UNIVERSITÉ Grenoble Alpes

Journée d'étude – Tragédie grecque et Numérique
15 décembre 2017

Élaboration et évaluation d'un modèle de Treetagger pour la lemmatisation et l'étiquetage du grec ancien

Olivier Kraif

Introduction

- Objectifs
  - utiliser des ressources existantes pour « entraîner » un étiqueteur morphosyntaxique (Treetagger)
  - intégrer l'analyse du grec ancien à Anatext
  - à terme : entraîner un parseur (analyse en dépendances) pour intégrer le corpus tragique au Lexicoscope

Introduction

- Méthodologie
  - faire le point sur les ressources disponibles
  - à partir de ces ressources, constituer un lexique de formes fléchies complet
  - constituer un corpus d'entraînement et un corpus de test (étiquetés)
  - lancer l'entraînement de Treetagger
  - évaluer les résultats sur le corpus de test

Ressources

Ressources

- Perseus Dependency Treebank
  - https://github.com/PerseusDL/treebank_data
  - Annotators : Giuseppe G. A. Celano, J. F. Gentile, Robert Gorman, Vanessa Gorman, Jordan Hawkesworth, Yoana Ivanova, Tovah Keynton, Florin Leonte, Alex Lessie, Daniel Lim Libatique, Meg Luthin, Francesco Mambrini, George Matthews, Jack Mitchell, Molly Miller, Jessica Nord, Sean Stewart, Anthony D. Yates, Polina Yordanova, and Sam Zukoff.
  - Guidelines : Celano, Giuseppe G. A. 2014. Guidelines for the annotation of the Ancient Greek Dependency Treebank 2.0. https://github.com/PerseusDL/treebank_data/edit/master/AGDT2/guidelines

Ressources

- Perseus Dependency Treebank

Aesop	Fables (1.1-1.50)
Aeschylus	Agamemnon
Athenaeus	The Deipnosophists (12-13)
Diodorus Siculus	Library (11)
Herodotus	Histories (1)
Hesiod	Shield of Heracles
Homer	Iliad
Lysias	Oration 1
Plato	Euthyphro
Plutarch	Alcibiades
Polybius	Histories (1)
Pseudo Apollodorus	Library (1.1.1-1.4.1)
Pseudo Homer	Hymn to Demeter
Sophocles	Ajax
Thucydides	Histories (1)

Ressources

```
<sentence id="2" document_id="urn:cts:greekLit:tlg0003.tlg001.perseus-grc1" subdoc="1.1.2">
  <word id="1" form="κίνησις" lemma="κίνησις" postag="n-s---fn-" relation="SBJ" head="18"/>
  <word id="2" form="γὰρ" lemma="γάρ" postag="d--------" relation="AuxY" head="18"/>
  <word id="3" form="αὕτη" lemma="οὗτος" postag="a-s---fn-" relation="ATR" head="1"/>
  <word id="4" form="μεγίστη" lemma="μέγας" postag="a-s---fn-" relation="PNOM" head="8"/>
  <word id="5" form="δὴ" lemma="δή" postag="d--------" relation="AuxZ" head="4"/>
  <word id="6" form="τοῖς" lemma="ὁ" postag="l-p---md-" relation="ATR" head="7"/>
  <word id="7" form="῞Ελλησιν" lemma="῞Ελλην" postag="n-p---md-" relation="ADV_CO" head="9"/>
  <word id="8" form="ἐγένετο" lemma="γίγνομαι" postag="v3saim---" relation="PRED_CO" head="18"/>
  <word id="9" form="καὶ" lemma="καί" postag="c--------" relation="COORD" head="4"/>
  <word id="10" form="μέρει" lemma="μέρος" postag="n-s---nd-" relation="ADV_CO" head="9"/>
  <word id="11" form="τινὶ" lemma="τις" postag="a-s---nd-" relation="ATR" head="10"/>
  <word id="12" form="τῶν" lemma="ὁ" postag="l-p---mg-" relation="ATR" head="13"/>
  <word id="13" form="βαρβάρων" lemma="βάρβαρος" postag="n-p---mg-" relation="ATR" head="10"/>
  <word id="14" form="," lemma="," postag="u--------" relation="AuxX" head="15"/>
  <word id="15" form="ὡς" lemma="ὡς" postag="c--------" relation="AuxC" head="23"/>
  <word id="16" form="δὲ" lemma="δέ" postag="d--------" relation="AuxZ" head="17"/>
  <word id="17" form="εἰπεῖν" lemma="εἶπον" postag="v--ana---" relation="ADV" head="15"/>
  <word id="18" form="καὶ" lemma="καί" postag="c--------" relation="COORD" head="0"/>
  <word id="19" form="ἐπὶ" lemma="ἐπί" postag="r--------" relation="AuxP" head="24"/>
  <word id="20" form="πλεῖστον" lemma="πλεῖστος" postag="a-s---na-" relation="ADV" head="19"/>
  <word id="21" form="ἀνθρώπων" lemma="ἄνθρωπος" postag="n-p---mg-" relation="ATR" head="20"/>
  <word id="22" form="." lemma="." postag="u--------" relation="AuxK" head="0"/>
  <word id="23" insertion_id="0022e" artificial="elliptic" relation="PRED_CO" lemma="γίγνομαι" postag="v3saim---" form="ἐγένετο"
head="18"/>
  <word id="24" insertion_id="0022f" artificial="elliptic" relation="PNOM" lemma="μέγας" postag="a-s---fn-" form="μεγίστη" head="23"/>
</sentence>
```

