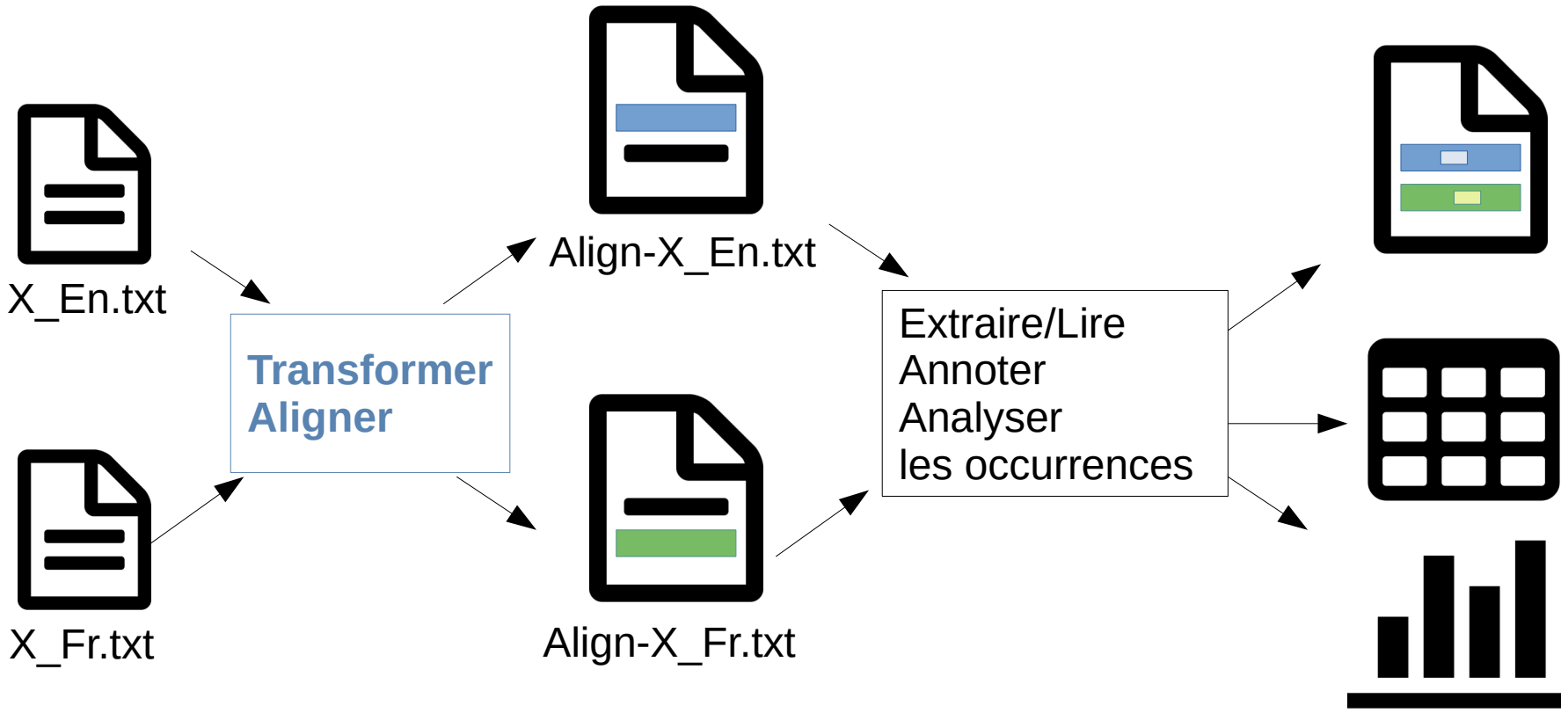
- et de ses représentations dans une autre langue
- pour des approches syntaxiques, sémantiques, lexicologiques, énonciatives et didactiques

Workflow (vue de loin)



Objectif : Maximiser le temps disponible pour l'annotation et l'analyse

Workflow (vue de loin)



Objectif : Maximiser le temps disponible pour l'annotation et l'analyse

Alinéa (input)

nouveau 1 | nouveau 2 | Comme un palimpseste.fr-it.txt | Comme un palimpseste.fr.txt

```
1  Comme un palimpseste
2
3      Je suis venu la première fois en Sicile, comme la
       plupart des voyageurs, à la recherche des vestiges
       de la civilisation grecque. Ségeste, Sélinonte,
       Agrigente, Syracuse, le théâtre en plein air de
       Taormine, les ruines désolées de Solonte : qui n' a
       rêvé à ces noms magiques ?  Entreprendre  de faire
       le tour de l'île, c'est acquitter la dette que tout
       homme occidental a contractée envers la patrie
       d'Archimède et d'Empédocle. Certes, je n'ai pas été
       déçu : en Grèce même on ne trouve aucun temple
       aussi bien conservé que celui de Ségeste, aucun
       ensemble aussi imposant que l'alignement des
       sanctuaires sur la colline d'Agrigente. Encore
       eût-il mieux valu, comme les Anciens eux-mêmes,
       aborder les côtes par la mer. La porte d'or
       incrustée d'ivoire, qui était scellée entre les
       colonnes doriques du grand temple d'Athéna à
       Syracuse, a disparu : les navigateurs, autrefois,
       apercevaient de loin son miroitement, et,
       lorsqu'ils reprenaient la mer, ils gardaient les
       yeux fixés sur ce carré d'or flamboyant, qui
       restait visible aussi longtemps que le bouclier de
       la déesse accroché au fronton du sanctuaire. Quand
       la dernière lueur s'était évanouie, ils versaient
       dans les flots une offrande de miel et d'encens qui
       mettrait leur voyage à l'abri des périls.
4      Les fables lues dans les manuels scolaires
       correspondent à une réalité encore vivante
       aujourd'hui. En Grèce on ne peut qu'admirer, en
       silence, les restes d'une civilisation morte à
       jamais ; en Sicile au contraire, il semble que tout
       ce qu'on raconte des Anciens demeure présent. Les
       paysans qui habitent sur les flancs de l'Etna, et
       qui voient des fumerolles s'élever par les
       anfractuosités de la roche noire, ou des torrents
       de lave dévaler en mugissant les pentes, comment ne
       croiraient-ils pas qu'Héphaïstos en personne
```

Comme un palimpseste.it.txt

```
1  Come un palinsesto
2
3      La prima volta sono venuto in Sicilia  come la
       maggior parte dei viaggiatori , in cerca delle
       vestigia  della civiltà  greca . Segesta ,
       Selinunte , Agrigento , Siracusa , il teatro
       all'aperto di Taormina , i ruderi desolati  di
       Solonte : chi non ha [mai] sognato  questi nomi
       magici ?  Accingersi  a fare il giro
       dell'isola, significa saldare il debito che
       tutti gli uomini occidentali hanno contratto
       nei confronti  della patria di Archimede e di
       Empedocle. Certo , non ho provato nessuna
       delusione, nella stessa Grecia  non si trova
       alcun tempio  così bene conservato  come
       quello di Segesta , nessun aggregato  tanto
       imponente quanto  l'infilata  dei santuari
       sulla collina  d'Agrigento . Anche se sarebbe
       stato preferibile, al pari degli  Antichi
       stessi, raggiungere  le coste dal mare . La
       porta aurea  incrostata  d'avorio, che era
       sigillata  tra le colonne doriche del grande
       tempio di Atena a Siracusa, è scomparsa : un
       tempo , i navigatori  scorgevano da lontano il
       suo  luccichio  e, quando riprendevano il mare,
       rimanevano con gli occhi puntati su quel
       quadrato d'oro fiammeggiante  che rimaneva
       visibile tanto a lungo quanto lo scudo della
       dea appeso al frontone  del santuario. Quando
       l'ultimo barlume svaniva, versavano sulle onde
       un'offerta di miele e d'incenso che avrebbe
       messo il loro viaggio al riparo dai pericoli.
4      Le favole  lette nei libri di testo
```

