BABELAO 10-11 (2022), p. 537-562

Analyse automatique du grec ancien par réseau de neurones. Évaluation sur le corpus *De Thessalonica Capta*

Par

Bastien Kindt*, Chahan Vidal-Gorène**, Saulo Delle Donne***

* *UCLouvain (Louvain-la-Neuve) et Peeters Publishers (Leuven)*
** *École Nationale des Chartes-PSL et Calfa (Paris)*
*** *Università del Salento (Lecce)*

Le corpus *De Thessalonica Capta* (désormais « DTC ») se compose de trois récits historiographiques écrits en grec à l'époque byzantine : 1) la *Narratio de Excidio Thessalonicensi* de Jean Kaminiatès (attaque des Sarrasins le 13 juillet 904 sous la conduite de Léon de Tripoli) ; 2) la *Narratio de Thessalonica Capta* d'Eustathe de Thessalonique (attaque des Normands de Guillaume II, roi de Sicile, le 24 août 1185) ; 3) la *Narratio de Extremo Thessalonicensi Excidio* de Jean Anagnostès (attaque des Turcs le 29 mars 1430, sous la conduite du sultan Mourad II)[1].

[1] Les titres latins sont repris aux éditions de E. Bekker (respectivement BEKKER 1838a, p. 481 ; BEKKER 1848, p. 365 et BEKKER 1838b, p. 484). Sur Jean Kaminiatès, voir KAZHDAN 1991, vol. II, p. 1098-1099, *s.v.* Kaminiates, John (A. Kazhdan), HUNGER 1978, p. 357-359, et ODORICO 2005, p. 11-13 ; sur Eustathe de Thessalonique, voir KAZHDAN 1991, vol. II, p. 754, *s.v.* Eustathios of Thessalonike (A. Kazhdan), HUNGER 1978, p. 426-429, et ODORICO, p. 24-28 ; pour Jean Anagnostès, voir KAZHDAN 1991, vol. II, p. 1056, *s.v.* John Anagnostes (A.-M. Talbot), HUNGER 1978, p. 484-485, et ODORICO 2005, p. 34-35. Sur Thessalonique à l'époque byzantine, voir KAZHDAN, vol. III, p. 2071-2073, *s.v.* Thessalonike (T.E. Gregory), et ODORICO 2005, p. 7-11. Les éditions utilisées

1. Description du corpus De Thessalonica Capta *et précisions méthodologiques*

P. Odorico synthétise en ces termes l'intérêt de cet ensemble littéraire :

« *Les trois textes présentés* […] *sont exceptionnels à plus d'un titre : d'abord, ils se distinguent du reste de la production byzantine par leur contenu, fortement autobiographique, ainsi que par la similitude de leur propos, qui porte sur la prise de Thessalonique, la deuxième ville de l'Empire. Ensuite, ils intriguent de par leur nature : ce sont en effet indubitablement – et c'est d'abord à ce titre qu'ils ont retenu l'attention – des sources historiques de première importance concernant les événements qu'ils relatent, qui se sont déroulés à trois moments forts différents de l'histoire de l'Empire d'Orient. Mais en même temps, ces textes ne sont pas à proprement parler des comptes rendus, mais plutôt des pièces à conviction soutenant des positions propres à chacun des auteurs, et qui visent à leur faire retirer quelque avantage d'un travail d'écriture. La contemporanéité des événements et de l'écriture, la volonté d'aménager l'histoire plutôt que de tenter d'en faire un sobre exposé, sont leurs caractères les plus remarquables, même si, dissimulés derrière des déclarations convenues sur l'importance du récit historique, ils s'intègrent comme naturellement dans le droit fil de l'historiographie traditionnelle. Car ces trois récits ont aussi une fonction toute personnelle : l'écriture sert à chaque narrateur à faire son deuil d'événements douloureux dont il a été victime, et leur permet de revenir sur ce qu'ils ont vécu pour le transformer en expression artistique, en pièces de littérature qui astreignent le réel à leurs propres impératifs* »[2].

Le projet GREgORI, spécialisé dans le traitement automatique du grec ancien et des langues de l'Orient chrétien, a analysé ce corpus[3]. Chaque mot du texte a reçu un lemme (une entrée de dictionnaire) et une catégorie morphosyntaxique (nom, adjectif, pronom, verbe, etc.). Ainsi, la forme κτῆσιν (Kaminiatès, I, 1 = BÖHLIG 1973, p. 3, l. 3) est-elle classée sous le lemme κτῆσις et est catégorisée comme nom commun (« N+Com »), la forme μεγάλου (Eustathe, *Tit.* = KYRIAKIDIS 1961, p. 3, l. 5) a reçu le lemme μέγας, adjectif (« A »), la forme ἡμῖν (Anagnostès, I = TSARAS 1958, p. 4, l. 10) figure sous le lemme ἡμεῖς, pronom personnel de la première personne du pluriel (« PRO+Per1p »), et, enfin, la forme μαρτυροῦσα (Kaminiatès, I, 1 = BÖHLIG 1973, p. 3, l. 4) sous le lemme verbal (« V ») μαρτυρέω[4].

---

sont les suivantes : BÖHLIG 1973 (Kaminiatès) ; KYRIAKIDIS 1961 (Eustathe) ; TSARAS 1958 (Anagnostès). Sur ce corpus, voir ODORICO 2005, avec introduction et traduction française.

[2] ODORICO 2005, p. 5. Les trois auteurs – témoins, acteurs et même victimes des événements – détournent à leur profit le genre littéraire de la *Narratio*. Le texte de Kaminiatès (déporté avec sa famille et retenu en détention en Orient) a des allures de plaidoyer pour sa libération (ODORICO 2005, p. 19-24), celui d'Eustathe (soupçonné de collusions avec les différentes puissances impliquées dans la vie politique de la ville dont il était l'évêque) a un caractère apologétique (ODORICO 2005, p. 34), et celui d'Anagnostès est un pamphlet à l'adresse des protagonistes de l'époque, byzantins, latins ou ottomans (ODORICO 2005, p. 40-41). Cela confère aux trois œuvres une originalité tout à fait particulière, sans pourtant rien enlever de leur intérêt historique. Sur la question de l'authenticité du texte de Kaminiatès, voir FRENDO, FOTIOU 2000, p. xxxvii-xxxix, ODORICO 2005, p. 13-25, et STRANO 2013.

[3] Le projet GREgORI – mené à l'Institut orientaliste de l'UCLouvain (Belgique) – a développé une expertise dans le domaine de l'analyse des textes écrits dans les principales langues de l'Orient chrétien, en l'occurrence le grec, l'arménien, le géorgien, le syriaque (ainsi qu'un dialecte néo-araméen, le turoyo) et l'arabe ; informations générales et bibliographie exhaustive sur le site WEB du projet à l'adresse https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html. Ils se concrétisent par la production régulière de concordances et d'index lemmatisés, monolingues ou bilingues. Les travaux du projet ont été à l'origine de la collection des concordances lemmatisées publiées dans les volumes du *Thesaurus Patrum Graecorum*. Les index complètent les éditions critiques de textes nouvellement publiés, voir, par exemple, CAPONE 2021 (index lemmatisé bilingue grec-latin, p. 25-99), STONE 2021 (index lemmatisé arménien, p. 129-220), PATARIDZE 2020 (index lemmatisé géorgien, p. 33-114) et SCHMIDT – KINDT 2021 (index et concordance lemmatisés syriaques), GRAND'HENRY 2020 (index lemmatisé arabe, p. 125-241).

[4] Le détail des étiquettes désignant les catégories morphosyntaxiques est fourni sur le site WEB (adresse ci-dessus, note 3) et sur les interfaces d'interrogation en ligne des corpus du projet, accessibles sous l'adresse

Le corpus DTC est donc un corpus numérique enrichi d'informations linguistiques. Il est mis à la disposition des chercheurs sous deux formes :

1) une copie au format PDF de la concordance lemmatisée complète du corpus DTC est accessible en ligne sur le site WEB du projet (cité sous la note 3)[5] ;

2) les versions numériques des trois textes du corpus DTC sont interrogeables sur les interfaces en ligne du projet (cité sous la note 4).

Les récits de Jean Kaminiatès, d'Eustathe de Thessalonique et de Jean Anagnostès rejoignent ainsi les autres œuvres historiographiques déjà traitées par le projet GREgORI, celles de Procope de Césarée, d'Agathias le Scholastique, de Ménandre le Protecteur, de Théophylacte de Simocatta, de Nicéphore le Patriarche, de Joseph Génésios ou de l'historien Ducas, sans oublier la chronique de Théophane le Confesseur. Tous ces travaux contribuent à dresser l'inventaire lexical de la langue des auteurs byzantins[6]. Mais, paradoxalement, il n'existe pas encore d'étude systématique de la langue des trois textes du corpus DTC. Les introductions des éditions n'évoquent le sujet que de manière très générale[7]. Eustathe de Thessalonique se démarque de Kaminiatès et Anagnostès par le recours à des citations tirées des classiques grecs (Homère, les tragiques, les orateurs, etc.) et par l'utilisation de nombreux mots empruntés, principalement au latin (59 emprunts au latin sur un total de 63 dans le corpus DTC), par exemple ἀδνούμιον, δομέστικος, δούξ, κάστρον, ῥήξ, etc. Kaminiatès et Anagnostès utilisent une langue byzantine atticisante plus simple, privilégiant les citations des Écritures et des auteurs chrétiens, tout en exploitant les procédés rhétoriques propres au genre.

De plus en plus de textes anciens sont disponibles en ligne. En grec, on peut citer le *Thesaurus Linguae Graecae* (depuis 1972), la section *Greek and Roman Materials* de la *Perseus Digital Library* (depuis 1985), le projet *First Thousand Years of Greek* (depuis 2016), ou le *Diosiris Ancient Greek Corpus* (depuis 2017)[8]. Cependant, pour les formes tardives des langues classiques et, davantage encore, pour les langues de l'Orient chrétien, les corpus numériques formatés dans un standard d'encodage interopérable, complètement analysés, mis librement à la disposition des chercheurs, et dont tous les éléments du texte sont interrogeables (formes, lemmes, catégories morphosyntaxiques, etc.) dans un environnement convivial, restent encore

---

https://www.gregoriproject.com. Des étiquettes précisant l'analyse morphologique des formes (cas, genre, nombre, voix, mode, temps, personne) sont également utilisées quand la nature d'un projet l'exige, voir par exemple les travaux menés en collaboration avec l'Université de Lausanne « Le devenir numérique d'un texte fondateur : l'Iliade et le Genavensis Græcus 44 » et l'*Iliadoscope* (voir sous l'adresse https://www2.unil.ch/iliade).

[5] DELLE DONNE, KINDT 2021.

[6] COULIE 1996.

[7] Voir cependant les différents index de l'édition de Kaminiatès dans BÖHLIG 1973, p. 69-95, et l'introduction à la traduction anglaise d'Eustathe dans MELVILLE JONES 1988, p. ix-x.

[8] Le *Thesaurus Linguae Graecae* (http://stephanus.tlg.uci.edu) réunit la majorité des textes grecs depuis Homère (VIII^e^ s. av. J.-C.) jusqu'à la chute de Constantinople (1453). La version abrégée du TLG est libre d'accès, sa version complète est payante. Pour le *Perseus Project*, voir http://www.perseus.tufts.edu. Le projet *First Thousand Years of Greek* (https://opengreekandlatin.github.io/First1Kgreek) propose un corpus de textes grecs (depuis Homère jusqu'en 250 ap. J.-C.) téléchargeable gratuitement au format XML-TEI ou consultable via l'interface *Scaife Viewer* de *Perseus* (https://scaife.perseus.org/) (MUELLNER 2019). Le *Disosiris Corpus* (https://figshare.com/articles/dataset/The_Diorisis_Ancient_Greek_Corpus/6187256) rassemble huit cent textes grecs depuis Homère jusqu'au V^e^ s. ap. J.-C., téléchargeables gratuitement au format XML-TEI et analysés automatiquement (VATRI, MCGILLIVRAY 2018).

rares[9]. Parallèlement, les outils et les ressources linguistiques utiles à la constitution de ces corpus numériques (et diffusables sous des formats libres et interopérables) font également défaut. Pourtant, l'élaboration de ces outils et ressources et la constitution de tels corpus revêtent une importance majeure pour assurer l'étude systématique et automatisée des textes écrits dans ces langues[10]. La plus-value est d'autant plus grande si le système d'analyse est explicitement décrit et clairement évalué[11].