Ressources

- Corpus LASLA

AESCAGAM.TXT	EUR_LYRI.TXT	ARISHA07.TXT	ISODIS18.TXT
AESCCHOE.TXT	EUR_MEDE.TXT	ARISHA08.TXT	ISODIS19.TXT
AESCEUME.TXT	EUR_ORES.TXT	ARISHA09.TXT	ISODIS20.TXT
AESCFRAG.TXT	EUR_PHOE.TXT	ARISHA10.TXT	ISODIS21.TXT
AESCPERS.TXT	EUR_RHES.TXT	ARISMETA.TXT	ISOLET01.TXT
AESCPROM.TXT	EUR_SUPP.TXT	ARISPOET.txt	ISOLET02.TXT
AESCSEPT.TXT	EUR_TROA.TXT	ARPHYSIQ.TXT	ISOLET03.TXT
AESCSUPP.TXT	ISODIS01.TXT	ARPOLITA.TXT	ISOLET04.TXT
AESCVARI.TXT	ISODIS02.TXT	ARPOLITB.TXT	ISOLET05.TXT
ANDTEX.TXT	ISODIS03.TXT	ARTOPI08.TXT	ISOLET06.TXT
ANTTEX.TXT	ISODIS04.TXT	EUR_ALCE.TXT	ISOLET07.TXT
ARANIMA.txt	ISODIS05.TXT	EUR_ANDR.TXT	ISOLET08.TXT
ARCATEGT.txt	ISODIS06.TXT	EUR_BACC.TXT	ISOLET09.TXT
ARGENAN1.TXT	ISODIS07.TXT	EUR_CYCL.TXT	LYSTEX.TXT
ARGENAN2.TXT	ISODIS08.TXT	EUR_ELEC.TXT	SOPHAJAX.TXT
ARGENAN3.TXT	ISODIS09.TXT	EUR_FRAG.TXT	SOPHANTI.TXT
ARGENAN4.TXT	ISODIS10.TXT	EUR_HECU.TXT	SOPHELEC.TXT
ARGENAN5.TXT	ISODIS11.TXT	EUR_HELE.TXT	SOPHFRAG.TXT
ARISHA01.TXT	ISODIS12.TXT	EUR_HERA.TXT	SOPHOEDC.TXT
ARISHA02.TXT	ISODIS13.TXT	EUR_HERC.TXT	SOPHOEDT.TXT
ARISHA03.TXT	ISODIS14.TXT	EUR_HIPP.TXT	SOPHOXFO.TXT
ARISHA04.TXT	ISODIS15.TXT	EUR_ION.TXT	SOPHPHIL.TXT
ARISHA05.TXT	ISODIS16.TXT	EUR_IPAU.TXT	SOPHTRAC.TXT
ARISHA06.TXT	ISODIS17.TXT	EUR_IPTA.TXT	TRMIN_AD.TXT