Normal text file | length : 2 944 lines : 7 | Ln : 1 Col : 1 Pos : 1 | Windows (CR LF) | ANSI | INS

Alinéa (ouput)

nouveau 1 | nouveau 2 | Comme un palimpseste.fr-it.txt | Comme un palimpseste.fr.txt

```
1  [1] Comme un palimpseste Je suis venu la première fois en Sicile , comme la plupart des voyageurs , à la recherche des
   vestiges de la civilisation grecque .
2  [1] Come un palinsesto La prima volta sono venuto in Sicilia come la maggior parte dei viaggiatori , in cerca delle
   vestigia della civiltà greca .
3
4  [2] Ségeste , Sélinonte , Agrigente , Syracuse , le théâtre en plein air de Taormine , les ruines désolées de Solonte :
   qui n' a rêvé à ces noms magiques ?
5  [2,3]   Segesta , Selinunte , Agrigento , Siracusa , il teatro all' aperto di Taormina , i ruderi desolati di Solonte :
   ¤ chi non ha [ mai ] sognato questi nomi magici ?
6
7  [3] Entreprendre de faire le tour de l' île , c' est acquitter la dette que tout homme occidental a contractée envers la
   patrie d' Archimède et d' Empédocle .
8  [4] Accingersi a fare il giro dell' isola , significa saldare il debito che tutti gli uomini occidentali hanno contratto
   nei confronti della patria di Archimede e di Empedocle .
9
10 [4] Certes , je n' ai pas été déçu : en Grèce même on ne trouve aucun temple aussi bien conservé que celui de Ségeste ,
   aucun ensemble aussi imposant que l' alignement des sanctuaires sur la colline d' Agrigente .
11 [5] Certo , non ho provato nessuna delusione , nella stessa Grecia non si trova alcun tempio così bene conservato come
   quello di Segesta , nessun aggregato tanto imponente quanto l' infilata dei santuari sulla collina d' Agrigento .
12
13 [5] Encore eût- -il mieux valu , comme les Anciens eux-mêmes , aborder les côtes par la mer .
14 [6] Anche se sarebbe stato preferibile , al pari degli Antichi stessi , raggiungere le coste dal mare .
15
16 [6] La porte d' or incrustée d' ivoire , qui était scellée entre les colonnes doriques du grand temple d' Athéna à
   Syracuse , a disparu : les navigateurs , autrefois , apercevaient de loin son miroitement , et , lorsqu' ils reprenaient
   la mer , ils gardaient les yeux fixés sur ce carré d' or flamboyant , qui restait visible aussi longtemps que le
   bouclier de la déesse accroché au fronton du sanctuaire .
17 [7,8]   La porta aurea incrostata d' avorio , che era sigillata tra le colonne doriche del grande tempio di Atena a
   Siracusa , è scomparsa :  ¤ un tempo , i navigatori scorgevano da lontano il suo luccichio e , quando riprendevano il
   mare , rimanevano con gli occhi puntati su quel quadrato d' oro fiammeggiante che rimaneva visibile tanto a lungo quanto
   lo scudo della dea appeso al frontone del santuario .
18
19 [7] Quand la dernière lueur s' était évanouie , ils versaient dans les flots une offrande de miel et d' encens qui
   mettrait leur voyage à l' abri des périls .
20 [9] Quando l' ultimo barlume svaniva , versavano sulle onde un' offerta di miele e d' incenso che avrebbe messo il loro
   viaggio al riparo dai pericoli .
21
22 [8] Les fables lues dans les manuels scolaires correspondent à une réalité encore vivante aujourd'hui .
23 [10]    Le favole lette nei libri di testo corrispondono ad una realtà ancora viva oggi
```

Normal text file | length : 6 244 lines : 46 | Ln : 3 Col : 1 Pos : 309 | Windows (CR LF) | ANSI | INS

Alinéa 3.52

2000-2006

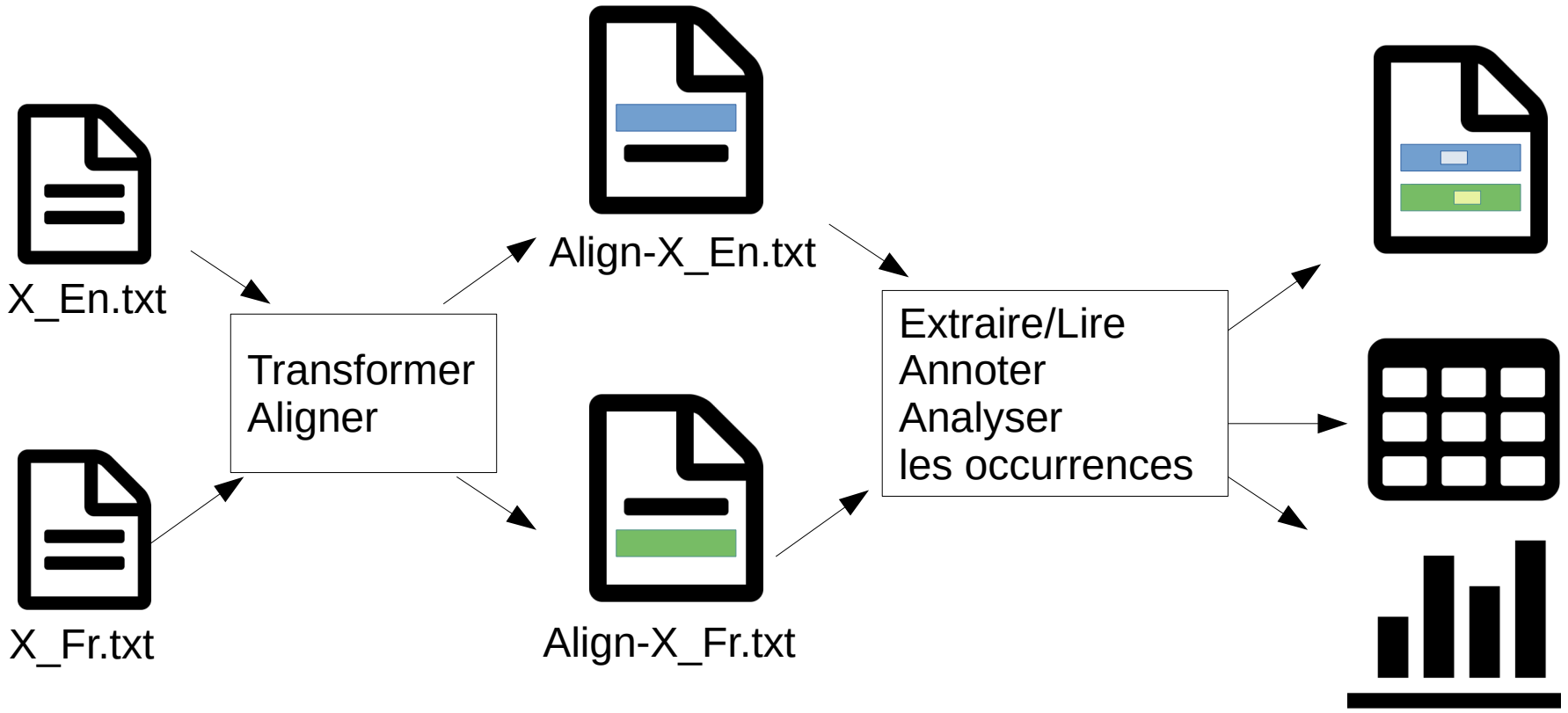
- Pr. Olivier Kraif : Laboratoire de linguistique et didactique des langues étrangères et maternelles (UGA)
- Alignement à la phrase (de deux textes au format TXT UTF-8)