Le corpus DTC compte 60.493 mots. Il a désormais rejoint le corpus grec complet du projet GREgORI, un ensemble littéraire qui, dans sa version en ligne, totalise à ce jour 753.990 mots-occurrences, tous assortis d'un lemme et d'une catégorie morphosyntaxique[12].

Le tableau 1 précise les effectifs des mots-occurrences, des formes différentes et des lemmes différents, tant pour chacun des trois textes que pour l'ensemble du corpus DTC.

	Mots-occurrences	Formes différentes	Lemmes différents
Kaminiatès (904)	22.365	6.834	3.384
Eustathe (1185)	28.515	9.261	4.890
Anagnostès (1430)	9.613	3.589	1.896
Total DTC	60.493	16.453	7.195

Tableau 1 : effectifs des mots-occurrences, des formes différentes et des lemmes différents du corpus DTC

Les analyses produites à l'aide des outils informatiques et des ressources linguistiques du projet GREgORI reposent sur une approche dite « par dictionnaires ». Cette méthode consiste à comparer le vocabulaire d'un texte à un lexique de référence, en l'occurrence un ensemble de ressources linguistiques propres au projet GREgORI et réunies dans différents dictionnaires électroniques[13]. Si ce *modus operandi* a fait ses preuves, son efficacité dépend aussi largement de l'adéquation entre les ressources linguistiques utilisées et le texte analysé. Trois conséquences en découlent.

1) Une forme inconnue des ressources linguistiques ne reçoit aucune analyse : les mots πολυμαθέστατε (Kaminiatès, I, 2 = BÖHLIG 1973, p. 3, l. 6), forme fléchie du lemme adjectival πολυμαθής, στρατοπεδαρχεῖν (Eustathe, XXI = KYRIAKIDIS 1961, p. 28, l. 9)[14], forme fléchie du lemme verbal στρατοπεδαρχέω, Μουράτης (Anagnostès, VII = TSARAS 1958, p. 18, l. 13),

[9] Pour l'arménien, voir le *Church Armenian Corpus – Concordance* du projet *Arak-29* (https://arak29.org). Pour le syriaque, *Simtho. The Syriac Thesaurus* (https://simtho.bethmardutho.org), qui propose un corpus de 13.080.848 mots-occurrences et un sous corpus de 6.337.520 mots-occurrences analysés automatiquement, ou le *Digital Syriac Corpus* (https://syriaccorpus.org). En géorgien, voir les banques de textes du *Thesaurus Indogermanischer Text- und Sprachmaterialien*, sous la rubrique *Caucasian languages* (https://titus.uni-frankfurt.de) et le *Georgian Language Corpus* (http://corpora.iliauni.edu.ge).

[10] VIDAL-GORÈNE, KINDT 2020, p. 22, et VIDAL-GORÈNE, DECOURS-PEREZ 2020.

[11] Le projet GREgORI s'est doté d'un standard de lemmatisation pour chacune des langues traitées, voir par exemple KINDT 2004 (pour le grec), COULIE, KINDT, KEPEKLIAN 2021 (pour l'arménien), COULIE, KINDT, PATARIDZE 2013 (pour le géorgien), KINDT, HAELEWYCK, SCHMIDT, ATAS 2018 (pour le syriaque) et TUERLINCKX 2004 (pour l'arabe).

[12] Les données lemmatisées du *Thesaurus Patrum Graecorum* (voir note 3) sont progressivement récupérées et reprises dans le corpus grec du projet GREgORI accessible en ligne.

[13] Voir KINDT, PIRARD 2016, KINDT 2018 et KINDT 2021.

[14] La numérotation des chapitres du texte d'Eustathe (absente de l'édition de S. Kyriakidis) est reprise à l'édition de J.R. Melville Jones, qui reproduit celle de l'édition due à T.L.F. Tafel (Frankfurt, 1832) ; voir MELVILLE JONES 1988, p. vii.

forme au nominatif du lemme homographe Μουράτης, ou enfin τοὐπίταγμα (Anagnostès, II = TSARAS 1958, p. 68, l. 21), crase constituée de l'article ὁ et du lemme nominal ἐπίταγμα, étaient encore absents des ressources linguistiques et sont restés sans analyse à l'issue de la comparaison[15].

2) Une forme susceptible de répondre à plusieurs analyses reçoit ces différentes analyses, car les traitements ne tiennent pas compte du contexte d'apparition des formes dans le texte : la forme ἀγαθῶν (Kaminiatès, I, 1 = BÖHLIG 1973, p. 3, l. 3) se voit donc attribuer, hors contexte, un lemme adjectival ἀγαθός et un lemme verbal ἀγαθόω, la forme ἡμέραν (Eustathe, XXXVI = KYRIAKIDIS 1961, p. 40, l. 18), un lemme nominal ἡμέρα mais aussi un lemme adjectival ἥμερος, la forme διηγήσει (Anagnostès, XXII = TSARAS 1958, p. 68, l. 4), un lemme nominal διήγησις et un autre, verbal, διηγέομαι.

3) Une forme dont l'analyse serait partielle ou incongrue dans ces mêmes ressources linguistiques se voit *a fortiori* attribuer une analyse non pertinente : la forme ἄνωθεν (Anagnostès, *Mon.* = TSARAS 1958, p. 74, l. 20) a été caractérisée et comme adverbe (« I+Adv ») et comme adverbe prépositionnel (« I+AdvPr »), car c'est ainsi qu'elle était enregistrée, erronément, dans les ressources linguistiques.

Le tableau 2 indique, tant pour chacun des trois textes que pour l'ensemble du corpus DTC, les effectifs et le pourcentage des mots restés sans analyse, des mots ayant reçu une seule analyse, et des mots répondant à plus d'une analyse.

	Effectifs des mots dont le nombre d'analyse = 0		Effectifs des mots dont le nombre d'analyse = 1		Effectifs des mots dont le nombre d'analyse > 1	
Kaminiatès	992	4,44%	18.701	83,62%	2.672	11,95%
Eustathe	2.248	7,88%	22.300	78,20%	3.969	13,92%
Anagnostès	373	3,88%	8.082	84,07%	1.164	12,11%
Total	3.613	5,97%	49.083	81,14%	7.805	12,90%

Tableau 2 : évaluation quantitative de la lemmatisation par « dictionnaires »

Le tableau 3, corollaire du tableau précédent, fait état des effectifs et du pourcentage des mots correctement analysés, tant au niveau du lemme que de la catégorie morphosyntaxique, quand une seule analyse est proposée.

	Lemmes corrects		Catégories morphosyntaxiques correctes	
Kaminiatès	18.689	99,93%	18.674	99,85%
Eustathe	22.257	99,80%	22.246	99,75%
Anagnostès	8.075	99,91%	8.066	99,80%
Total	49.021	99,87%	48.896	99,61%

Tableau 3 : évaluation qualitative de la lemmatisation par « dictionnaires » pour les mots répondant à une seule analyse

[15] Les formes présentant une coquille typographique pourraient être citées ici car elles ne peuvent être reconnues lors de l'analyse « par dictionnaires ». Par exemple, la forme ἐνθυμάτων (*sic* dans l'édition, coquille pour ἐνθυμημάτων) est restée sans analyse avant d'être corrigée (Anagnostès, III = TSARAS 1958, p. 10, l. 3). L'analyse du corpus DTC a permis d'identifier et de corriger trente-six formes. Elles sont consignées dans l'introduction de la concordance au format PDF disponible sur le site WEB du projet GREgORI. En ce sens, la lemmatisation améliore la qualité des textes édités. Ce fait souligne aussi l'intérêt de traiter un texte avant l'impression définitive de son édition. Sur cette notion, voir aussi COULIE 1996, p. 41-42.

Les informations consignées dans les tableaux 2 et 3 illustrent trois choses.

1) D'abord, les ressources linguistiques mises en œuvre fournissent d'excellents résultats pour les mots susceptibles de n'avoir qu'un seul lemme et qu'une seule catégorie grammaticale. Pour les lemmes, le taux d'exactitude atteint 99,87%, pour les catégories morphosyntaxiques, 99,61% (tableau 3).

2) Ensuite, une proportion de 5,97% des mots n'a pas reçu d'analyse (tableau 2). Il s'agit majoritairement de mots nouveaux, jamais rencontrés au fil des analyses précédentes ou jamais enregistrés dans les ressources linguistiques. Pour transformer le corpus DTC en un corpus numérique abouti, il a fallu mettre à jour, manuellement, ces 3.613 analyses manquantes.

3) Enfin, une proportion de 12,90% des mots a reçu plus d'une analyse (tableau 2). Pour obtenir un corpus numérique abouti, il a fallu désambiguïser, manuellement, les 7.805 formes susceptibles, hors contexte, d'être attachées à plus d'un lemme.

Ces phases de révision demeurent indispensables pour fournir aux chercheurs des corpus entièrement analysés. Mais, réalisé manuellement, cet exercice est toujours chronophage et donc coûteux. Une approche par apprentissage machine, via des réseaux de neurones, constitue une alternative efficace pour surmonter ces deux aspects. Dans la suite de l'article nous illustrons le bénéfice de l'utilisation des réseaux de neurones pour le traitement et la révision de corpus.

Un réseau de neurones est un système informatique que l'on peut entraîner pour accomplir une tâche spécifique, en l'occurrence une analyse de texte. Entraîné avec de larges corpus de textes déjà analysés (lemmes et catégories morphosyntaxiques), le réseau de neurones construit un modèle de langue interne afin de produire des analyses contextuelles adéquates pour des formes ambiguës et de déduire des lemmes pour des formes inédites. En règle générale, plus le volume et la qualité du corpus des textes déjà analysés et utilisés lors de la phase d'entraînement sont grands, plus l'apprentissage s'avère efficace. Ces données sont appelées la vérité de terrain[16]. Par ailleurs, cette approche permet des adaptations technologiques rendant possible l'analyse de langues peu dotées, c'est-à-dire les langues pour lesquelles peu de ressources linguistiques sont disponibles, ce qui est le cas pour les langues de l'Orient chrétien[17].

Dans le cas présent, la vérité de terrain est constituée de différents corpus grecs du projet GREgORI, un ensemble de 1.786.629 mots-occurrences (164.621 formes différentes, 38.070 lemmes), tous munis d'un lemme et d'une catégorie morphosyntaxique[18]. Le modèle d'analyse produit a été utilisé pour assurer une seconde analyse du corpus DTC. La comparaison des résultats obtenus par les deux approches, celle « par dictionnaires » et celle par réseau de neurones, permet dès lors d'évaluer la fiabilité de cette dernière.

---

[16] Pour une description de cette approche, voir PATEL, THAKKAR 2020 ; pour les langues anciennes, DEREZA 2018 ; en arménien, géorgien et syriaque, VIDAL-GORÈNE, KINDT 2020.

[17] VIDAL-GORÈNE, KINDT 2022.

[18] La vérité de terrain rassemble des textes patristiques (Grégoire de Nazianze, Grégoire de Nysse, Basile de Césarée, etc.), des textes ascétiques (Isaac de Ninive, Philoxène de Mabboug, Jean de Dalyatha, etc.), les textes historiographiques cités plus haut (Procope de Césarée, Agathias le Scholastique, Ménandre le Protecteur, Théophylacte de Simocatta, Nicéphore le Patriarche, Joseph Génésios, Ducas, Théophane le Confesseur, etc.) et d'autres œuvres (chroniques, récits hagiographiques, homélies diverses, etc.). La majorité de ces textes ont fait l'objet d'une concordance lemmatisée dans le *Thesaurus Patrum Graecorum* (voir note 3). Cette énumération n'est pas exhaustive. Les données de la vérité de terrain ont été scindées en deux sous-corpus, un corpus d'entraînement et de développement (1.611.206 mots-occurrences, 154.533 formes différentes, 35.795 lemmes), d'une part, et un corpus de test (175.423 mots-occurrences, 37.062 formes différentes, 13.532 lemmes), d'autre part.

