ANR-15-CE38-0008

DEscription et MOdélisation des Chaînes de Référence : outils pour l'**A**nnotation de corpus (en diachronie et en langues comparées) et le **T**raitement automatique

Cultures, patrimoines, création (DS0805) – Edition 2015

LIVRABLE L3.2 : « Outils de détection automatique de chaînes de coréférences »

février 2020

Editeur : Frédéric Landragin (Lattice, coordinateur du projet)

Principaux contributeurs : Loïc Grobol (Lattice), Bruno Oberlé (LiLPa), Marco Dinarelli (anciennement Lattice, désormais au LIG – Laboratoire d'Informatique de Grenoble), Frédéric Landragin (Lattice)

Acronyme du projet	DEMOCRAT
Titre du projet	DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique
Coordinateur du projet (société/organisme)	Frédéric Landragin (Lattice-ENS-CNRS-Université Sorbonne Nouvelle – Paris 3)
Date de début du projet Date de fin du projet	1er mars 2016 (*T0 scientifique*) 29 février 2020 (*Tfinal scientifique*)
Labels et correspondants des pôles de compétitivité (pôle, nom et courriel du corresp.)	–
Site web du projet, le cas échéant	http://www.lattice.cnrs.fr/democrat/

Rédacteur de ce rapport	
Civilité, prénom, nom	Frédéric Landragin
Téléphone	01 58 07 66 20
Courriel	frederic.landragin@ens.fr ou (plus récent) frederic.landragin@ens.psl.eu
Date de rédaction	Février 2020
Période faisant l'objet du travail réalisé	Du 1er mars 2016 au 29 février 2020

1. Introduction

Ce livrable correspond à la « mise à disposition de l'outil de détection automatique de chaînes de coréférences » dont le développement et la mise au point constituent le principal objectif TAL (Traitement Automatique des Langues) du projet Democrat. De fait, ce sont deux systèmes et non un seul qui sont ici livrés et décrits :

1. Le premier est appelé COFR et correspond grosso modo à une adaptation pour la langue française – avec entraînement sur le corpus Democrat qui avait fait l'objet du livrable L1 – d'un système (extérieur à Democrat) conçu initialement pour l'anglais, le système de Kantor et Globerson : B. Kantor & A. Globerson (2019) « Coreference Resolution with Entity Equalization », In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, pp. 673–677, article disponible en ligne ici : https://www.aclweb.org/anthology/P19-1066/
2. Le second est appelé DeCOFR et correspond à l'application des recherches opérées dans Democrat sur une nouvelle architecture de réseau de neurones artificiels. Ce travail ayant démarré dès le début du projet, donc avant que le corpus Democrat ne soit livré, tous les entraînements ont été faits sur un corpus alternatif, extérieur à Democrat, le corpus ANCOR : J. Muzerelle, A. Lefeuvre, E. Schang, J.-Y. Antoine, A. Pelletier, D. Maurel, I. Eshkol et J. Villaneau (2013) « Ancor-Centre corpus ». Distribué sur Ortolang : https://www.ortolang.fr/market/corpora/ortolang-000903

S'y ajoutent d'autres objectifs TAL, qui ne faisaient pas partie de la liste initiale des objectifs, mais dont l'exploration a été nécessaire pour aboutir à des systèmes finalisés de qualité. Ces autres objectifs relèvent de recherches fondamentales sur l'architecture des réseaux de neurones artificiels et sur les spécificités de ces architectures pour le traitement d'objets aussi complexes que des chaînes de coréférences. Ces recherches fondamentales ont été publiées et permettent de contribuer aux avancées – de la communauté mondiale – sur l'apprentissage profond pour le TAL. Les publications concernées sont d'ailleurs les plus citées de l'ensemble des publications du projet Democrat (voir le rapport final du projet). Elles sont toutes disponibles en accès libre, et la section suivante indique les liens pour y accéder.