  1. Extraction de points d'ancrage (pré-alignement), tokens susceptibles d'entretenir une relation d'équivalence biunivoque entre les deux textes (Nombres, Noms propres, Cognats, tout couple de tokens monosémiques)

  2. Algorithme récursif basées sur la recherche de la meilleure suite de transitions (longueur de segment), comme celles de Gale & Church (1993) ou de Brown et al. (1991)

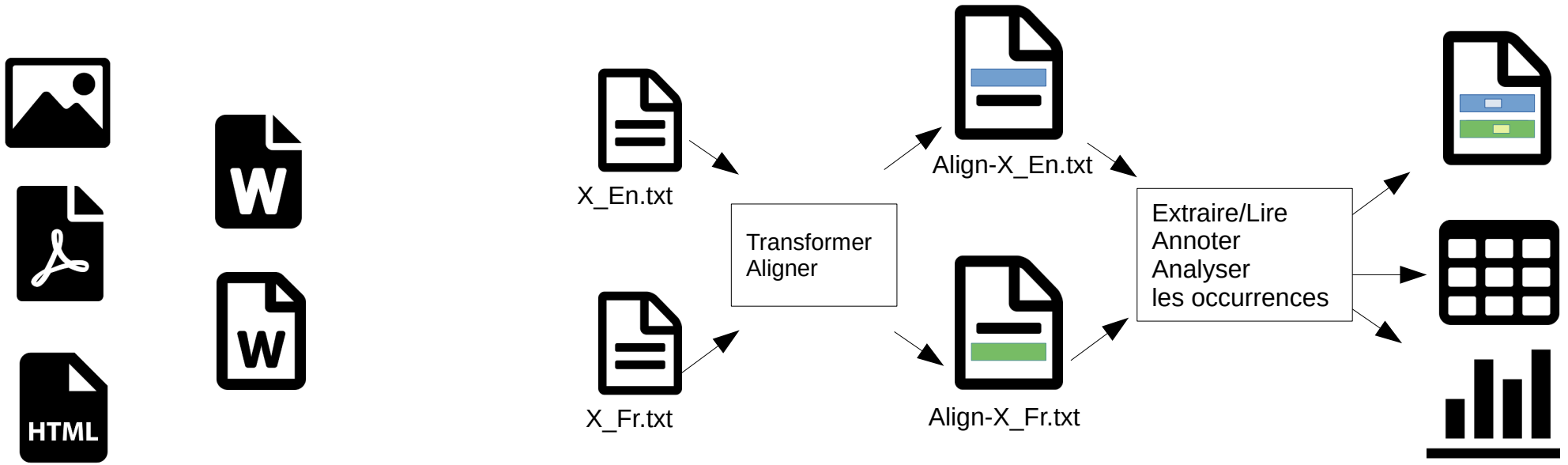
- Produit en sortie un texte
  - Fusionné, reformaté en couples de segments, avec une introduction d’espaces entre tokens

Workflow (vue de loin rappel)



Objectif : Maximiser le temps disponible pour l'annotation et l'analyse

Workflow (le début)

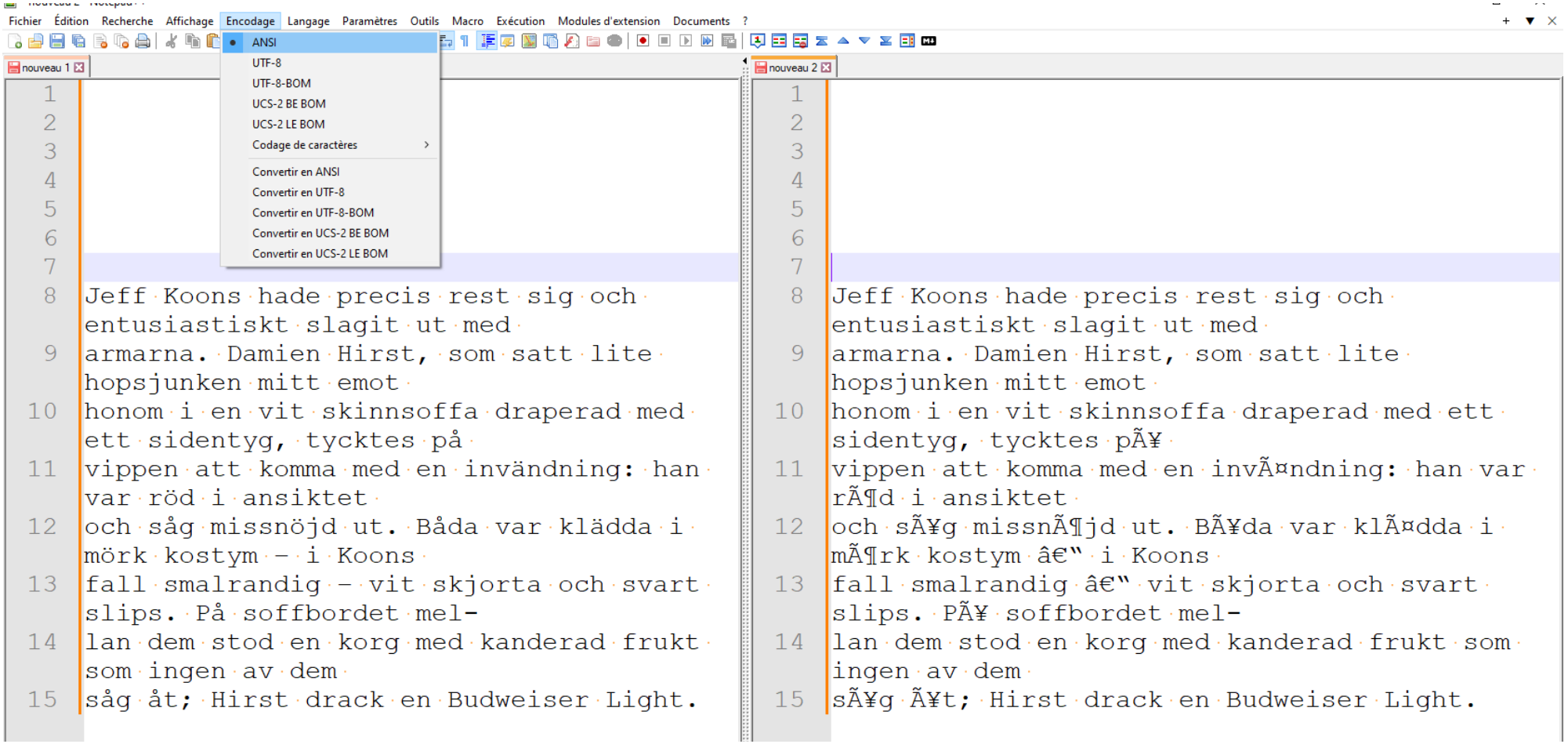


Objectif : Régulariser les sources malgrè leurs variétés, spécificités et et schémas de codage des caractères