Ressources

- Corpus LASLA

```
&*#EK$ISSZ          E*#IK$ISSOVSIM          3   11085  1 14 5885 :BF
&SJIRU$AZ           SJIRU@,A                3   11085  2 15 5886  BF
&D$E                D#                      3   11085  3 16 5887  BL
&.AMELOT            #AM$ELZM               03   11085  4 17 5888  BB
&QME@VLA            QME$VLAUA               3   11086  1 18 5889  BB
&Q@AT               Q$AMUZM                 3   11086  2 19 5890  BE
&E#IT               E#IT                    3   11086  3 20 5891  BH
&#AKK$GKZM          .AKKGKA                03   11086  4 21 5892  BE
&SU$ASIT            SU$ASIM                 3   11087  1 22 5893  BB
&#AMU$IQMOOT        #AMU$IQMOVM             3   11087  2 23 5894  BC
&#AQODE$IJMVLI      #AQODEIJM$VLEMA        03   11087  3 24 5895C:BF
&SVMUAR$ASSZ        NVMUEU$ARAJUAI          3   11088  1 25 5896  BF
&D$E                D#                      3   11088  2 26 5897  BL
&A#IH$GR            A#IH*$GR                3   11088  3 27 5898  BB
&Q$OMUOT            Q$OMU,Z                 S3   11088  4 28 5899 .BB
&UOI$OSDE           UOI$AD#                 3   11089  1  1 5900  BE
&#EQ$I              #EQ#                    3   11089  2  2 5901  BH
&#EC$Z              #ELO*$I                 3   11089  3  3 5902  BE
&*#RIQ$G            *#RIQ*$G                3   11089  4  4 5903  BB
&*FE$VT             *DI$OHEM               03   11089  5  5 5904  BB
&UE$VXZ             UE$VXOVSA               3   11090  1  6 5905  BF
&W$OBOT             W$OBOM                  3   11090  2  7 5906  BB
&SUE$IXZ            SUE$IXEI                3   11090  3  8 5907  BF
```

Ressources

- Corpus LASLA

&χθών	χθονὸς	3	1	1	1	1	0B
&μέν	μὲν	3	1	1	2	2	0L
&εἰς	ἐς	3	1	1	3	3	0HM
&τηλουρός	τηλουρὸν	3	1	1	4	4	0C
&ἥκω	ἥκομεν	3	1	1	5	5	0F
&πέδον	πέδον	03	1	1	6	6	,0B
&Σκύθης	Σκύθην	3	1	2	1	7	0C
&εἰς	ἐς	3	1	2	2	8	0HN
&οἶμος	οἶμον	3	1	2	3	9	,0B
&ἄβροτος	ἄβροτον	3	1	2	4	10	0C
&εἰς	εἰς	3	1	2	5	11	0H
&ἐρημία	ἐρημίαν	S3	1	2	6	12	.0B
&Ἥφαιστος	Ἥφαιστε	3	1	3	1	1	13	,0B
&σύ	σοὶ	3	1	3	2	2	14	0E
&δέ	δὲ	3	1	3	3	3	15	0L
&χρή	χρὴ	3	1	3	4	4	16	0F
&μέλω	μέλειν	3	1	3	5	5	17	0F
&ἐπιστολή	ἐπιστολὰς	03	1	3	6	6	18	0B
&ὅς	ἅς	3	1	4	1	7	19	0E
&σύ	σοι	3	1	4	2	8	20	0E
&πατήρ	πατὴρ	3	1	4	3	9	21	0B
&ἐφίημι	ἐφεῖτο	3	1	4	4	10	22	,0F
&ὅδε	τόνδε	3	1	4	5	11	23	0E
&πρός	πρὸς	3	1	4	6	12	24	0H
&πέτρα	πέτραις	03	1	4	7	13	25	0B
&ὑψηλόκρημνος	ὑψηλοκρήμνοις	3	1	5	1	14	26	0C

Ressources

- **Eulexis**
  - http://outils.biblissima.fr/fr/eulexis/
  - lemmatiseur
  - lexiques
    - Bailly.Abr.5.1.txt (57 000 entrées)

ἀβριθής ής, ές :
qui ne pèse pas.
Étym. ἀ, βρίθω.

ἁβροβάτης ου;
adj. m.
d’allure efféminée.
Étym. ἁβρός, βαίνω.

ἁβρόβιος ος, ον :
efféminé.
Étym. ἁβρός, βίος.

Ressources

- **Eulexis**
  - LSJ-6.5.1.txt (127 600 entrées)

ἀβαρταί = πτηναί (Cypr.), Hsch.

ἀβαρύ = ὀρίγανον (Maced.), Hsch.

ἀβάς εὐήθης; also = ἱερὰ νόσος (Tarent.), Hsch.

ἀβάσαι ἀριστῆσαι, καὶ ἀρθῆναι, Hsch.