Dans la pratique, la préparation et la mise en œuvre du modèle d'analyse est réalisée par Calfa, une entreprise spécialisée dans l'analyse des documents en arménien et dans le traitement automatique des langues orientales[19]. L'approche neuronale repose sur PIE[20], outil déjà utilisé et évalué par GREgORI et Calfa pour le traitement conjoint de l'arménien, du syriaque et du géorgien[21].

Les résultats produits par le modèle d'analyse sont le fruit d'une approche statistique et probabiliste. Une révision reste utile. Mais le modèle d'analyse tient compte du contexte d'apparition des mots dans le texte traité. Il peut donc proposer une solution d'analyse pour les mots ambigus, ce qui correspond à 5,97% du vocabulaire du corpus DTC (tableau 2). Conçu suite à l'examen de données massives et variées – relativement représentatives de la langue traitée –, le modèle d'analyse peut aussi prédire des formes inédites absentes de la vérité de terrain, ce qui correspond à 12,90% du vocabulaire du corpus DTC (tableau 2). Ces deux aspects sont des avantages importants de l'utilisation d'un réseau de neurones par rapport à la méthode dite « par dictionnaires » et sont susceptibles de réduire sensiblement la phase de révision manuelle des données quand le modèle d'analyse aura atteint, au fil des expériences plusieurs fois renouvelées, une efficacité optimale.

2. Évaluation

Les tableaux 4 et 5 fournissent les résultats de l'utilisation du modèle d'analyse sur le corpus de développement et sur le corpus DTC. Pour la lemmatisation, sur l'ensemble des mots, le taux de résultats corrects (Accuracy) atteint 97,94%. Pour les mots connus du corpus d'entraînement, ce taux s'élève même à 99,60%. Une baisse de rendement est observée pour l'analyse des mots inconnus du corpus d'entraînement, les résultats obtenus étant corrects ici dans 83,91% des cas. En revanche, le modèle d'analyse se montre plus efficace pour le traitement des lemmes ambigus, avec un score de 97,94% de résultats exacts.

[19] Sur Calfa, cfr https://calfa.fr. Calfa propose des solutions sur mesure pour l'analyse des textes orientaux, manuscrits ou imprimés, depuis la saisie des textes à l'aide d'outils de reconnaissance optique de caractères – optical characters recognition (OCR) pour les imprimés, handwritten text recognition (HTR) pour les textes manuscrits –, jusqu'à leur analyse linguistique (lemmatisation, analyse catégorielle, analyse flexionnelle, etc.).

[20] MANJAVACAS, KÁDÁR, KESTEMONT 2019.

[21] VIDAL-GORÈNE, KINDT 2020. L'architecture de PIE utilisée pour le traitement du corpus DTC est identique à celle mise en œuvre en 2020, avec cependant un « batch size » de 128 au lieu de 25.

Lemmatisation				
Corpus de test[22]	Tous les mots	Mots connus du corpus d'entraînement	Mots inconnus du corpus d'entraînement	Mots ambigus dans le corpus d'entraînement
F1-score[23]	73,19%	95,99%	58,14%	60,90%
Accuracy	98,00%	99,43%	79,81%	96,31%
Corpus DTC				
F1-score	79,38%	96,84%	66,48%	65,27%
Accuracy	97,94%	99,60%	83,91%	97,94%

Tableau 4 : évaluation de la lemmatisation

Pour la catégorisation morphosyntaxique, le taux des résultats corrects est de 98,59% pour l'ensemble des mots, 99,27% pour les mots connus du corpus d'entraînement et de 97,57% pour les mots ambigus. Pour les mots inconnus, le score descend cependant à 92,75%.

Catégorisation morphosyntaxique				
Corpus de test	Tous les mots	Mots connus du corpus d'entraînement	Mots inconnus du corpus d'entraînement	Mots ambigus dans le corpus d'entraînement
F1-score	48,39%	56,29%	37,47%	73,59%
Accuracy	98,31%	99,28%	86,04%	97,29%
Corpus DTC				
F1-score	71,04%	79,86%	44,70%	76,33%
Accuracy	98,59%	99,27%	92,75%	97,57%

Tableau 5 : évaluation de la catégorisation morphosyntaxique

Ces résultats sont encourageants. L'évaluation antérieure portant sur l'arménien, le géorgien et le syriaque[24], avait fourni, pour la lemmatisation et sur l'ensemble des mots, une Accuracy de 90,44% en arménien, de 96,28% en géorgien et de 88,17 en syriaque. Pour la catégorisation morphosyntaxique, les résultats étaient de 92,38%, 97,18% et de 88,13%, pour ces mêmes langues. Dans les trois cas, les corpus d'entraînement étaient bien inférieurs en termes de nombre de mots, 66.812 mots pour l'arménien, 150.869 pour le géorgien et 10.612 pour le syriaque, contre 1.764.555 mots pour l'expérience menée ici, en grec. Ce fait explique déjà partiellement les bons résultats obtenus.

Si le modèle fournit souvent des résultats attendus, il commet donc aussi des erreurs. Les exemples qui suivent illustrent quelques cas de figures. Les mots des différents segments de phrase présentés dans ces exemples sont accompagnés des lemmes et des catégories

[22] Le corpus de test est un sous-ensemble de la vérité de terrain qui n'a pas été utilisé pour l'entraînement et l'apprentissage du réseau de neurones. Les données qu'ils présentent servent de repères pour évaluer la performance globale du modèle d'analyse et la performance attendue sur le corpus DTC ; voir note 18.

[23] Le « F1-score » est une des métriques utilisées pour évaluer de tels traitements. L'« accuracy » est une métrique indiquant le taux des résultats corrects. Pour une définition de ces métriques, voir DERCZYNSKI 2016.

[24] VIDAL-GORÈNE, KINDT 2020 (les modèles d'analyses avaient été développés au printemps 2019).

morphosyntaxiques, les analyses « par dictionnaires » (GREgORI), celles du modèle d'analyse (RNN), et enfin celles vérifiées par les experts humains (DTC).

1.	**καὶ**	**πάντων**	**τῶν**	**ἐν**	**τῇ**	**πόλει,**	**ἀνδρῶν,**	**γυναικῶν,**	**παίδων**
GREgORI	καί I+Part	πᾶς A	ὁ DET	ἐν I+Prep	ὁ DET	πόλις πολέω N+Com V	ἀνήρ ἀνδρόω N+Com V	γυνή γυναικόω N+Com V	παῖς N+Com
RNN	καί I+Part	πᾶς A	ὁ DET	ἐν I+Prep	ὁ DET	πόλις N+Com	ἀνήρ N+Com	γυνή N+Com	παῖς N+Com
DTC	καί I+Part	πᾶς A	ὁ DET	ἐν I+Prep	ὁ DET	πόλις N+Com	ἀνήρ N+Com	γυνή N+Com	παῖς N+Com

En 1 (Anagnostès, IV = TSARAS 1958, p. 10, l. 22), les formes πόλει, ἀνδρῶν et γυναικῶν sont toutes les trois susceptibles d'avoir deux lemmes, l'un nominal (respectivement πόλις, ἀνήρ, γυνή), l'autre verbal (respectivement πολέω, ἀνδρόω, γυναικόω). Dans les trois cas, le modèle d'analyse prédit, à raison, les lemmes nominaux.

2.	**ὦ**	**σοφώτατε**	**ἀνδρῶν**	**καὶ**	**φιλομαθέστατε**	**Γρηγόριε**
GREgORI	εἰμί ὦ (ᾱ̓́) ὦ (τό) V I+Intj N+Lettre	σοφός A	ἀνδρόω ἀνήρ V N+Com	καί I+Part	*inconnu* A	Γρηγόριος N+Ant
RNN	ὦ (ᾱ̓́) I+intj	σοφός A	ἀνήρ N+Com	καί I+Part	φιλομαθής A	Γρηγόριος N+Ant
DTC	ὦ (ᾱ̓́) I+intj	σοφός A	ἀνήρ N+Com	καί I+Part	φιλομαθής A	Γρηγόριος N+Ant

En 2 (Kaminiatès, LXXIV, 10 = BÖHLIG 1973, p. 64, l. 59), la forme ὦ est ambiguë et la forme φιλομαθέστατε est inconnue des ressources linguistiques. Dans les deux cas, le modèle prédit la bonne analyse, respectivement le lemme ὦ (ᾱ̓́), l'interjection interpellative et emphatique, et l'adjectif φιλομαθής.

3.	**ὁ**	**μὲν**	**πεζὸς**	**λαὸς**	**περιεκάθισε**	**καθ΄**	**ὥραν**	**ἀρίστου**
GREgORI	ὁ DET	μέν I+Part	πεζός A	λαός N+Com	*inconnu* *inconnu*	κατά I+Prep	ὥρα N+Com	ἀγαθός ἄριστον A N+Com
RNN	ὁ DET	μέν I+Part	πεζός A	λαός N+Com	περικαθίζω V	κατά I+Prep	ὥρα N+Com	ἀγαθός A
DTC	ὁ DET	μέν I+Part	πεζός A	λαός N+Com	περικαθίζω V	κατά I+Prep	ὥρα N+Com	ἄριστον N+Com

En 3 (Eustathe, LV = KYRIAKIDIS 1961, p. 66, l.19), la forme verbale περιεκάθισε inconnue des ressources linguistiques du projet GREgORI est correctement prédite par le modèle d'analyse. Mais la forme ἀρίστου, relevant des deux lemmes ἀγαθός (« bon ») ou ἄριστον (« le déjeuner ») est erronément analysée comme une forme fléchie du lemme adjectival, alors que le contexte impose le lemme nominal ἄριστον.

4.	**Λέων**	**ἦν**	**ἐκεῖνος**	**ὁ**	**Μαζιδᾶς**
GREgORI	Λέων λέων N+Ant N+Com	εἰμί V	ἐκεῖνος PRO+dem	ὁ DET	*inconnu* *inconnu*
RNN	λέων N+Com	εἰμί V	ἐκεῖνος PRO+dem	ὁ DET	μαζιδᾶ N+Ant
DTC	Λέων N+Ant	εἰμί V	ἐκεῖνος PRO+dem	ὁ DET	Μαζιδᾶς N+Prop

En 4 (Eustathe, LXII = KYRIAKIDIS 1961, p. 76, l. 34), le modèle d'analyse propose, à tort, un lemme nominal λέων pour l'anthroponyme Λέων (« lion » *vs* « Léon »), et un lemme anthroponymique μαζιδᾶ pour le nom propre Μαζιδᾶς, qui détermine ce personnage[25].

5.	**Θεὸν**	**γὰρ**	**ἐξιλεοῦν**	**ἀσύγγνωστα**	**πταίσαντας**	**[…]**	**χρή**
GREgORI	θεός N+Com	γάρ I+Part	ἐξιλεόω V	ἀσύγγνωστος A	πταίω V		χρή V
RNN	θεός N+Com	γάρ I+Part	ἐξιλεόω V	*ἀσυγγύσνωστος *A	πταίω V		χρή V
DTC	θεός N+Com	γάρ I+Part	ἐξιλεόω V	ἀσύγγνωστος A	πταίω V		χρή V

En 5 (Anagnostès, *Monodie* = TSARAS 1958, p. 76, l. 13), le modèle propose un lemme fantaisiste *ἀσυγγύσνωστος pour la forme ἀσύγγνωστα (lemme ἀσύγγνωστος) connue des ressources linguistiques du projet GREgORI.

6.	**ὁ**	**Μουράτης**	**[...]**	**γράμματα**	**τοῖς**	**βέλεσι**	**πεπομφώς**
GREgORI	ὁ DET	*inconnu* *inconnu*		γράμμα N+Com	ὁ DET	βέλος N+Com	πέμπω V
RNN	ὁ DET	Μουράτος N+Ant		γράμμα N+Com	ὁ DET	βέλος N+Com	*πομφόω *V
DTC	ὁ DET	Μουράτης N+Ant		γράμμα N+Com	ὁ DET	βέλος N+Com	πέμπω V

En 6 (Anagnostès, IX = TSARAS 1958, p. 24, l. 6), le modèle propose un lemme anthroponymique Μουράτος à la place du lemme attendu, Μουράτης[26] et un lemme verbal fantaisiste *πομφόω pour la forme πεπομφώς[27].