Etapes :

- OCR/HTR, Transcodage
- Suppression entête/pied/numéro de page, commentaires…

Unicode UTF-8 / ANSI CP1252



Textes extraits d'un PDF en suedois

Schémas de codage des caractères

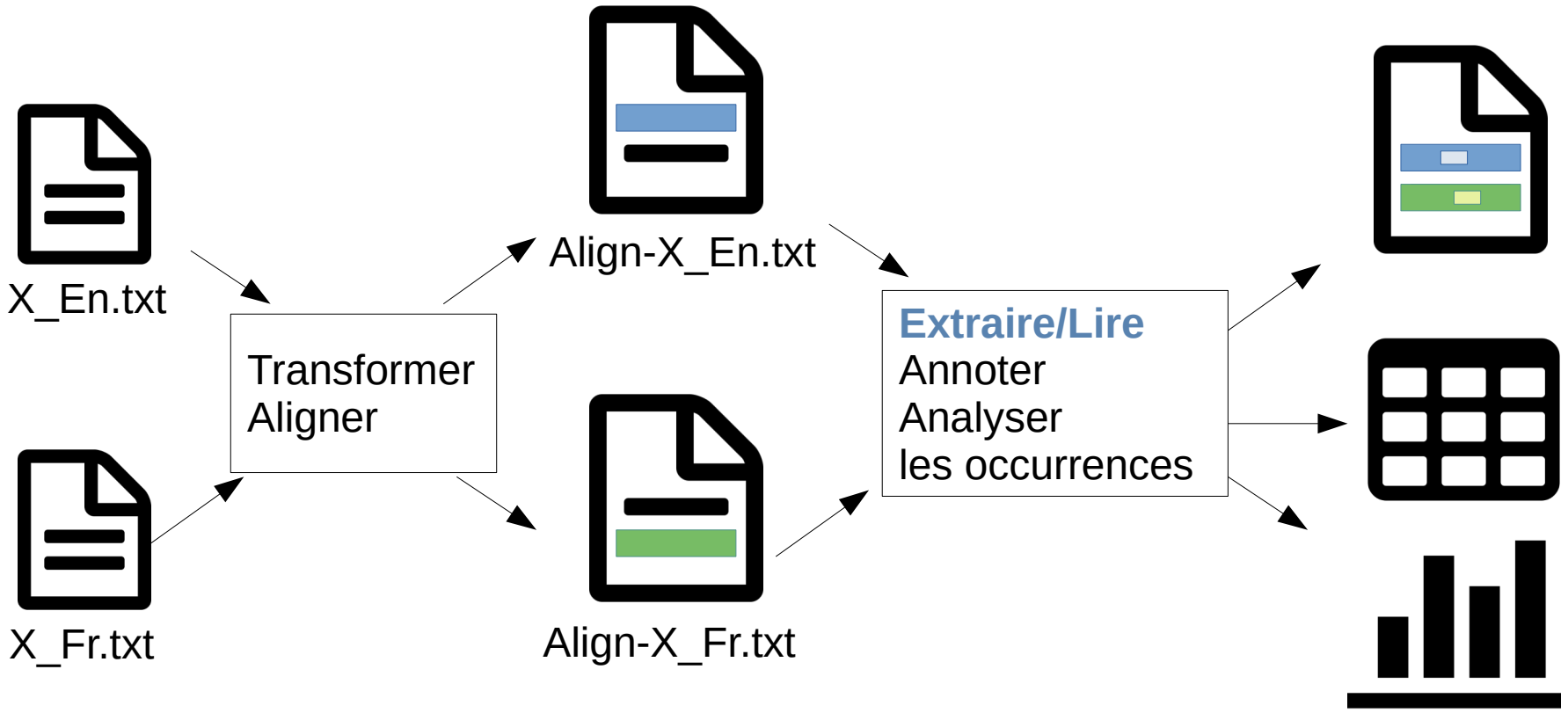
0000 = a
0001 = b
0010 = c
.
.
1111 = £

Schéma UL42

0000 = a
0001 = à
0010 = A
.
.
1111 = ç

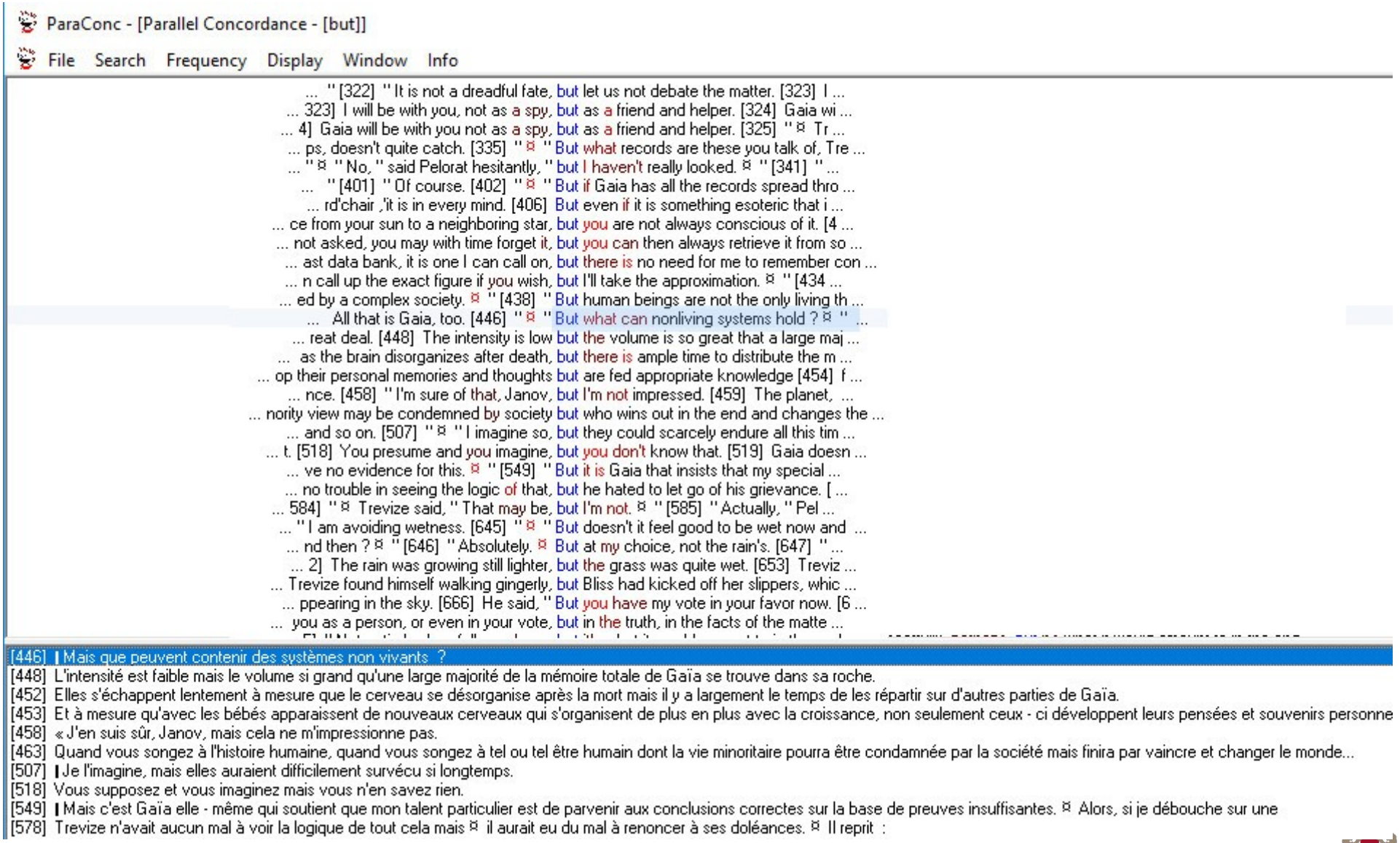
Schéma UP86

Workflow (vue de loin)



