Rapport final du projet ANR Democrat, "Description et modélisation des chaînes de référence : outils pour l'annotation de corpus et le traitement automatique"

Frédéric Landragin

AGENCE NATIONALE DE LA RECHERCHE ANR	**Compte-rendu de fin de projet**	

Projet ANR-15-CE38-0008

DEMOCRAT

Programme DS0805 2015

A IDENTIFICATION

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, CNRS, ENS Paris, PSL Research University, Université de Paris 3)
Période du projet (date de début – date de fin)	1er mars 2016 (T_0 *scientifique*) – 29 février 2020 (T_{final} *scientifique*)
Site web du projet, le cas échéant	http://www.lattice.cnrs.fr/democrat/

Rédacteur de ce rapport	
Civilité, prénom, nom	Mr Frédéric Landragin
Téléphone	01 58 07 66 20
Adresse électronique	frederic.landragin@ens.psl.eu
Date de rédaction	Février 2020

Si différent du rédacteur, indiquer un contact pour le projet	
Civilité, prénom, nom	-
Téléphone	-
Adresse électronique	-

Liste des partenaires présents à la fin du projet (société/organisme et responsable scientifique)	1. Laboratoire Lattice (Langues, Textes, Traitements Informatiques, Cognition), responsable Frédéric Landragin 2. Laboratoire LiLPa (Linguistique, Langues et Parole), responsable Catherine Schnedecker 3. Laboratoires ICAR (Interactions, Corpus, Apprentissages, Représentations) et IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités), relevant tous les deux de l'ENS de Lyon, responsable Céline Guillot-Barbance

B RESUME CONSOLIDE PUBLIC

Ce résumé est destiné à être diffusé auprès d'un large public pour promouvoir les résultats du projet, il ne fera donc pas mention de résultats confidentiels et utilisera un vocabulaire adapté mais n'excluant pas les termes techniques. Il en sera fourni une version française et une version en anglais. Il est nécessaire de respecter les instructions ci-dessous.

B.1 INSTRUCTIONS POUR LES RESUMES CONSOLIDES PUBLICS

Les résumés publics en français et en anglais doivent être structurés de la façon suivante.

Titre d'accroche du projet *(environ 80 caractères espaces compris)*
Titre d'accroche, si possible percutant et concis, qui résume et explicite votre projet selon une logique grand public : il n'est pas nécessaire de présenter exhaustivement le projet mais il faut plutôt s'appuyer sur son aspect le plus marquant.

Les deux premiers paragraphes sont précédés d'un titre spécifique au projet rédigé par vos soins.

Titre 1 : situe l'objectif général du projet et sa problématique *(150 caractères max espaces compris)*
Paragraphe 1 : *(environ 1200 caractères espaces compris)*
Le paragraphe 1 précise les enjeux et objectifs du projet : indiquez le contexte, l'objectif général, les problèmes traités, les solutions recherchées, les perspectives et les retombées au niveau technique ou/et sociétal

Titre 2 : précise les méthodes ou technologies utilisées *(150 caractères max espaces compris)*
Paragraphe 2 : *(environ 1200 caractères espaces compris)*
Le paragraphe 2 indique comment les résultats attendus sont obtenus grâce à certaines méthodes ou/et technologies. Les technologies utilisées ou/et les méthodes permettant de surmonter les verrous sont explicitées (il faut éviter le jargon scientifique, les acronymes ou les abréviations).

Résultats majeurs du projet *(environ 600 caractères espaces compris)*
Faits marquants diffusables en direction du grand public, expliciter les applications ou/et les usages rendus possibles, quelles sont les pistes de recherche ou/et de développement originales, éventuellement non prévues au départ.
Préciser aussi toute autre retombée= partenariats internationaux, nouveaux débouchés, nouveaux contrats, start-up, synergies de recherche, pôles de compétitivités, etc.

Production scientifique et brevets depuis le début du projet *(environ 500 caractères espaces compris)*
Ne pas mettre une simple liste mais faire quelques commentaires. Vous pouvez aussi indiquer les actions de normalisation

Illustration
Une illustration avec un schéma, graphique ou photo et une brève légende. L'illustration doit être clairement lisible à une taille d'environ 6cm de large et 5cm de hauteur. Prévoir une résolution suffisante pour l'impression. Envoyer seulement des illustrations dont vous détenez les droits.