[25] « Τις τῶν τῆς στρατιᾶς, οὐκ ἀφανής, Λέων ἦν ἐκεῖνος ὁ Μαζιδᾶς », « un soldat assez connu, le fameux Léon Mazidas » (ODORICO 2005, p. 197), mais totalement inconnu par ailleurs. L'étiquette « N+Prop » identifie comme nom propre les lemmes qui ne sont ni anthroponyme (« N+Ant »), ni toponyme (« N+Top »). Les noms de famille sont classés sous cette étiquette. En l'absence d'information supplémentaire, elle caractérise aussi le lemme Μαζιδᾶς.

[26] Le modèle d'analyse propose, tout en « rigueur », le lemme Μουράτος pour l'ensemble des vingt occurrences du corpus DTC, toutes attestées chez Anagnostès. Un expert humain, par contre, reconnaissant intuitivement le paradigme auquel attribuer les différentes formes fléchies du mot, n'aurait bien évidemment pas fait cette erreur.

[27] Parmi ces lemmes fantaisistes – aussi inattendus qu'inadmissibles pour les hellénistes – figurent, par exemple, καίπειπερ (pour κἀπειδήπερ, crase constituée des formes καί et ἐπειδήπερ ; Anagnostès, I = TSARAS 1958, p. 4, l. 18), προσεπιβατάπλαι (pour la forme προσεπικαταβάλλει, lemme προσεπικαταβάλλω ; Eustathe, XLVI = KYRIAKIDIS 1961, p. 54, l. 19) ou εἰκοναχμάω (pour la forme εἰκονομάχον, lemme εἰκονομάχος ; Eustathe, XCIV = KYRIAKIDIS 1961, p. 108, l. 27).

7.	**μόνους**	**ἡμῖν**	**τοὺς**	**τέσσαρας**	**νεώς,**	**[...]**	**καταλελοιπώς**
GREgORI	μόνος A	ἡμεῖς PRO+Per1p	ὁ DET	τέσσαρες NUM+Car	ναός N+Com		καταλείπω V
RNN	μόνος A	ἡμεῖς PRO+Per1p	ὁ DET	τέσσαρες NUM+Car	ναῦς N+Com		καταλείπω V
DTC	μόνος A	ἡμεῖς PRO+Per1p	ὁ DET	τέσσαρες NUM+Car	ναός N+Com		καταλείπω V

En 7, la forme νεώς, peut, hors contexte, avoir pour lemme ναῦς « le navire » ou ναός « l'église ». Le modèle lui attribue le lemme ναῦς, ce qui n'est pas correct[28]. Dans cet extrait, on remarquera aussi que le modèle analyse correctement la forme καταλελοιπώς, sous le lemme καταλείπω, alors qu'il ne résout pas la forme πεπομφώς, bien que ces deux formes relèvent de la même analyse grammaticale (actif participe parfait, nominatif masculin singulier).

Pour terminer, le tableau 6 offre un échantillon d'analyses tiré du texte de Jean Kaminiatès. Les tableaux 8 et 9, placés en annexe, fournissent un échantillon d'analyses pour les textes d'Eustathe et d'Anagnostès.

[28] La forme νεώς (attestée neuf fois dans le corpus DTC) est un doublet atticisant de ναούς (attestée cinq fois dans le corpus DTC). Ici encore le modèle d'analyse se montre « rigoureux », puisque toutes les formes νεώς reçoivent, même erronément, le lemme ναῦς. En grec et dans le corpus DTC, l' « église » est aussi désignée par les termes ἐκκλησία (dont le sens premier est l' « assemblée » ; quinze occurrences dans DTC) ou encore le classicisant οἶκος (dont le sens premier est la « maison » ; trente-trois occurrences dans DTC) et l'archaïsant τέμενος (« portion de territoire avec un autel ou un temple ; sanctuaire »). Les expressions « ὁ τῆς παντουργοῦ καὶ θείας τοῦ ὑπερουσίου λόγου σοφίας οἶκος » (Kaminiatès, XI = BÖHLIG 1973, p. 12, l. 18) et τὸ τῆς μεγίστης τοῦ Θεοῦ Σοφίας εὐαγέστατον τέμενος (Eustathe, XVI = KYRIAKIDIS 1961, p. 22, l. 25), « le temple de la Divine-Sagesse du Verbe éternel » (ODORICO 2005, p. 69), désignent l'une et l'autre, tout simplement, l'église Sainte-Sophie à Thessalonique. Ces seuls exemples rappellent la variété lexicale et sémantique du grec à laquelle les héllenistes en général et les byzantinistes en particulier sont accoutumés.

Kaminiatès, XXVIII, 1 (= BÖHLIG 1973, p. 26, l. 23-27) – « Après nous avoir attaqué à plusieurs reprises pendant toute cette journée, les barbares reculèrent, car ils avaient été défaits toujours davantage. Au signal reçu, ils retirèrent leurs vaisseaux, abandonnant le combat par mer, et mouillèrent près de la plage du côté oriental [de la ville] » (traduction ODORICO 2005, p. 86).

		Analyses « par dictionnaires »		**Analyses par réseau de neurones**	
	Formes du texte	**Lemmes**	**Catégories morphosyntaxiques**	**Lemmes**	**Catégories morphosyntaxiques**
1	Οἱ	ὁ	DET	ὁ	DET
2	βάρβαροι	βάρβαρος	A	βάρβαρος	A
3	δὲ	δέ	I+Part	δέ	I+Part
4	οὐχ	οὐ	I+Neg	οὐ	I+Neg
5	ἅπαξ	ἅπαξ	I+Adv	ἅπαξ	I+Adv
6	μόνον	μόνος	A	μόνος	A
7	ἀλλὰ	ἀλλά	I+Part	ἀλλά	I+Part
8	καὶ	καί	I+Part	καί	I+Part
9	πολλάκις	πολλάκις	I+Adv	πολλάκις	I+Adv
10	διὰ	διά	I+Prep	διά	I+Prep
11	πάσης	πᾶς	A	πᾶς	A
12	ἐκείνης	ἐκεῖνος	PRO+Dem	ἐκεῖνος	PRO+Dem
13	τῆς	ὁ	DET	ὁ	DET
14	ἡμέρας	ἡμέρα	N+Com	ἡμέρα	N+Com
15	συνεφορμήσαντες,	συνεφορμάω	V	συνεφορμέω	V
16	μᾶλλον	μάλα	I+Adv	μάλα	I+Adv
17	ἢ	ἤ (καί)	I+Part	ἤ (καί)	I+Part
18	πρότερον	πρότερος	A	πρότερος	A
19	πεπληγμένοι	πλήσσω	V	πλήσσω	V
20	πρὸς	πρός	I+Prep	πρός	I+Prep
21	τοὐπίσω	ὁ@ὀπίσω	DET@I+Adv	ὁ@ὀπίσω	DET@I+Adv
22	χωρήσαντες,	χωρέω	V	χωρέω	V
23	ὑφ΄	ὑπό	I+Prep	ὑπό	I+Prep
24	ἑνὶ	εἷς	NUM+Car	εἷς	NUM+Car
25	συνθήματι	σύνθημα	N+Com	σύνθημα	N+Com
26	τὴν	ὁ	DET	ὁ	DET
27	[πρὸς]	πρός	I+Prep	πρός	I+Prep
28	καθ΄	κατά	I+Prep	κατά	I+Prep
29	ὕδατα	ὕδωρ	N+Com	ὕδωρ	N+Com
30	λιπόντες	λείπω	V	λείπω	V
31	μάχην	μάχη	N+Com	μάχη	N+Com
32	ὑπανεχώρουν	ὑπαναχωρέω	V	ὑπανωχέω	V
33	ταῖς	ὁ	DET	ὁ	DET
34	ναυσί,	ναῦς	N+Com	ναῦς	N+Com
35	καί	καί	I+Part	καί	I+Part
36	τινι	τις	PRO+Ind	τις	PRO+Ind
37	πρὸς	πρός	I+Prep	πρός	I+Prep
38	ἀνατολὰς	ἀνατολή	N+Com	ἀνατολή	N+Com
39	ὄντι	εἰμί	V	εἰμί	V
40	τῆς	ὁ	DET	ὁ	DET

41	πόλεως	πόλις	N+Com	πόλις	N+Com
42	αἰγιαλῷ	αἰγιαλός	N+Com	αἰγιαλός	N+Com
43	προσωρμίσθησαν.	προσορμίζω	V	προσορμίζω	V

Tableau 6 : comparaison des résultats sur un extrait de Kaminiatès (904)

Dans cet extrait, les deux lemmes συνεφορμέω (ligne 15) et ὑπανωχέω (ligne 32) proposés par le modèle d'analyse sont des prédictions erronées (et fantaisistes). En revanche, les catégories morphosyntaxiques prédites sont correctes.

L'examen des résultats prédits par le modèle d'analyse n'est pas encore achevé. Des investigations complémentaires seront réalisées afin de mieux comprendre l'origine des prédictions erronées proposées par le modèle d'analyse et d'y remédier. Les données du tableau 9, basées cette fois sur les catégories morphosyntaxiques, apportent déjà une partie de la réponse.

	Nombre d'occurrences	**Nombre d'analyses correctes**	**Pourcentage**
Adjectifs	5.873	5.520	93,99%
Noms	10.347	9.734	94,08%
Noms communs	9.457	9.063	95,83%
Anthroponymes	337	240	71,22%
Toponymes	163	108	66,26%
Noms propres	86	29	33,72%
Pronoms	6.379	6.316	99,01%
Verbes	11.618	10.825	93,17%
Emprunts au latin	64	33	51,56%

Tableau 7 : effectifs des prédictions correctes classées selon leurs catégories morphosyntaxiques

Le modèle d'analyse est moins performant sur les anthroponymes (71,22% de résultats corrects)[29], sur les toponymes (66,26% de résultats corrects) et sur les autres types de noms propres (33,72%) que sur les autres catégories morphosyntaxiques, tels les adjectifs, les noms communs, les pronoms et les verbes, dont les scores sont toujours supérieurs à 93%. Ce n'est pas une surprise dans la mesure où les noms propres de personnes et de lieux, dans les sources historiques byzantines, peuvent être nombreux et variés. Le modèle d'analyse se montre également moins performant pour identifier les mots empruntés aux latins (51,56%), présents dans les langues grecques tardive et byzantine, mais sans doute moins systématisés que les autres mots au niveau de leur intégration dans la langue.

En revanche, le modèle a correctement analysé onze lemmes anthroponymiques (Ἀλεξίῳ *s.l.* Ἀλέξιος, Ἄρην *s.l.* Ἄρης, Διομήδει *s.l.* Διομήδης, Διονύσιοι *s.l.* Διονύσιος (ὁ), Ἠσαΐᾳ *s.l.* Ἠσαΐας, Ἰερεμίαν *s.l.* Ἰερεμίας, ΙΩΑΝΝΟΥ *s.l.* Ἰωάννης, Ῥογέριος *s.l.* Ῥογέριος, Φαλάριδες *s.l.* Φάλαρις, Φανίου *s.l.* Φάνιος, Χάρωνος *s.l.* Χάρων), quatre lemmes toponymiques (Δυρραχίῳ *s.l.* Δυρράχιον, Νίκαια *s.l.* Νίκαια (τοπ), Πανόρμου *s.l.* Πάνορμος, Προῦσα *s.l.* Προῦσα) et cinq lemmes de noms propres (Δαλασηνός *s.l.* Δαλασηνός, Ἐριννύς et Ἐριννύσιν *s.l.* Ἐρινύς, Κομνηνός et Κομνηνῷ *s.l.* Κομνηνός, Χοῦμνον, Χοῦμνος et Χούμνου *s.l.* Χοῦμνος), alors que les formes citées ici sont absentes de la vérité de terrain, ainsi que deux lemmes empruntés au latin (forme κουρίαν *s.l.* κουρία et forme σίγνου *s.l.* σίγνον, deux lemmes nominaux).

[29] Comme l'illustrent les exemples 1 et 3, ci-dessus.