Objectif : Maximiser le temps disponible pour l'annotation et l'analyse

ParaConc (beta 1.0 - 2003)



Capture d'écran ParaConc

ParaConc (input) ANSI CP1252

Beigbeder-Francs_ES-ANSI.txt

```
7  [10] ⟶ Todo es provisional :
8  [11] ⟶ el amor, el arte, el planeta Tierra, vosotros, yo.
9  [12] ⟶ La muerte es algo tan ineludible que pilla a todo el mundo por sorpresa.
10 [13] ⟶ ¿Cómo saber si este día no será el último ?
11 [14] ⟶ Creemos tener tiempo.
12 [15] ⟶ Y luego, de repente, ya está, nos ahogamos, fin del tiempo reglamentario.
13 [16] ⟶ La muerte es la única cita que no está anotada en nuestra agenda.
14 [17] ⟶ Todo se compra :
15 [18] ⟶ el amor, el arte, el planeta Tierra, vosotros, yo.
16 [19] ⟶ Escribo este libro para que me echen del trabajo.
17 [20] ⟶ Si me fuese, me quedaría sin indemnización.
18 [21] ⟶ Necesito serrar la rama sobre la que se asienta mi comodidad.
19 [22] ⟶ Mi libertad se llama subsidio de desempleo.
20 [23] ⟶ Prefiero ser despedido por una empresa que por la vida.
21 [24] ⟶ PORQUE TENGO MIEDO.
```

Beigbeder-Francs_FR-ANSI.txt

```
7  [10] ⟶ Tout est provisoire :
8  [11] ⟶ l'amour, l'art, la planète Terre, vous, moi.
9  [12] ⟶ La mort est tellement inéluctable qu'elle prend tout le monde par surprise.
10 [13] ⟶ Comment savoir si cette journée n'est pas la dernière ?
11 [14] ⟶ On croit qu'on a le temps.
12 [15] ⟶ Et puis, tout d'un coup, ça y est, on se noie, fin du temps réglementaire.
13 [16] ⟶ La mort est le seul rendez - vous qui ne soit pas noté dans votre organizer.
14 [17] ⟶ Tout s'achète :
15 [18] ⟶ l'amour, l'art, la planète Terre, vous, moi.
16 [19] ⟶ J'écris ce livre pour me faire virer.
17 [20] ⟶ Si je démissionnais, je ne toucherais pas d'indemnités.
18 [21] ⟶ Il me faut scier la branche sur laquelle mon confort est assis.
19 [22] ⟶ Ma liberté s'appelle assurance chômage.
20 [23] ⟶ Je préfère être licencié par une entreprise que par la vie.
21 [24] ⟶ CAR J'AI PEUR.
22 [25] ⟶ Autour de moi, les collègues
```

Alinéa (ouput rappel) Unicode UTF-8

nouveau 1 | nouveau 2 | Comme un palimpseste.fr-it.txt | Comme un palimpseste.fr.txt

```
1  [1] Comme un palimpseste Je suis venu la première fois en Sicile , comme la plupart des voyageurs , à la recherche des
   vestiges de la civilisation grecque .
2  [1] Come un palinsesto La prima volta sono venuto in Sicilia come la maggior parte dei viaggiatori , in cerca delle
   vestigia della civiltà greca .
3
4  [2] Ségeste , Sélinonte , Agrigente , Syracuse , le théâtre en plein air de Taormine , les ruines désolées de Solonte :
   qui n' a rêvé à ces noms magiques ?
5  [2,3]   Segesta , Selinunte , Agrigento , Siracusa , il teatro all' aperto di Taormina , i ruderi desolati di Solonte :
   ¤ chi non ha [ mai ] sognato questi nomi magici ?
6
7  [3] Entreprendre de faire le tour de l' île , c' est acquitter la dette que tout homme occidental a contractée envers la
   patrie d' Archimède et d' Empédocle .
8  [4] Accingersi a fare il giro dell' isola , significa saldare il debito che tutti gli uomini occidentali hanno contratto
   nei confronti della patria di Archimede e di Empedocle .
9
10 [4] Certes , je n' ai pas été déçu : en Grèce même on ne trouve aucun temple aussi bien conservé que celui de Ségeste ,
   aucun ensemble aussi imposant que l' alignement des sanctuaires sur la colline d' Agrigente .
11 [5] Certo , non ho provato nessuna delusione , nella stessa Grecia non si trova alcun tempio così bene conservato come
   quello di Segesta , nessun aggregato tanto imponente quanto l' infilata dei santuari sulla collina d' Agrigento .
12
13 [5] Encore eût- -il mieux valu , comme les Anciens eux-mêmes , aborder les côtes par la mer .
14 [6] Anche se sarebbe stato preferibile , al pari degli Antichi stessi , raggiungere le coste dal mare .
15
16 [6] La porte d' or incrustée d' ivoire , qui était scellée entre les colonnes doriques du grand temple d' Athéna à
   Syracuse , a disparu : les navigateurs , autrefois , apercevaient de loin son miroitement , et , lorsqu' ils reprenaient
   la mer , ils gardaient les yeux fixés sur ce carré d' or flamboyant , qui restait visible aussi longtemps que le
   bouclier de la déesse accroché au fronton du sanctuaire .
17 [7,8]   La porta aurea incrostata d' avorio , che era sigillata tra le colonne doriche del grande tempio di Atena a
   Siracusa , è scomparsa :  ¤ un tempo , i navigatori scorgevano da lontano il suo luccichio e , quando riprendevano il
   mare , rimanevano con gli occhi puntati su quel quadrato d' oro fiammeggiante che rimaneva visibile tanto a lungo quanto
   lo scudo della dea appeso al frontone del santuario .
18
19 [7] Quand la dernière lueur s' était évanouie , ils versaient dans les flots une offrande de miel et d' encens qui
   mettrait leur voyage à l' abri des périls .
20 [9] Quando l' ultimo barlume svaniva , versavano sulle onde un' offerta di miele e d' incenso che avrebbe messo il loro
   viaggio al riparo dai pericoli .
21
22 [8] Les fables lues dans les manuels scolaires correspondent à une réalité encore vivante aujourd'hui .
23 [10]    Le favole lette nei libri di testo corrispondono ad una realtà ancora viva oggi
```

Normal text file | length : 6 244 lines : 46 | Ln : 3 Col : 1 Pos : 309 | Windows (CR LF) | ANSI | INS