Informations factuelles
Rédiger une phrase précisant le type de projet (recherche industrielle, recherche fondamentale, développement expérimental, exploratoire, innovation, etc.), le coordonnateur, les partenaires, la date de démarrage effectif, la durée du projet, l'aide ANR et le coût global du projet, par exemple « Le projet XXX est un projet de recherche fondamentale coordonné par xxx. Il associe aussi xxx, ainsi que des laboratoires xxx et xxx). Le projet a commencé en juin 2006 et a duré 36 mois. Il a bénéficié d'une aide ANR de xxx € pour un coût global de l'ordre de xxx € »

B.2 RESUME CONSOLIDE PUBLIC EN FRANÇAIS

Description, modélisation et détection automatique des chaînes de référence en français

1. Enjeux et objectifs pour l'étude des expressions référentielles et des chaînes de référence

Début 2016, lorsque le projet Democrat s'est mis en place, il n'existait pas : (i) de description intégrée permettant la modélisation des chaînes de référence, ni de prédictions sur leur typologie ou leurs comportements textuels ; (ii) de corpus permettant d'apprécier l'évolution historique de leur composition ; (iii) d'outil permettant de visualiser et d'explorer les chaînes de référence ; (iv) de système de traitement automatique des langues (TAL) capable de traiter du texte tout-venant, écrit en français, pour en extraire les expressions référentielles et les chaînes de référence. Democrat s'est donné pour ambition d'apporter de nouveaux résultats sur ces 4 aspects, qui constituent les 4 volets et les 4 livrables du projet. En apportant de nouvelles données et analyses sur la langue, il permet : (i) de nourrir l'ensemble des applications de TAL grâce à un corpus d'envergure adapté aux besoins de l'apprentissage

artificiel ; (ii) de renforcer la place du français dans le monde ; (iii) d'apporter de nouvelles connaissances à toutes les disciplines connexes à la linguistique, comme la psycho-linguistique et l'enseignement des langues.

2. De la linguistique théorique et descriptive à la linguistique de corpus outillée et au traitement automatique des langues (TAL)

Les membres du projet Democrat ont annoté manuellement un corpus écrit, mis à disposition librement en juin 2019 sur la plateforme Ortolang. Il s'agit d'un ensemble de 58 textes totalisant 689.000 mots et représentant diverses périodes, pour moitié littéraires et moitié non littéraires et non narratifs. Tous les siècles du 12e au 21e sont couverts, ce qui autorise des analyses diachroniques. Le nombre d'annotations – 198.000 expressions référentielles annotées – a été envisagé de manière à permettre des analyses statistiques et à autoriser des exploitations TAL. De nouvelles modalités d'analyse qualitative et quantitative ont été expérimentées, avec des mesures adaptées aux chaînes de référence et des visualisations dédiées – concordanciers, diagrammes de progression – à l'origine d'une évolution majeure de l'interface graphique et de la bibliothèque d'outils spécifiques de la plateforme TXM. Deux systèmes de TAL ont été développés pour détecter automatiquement les chaînes de référence dans du texte tout-venant, l'un entraîné directement sur le corpus Democrat, l'autre explorant de nouvelles architectures de réseaux neuronaux artificiels pour le faire plus efficacement.

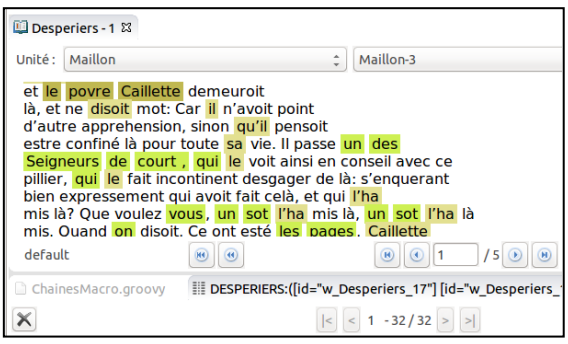
Résultats majeurs du projet

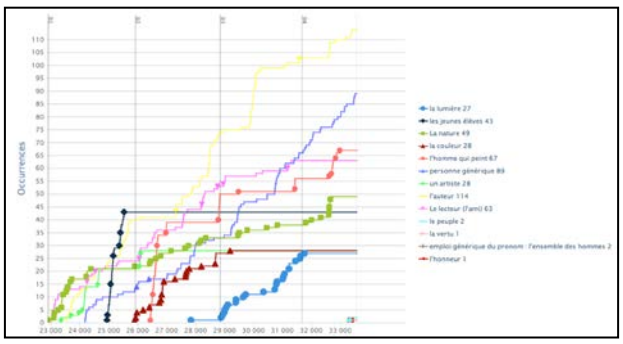
Début 2020, lors de la finalisation de Democrat, il existe désormais : (i) une description intégrée, discursive, diachronique et inter-genres des chaînes de référence ; (ii) un corpus de français écrit annoté en chaînes de référence ; (iii) plusieurs outils permettant de visualiser et d'explorer les chaînes de référence ; (iv) deux systèmes de traitement automatique des langues (TAL) capables de traiter du texte tout-venant, écrit en français, pour en extraire les expressions référentielles et les chaînes de référence – qui ont de plus apporté des innovations au domaine du *deep learning*.