L'ensemble des données du corpus DTC se composent actuellement de trois types de résultats : 1) les analyses produites par l'approche dite « par dictionnaires » réalisées en mettant en œuvre les ressources linguistiques du projet GREgORI ; 2) celles produites par le modèle d'analyse issu du réseau de neurones mis en place par Calfa ; 3) les données finales révisées et corrigées manuellement. Ces trois résultats peuvent donc se prêter à des comparaisons. Les figures 1 et 2 en fournissent une représentation graphique, d'abord pour la lemmatisation, ensuite pour la caractérisation morphosyntaxique.

La figure 1, représente les lemmes classés selon leurs catégories morphosyntaxiques (en ordonnée)[30]. On remarque d'emblée que, pour la majorité des mots outils, tels les pronoms (étiquettes « PRO+ »), les propositions de GREgORI et les prédictions du réseau présentent clairement des similitudes. Il en est de même pour les adverbes (I+Adv), les adverbes prépositionnels (I+AdvPr), les conjonctions (I+Conj), les interjections (I+Intj), les négations (I+Neg) et les particules (I+Part). *A contrario*, les noms communs (N+Com) empruntés à d'autres langues montrent des résultats plus disparates. Les noms communs empruntés à l'italien (N+Com+eItal) ou au latin (N+Com+eLat), sont moins proches des réponses attendues et le réseau de neurones prédit un meilleur lemme ou une meilleure catégorie morphosyntaxique que ce qui est proposé par GREgORI. Cela vaut ici pour la lemmatisation comme pour la caractérisation morphosyntaxique. Pour les déterminants numériques cardinaux (NUMA+Car), la lemmatisation par GREgORI est plus efficace que les prédictions, mais c'est l'inverse pour la catégorisation morphosyntaxique, les prédictions du réseau étant alors meilleures. La représentation des anthroponymes (N+Ant), des toponymes (N+Top) et des noms propres (N+Prop) correspond aux données du tableau 7. La figure 2 représente de la même manière les résultats de la catégorisation morphosyntaxique.

Les figures 3, 4, 5, 6, 7 et 8, placées en annexes, produisent ces données texte par texte.

[30] Les lemmes proposés par GREgORI sont symbolisés au moyen de ronds rouges, ceux prédits par le réseau de neurones au moyen de triangles verts. Ces éléments sont positionnés sur une grille (en abscisse) selon leur différence ou leur similitude avec les résultats attendus (les données finales révisées manuellement). Plus un élément s'approche de zéro (à gauche de la figure), plus sa similitude avec l'élément correspondant dans les données finales est grande (distance de Levenshtein). Plus un élément s'éloigne de l'élément correspondant dans les données finales, plus il s'écarte de la valeur zéro (vers la droite de la figure). Les lignes verticales verte (proche de la valeur zéro) et rouge (entre zéro et deux) figurent les moyennes pondérées pour le réseau de neurones, d'abord, et pour GREgORI, ensuite.

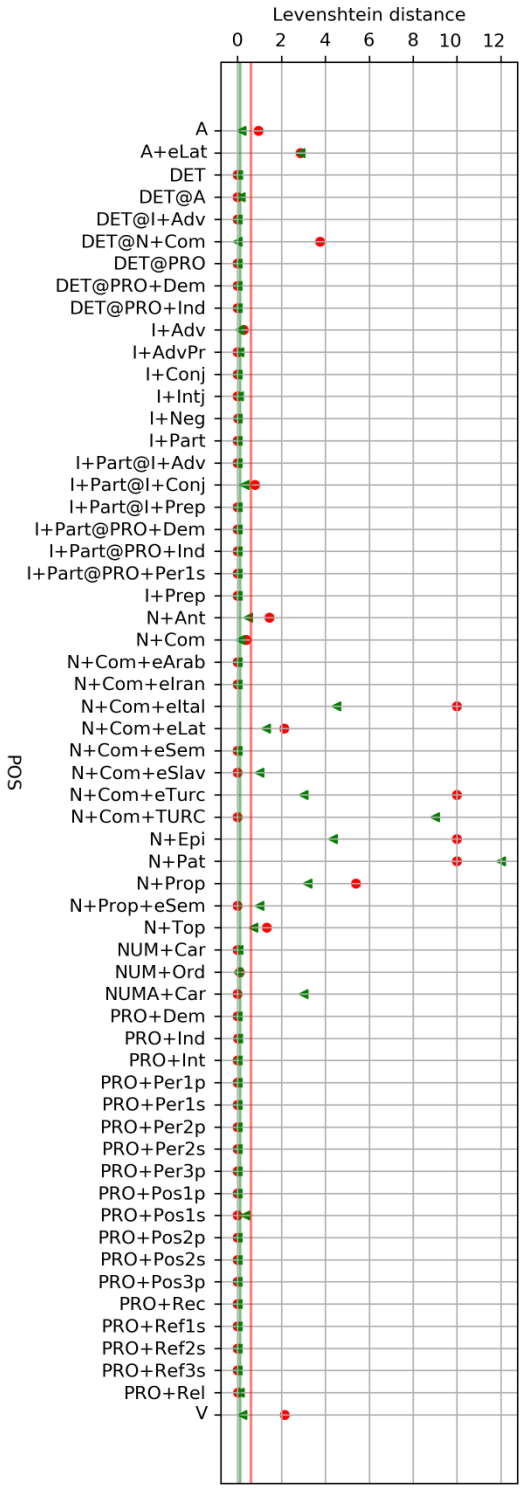

Figure 1 : représentation graphique des résultats de la lemmatisation dans chaque catégorie morphosyntaxique (corpus DTC)

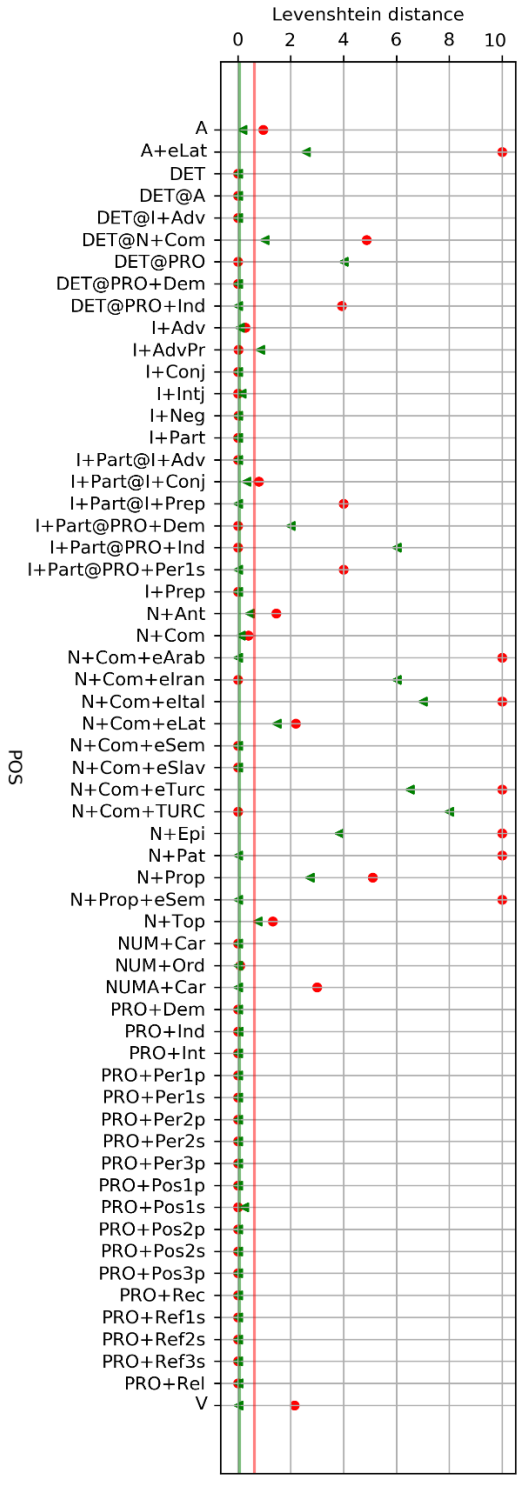

Figure 2 : représentation graphique des résultats de la catégorisation morphosyntaxique (corpus DTC)

3. Conclusions et perspectives

Les *Narrationes* du corpus DTC sont des pièces remarquables de l'historiographie byzantine. Même si elles ne sont pas très longues (voir le tableau 1 : 22.365 mots-occurrences pour Jean Kaminiatès, 28.515 pour Eustathe de Thessalonique et 9.613 pour Jean Anagnostès ; l'œuvre de Procope de Césarée compte à elle seule 292.552 mots-occurrences), elles se distinguent par leur contenu et par les intentions, explicites ou implicites, de leurs auteurs qui détournent les codes du genre au profit de leurs propres intérêts ; Paolo Odorico qualifie ces œuvres d'exceptionnelles.

Leur intégration dans les corpus du projet GREgORI allait de soi. En dresser l'inventaire lexical ouvre la voie à d'autres recherches sur la langue et le style de ces textes (sujet encore peu exploré, voir note 7).

Mais à cette occasion, une analyse dite « par dictionnaires », méthode jusqu'alors privilégiée dans le cadre du projet, a été complétée par une approche reposant sur l'utilisation de réseaux de neurones. Cette démarche correspond aux tendances actuelles dans le domaine des humanités numériques. La valider sur des textes grecs d'époque byzantine, mais aussi sur des sources en arménien, en géorgien ou en syriaque, est une opération intéressante et, à terme, féconde[31]. La collaboration avec Calfa permet opportunément de mettre en œuvre de telles expérimentations. Les données de la première analyse, vérifiées manuellement, servent de données de références stables, informations indispensables, pour évaluer la qualité des résultats produits par la seconde analyse.

L'analyse par réseau de neurones fournit des résultats encourageants. Pour la lemmatisation, sur l'ensemble des mots, 97,94% des prédictions sont correctes (tableau 4). Pour la catégorisation morphosyntaxique, 98,31% des prédictions sont correctes (tableau 5). Ces mesures sont des seuils. Les analyses et les développements ultérieurs permettront de les améliorer. Il s'agira d'accroître le volume des données d'apprentissage et d'en affiner la qualité, tant en grec que dans les différentes langues.

Parallèlement, une analyse détaillée des résultats obtenus sera aussi instructive (figures 1-8). Cela permettra, par exemple, de déterminer quels éléments du lexique de chaque langue nécessitent la création de ressources linguistiques plus adéquates ou réclament des phases d'apprentissage plus spécifiques.

À ce stade, pour optimiser les analyses et minimiser les interventions humaines encore indispensables pour la validation des résultats, GREgORI et Calfa privilégient une méthodologie hybride combinant l'approche « par dictionnaires » et celle par réseau de neurones[32]. L'approche « par dictionnaires » s'avère très efficace pour les mots répondant à une seule analyse, un lemme et une catégorie morphosyntaxique (tableaux 2 et 3). Mais elle impose un travail de révision pour les mots inconnus et les mots répondant à plusieurs analyses. L'approche par réseau de neurones prédit des analyses fiables, comme le démontre la position de la moyenne pondérée pour cette approche (sur les figures 1-8), même pour les mots inconnus et les mots répondant à plusieurs analyses, pour autant que les apprentissages reposent sur des données volumineuses et bien analysées.

---

[31] STRASSEL *et alii* 2011, p. 1 : « All human language technology demands substantial quantities of data for system training and development, plus stable benchmark data to measure ongoing progress. While creation of high quality linguistic resources is both costly and time consuming, such data has the potential to profoundly impact not just a single evaluation program but language technology research in general ».

[32] Voir VIDAL-GORÈNE, KINDT 2022.

ANNEXES

Annexe 1. *Tableau 8 : échantillon d'analyses pour le texte d'Eustathe de Thessalonique*

Eusthate, LIX (= KYRIAKIDIS 1961, p. 72, l. 28-33) – « En revanche les forces ennemies, qui venaient par terre et par mer, se préparaient méthodiquement à la bataille. Ceux qui se trouvaient à l'ouest construisaient des machines de guerre plus modernes que celles qu'on voit d'habitude, et que leur taille rendait difficiles à manœuvrer. Finalement, elles ne donnèrent pas de grands résultats » (traduction ODORICO 2005, p. 194).