Script de tranformation

- Permettre la compatibilité entre
  - la sortie d' Alinéa et
  - les entrées de ParaConc

```
193  def transformAndConvertFromAlinea():
194      # 1-nettoyage des espaces en trop
195      dirAlignMerge = './alignéFusionné/'
196      dirAlignMergeSpaceClean = './alignéFusionné/spaceCleaning/'
197      massiveTxtReplace(dirAlignMerge, dirAlignMergeSpaceClean)
198      #-----------------------------------------------------------
199
200      # 2-separation en plusieurs fichiers
201      # on parcours tous les fichiers nettoyés
202      dirAlignSplit = './alignéSéparé/TXT-utf8/'
203      listPathTXT = glob.glob(dirAlignMergeSpaceClean+'*.txt')
204
205
206      for txtCur in listPathTXT:
207          head, filename = os.path.split(txtCur)
208
209          # on decoupe le nom du fichier fusionné pour obtenir 2 noms de fichiers
210          aa, bb = filename.split('English-')
211          print(filename, ' >> ',aa+'English.txt', ' & ', bb)
212          outfile1 = dirAlignSplit+aa+'English.txt'
213          outfile2 = dirAlignSplit+bb
214
215          csplit.splitAlign(txtCur, outfile1, outfile2)
216
217      #-----------------------------------------------------------
218      # 3-Changement de formats (txt-utf8 vers txt-ansi (pour paraconc) et
219      # txt-utf8 vers docx pour utilisateurs inconditionnelles de word
220      dirAlignSplitAnsi = "./alignéSéparé/TXT-ansi/"
221      listPathTXT = glob.glob(dirAlignSplit+'*.txt')
222
223      dirAlignSplitDocx = "./alignéSéparé/DOCX/"
224
225      for txtCur in listPathTXT:
226          head, filename = os.path.split(txtCur)
227
228          outputfile = dirAlignSplitAnsi+filename
229          convertTxtUtf8ToANSI(txtCur, outputfile)
230
231          outputfileDocx = dirAlignSplitDocx+filename.replace('.txt','.docx')
232          convertTxtUtf8ToDocxUtf8(txtCur, outputfileDocx)
```

ParaConc (limites)

- Format d'entrée ANSI CP1252
  - Nouvelle version (UTF-16)
- Pas d'information sur le fichier source lors du parcours des occurrences (ni à l'export)
- Bug/Limite sur la taille du contexte aligné
- Pas de possibilité d'annoter directement les occurrences relevées
- Format d'export adapté à la lecture mais peu adapté à l'annotation
- Code source fermé