ἀβασάνιστος ον, not tortured, ἀ. θνῄσκειν J. BJ 1.32.3, cf. Plu. 2.275c; κημοῖς ὑπερῴαν ἀ. Ael. NA 13.9. Adv. -τως without pain, βλέπειν τὸν ἥλιον ib. 10.14.|| untried, unexamined, ἀ. τι ἐᾶσαι Antipho 1.13; ἀπολιπεῖν Plb. 4.75.3; παραλείπειν Plu. 2.59c.
Adv. -τως without due examination, Th. 1.20, Plu. 2.28b.

ἀβασίλευτος ον, not ruled by a king, Th. 2.80, X. HG 5.2.17; generally, free from rule, Plu. 2.1125d, Artem. 1.8.

Ressources

- **Eulexis**
  - Pape.3.1.txt (102 800 entrées)

ἀβαρής ές (βάρος), nicht schwer, leicht, Arist. coel. 1.8; Luc. Dial.Mort. 10.5; überhaupt nicht lästig, Mel. 121 (VII.461) und NT, wie 2 Cor. 11.9; τινί.

ἀβασάνιστος nicht gefoltert, Plut. qu.Rom. 44, nicht durch die Folter erforscht, Antiph. 1.13, σιωπώμενον καὶ ἀβ. ἐᾶν; überh. unersrtert, Plut. u. Sp. ἀβ. τι παραλείπειν.
Bei K.S. auch ungesucht, natürlich.
• Adv. ἀβασανίστως, οἳ ἄνθρωποι τὰς ἀκοὰς ἀβ. δέχονται, ohne genaue Prüfung, Thuc. 1.20.

ἀβασίλευτος nicht von Königen regiert, χάονες Thuc. 2.80; θρᾷκες Xen. Hell. 5.2.12; Plut. Alc. 36; – ohne König, Herodian. 4.14.1; unabhängig, πολιτεία ἀβ. καὶ αὐτόνομος Plut. Rom. 27.

ἀβασκάνιστος Plat. amat. 13, muß ἀβασάνιστος oder ἀβάσκαντος heißen.

ἀβάσκανος neidlos, Teles. Stob. 163.83; Sp. auch adv.

ἀβάσκαντον τό, Amulet gegen den Neid, Diosc..

Ressources

- **LemGreek**
  - http://www.sourceschretiennes.mom.fr/outils-recherche/lemmatiseur-textes-en-grec-ancien-lemgreek
  - association formes lemmes issues de
    - Perseus
    - https://www.bibleworks.com/

Entraînement

Entraînement

- Corpus Perseus
  - Perseus : 33 textes, 33 555 phrases, 548 611 tokens
- Corpus Lasla (étiquettes Perseus)
  - Lasla : 96 textes, 864 188 tokens

Entraînement

- Utilisation du corpus Perseus
- Corpus Lasla n'apporte pas de nouvelles formes fléchies / Perseus
    - ajoute de nouveaux lemmes pour des combinaisons forme fléchies + catégories
    - augmente l'ambiguïté

    → ressources mise de côté

Entraînement

- Corpus d'entraînement
  - 545 775 tokens
- Corpus test
  - 2 836 tokens
- Lexique de formes fléchies
  - 140 145 formes enregistrées

Entraînement

- Corpus d’entraînement

• χρόνῳ	n-s---md-
• δ᾽	g--------
• ἐν	r--------
• ὑστέρῳ	a-s---md-
• μέν	g--------
• ,	u--------
• ἀσμένῃ	a-s---fd-
• δέ	g--------
• μοι	p-s---fd-
• ,	u--------
• ὁ	l-s---mn-
• κλεινὸς	a-s---mn-
• ἦλθε	v3saia---
• Ζηνὸς	n-s---mg-
• Ἀλκμήνης	n-s---fg-
• τε	g--------
• παῖς	n-s---mn-
• .	SENT

Entraînement

- Lexique de formes fléchies
  - Αὐταρίτου n-s---mg- Αὐτά ριτος
  - Αὐτοκλέης n-s---mn- Αὐτοκλῆς
  - Αὐτολύκοιο n-s---mg- Αὐτόλυκος
  - Αὐτολύκου n-s---mg- Αὐτόλυκος
  - Αὐτομέδοντα n-s---ma- Αὐτομέδων
  - Αὐτομέδοντι n-s---md- Αὐτομέδων
  - Αὐτομέδοντος n-s---mg- Αὐτομέδων
  - Αὐτομέδων n-s---mn- Αὐτομέδων
  - Αὐτονόη n-s---fn- Αὐτονόη
  - Αὐτονόην n-s---fa- Αὐτονόη
  - Αὐτοφόνοιο n-s---mg- Αὐτόφονος
  - Αὐτόλυκ᾽ n-s---mv- Αὐτόλυκος
  - Αὐτόλυκον n-s---ma- Αὐτόλυκος
  - Αὐτόλυκος n-s---mn- Αὐτόλυκος
  - Αὐτόλυκόν n-s---ma- Αὐτόλυκος
  - Αὐτόλυκός n-s---mn- Αὐτόλυκος