2. Accès aux publications

Toutes ont été déposées sur HAL, conformément aux objectifs du projet Democrat. La liste suivante est chronologique, en commençant par la publication la plus récente (qui a été acceptée mais n'est pas encore parue). Les publications numéros 1 et 2 reflètent directement les systèmes faisant l'objet de ce livrable (publication 1 pour le système COFR, publication 2 pour le système DeCOFR). Les autres publications en sont parfois des prémices, et relèvent parfois de recherches plus fondamentales.

1. Wilkens, R., Oberle, B., Landragin, F. & Todirascu, A. (2020) « French coreference for spoken and written language ». In : *Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, https://hal.archives-ouvertes.fr/hal-02476902
   C'est la publication à citer pour toute exploitation du système COFR. Il est à noter que ce système a été développé dans le cadre d'une collaboration entre l'ANR Democrat et l'ANR Alector (Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques), ANR-16-CE28-0005. La nature de la collaboration est la suivante : COFR, le système de Democrat, a bénéficié de travaux communs menés avec deux membres d'Alector, Rodrigo Wilkens et Amalia Todirascu (qui fait également partie du projet Democrat), en particulier pour l'état de l'art et une partie des expérimentations d'apprentissage.

2. Grobol, L. (2019) « Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French ». In : *Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19 - NAACL)*, Jun 2019, Minneapolis, United States, https://hal.inria.fr/hal-02151569v2
   C'est la publication à citer pour toute exploitation du système DeCOFR. Il est à noter que ce système a été développé dans le cadre d'une collaboration entre l'ANR Democrat et le Labex EFL, « Empirical Foundations of Linguistics », ANR-10-LABX-0083. La nature de la collaboration est la suivante : DeCOFR est le système de Democrat, et les expérimentations d'apprentissage ont de fait été effectuées sur du matériel informatique acheté sur crédits Democrat, mais son élaboration relève pour l'essentiel des travaux de thèse de Loïc Grobol, dont le contrat doctoral est financé par le Labex EFL.

3. Dinarelli, D. & Grobol, L. (2019) « Seq2Biseq: Bidirectional Output-wise Recurrent Neural Networks for Sequence Modelling ». *In : 20th International Conference on Computational*

Linguistics and Intelligent Text Processing (CICLing 2019), La Rochelle, France, https://hal.inria.fr/hal-02085093

4. Dinarelli, D. & Grobol, L. (2019) « Modèles neuronaux hybrides pour la modélisation de séquences : le meilleur de trois mondes ». In : Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2019), Toulouse, France, https://hal.archives-ouvertes.fr/hal-02157160v2

5. Oberle, B. (2019) « Détection automatique de chaînes de coréférence pour le français écrit : règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques ». In : *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL 2019*), Toulouse, France, https://halshs.archives-ouvertes.fr/halshs-01793477

6. Landragin, F. & Oberle, B. (2018) « Identification automatique de chaînes de coréférences : vers une analyse des erreurs pour mieux cibler l'apprentissage », In : *Journée commune AFIA-ATALA sur le Traitement Automatique des Langues et l'Intelligence Artificielle, Onzième édition de la plate-forme Intelligence Artificielle (PFIA 2018)*, Nancy, https://hal.archives-ouvertes.fr/hal-01819602

7. Dinarelli, M. & Grobol, L. (2018) « Modélisation d'un contexte global d'étiquettes pour l'étiquetage de séquences dans les réseaux neuronaux récurrents », In : *Journée commune AFIA-ATALA sur le Traitement Automatique des Langues et l'Intelligence Artificielle, Onzième édition de la plate-forme Intelligence Artificielle (PFIA 2018)*, Nancy, https://hal.archives-ouvertes.fr/hal-02002111

8. Dinarelli, M. & Dupont, Y. (2017) « Modélisation de dépendances entre étiquettes dans les réseaux neuronaux », *Traitement Automatique des Langues*, 58(1), pp. 13-37, https://hal.archives-ouvertes.fr/hal-01579114