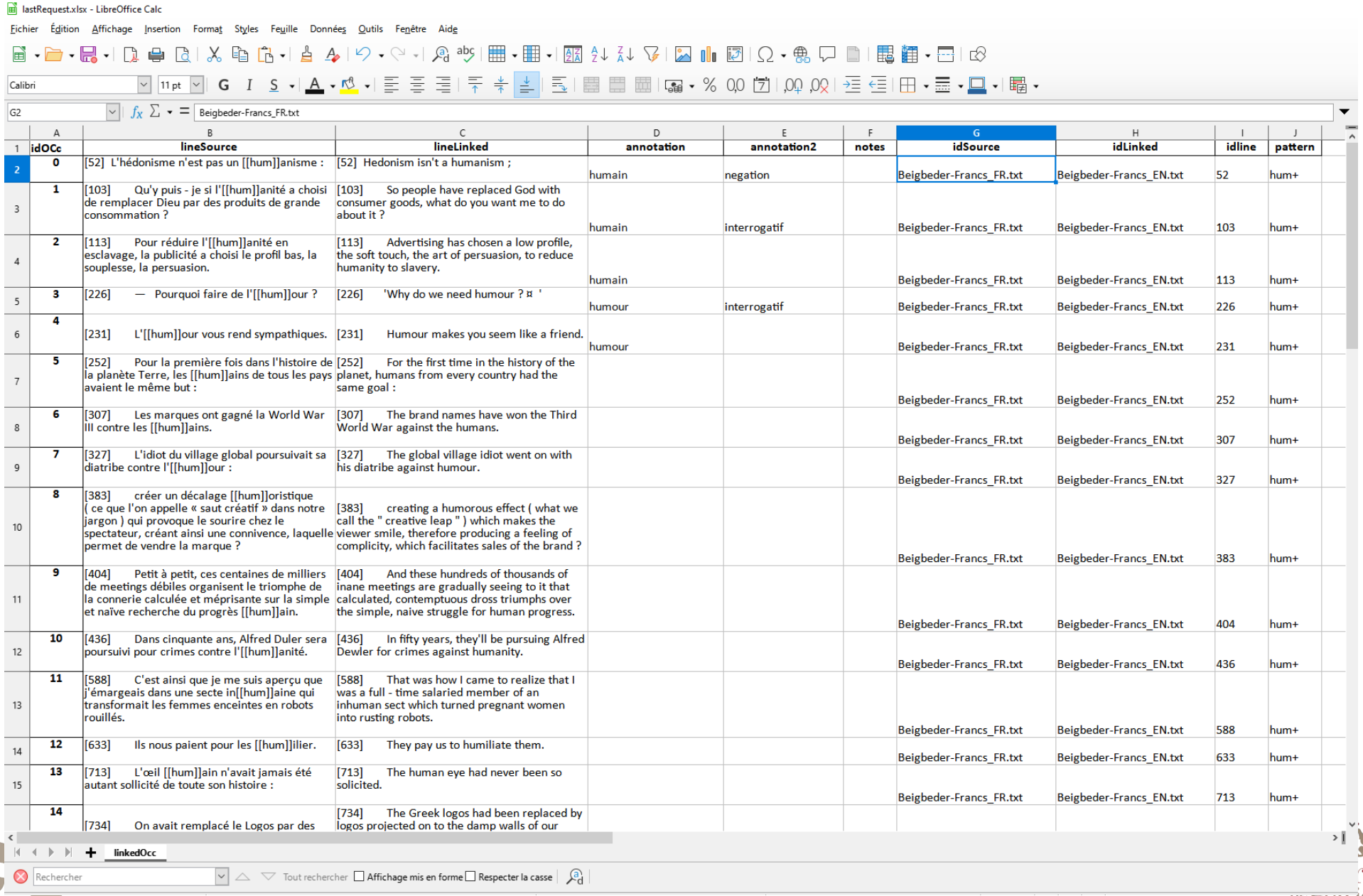
https://paraconc.com/

ParaConc (export)

```
10 [48]  Or [[on]] arrive à conquérir la clarté, à se comprendre, et c'est cela l'activité de langage.
11 [48]  If we were "pre - harmonized", there would be no problem ;
12
13 [55]  [[On]] distinguera les sciences de la vie, comme la neurologie, des sciences de la société  ; ¤  des
   sciences de la structure, de la matière  :
14 [55]  We shall distinguish the life sciences, like neurology, and the social sciences and sciences of structure,
   of matter.
15
16 [59]  si à un moment donné, il y a une faille, [[on]] est obligé de se dire  :
17 [59]  if at any time a flaw appears, one must say :
18
19 [75]  "Moi, à partir de certaines manipulations d'énoncés, je ne sais plus de quoi vous parlez, personne ne
   dirait jamais cela", donc ça ne le satisfait pas et vous voyez que de ce point de vue là, [[on]] n'est vraiment
   pas dans la situation de quelqu'un qui à un moment donné dirait  :
20 [75]  It does not satisfy him and we can see that from this point of view, we really are not in the same
   position as someone who, at some time might say :
21
22 [76]  "J'ai 'capté' comme [[on]] peut capter un phénomène à un moment donné, j'ai analysé, je regarde une
   photographie et je suis satisfait de ce que me donne à un moment donné mon observation  ;
23 [76]  "I've 'captured it', like a phenomenon can be captured, I've analyzed it, I'm looking at a snapshot and am
   satisfied with what my observation has given me at a given moment ;
24
25 [80]  Mais ce qu'il y a de remarquable aussi, et c'est une sorte de contrepartie, c'est que, sur des phénomènes
   dont [[on]] aurait pu penser qu'ils avaient déjà été bien étudiés en détail, on s'aperçoit très souvent à partir
   d'une démarche théorisante que des tas de phénomènes que personne n'avait jamais observés apparaissent
   brusquement  ;
26 [80]  But just as remarkable is a sort of compensation : ¤  when applied to phenomena we might have thought
   studied in the fullest detail, we very often realize that a modelizing procedure makes a good many hitherto
   unobserved phenomena suddenly appear.
27
28 [80]  Mais ce qu'il y a de remarquable aussi, et c'est une sorte de contrepartie, c'est que, sur des phénomènes
   dont on aurait pu penser qu'ils avaient déjà été bien étudiés en détail, [[on]] s'aperçoit très souvent à partir
```

Université de Poitiers

Exemple de tableur d'annotation attendu



lastRequest.xlsx - LibreOffice Calc

Fichier Édition Affichage Insertion Format Styles Feuille Données Outils Fenêtre Aide

Calibri 11 pt

G2 Beigbeder-Francs_FR.txt

	A	B	C	D	E	F	G	H	I	J
1	idOCc	lineSource	lineLinked	annotation	annotation2	notes	idSource	idLinked	idline	pattern
2	0	[52] L'hédonisme n'est pas un [[hum]]anisme :	[52] Hedonism isn't a humanism ;	humain	negation		Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	52	hum+
3	1	[103] Qu'y puis - je si l'[[hum]]anité a choisi de remplacer Dieu par des produits de grande consommation ?	[103] So people have replaced God with consumer goods, what do you want me to do about it ?	humain	interrogatif		Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	103	hum+
4	2	[113] Pour réduire l'[[hum]]anité en esclavage, la publicité a choisi le profil bas, la souplesse, la persuasion.	[113] Advertising has chosen a low profile, the soft touch, the art of persuasion, to reduce humanity to slavery.	humain			Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	113	hum+
5	3	[226] — Pourquoi faire de l'[[hum]]our ?	[226] 'Why do we need humour ? ¤ '	humour	interrogatif		Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	226	hum+
6	4	[231] L'[[hum]]our vous rend sympathiques.	[231] Humour makes you seem like a friend.	humour			Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	231	hum+
7	5	[252] Pour la première fois dans l'histoire de la planète Terre, les [[hum]]ains de tous les pays avaient le même but :	[252] For the first time in the history of the planet, humans from every country had the same goal :				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	252	hum+
8	6	[307] Les marques ont gagné la World War III contre les [[hum]]ains.	[307] The brand names have won the Third World War against the humans.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	307	hum+
9	7	[327] L'idiot du village global poursuivait sa diatribe contre l'[[hum]]our :	[327] The global village idiot went on with his diatribe against humour.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	327	hum+
10	8	[383] créer un décalage [[hum]]oristique (ce que l'on appelle « saut créatif » dans notre jargon) qui provoque le sourire chez le spectateur, créant ainsi une connivence, laquelle permet de vendre la marque ?	[383] creating a humorous effect (what we call the " creative leap ") which makes the viewer smile, therefore producing a feeling of complicity, which facilitates sales of the brand ?				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	383	hum+
11	9	[404] Petit à petit, ces centaines de milliers de meetings débiles organisent le triomphe de la connerie calculée et méprisante sur la simple et naïve recherche du progrès [[hum]]ain.	[404] And these hundreds of thousands of inane meetings are gradually seeing to it that calculated, contemptuous dross triumphs over the simple, naive struggle for human progress.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	404	hum+
12	10	[436] Dans cinquante ans, Alfred Duler sera poursuivi pour crimes contre l'[[hum]]anité.	[436] In fifty years, they'll be pursuing Alfred Dewler for crimes against humanity.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	436	hum+
13	11	[588] C'est ainsi que je me suis aperçu que j'émargeais dans une secte in[[hum]]aine qui transformait les femmes enceintes en robots rouillés.	[588] That was how I came to realize that I was a full - time salaried member of an inhuman sect which turned pregnant women into rusting robots.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	588	hum+
14	12	[633] Ils nous paient pour les [[hum]]ilier.	[633] They pay us to humiliate them.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	633	hum+
15	13	[713] L'œil [[hum]]ain n'avait jamais été autant sollicité de toute son histoire :	[713] The human eye had never been so solicited.				Beigbeder-Francs_FR.txt	Beigbeder-Francs_EN.txt	713	hum+
	14	[734] On avait remplacé le Logos par des	[734] The Greek logos had been replaced by logos projected on to the damp walls of our							