Entraînement

- Codage des traits sur 7 caractères, p.ex. : n-s---mn-
  - Partie du discours : l, n, a, p, v, d, r, c, m, i, g, u, b, k
  - personne : 1, 2, 3
  - nombre : s, p, d
  - temps : p, i, f, a, ...
  - mode : p, o, i, n, ...
  - voix : a, p, e, m
  - genre : m, f, n
  - cas : n, a, g, d, v
  - superlatif : s ou -

Entraînement

- 14 étiquettes de POS (parties du discours) :

l	Article
n	Nom
a	Adjectif
p	Pronom
v	Verbe
d	Adverbe
r	Apposition
c	Conjonction
m	Numéral
i	Interjection
g	Particule
u	Ponctuation
b	Aucune
k	Inconnue

Problèmes rencontrés

Problèmes rencontrés

- Normalisation des caractères
  - diacritiques : tónos vs oxia
    - remplacement de ί par ί (U+03AF par U+1F77)
    - remplacement de ώ par ώ
    - remplacement de ά par ά
  - point médian · remplacé par .
  - apostrophe ‘ remplacée par diacritique virgule en chef (U+0313) :·

Problèmes rencontrés

- Tokenisation ad hoc
  - création d'un lexique de formes composées (9193 entrées) :
    - Αἰγιά λεια
    - Αἰγά ς
    - Αἰθά λειαν
    - Αἰθίοπά ς
    - Αἰά ντεσά
    - Αἰά ντεσσι
    - Αὐγηϊά δαο
    - Αὐδά ταν
    - Αὐτά ριτον
    - Αὐτά ριτος
    - Βαγαβά ζου
    - Βισά λται
    - …..

Évaluation

Evaluation

- Premier test: prise en compte de la partie du discours seulement
  - Tokenisation : précision = 100 %
  - Etiquettes : précision=92,8 %, rappel=92,8 %
  - Lemmatisation : précision=99 % rappel=98,7 %, 24 % de lemmes ambigus
    - p.ex. : φρενός n πρήν/φρήν

Evaluation

- Catégories plus problématiques
  - Adverbes : 61 mal étiquetés (33 particules, 20 conjonctions, 4 pronoms, …)

    p.ex. : δὲ, γὰρ, ...
  - Pronoms : 29 mal étiquetés (17 adjectifs, 12 articles)

    p. ex. : αὐτέων, αὐτοὺς, ...
  - Particules : 26 mal étiquetés (14 adverbes, 6 conjonctions, 4 pronoms, …)
  - Adjectifs : 20 mal étiquetés (13 pronoms, 6 noms,…)

    p.ex. : ταῦτα, τούτῳ

Evaluation

- Deuxième test: prise en compte de l'ensemble des traits lors de l'apprentissage
  - Tokenisation : précision = 100 %
  - Étiquettes :
    - Identité complète (tous les traits) : précision 87,2 % rappel=87,2 %
  - Lemmatisation : précision=98 % rappel=97,5 %, 17% de lemmes ambigus

Evaluation

- Évaluation par traits : cf.

/home/kraifo/Documents/Recherches/Publis/Communications/2018 Journée d'étude - tragédie/erreur.traits.log

	P	R	F
Partie du discours	93,0%	93,0%	93,0%
personne	93,3%	96,0%	94,6%
nombre	97,3%	97,9%	97,6%
mode	95,5%	97,1%	96,3%
temps	96,1%	97,8%	96,9%
voix	95,7%	98,9%	97,3%
genre	91,9%	92,8%	92,3%
cas	93,4%	94,1%	93,7%
superlatif	66,7%	80,0%	72,7%