Productions scientifiques et brevets depuis le début du projet

Le projet Democrat a produit environ 15 publications relevant d'études linguistiques des chaînes de référence. Les aspects méthodologiques regroupant la procédure d'annotation, les mesures et les techniques de visualisation expérimentées, ont fait l'objet de 15 publications supplémentaires. Les applications de traitement automatique se sont matérialisées dans 13 publications, d'une part sur le *deep learning*, d'autre part sur les systèmes livrés. Tout a été mis à disposition librement, logiciel compris.

Illustration : Interface d'annotation et diagramme de progression dans TXM





Le projet Democrat est un projet de recherche fondamentale coordonné par Frédéric Landragin. Il associe les laboratoires Lattice, LiLPa, ICAR et IHRIM. Le projet a commencé en mars 2006 et a duré 48 mois. Il a bénéficié d'une aide ANR de 390.000 €.

B.3 RESUME CONSOLIDE PUBLIC EN ANGLAIS

Suivre impérativement les instructions ci-dessus.

Description, modelling and automatic detection of reference chains in French

1. Issues and objectives for the study of referring expressions and reference chains

At the beginning of 2016, when the Democrat project was set up, there did not exist: (i) any integrated description allowing the modelling of coreference chains, nor any predictions about their typology and their textual behaviour; (ii) any corpus to apprehend the historical evolution of their composition; (iii) any tool for visualizing and exploring coreference chains; (iv) any natural language processing (NLP) software able to process raw texts written in French so as to extract the referring expressions and the coreference chains. Democrat's ambition is to bring new results on these 4 aspects, which constitute the 4 work-packages and the 4 deliverables of the project. Providing new data and analyses on the French language are intended to: (i) feed all NLP applications thanks to a large corpus adapted to the needs of machine learning; (ii) consolidate the role of the French language in the world; (iii) provide new knowledge to linguistics-related subjects such as psycholinguistics and language teaching.

2. From theoretical and descriptive linguistics to tooled corpus linguistics and natural language processing (NLP)

The members of the Democrat project have manually annotated a corpus, which is freely available since June 2019 on the Ortolang platform. It consists of a set of 58 texts totalling 689,000 words and representing various periods, half of them literary and half of non-narrative textual genres. All the centuries from the 12th to the 21st are covered, allowing diachronic analyses. The number of annotations – 198,000 annotated referring expressions – has been envisaged in order to allow statistical analyses and to authorize NLP exploitations. New methods of qualitative and quantitative analysis have been tested, with measurements adapted to the coreference chains and dedicated visualisations – concordancers, progression graphs – which have led to a major evolution of the user interface and the toolbox of the TXM platform. Two NLP systems have been developed to automatically detect coreference chains in raw text, one trained directly on the Democrat corpus, the other exploring new artificial neural network architectures to do so more efficiently.

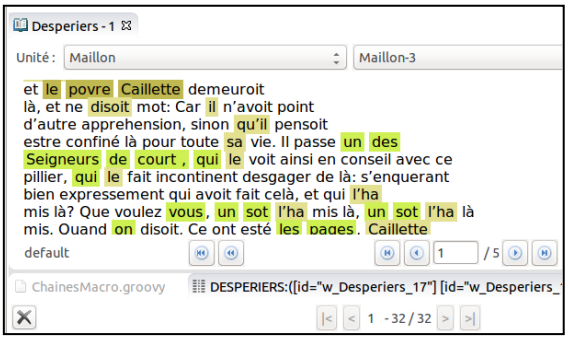
Main results of the project

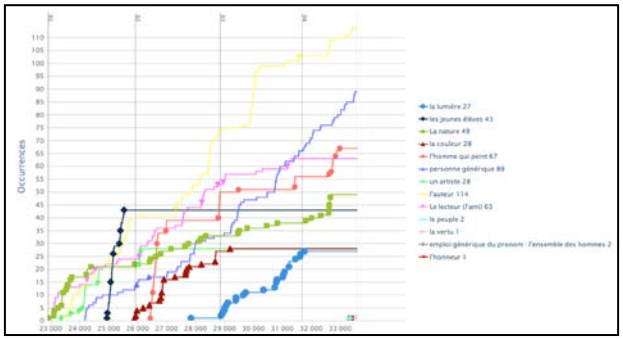
At the beginning of 2020, when Democrat was finalised, there is now: (i) an integrated, discursive, diachronic and inter-genre description of coreference chains; (ii) a corpus of written French texts with annotated coreference chains; (iii) several tools for visualizing and exploring the reference chains; (iv) two natural language processing systems that are able to process raw text written in French and to extract referring expressions as well as coreference chains – which have also brought innovations to the field of deep learning.