		Analyses « par dictionnaires »		**Analyses par réseau de neurones**	
	Formes du texte	**Lemmes**	**Catégories morphosyntaxiques**	**Lemmes**	**Catégories morphosyntaxiques**
1	Τὸ	ὁ	DET	ὁ	DET
2	δὲ	δέ	I+Part	δέ	I+Part
3	πολέμιον	πολέμιος	A	πολέμιος	A
4	ἐπελθόν,	ἐπέρχομαι	V	ἐπέρχομαι	V
5	τό	ὁ	DET	ὁ	DET
6	τε	τε	I+Part	τε	I+Part
7	ξηρᾶθεν	ξηρᾶθεν	I+Adv	ξηραθέω	I+Adv
8	καὶ	καί	I+Part	καί	I+Part
9	τὸ	ὁ	DET	ὁ	DET
10	ἐκ	ἐκ	I+Prep	ἐκ	I+Prep
11	θαλάσσης,	θάλασσα	N+Com	θάλασσα	N+Com
12	ἐξηρτύετο	ἐξαρτύω	V	ἐξαρτύω	V
13	κατ΄	κατά	I+Prep	κατά	I+Prep
14	ἐπιστήμην	ἐπιστήμη	N+Com	ἐπιστήμη	N+Com
15	εἰς	εἰς	I+Prep	εἰς	I+Prep
16	μάχην.	μάχη	N+Com	μάχη	N+Com
17	Καὶ	καί	I+Part	καί	I+Part
18	οἱ	ὁ	DET	ὁ	DET
19	μὲν	μέν	I+Part	μέν	I+Part
20	ἐκ	ἐκ	I+Prep	ἐκ	I+Prep
21	τῶν	ὁ	DET	ὁ	DET
22	δυσμικῶν	δυσμικός	A	δυσμικός	A
23	ἄλλα	ἄλλος	PRO+Ind	ἄλλος	PRO+Ind
24	ἐποίουν	ποιέω	V	ποιέω	V
25	καινά	καινός	A	καινός	A
26	τινα	τις	PRO+Ind	τις	PRO+Ind
27	κατὰ	κατά	I+Prep	κατά	I+Prep
28	νόμους	νόμος	N+Com	νόμος	N+Com
29	ἑλεπόλεων,	ἑλέπολις	A	ἑλέπολις	A
30	αἷς	ὅς ἥ ὅ	PRO+Rel	ὅς ἥ ὅ	PRO+Rel
31	διὰ	διά	I+Prep	διά	I+Prep
32	τὸ	ὁ	DET	ὁ	DET
33	ἐκ	ἐκ	I+Prep	ἐκ	I+Prep
34	μεγέθους	μέγεθος	N+Com	μέγεθος	N+Com
35	δυσμεταχείριστον	δυσμεταχείριστος	A	δυσχεταγίσιρ	A
36	οὐδὲ	οὐδέ	I+Part	οὐδέ	I+Part
37	ἐνέλαμψέ	ἐλλάμπω	V	ἐλλάμπω	V

38	τις	τις	PRO+Ind	τις	PRO+Ind
39	ἐνέργεια,	ἐνέργεια	N+Com	ἐνέργεια	N+Com

Tableau 7 : comparaison des résultats sur un extrait d'Eustathe (1185)

Les lemmes ξηραθέω (ligne 7) et δυσχεταγίσιρ (ligne 35) sont des prédictions erronées. Les catégories morphosyntaxiques prédites sont correctes, malgré l'étonnante catégorisation comme adverbe (« I+Adv ») de ξηραθέω.

Annexe 2. ***Tableau 9 : échantillon d'analyses pour le texte de Jean Anagnostès***

Anagnostès (= TSARAS 1958, p. 18, l. 20-27) – « Ils s'approchèrent de la ville, plantèrent leurs tentes comme à l'accoutumée, et ils se mirent à rôder tout autour, comme s'ils en étaient les sentinelles : on avait l'impression qu'il n'y avait pas un seul endroit vide d'hommes. Mourad rangea ses commandants en position de combat, chacun devant un endroit précis de la ville, et fit planter ses tentes juste en face de l'Acropole, pour pouvoir bien observer tous les combattants, ainsi que tout l'intérieur de la ville, d'une position élevée » (traduction ODORICO 2005, p. 266).

		Analyses « par dictionnaires »		**Analyses par réseau de neurones**	
	Formes du texte	**Lemmes**	**Catégories morphosyntaxiques**	**Lemmes**	**Catégories morphosyntaxiques**
1	Τῇ	ὁ	DET	ὁ	DET
2	πόλει	πόλις	N+Com	πόλις	N+Com
3	δὲ	δέ	I+Part	δέ	I+Part
4	προσεγγίσαντες	προσεγγίζω	V	προσεγγίζω	V
5	καὶ	καί	I+Part	καί	I+Part
6	τὰς	ὁ	DET	ὁ	DET
7	σκηνάς,	σκηνή	N+Com	σκηνή	N+Com
8	ὥσπερ	ὥσπερ	I+Conj	ὥσπερ	I+Conj
9	ἔθος,	ἔθος	N+Com	ἔθος	N+Com
10	πηξάμενοι,	πήγνυμι	V	πήγνυμι	V
11	δίκην	δίκη	N+Com	δίκη	N+Com
12	φρουρίου	φρούριον	N+Com	φρούριον	N+Com
13	πᾶσαν	πᾶς	A	πᾶς	A
14	περιέλαβον	περιλαμβάνω	V	περιλαμβάνω	V
15	ταύτην,	οὗτος	PRO+Dem	οὗτος	PRO+Dem
16	ὡς	ὡς (ὅς)	I+Conj	ὡς (ὅς)	I+Conj
17	μηδαμοῦ	μηδαμός	A	μηδαμός	A
18	σχεδὸν	σχεδόν	I+Adv	σχεδόν	I+Adv
19	κενὸν	κενός	A	κενός	A
20	ἀνθρώπων	ἄνθρωπος	N+Com	ἄνθρωπος	N+Com
21	φαίνεσθαι	φαίνω	V	φαίνω	V
22	τόπον.	τόπος	N+Com	τόπος	N+Com
23	Εἶτα	εἶτα	I+Adv	εἶτα	I+Adv
24	τοὺς	ὁ	DET	ὁ	DET
25	στρατηγοὺς	στρατηγός	N+Com	στρατηγός	N+Com
26	ἕκαστον	ἕκαστος	PRO+Ind	ἕκαστος	PRO+Ind
27	ἐν	ἐν	I+Prep	ἐν	I+Prep
28	ὡρισμένῳ	ὁρίζω	V	ὁρίζω	V
29	τῆς	ὁ	DET	ὁ	DET

30	πόλεως	πόλις	N+Com	πόλις	N+Com
31	μέρει	μέρος	N+Com	μέρος	N+Com
32	πρὸς	πρός	I+Prep	πρός	I+Prep
33	τὸ	ὁ	DET	ὁ	DET
34	πολεμεῖν	πολεμέω	V	πολεμέω	V
35	ἀποτάξας,	ἀποτάσσω	V	ἀποτάσσω	V
36	αὐτὸς	αὐτός	PRO+Dem	αὐτός	PRO+Dem
37	τὰς	ὁ	DET	ὁ	DET
38	ἰδίας	ἴδιος	A	ἴδιος	A
39	σκηνὰς	σκηνή	N+Com	σκηνή	N+Com
40	ἀντικρὺ	ἄντικρυ	I+AdvPr	ἀντικρύ	I+AdvPr
41	τῆς	ὁ	DET	ὁ	DET
42	ἀκροπόλεως	ἀκρόπολις	N+Com	ἀκρόπολις	N+Com
43	πήγνυσιν,	πήγνυμι	V	πήγνυμι	V
44	ὡς	ὡς (ὅς)	I+Conj	ὡς (ὅς)	I+Conj
45	ἂν	ἄν	I+Part	ἄν	I+Part
46	πάντας	πᾶς	A	πᾶς	A
47	ἐκ	ἐκ	I+Prep	ἐκ	I+Prep
48	μετεώρου	μετέωρος	A	μετέωρος	A
49	καλῶς	καλῶς	I+Adv	καλῶς	I+Adv
50	ὁρᾶν	ὁράω	V	ὁράω	V
51	ἔχοι	ἔχω	V	ἔχω	V
52	καὶ	καί	I+Part	καί	I+Part
53	πᾶσαν	πᾶς	A	πᾶς	A
54	τὴν	ὁ	DET	ὁ	DET
55	πόλιν	πόλις	N+Com	πόλις	N+Com
56	ἐντός.	ἐντός	I+AdvPr	ἐντός	I+AdvPr

Tableau 8 : comparaison des résultats sur un extrait d'Anagnostès (1430)

Dans ce troisième extrait, le modèle d'analyse propose à raison un lemme ἀντικρύ (ligne 40)[33] là où les ressources linguistiques du projet GREgORI ont ἄντικρυ. Ici aussi, pour ces mots, les catégories morphosyntaxiques prédites sont correctes.

[33] Voir BAILLY 2000, p. 180 ou LIDDELL-SCOTT 1977, p. 157.

Annexe 3. *Figures 3-4*

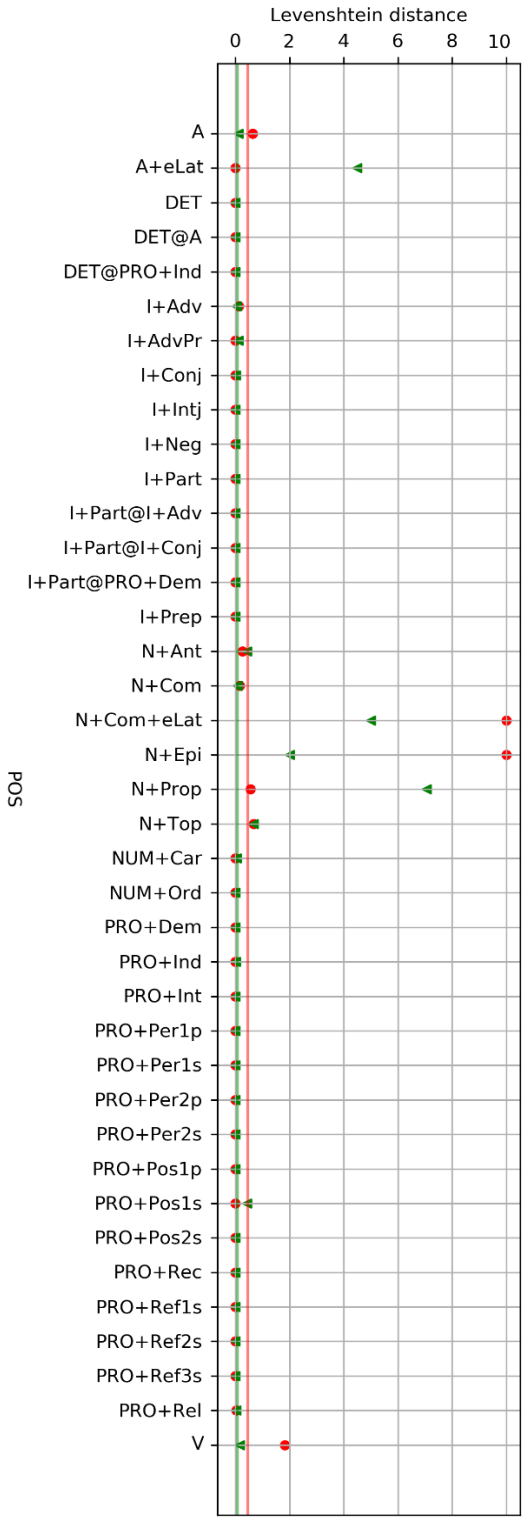

Figure 3 : représentation graphique des résultats de la lemmatisation dans chaque catégorie morphosyntaxique (Jean Kaminiatès)