9. Dinarelli, M., Vukotic, V. & Raymond, C. (2017) « Label-dependency coding in Simple Recurrent Networks for Spoken Language Understanding », In: *Proceedings of The 18th Annual Conference of the International Speech Communication Association (Interspeech 2017)*, Stockholm, Sweden, https://hal.archives-ouvertes.fr/hal-01553830v1

10. Dupont, Y., Dinarelli, M. & Tellier, I. (2017) « Label-Dependencies Aware Recurrent Neural Networks », In: *International Conference on Intelligent Text Processing and Computational Linguistics (CICling 2017)*, Budapest, Hungary, https://hal.archives-ouvertes.fr/hal-01579071
Note : cet article a gagné le premier prix « Best verifiability, reproducibility and working description award » de la conférence.

11. Grobol, L., Tellier, I., de la Clergerie, É., Dinarelli, M. & Landragin, F. (2017) « Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral », In: *Vingt-quatrième Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans, pp. 200-208, https://hal.inria.fr/hal-01558711

12. Dupont, Y., Dinarelli, M. & Tellier, I. (2017) « Réseaux neuronaux profonds pour l'étiquetage de séquences », In: *Vingt-quatrième Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans, pp. 19-27, https://hal.archives-ouvertes.fr/hal-01579192

13. Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y. & Dinarelli, M. (2016) « Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR », In: *Seventeenth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, Konya, Turquie, https://hal.archives-ouvertes.fr/hal-01344977v1

3. Accès aux outils

Conformément aux objectifs du projet Democrat, les outils développés dans le cadre du projet sont diffusés gratuitement sous licence ouverte. L'ensemble des fichiers, notamment les codes sources, ont ainsi été déposés sur des sites de type Github.

Lien vers l'outil COFR : https://github.com/boberle/cofr

Lien vers l'outil DeCOFR : https://github.com/LoicGrobol

4. Descriptif succinct de l'outil COFR : « COreference resolution tool for FRench »

COFR est un système bout-en-bout – capable de traiter du texte brut tout-venant – qui n'utilise aucune autre ressource que des plongements de mots. Il s'agit d'une adaptation du système à base de réseaux de neurones de Kantor et Globerson (« Coreference Resolution with Entity Equalization », In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, pp. 673–677, 2019). Puisque le corpus Democrat dispose de singletons annotés, contrairement au corpus de référence en langue anglaise (CoNLL-2012), COFR a été adapté pour détecter l'ensemble des mentions, qu'elles soient singletons ou coréférentes. Cela nous a conduit à diviser le système original en deux modules spécialisés, chacun avec un modèle entraîné séparément : un détecteur de mentions et un résolveur de coréférences. Le score CoNLL – métrique standard d'évaluation des systèmes de résolution de la coréférence – obtenu par COFR est de 75.00 %.

Le système peut être téléchargé à partir de https://github.com/boberle/cofr. Il nécessite Python 3 et TensorFlow 1. Les instructions pour télécharger le corpus et les modèles pré-entraînés sont indiquées sur le site, ainsi que les commandes à exécuter pour reproduire les résultats que nous avons publiés, prédire la coréférence pour des textes nouveaux, et entraîner d'autres modèles avec d'autres corpus ou d'autres paramètres.

Le système accepte en entrée un format spécifique de type "json" défini par le système original anglais. Nous proposons des scripts de conversion, afin de pouvoir utiliser le système à partir de textes tout-venants et de textes au format CoNLL. La sortie peut également être convertie au format CoNLL, qui est le format standard pour l'évaluation des systèmes automatiques de détection de coréférences.