linkedOcc

Rechercher Tout rechercher Affichage mis en forme Respecter la casse

"Manipuler" pour comprendre

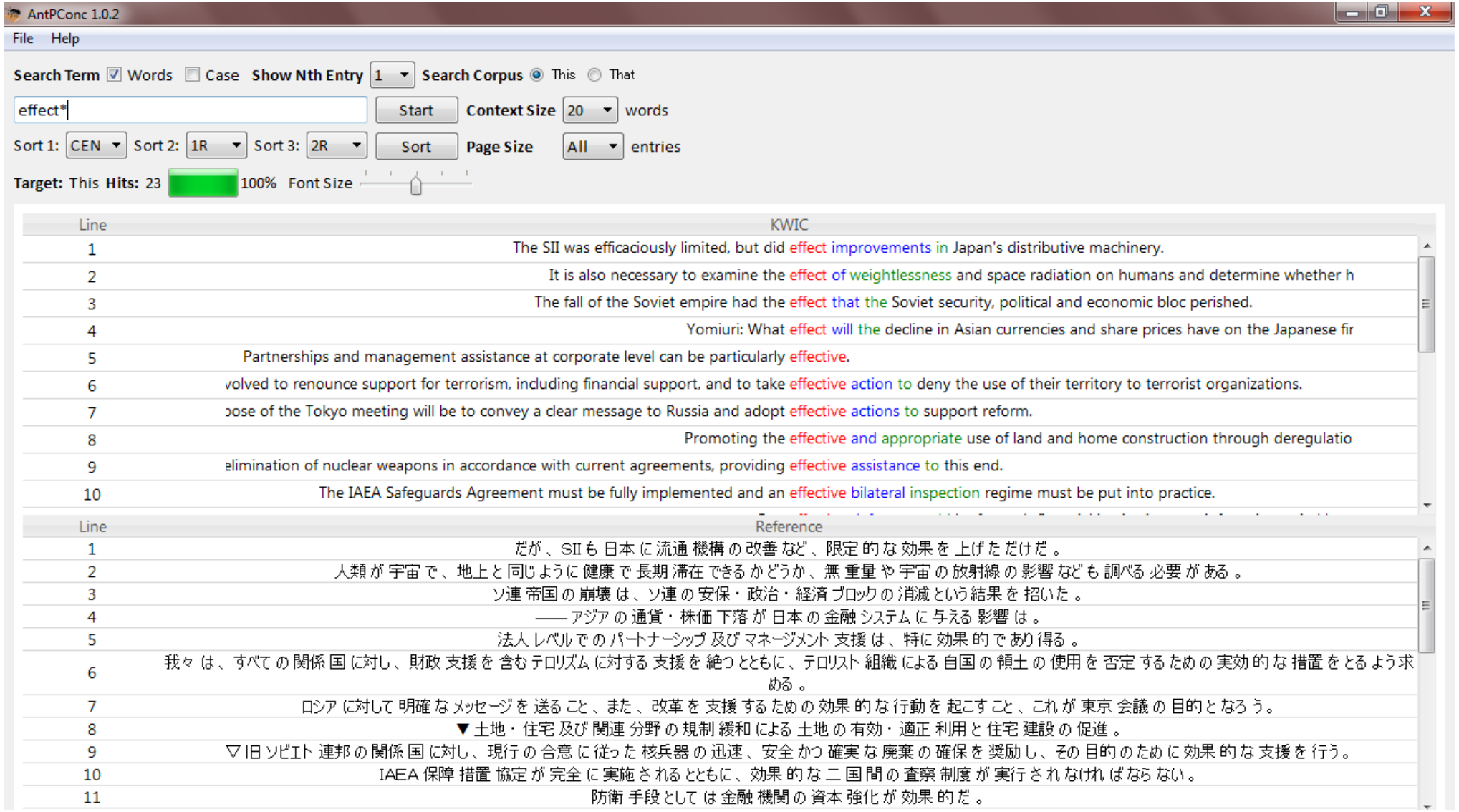
- Une matrice de donnée facilite les opérations :
  - **Grouper**/Filtrer
    - Rassembler les occurrences qui partagent une propriété commune
  - **Ordonnancer**/Trier
    - Ajouter aux occurrences une notion d'ordre/antériorité-postérité
  - **Compter**
    - Additionner les occurrences d'un groupe
    - Mesurer la corrélation de deux variables, etc...

- La granularité des variables influence la perceptibilité d'un phénomène

" Manipuler " les matrices

- Excel ou LibreOffice Calc
  - Opération simple, one shoot, quelques milliers de lignes
- OpenRefine
  - Opération de curation, contrôle qualité
- R ou Pandas (script-programmation)
  - Opération complexe, exploratoire, générative, reproductible, plusieurs centaines de milliers de lignes
- Base de données relationnel (Access, Sqlite, MySql, Oracle...)
  - Opération complexe (mais anticipée), plusieurs millions de lignes, accès concurrents

Alternative : AntPConc



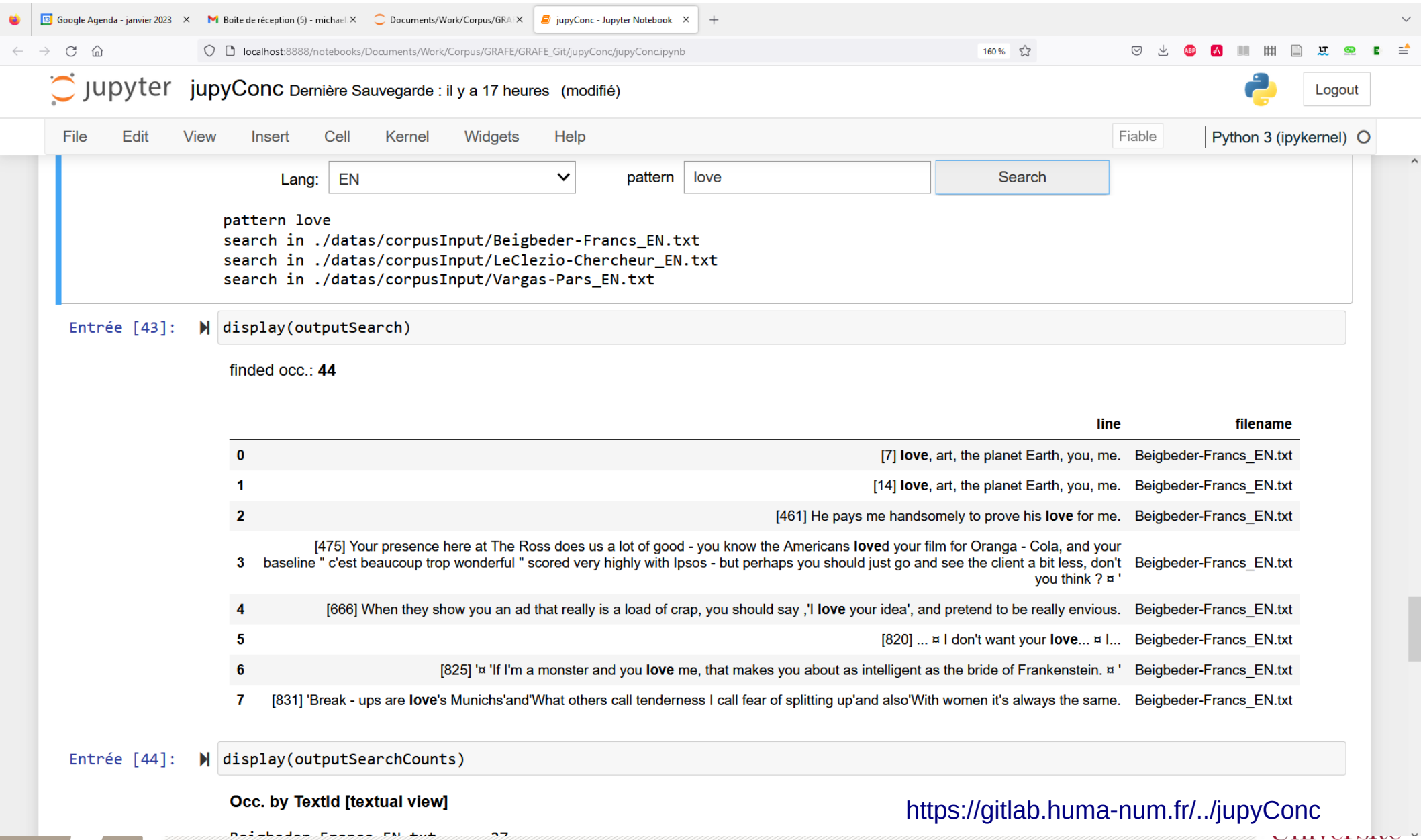
Université de Poitiers

AntPConc (limites)

- Format d'entrée ~~ANSI CP1252~~ UTF-8
- Pas d'information sur le fichier source lors du parcours des occurrences (ni à l'export)
- ~~Bug/Limite sur la taille du contexte aligné~~
- Pas de possibilité d'annoter les occurrences relevées
- Format d'export adapté à la lecture mais peu adapté à l'annotation
- Code source fermé

https://www.laurenceanthony.net/software/antpconc/

Surprise : jupyConc
(alternative lowcost maison)



https://gitlab.huma-num.fr/../jupyConc

Conclusion

- Régularisation des sources et des formats
- Alignement par Alinéa
- Transformations par scripts python
- L’outil d’exploration choisi impose des transformations additionnelles à effectuer sur les textes alignées