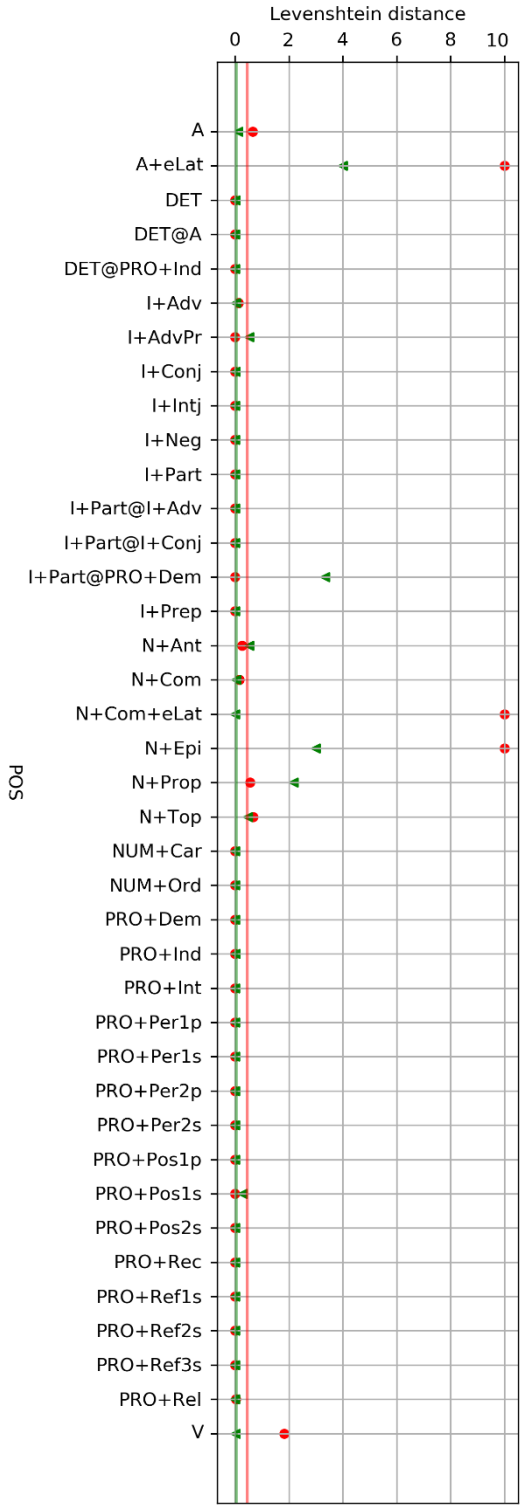

Figure 4 : représentation graphique des résultats de la catégorisation morphosyntaxique (Jean Kaminiatès)

Annexe 4. *Figures 5-6*

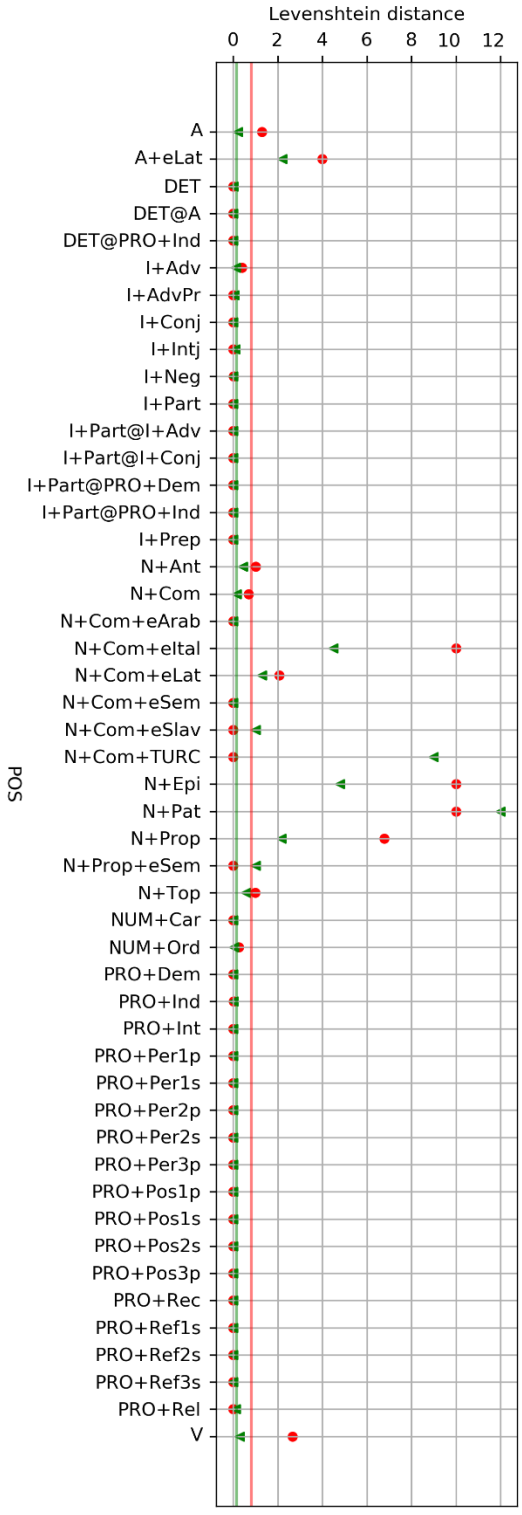

Figure 5 : représentation graphique des résultats de la lemmatisation dans chaque catégorie morphosyntaxique (Eustathe de Tessalonique)

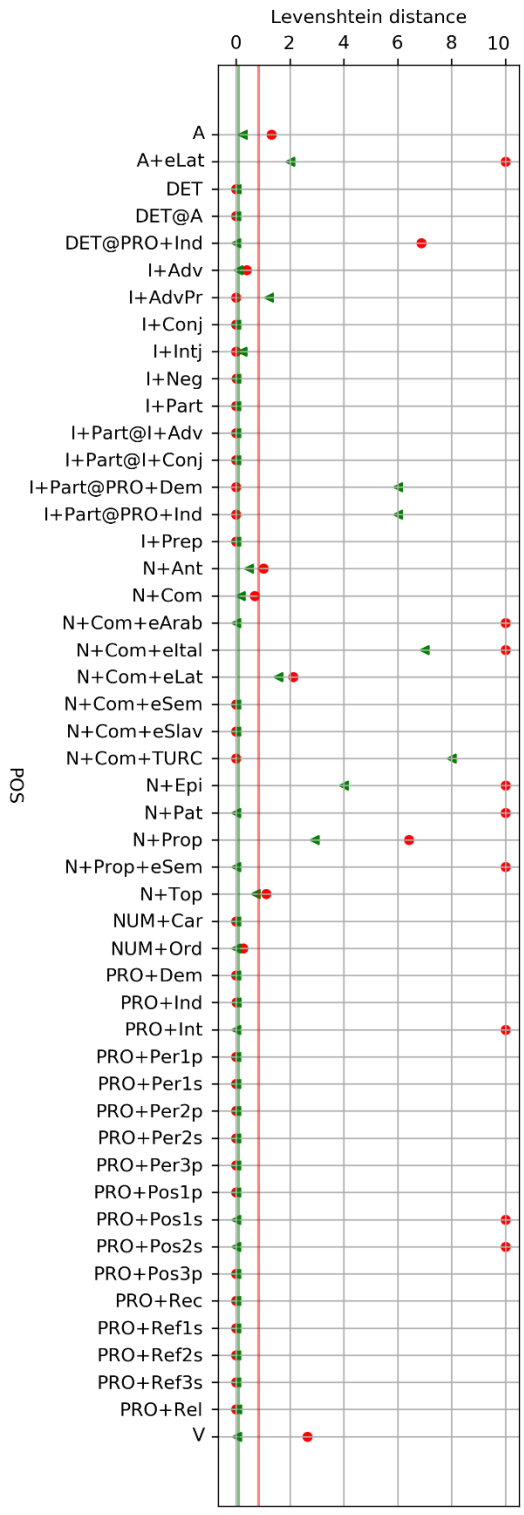

Figure 6 : représentation graphique des résultats de la catégorisation morphosyntaxique (Eustathe de Tessalonique)

Annexe 5. *Figures 7-8*

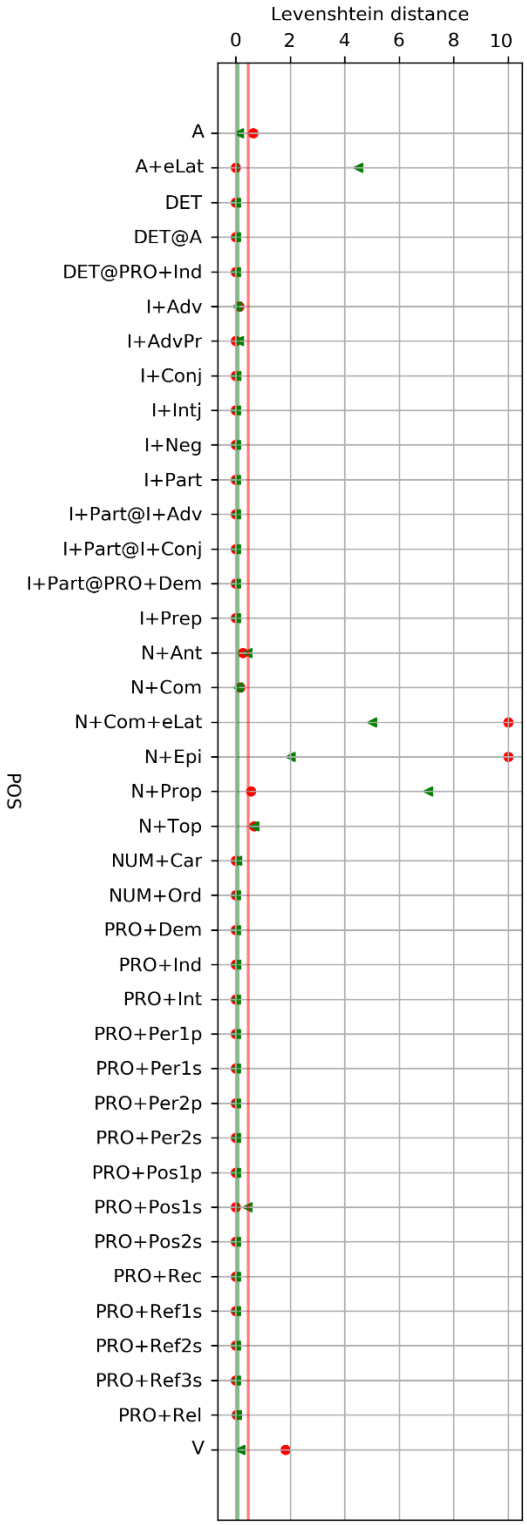

Figure 7 : représentation graphique des résultats de la lemmatisation dans chaque catégorie morphosyntaxique (Jean Anagnostès)

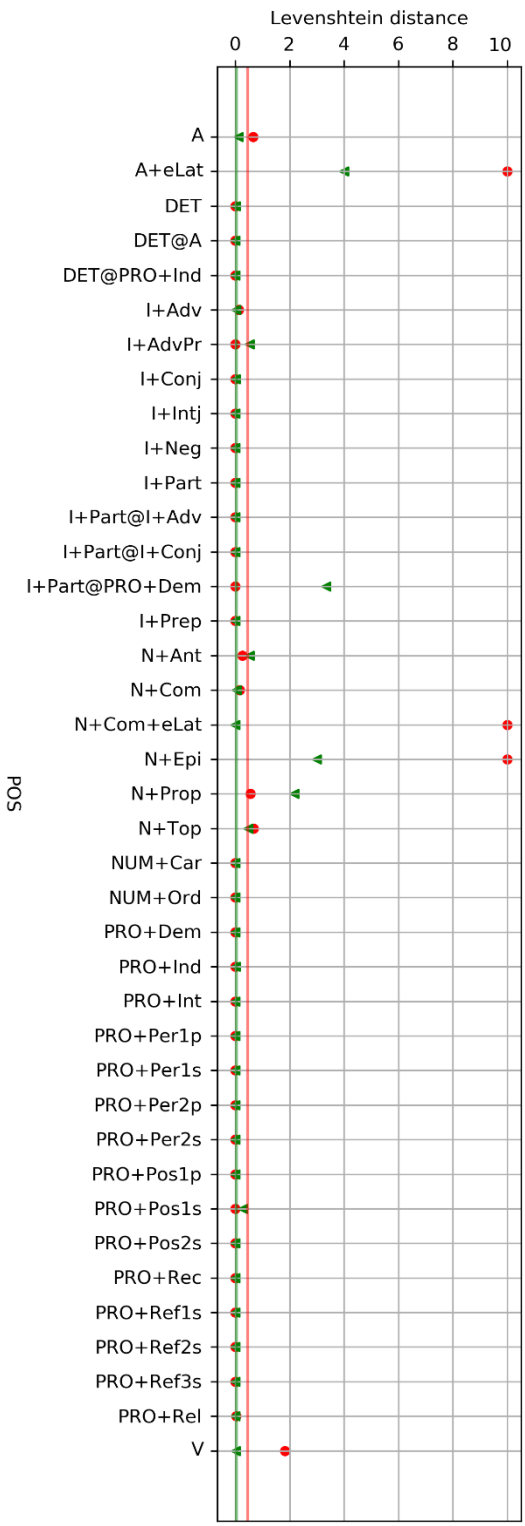

Figure 8 : représentation graphique des résultats de la catégorisation morphosyntaxique (Jean Anagnostès)

BIBLIOGRAPHIE

BAILLY, A., 2000 : *Dictionnaire Grec Français*, rédigé avec le concours de E. EGGER, 26e éd. revue et corrigée par L. SECHAN et P. CHANTRAINE, Paris, 1963 (réimpr. 2000).