5. Descriptif succinct de l'outil DeCOFR

DeCOFR est une adaptation du système de Lee *et al*. 2018 (Kenton Lee, Luheng He, and Luke Zettlemoyer, « Higher-Order Coreference Resolution with Coarse-to-Fine Inference », In : Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, ACL, New Orleans, Louisiana, Vol. 2, pages 687–692, 2018) pour le rendre plus adapté à d'autres paradigmes. La première raison de l'adaptation réalisée est que le système de Lee *et al*. opère systématiquement au niveau d'un document entier, ce qui paraît raisonnable au vu de la nature discursive des chaînes de coréférences, mais pose un problème de taille de mémoire : le document entier doit être gardé en mémoire, ce qui entraîne des besoins de calculs potentiellement très élevés. Lee *et al*. proposent de compenser ce problème en effectuant à chaque étape une série d'élagages un peu brutaux, mais le revers de la médaille est que cela complexifie la mise en œuvre et rend le processus d'apprentissage moins efficace. Au final, l'apprentissage est toujours très gourmand en mémoire et en calculs. La deuxième raison de l'adaptation réalisée pour DeCOFR est le fait que le système de Lee *et al*. ne fait pas de distinction entre des expressions référentielles et des expressions non référentielles, c'est-à-dire ne détecte que les mentions susceptibles d'appartenir à des chaînes de coréférences, et pas les mentions qui restent isolées – les singletons. Ce n'est pas un problème pour son entraînement avec le corpus CoNLL-2012 (cadre dans lequel le système de Lee *et al*. a été développé), mais c'en est un quand on considère un corpus comprenant des singletons. Ce qui est le cas du corpus ANCOR – et aussi du corpus Democrat. Pour pallier ces problèmes, DeCOFR cible en tenant compte du contexte immédiat plutôt que du document entier, et opère une détection des mentions, en tant que telle, avec prise en compte des singletons, en préalable à la détection des coréférences.

Comme pour le système COFR, DeCOFR accepte en entrée un format spécifique de type "json" défini par le système original anglais.

6. Bilan et exemples d'exécution

A titre de bilan, soulignons que les recherches entreprises dans le projet Democrat sont initiatrices et ouvrent la voie à la détection automatique des chaînes de coréférences pour la langue française. Avant Democrat, il existait – pour le traitement du français – principalement des systèmes à bases de règles, par définition très peu voués à évoluer (car cela nécessite de reprendre tout ou bonne partie du système), le plus récent – et probablement le meilleur – étant le système ODACR (Outil de Détection Automatique des Chaînes de Référence, https://halshs.archives-ouvertes.fr/hal-01837101/), développé par Bruno Oberlé au début de Democrat. Parmi les autres systèmes, à savoir les systèmes fondés sur de l'apprentissage artificiel, citons CROC (voir publication n° 13 dans la liste ci-dessus), dont l'exécution

nécessite de partir d'un texte déjà annoté en mentions. La tâche réalisée par CROC se limite ainsi à l'une des deux principales étapes de traitement, celle consistant à apparier les mentions coréférentes, et il ne s'agit donc pas d'un système bout-en-bout.

Le projet Democrat livre et rend publics deux systèmes bout-en-bout fondés sur les techniques les plus récentes d'apprentissage artificiel, à savoir les réseaux de neurones artificiels. Il permet ainsi à la communauté internationale de disposer d'équivalents des systèmes récemment développés pour la langue anglaise, espagnole ou polonaise. Avec la livraison effectuée en juin 2019 du corpus Democrat, chaque chercheur peut ainsi tester la détection automatique de coréférences sur le français et développer son propre système. De plus, le projet Democrat fournit des avancées significatives sur les architectures de réseaux de neurones artificiels adaptées à la détection des chaînes de coréférences, pour le français comme pour d'autres langues, et même pour des tâches plus générales telles que l'étiquetage de séquences (voir les publications n° 3, 4, 7, 8, 9, 10 et 12 dans la liste ci-dessus).

Enfin, à titre d'illustration des apports d'un système bout-en-bout, nous présentons ci-dessous deux exemples d'exécution avec COFR : l'un issu du corpus Democrat, et l'autre issu d'un texte littéraire qui ne fait pas partie du corpus. Ces exemples sont donnés ici à titre d'illustration. Les systèmes COFR et DeCOFR ont été finalisés en février 2020, soit (comme cela était prévu) à la toute fin du projet. Des analyses des erreurs et des comparaisons avec d'autres systèmes ont été abordées et sont actuellement à l'étude, mais le projet Democrat ne proposera pas de bilan de ce chantier en cours : c'est un travail de recherche en soi, qui inclut une phase d'analyse linguistique des erreurs commises par les systèmes, et donc une phase de mise en œuvre d'une typologie d'erreurs (ainsi que de nombreuses expérimentations).