Conclusion

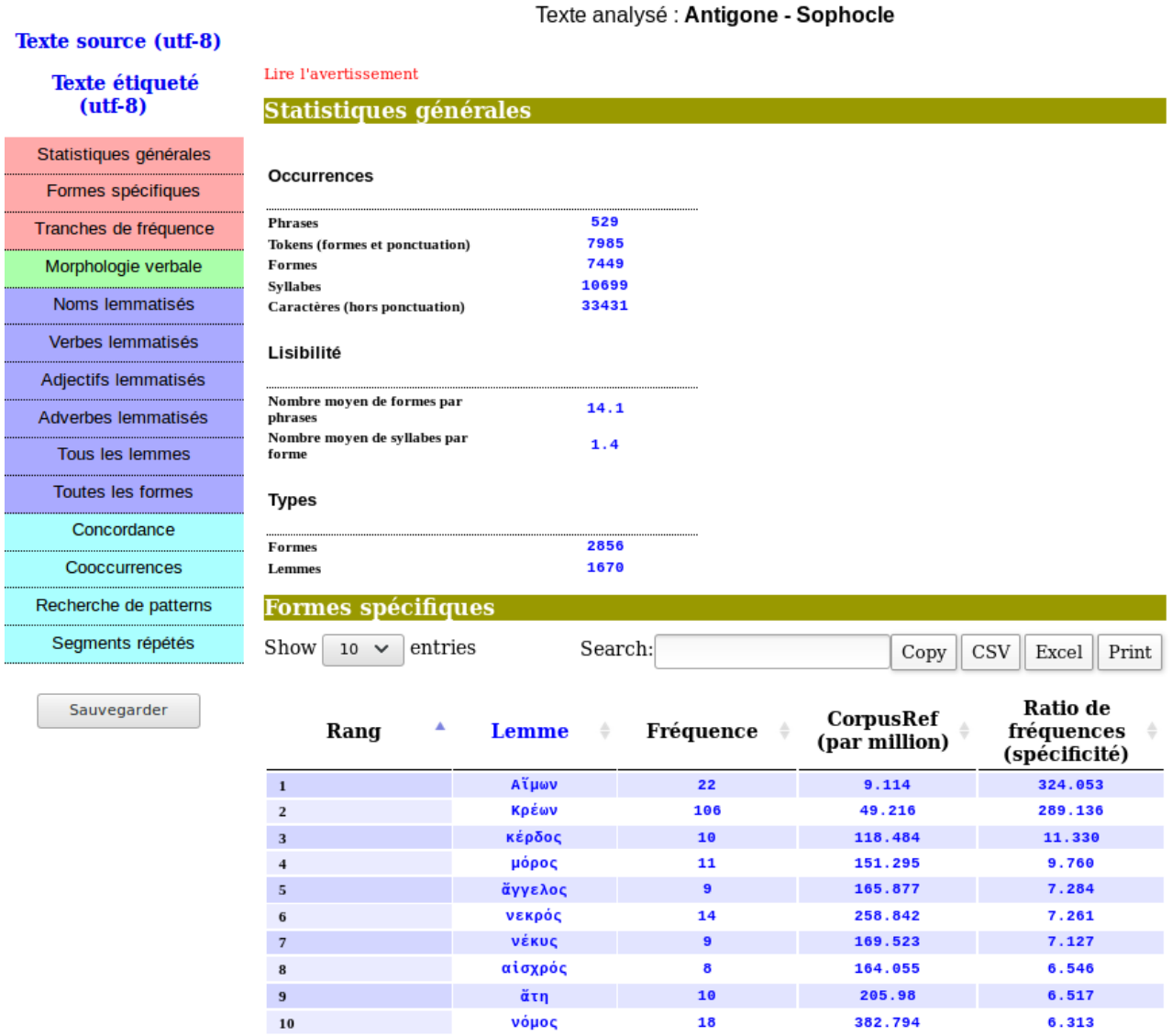
Conclusion

- Le corpus du projet Perseus (AGLDT) semble suffisamment grand et diversifié pour entraîner un étiqueteur « générique » de bonne qualité

- Pour estimer la couverture, il faudrait l'évaluer sur l'ensemble du corpus tragique.
  - ex. : Pour *Antigone* de Sophocle, 243 formes inconnues sur 7985, soit un rappel de 97 %.

Conclusion

- Le grec ancien est désormais sur Anatext !



εὐφίλητε	a-s---mv-	εὐφίλητος
εὐχή	n-s---fn-	εὐχή
εὐχαριστίας	n-p---fa-	εὐχαριστία
εὐχαρίστως	d--------	εὐχά ριστος
εὐχαρίστω !	v1spia--	
εὐχαῖς	n-p---fd-	εὐχή
εὐχαῖσι	n-p---fd-	εὐχή
εὐχερεί	---------	εὐχερεί