BEKKER, E. (éd.), 1838a : *Theophanes Continuatus, Ioannes Cameniata, Symeon Magister, Georgius Monachus* (Corpus Scriptorum Historiae Byzantinae), Bonn.

— 1838b : *Georgius Phrantzes, Ioannes Cananus, Iaonnes Anagnostes* (Corpus Scriptorum Historiae Byzantinae), Bonn.

— 1848 : *Leo Grammaticus, Eustathius* (Corpus Scriptorum Historiae Byzantinae), Bonn.

BÖHLIG, G. (éd.), 1973 : *Ioannis Caminiatae de expugnatione Thessalonicae* (Corpus Fontium Historiae Byzantinae, Series Berolinensis, 4), Berlin.

CAPONE, A. (éd.), 2021 : *Sancti Gregorii Nazianzeni Opera*. Versio Latina, I. *Epistulae 102-101* cum indice verborum a B. KINDT et B. COULIE confecto (Corpus Christianorum. Series Graeca, 98. Corpus Nazianzenum, 31), Turnhout.

COULIE, B., 1996 : « La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs », *Byzantion* 66, p. 35-54.

COULIE, B., KINDT, B. , KEPEKLIAN, G., 2021 : « Étiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l'arménien ancien », *Études Arméniennes Contemporaines* (sous presse).

COULIE, B., KINDT, B., PATARIDZE, T., 2013 : « Lemmatisation automatique des sources en géorgien ancien », *Le Muséon* 126, p. 161-201.

DELLE DONNE, S., KINDT, B., 2021 : *De Thessalonica Capta. John Caminiates, The capture of Thessaloniki - Eustathios of Thesssaloniki, The conquest of Thessalonica - John Anagnostes, Account of the Last Capture of Thessalonica – Greek Concordance* (*UCLouvain*, *GREgORI Project*), Louvain-la-Neuve.

DERCZYNSKI, L., 2016 : « Complementarity, F-score, and NLP Evaluation », dans N. CALZOLARI *et alii.* (éd.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016, Portorož), Paris, p. 261-266.

DEREZA, O., 2018 : « Lemmatization for Ancient Languages : Rules or Neural Networks? », dans D. USTALOV, A. FILCHENKOV, L. PIVOVAROVA, J. ŽIŽKA (éd.), *7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings*, New York, Dordrecht, Heidelberg, London, p. 35–47.

FRENDO, D., FOTIOU, A., 2000 : *John Kaminiates, The Capture of Thessaloniki. Translation, introduction and notes* (Byzantina Australiensia, 12), Canberra.

GRAND'HENRY, J., 2020 : *Sancti Gregorii Nazianzeni Opera*. Versio Arabica Antiqua, V. Oratio XLII (arab. 14) (Corpus Christianorum. Series Graeca, 99. Corpus Nazianzenum, 32), Turnhout.

HUNGER, H., 1978 : *Die hochspracheliche profane Literatur der Byzantiner*. Erster Band. *Philosophie, Rhetorik, Epistolographie, Geschichtsschreibung, Geographie* (*Byzantinisches Handbuch*, XII, 5, 1), Munich.

KAZHDAN, A.P., 1991 : *The Oxford Dictionary of Byzantium*, 3 vol., Oxford.

KINDT, B., 2004 : « La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien. Les principes de formulation des lemmes du Dictionnaire Automatique Grec », *Byzantion*, 74, p. 213-272.

— 2018 : « Processing Tools for Greek and Other Languages of the Christian Middle East », *Journal of Data Mining and Digital Humanities, Episciences.org, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.*
(en ligne : https://jdmdh.episciences.org/4184/pdf)

— 2021 : « Du texte à l'index. L'étiquetage lexical du *De Septem Orbis Spectaculis* de Philon le Paradoxographe : méthode et finalité », dans G. LABARRE, *Sources, Histoire et Éditions. Les outils de la recherche. Formation et recherche en sciences de l'Antiquité* (*Institut des sciences et techniques de l'Antiquité*), Besançon, p. 175-218.

KINDT, B., HAELEWYCK, J.-C., SCHMIDT, A.B., ATAS, N., 2018 : « La concordance bilingue grecque-syriaque des Discours de Grégoire de Nazianze », *BABELAO* 7, p. 51-80.

KINDT, B., PIRARD, M., 2016 : « De Nazianze à Ninive. La couverture lexicale du Dictionnaire Automatique Grec », dans V. SOMERS, P. YANNOPOULOS (éd.), *Philokappadox*. *In memoriam Justin Mossay* (Orientalia Lovaniensia Analecta, 251. Bibliothèque de *Byzantion*, 14), Louvain, Paris, Bristol CT, p. 49-77.

KYRIAKIDIS, S. (éd.), 1961 : *Eustazio di Tessalonica, La espugnazione di Tessalonica* (Testi e Monumenti, 5), Palerme.

LIDDELL–SCOTT 1977 = H.G. LIDDELL, R. SCOTT, H.S. JONES, *A Greek-English Lexicon*, 9e éd., Oxford, 1940 (réimpr. 1977).

MANJAVACAS, E., KÁDÁR, Á., KESTEMONT, M., 2019 : « Improving lemmatization of non-standard languages with joint learning », dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, p. 1493–1503.

MELVILLE JONES, J.R., 1988 : *Eustathios of Thessaloniki, The Capture of Thessaloniki. A translation with introduction and commentary* (Byzantina Australiensia, 8), Canberra.

MUELLNER, L., 2019 : « The Free First Thousand Years of Greek », dans M. BERTI (éd.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, Berlin, Boston, p. 7-18

ODORICO, P., 2005 : *Thessalonique. Chroniques d'une ville prise. Jean Caminiatès, Eustathe de Thessalonique, Jean Anagnostès*, textes présentés et traduits par Paolo ODORICO, Toulouse, 2005.

PATARIDZE, T. (éd.), 2020 : *Vie et conduite des Bienheureux Justes-nus et de notre saint Père Zosime : trois traductions géorgiennes* (Corpus Scriptorum Christianorum Orientalium, 686-687. Scriptores Iberici, 25-26), Leuven.

PATEL, P.P., THAKKAR, A.R., 2020 : *A Journey From Neural Networks to Deep Networks: Comprehensive Understanding for Deep Learning*, dans S. SUMATHI, M. JANANI, *Neural Networks for Natural Language Processing*, Hershey, 2020, p. 31-62.

SCHMIDT, A.B., KINDT, B., 2021 : « Eine syrische Amulettrolle mit Beschwörungen für Frauen : Erevan, Matenadaran, rot. syr. 72. Teil II. Wortindex », dans E.A. ISHAC, Th. CSANÁDY, Th. ZAMMIT LUPI (éd.) (with an Introduction by R.A. KITCHEN), *Tracing Written Heritage in a Digital Age*, Wiesbaden, p. 59-76.

STONE 2021, M. (éd.) : *The Genesis Commentary by Step'anos of Siwnik' (Dub.)* (Corpus Scriptorum Christianorum Orientalium, 695. Scriptores Armeniaci, 32), Leuven.

STRANO, G., 2013 : « Storia e modelli letterari nella Presa di Tessalonica di Giovanni Caminiata », dans A. RIGO, A. BABUIN, M. TRIZIO (éd.), *Vie per Bisanzio*. *VII Congresso Nazionale dell'Associazione Italiana di Studi Bizantini, Venezia, 25-28 novembre 2009*, vol. I, Bari, p. 61-74.

STRASSEL, S., *et alii*, 2011 : *Data Acquisition and Linguistic Resources*, dans J. OLIVE, C. CHRISTIANSON, J. MCCARY (éd.), *Handbook of Natural Language Processing and Machine Translation*, New York, Dordrecht, Heidelberg, London, 2011, p. 1-131.

TSARAS, G. (éd.), 1958 : *Ἰωάννου Ἀναγνώστου διήγησις περὶ τῆς τελευταίας ἁλώσεως τῆς Θεσσαλονίκης, μονῳδία ἐπὶ τῇ ἁλώσει τῆς Θεσσαλονίκης*, Thessalonique.

TUERLINCKX, L., 2004 : « La lemmatisation de l'arabe non classique », dans A. DISTER, C. FAIRON, G. PURNELLE (éd*.), Le poids des mots. 7es Journées internationales d'Analyse statistique des Données Textuelles, 10-12 mars 2004, Louvain-la-Neuve*, Louvain-la-Neuve, vol. II, p. 1069-1078.

VATRI, A., MCGILLIVRAY, B., 2018 : « The Diorisis Ancient Greek Corpus », *Research Data Journal for the Humanities and Social Sciences* 3, p. 55-65.

VIDAL-GORÈNE, C., DECOURS-PEREZ, A., 2020 : « Languages Resources for Poorly Endowed Languages : The Case Study of Classical Armenian », dans R. SPRUGNOLI, M. PASSAROTTI (éd.), *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020, Marseille),* Paris, p. 3145–3152.

VIDAL-GORÈNE, C., KINDT, B., 2020 : « Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian and Syriac », dans R. SPRUGNOLI, M. PASSAROTTI (éd.), *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020, Marseille)*, Paris, p. 22-27.

— 2022 : « From manuscript to tagged corpora. An automated process for Ancient Armenian or other under resourced languages of the Christian East » (article en soumission).

RÉSUMÉ

Le corpus DTC réunit des textes historiographiques grecs d'époque byzantine. Ces textes ont été analysés semi-automatiquement (lemmatisation et catégorisation morphosyntaxique) avec les outils informatiques et les ressources linguistiques du projet GREgORI (UCLouvain, Louvain-la-Neuve, Belgique) spécialisé dans le traitement automatique du grec et des langues de l'Orient chrétien. Une seconde analyse a été menée en collaboration avec l'entreprise Calfa (Paris, France) spécialisée dans le traitement de l'arménien et la mise en œuvre d'approches basées sur l'intelligence artificielle. Cette seconde analyse est réalisée par un réseau de neurones. Cette étude compare et évalue les résultats produits par les deux méthodes et propose une approche hybride pour le traitement automatique des langues concernées.

ABSTRACT

The DTC corpus brings together historical texts written in Greek during the Byzantine period. These texts were analyzed semi-automatically (lemmatization and POS-tagging) by using computer tools and linguistic resources of the GREgORI project (UCLouvain, Louvain-la-Neuve, Belgium) specialized in the NLP of Greek and the languages of the Christian East. A second analysis was carried out in collaboration with the company Calfa (Paris, France) developping NLP tools for Armenian and implementing approach relating to artificial intelligence. This second analysis is performed by a neural network. This study compares and evaluates the results produced by the two methods and proposes a hybrid approach for the processing of the languages concerned.

MOTS-CLEFS

1. Traitement automatique des langues (TAL)
2. Lemmatisation
3. Étiquetage morphosyntaxique
4. Grec ancien
5. Jean Anagnostès
6. Eustathe de Thessalonique
7. Jean Kaminiatès

KEYWORDS

1. Natural Language Processing (NLP)
2. Lemmatisation
3. POS-tagging
4. Ancient Greek
5. John Anagnsotes
6. Eusthatios of Thessalonike
7. John Kaminiates