Une comparaison des systèmes COFR et DeCOFR, ainsi que des analyses des spécificités de chacun des systèmes en fonction de ses points forts et de ses points faibles, dépassent largement les objectifs initiaux du projet Democrat, et en constituent donc des perspectives.

Exemple 1 : il s'agit du début de la page wikipédia « Singe », qui constitue l'un des textes de Democrat. Le texte annoté qui suit est obtenu en sortie de COFR, l'entrée étant le texte brut, sans aucune annotation. Le résultat comporte 20 chaînes de coréférences, chaque référent concerné étant indiqué par un indice (de 1 à 20) et un code couleur. Les mentions à ces référents sont mises entre crochets et en caractères gras. Les singletons – très nombreux – sont mis entre crochets mais ne comportent aucun indice.

> **[Les singes]$_1$** sont [des mammifères de [l' ordre de **[les primates]$_2$**]] , généralement arboricoles , à [la face souvent glabre] et caractérisés par [un encéphale développé] et de longs membres terminés par [des doigts] . Bien que [**[leur]$_1$** ressemblance avec **[l' Homme]$_3$**] ait toujours frappé [les esprits] , [la science] a mis de nombreux siècles à prouver **[le lien étroit]$_4$** **[qui]$_4$** existe entre [**[ces animaux]$_1$** et **[l' espèce humaine]$_5$**] . Au sein **[des primates]$_2$** , **[les singes]$_1$** forment **[un infra-ordre monophylétique]$_6$** , si l' [on] **[y]$_6$** inclut **[le genre Homo]$_7$** , nommé **[Simiiformes]$_6$** et **[qui]$_7$** se divise entre **[les singes de [le « Nouveau Monde]$_8$]$_1$** »

([Amérique centrale et méridionale]) et [ceux de **[l' « Ancien Monde]$_{9}$**] » ([Afrique] et [Asie tropicales]) . **[Ces derniers]$_{1}$** comprennent **[les hominoïdes]$_{10}$** , également appelés « [grands singes] » , **[dont]$_{10}$** fait partie [**[Homo sapiens]$_{11}$** et [**[ses]$_{11}$** ancêtres les plus proches]] . Même s' il ne fait plus de doute aujourd'hui que « **[l' Homme]$_{3}$** est [un singe] comme [les autres] » , **[le terme]$_{12}$** est majoritairement utilisé pour parler [des animaux sauvages] et **[évoque]$_{12}$** **[[un référentiel culturel] , littéraire et artistique]$_{13}$** **[qui]$_{13}$** exclut **[l' espèce humaine]$_{5}$** . [Dénominations] [Étymologie] **[Le terme]$_{12}$** viendrait de [le latin impérial simius] , plutôt que de [le latin classique simia] . [Les adjectifs] se rapportant à **[le singe]$_{3}$** sont [simien] et [simiesque] . [Noms vernaculaires] **[Les « singes de [le Nouveau Monde]$_{8}$]$_{1}$** » et **[les « singes de [l' Ancien Monde]$_{9}$]$_{1}$** » sont regroupés par [la classification phylogénétique] dans [l' infra-ordre de **[les Simiiformes]$_{6}$**] . **[Le terme de « [grand singe]$_{3}$]$_{12}$** » désigne [toutes les espèces] faisant partie de **[les hominidés]$_{10}$** , c'est-à-dire [les espèces actuelles de [gorilles] , [chimpanzés communs] ou [bonobos] , [orangs-outans] et [hommes]] , ainsi que [les espèces intermédiaires aujourd'hui éteintes] . En **[français]$_{14}$** , [les différentes sortes de singes] sont désignées par [[des noms plus ou moins précis] comme [babouin] , [chimpanzé] , [gibbon] , [gorille] , [macaque] , [orang-outan] , [ouistiti]] , etc. Contrairement à [les oiseaux] , il n' existe pas , en **[français]$_{14}$** , d' **[organisme reconnu]$_{15}$** **[qui]$_{15}$** propose des noms uniques pour [les espèces de **[singe]$_{3}$**] . De ce fait , [de nombreux singes] , particulièrement en [Amérique de le Sud] , possèdent **[plusieurs noms communs]$_{16}$** , à [le sens « [nom de vulgarisation scientifique]] » en **[français]$_{14}$** . **[Les noms]$_{16}$** peuvent être calqués sur [les noms scientifiques] comme [les Lagotriche] ou sur [les noms vernaculaires locaux] comme [Sapajou] . En outre , de le fait de [la ressemblance morphologique entre [espèces]] , **[beaucoup de noms vernaculaires]$_{17}$** désignent de fait [plusieurs espèces] , [la progression de [les connaissances]] ayant permis ultérieurement de faire la différence entre **[elles]$_{17}$** . De plus , [l' usage de **[les noms vernaculaires]$_{16}$**] a varié à le cours de [le temps] . Ainsi **[le terme chimpanzé]$_{18}$** , quand **[il]$_{18}$** a été adopté en **[français]$_{14}$** , désignait indistinctement **[deux espèces]$_{19}$** , **[qui]$_{19}$** , après qu' **[elles]$_{19}$** furent différenciées , ont été nommées dans un premier temps « [**[chimpanzé commun]$_{20}$** » et « [chimpanzé nain]] » , puis « **[chimpanzee commun]$_{20}$** » et « [bonobo] ».

Note : cette mise en forme des sorties a été obtenue automatiquement à partir du fichier de sortie de COFR, qui est en format CoNLL, c'est-à-dire en format tabulaire comme le montre la toute première phrase de l'extrait :

```
1   Les             (1
2   singes          1)
3   sont            -
4   des             (2
5   mammifères      -
6   de              -
7   l'              (3
8   ordre           -
9   de              -
10  les             (4
11  primates        2)3)4)
12  ,               -
13  généralement    -
14  arboricoles     -
15  ,               -
```

```
16  à             -
17  la            (5
18  face          -
19  souvent       -
20  glabre        5)
21  et            -
22  caractérisés  -
23  par           -
24  un            (6
25  encéphale     -
26  développé     6)
27  et            -
28  de            -
29  longs         -
30  membres       -
31  terminés      -
32  par           -
33  des           (7
34  doigts        7)
35  .             –
```

Exemple 2 : il s'agit du début du deuxième chapitre de « La Chartreuse de Parme » de Stendhal. Ce texte ne fait pas partie du corpus Democrat, autrement dit le système ne le connaît pas du tout. Les 20 chaînes de coréférences identifiées, ainsi que les singletons délimités, donnent une idée de ses performances.

> **[Le marquis]**$_1$ professait [une haine vigoureuse] pour [les lumières] ; ce sont **[les idées]**$_2$, disait **[-il]**$_1$, **[qui]**$_2$ ont perdu [l' Italie] ; **[il]**$_1$ ne savait trop comment concilier [cette sainte horreur de [l' instruction]] , avec le désir de voir **[[son]**$_1$ **fils Fabrice]**$_3$ perfectionner [l' éducation si brillamment commencée chez [les jésuites]] . Pour courir [le moins de risques possible] , **[il]**$_1$ chargea **[le bon abbé Blanès]**$_4$, curé de [Grianta] , de faire continuer **[Fabrice]**$_3$ **[ses]**$_3$ études en **[latin]**$_5$. Il eût fallu que **[le curé lui-même]**$_4$ sût **[cette langue]**$_6$; or **[elle]**$_6$ était [l' objet de [**[ses]**$_3$ mépris]] ; [**[ses]**$_3$ connaissances en [ce genre]] se bornaient à réciter , par cœur , [les prières de **[[son]**$_3$ **missel]**$_7$] , **[dont]**$_7$ **[il]**$_3$ pouvait rendre à peu près [le sens] à [**[ses]**$_3$ ouailles] . Mais **[ce curé]**$_4$ n' en était pas moins fort respecté et même redouté dans [le canton] ; **[il]**$_4$ avait toujours dit que ce n' était point en [treize semaines] ni même en [treize mois] , que l' on verrait s' accomplir [la célèbre prophétie de [saint Giovita]] , le patron de [Brescia] . **[Il]**$_4$ ajoutait , quand **[il]**$_4$ parlait à [des amis sûrs] , que [ce nombre treize] devait être interprété d' **[une façon]**$_8$ **[qui]**$_8$ étonnerait bien de **[le monde]**$_9$, s' il était permis de tout dire ([1813]) . [Le fait] est que **[l' abbé Blanès]**$_4$, personnage d' [[une honnêteté] et d' [une vertu primitives]] , et de plus homme d' esprit , passait [toutes les nuits] à [le haut de **[[son]**$_4$ **clocher]**$_{10}$] ; **[il]**$_4$ était fou d' [astrologie] . Après avoir usé [**[ses]**$_4$ journées] à calculer **[[des conjonctions] et [des positions d' étoiles]]**$_{11}$, **[il]**$_4$ employait [la meilleure part de [**[ses]**$_4$ nuits]] à **[les]**$_{11}$ suivre dans [le ciel] . Par suite de [**[sa]**$_4$ pauvreté] , **[il]**$_4$ n' avait d' [autre instrument] qu' [une longue lunette à [tuyau de carton]] . **[On]**$_{12}$ peut juger de **[le mépris]**$_{13}$ **[qu']**$_{13}$ avait pour [l' étude de [les langues]] **[un homme]**$_{14}$ **[qui]**$_{14}$ passait [**[sa]**$_{14}$ vie] à découvrir [l' époque précise de **[la chute de [les empires] et de [les révolutions]]**$_{15}$] **[qui]**$_{15}$ changent [la face de **[le monde]**$_9$] . Que sais **[-je]**$_{16}$ de plus sur **[un cheval]**$_{17}$, disait **[-il]**$_{16}$ à **[Fabrice]**$_3$, depuis qu' **[on]**$_{12}$ **[m']**$_{16}$ a appris qu' en **[latin]**$_5$ **[il]**$_{17}$ s' appelle [equus] ? **[Les**

paysans]$_{18}$ redoutaient [**l' abbé Blanès**]$_{4}$ comme [un grand magicien] : pour [**lui**]$_{4}$, à l' aide de [**la peur**]$_{19}$ [**qu'**]$_{19}$ inspiraient [[**ses**]$_{4}$ stations] dans [**le clocher**]$_{10}$, [**il**]$_{4}$ [**les**]$_{18}$ empêchait de voler . [[**Ses**]$_{4}$ confrères les curés de [les environs]] , fort jaloux de [[**son**]$_{4}$ influence] , [**le**]$_{4}$ détestaient ; [**le marquis del Dongo**]$_{20}$ [**le**]$_{4}$ méprisait tout simplement , parce qu' [**il**]$_{20}$ raisonnait trop pour [un homme de si bas étage] . [**Fabrice**]$_{3}$ [**l'**]$_{4}$ adorait ; pour [**lui**]$_{4}$ plaire [**il**]$_{3}$ passait quelquefois [des soirées entières] à faire [des additions] ou [des multiplications énormes] . Puis [**il**]$_{3}$ montait à [**le clocher**]$_{10}$: c' était [une grande faveur] et que [**l' abbé Blanès**]$_{4}$ n' avait jamais accordée à personne ; mais [**il**]$_{4}$ aimait [**cet enfant**]$_{4}$ pour [[**sa**]$_{4}$ naïveté] . Si [**tu**]$_{4}$ ne deviens pas hypocrite , [**lui**]$_{4}$ disait [**-il**]$_{4}$, peut-être [**tu**]$_{4}$ seras [un homme] .