Scientific productions and patents since the beginning of the project

The Democrat Project has produced about 15 publications on linguistic studies of coreference chains. Methodological aspects, including the annotation procedure, measurements and visualisation techniques, were the object of 15 additional publications. Natural language processing applications materialised in 13 publications, some on deep learning, and others on Democrat systems. Everything has been made freely available, including the software.

Illustration: Annotation user interface and progression graph from TXM





Factual information

The Democrat project is a fundamental research project coordinated by Frédéric Landragin. It associates the Lattice, LiLPa, ICAR and IHRIM laboratories. The project began in March 2006 and lasted 48 months. It received an ANR aid of € 390,000.

C MEMOIRE SCIENTIFIQUE

***Maximum 5 pages**. On donne ci-dessous des indications sur le contenu possible du mémoire. Ce mémoire peut être accompagné de rapports annexes plus détaillés.*

Le mémoire scientifique couvre la totalité de la durée du projet. Il doit présenter une synthèse auto-suffisante rappelant les objectifs, le travail réalisé et les résultats obtenus mis en perspective avec les attentes initiales et l'état de l'art. C'est un document d'un format semblable à celui des articles scientifiques ou des monographies. Il doit refléter le caractère collectif de l'effort fait par les partenaires au cours du projet. Le coordinateur prépare ce rapport sur la base des contributions de tous les partenaires. Une version préliminaire en est soumise à l'ANR pour la revue de fin de projet.

Un mémoire scientifique signalé comme confidentiel ne sera pas diffusé. Justifier brièvement la raison de la confidentialité demandée. Les mémoires non confidentiels seront susceptibles d'être diffusés par l'ANR, notamment via les archives ouvertes http://hal.archives-ouvertes.fr.

Mémoire scientifique confidentiel : non

C.1 RESUME DU MEMOIRE

Le projet Democrat a développé des recherches sur la langue et sur la structuration textuelle du français *via* l'analyse détaillée et contrastive des chaînes de référence (instanciations successives d'une même entité extralinguistique : « Pierre », « le jeune homme », « il », « il ») dans un corpus diachronique de textes écrits entre le 12e et le 21e siècle, de genres textuels variés, narratifs et non narratifs. Le projet a mis à disposition de la communauté scientifique :

1. Un modèle intégré et discursif de la référence et de la composition des chaînes de référence en français (sous la forme de publications), fin d'un point de vue linguistique et suffisamment large pour explorer des voies comme la saillance ou le flou référentiel.

2. Un corpus annoté qui puisse servir de corpus de référence et de corpus d'apprentissage pour les campagnes d'évaluation internationales portant sur la coréférence. Ce corpus, annoté et vérifié manuellement, a une taille de 689.000 mots et comporte 198.000 expressions référentielles délimitées et annotées (selon les consignes d'un manuel d'annotation de 30 pages), ce qui représente 20.000 chaînes de coréférence(s) comportant 2 maillons ou plus – dont 9.000 chaînes de référence comportant 3 maillons ou plus.

3. Un outil d'annotation et de manipulation des données annotées, développé sur la plateforme TXM à partir des fonctionnalités d'annotation d'ANALEC et d'expérimentations d'annotation et d'analyse de maillons et de chaînes de référence à l'aide de plusieurs outils (SACR, par exemple). Le projet Democrat a porté l'extension URS (unité, relation, schéma) de TXM, permettant d'annoter des objets aussi complexes que des chaînes de référence.

4. Deux systèmes TAL de détection automatique des coréférences, capables de traiter du texte tout venant pour en extraire les expressions référentielles et les chaînes de référence. Plusieurs techniques ont été testées en parallèle, ce qui a permis de faire des avancées significatives dans le domaine du *deep learning*.

C.2 ENJEUX ET PROBLEMATIQUE, ETAT DE L'ART

Les constats et enjeux sur lesquels s'appuyait la soumission du projet Democrat étaient :

1. Au niveau linguistique, il existe de multiples études de la référence, des expressions référentielles (Charolles, 2002) et des chaînes de référence en français (Corblin, 1995 ; Schnedecker, 1997 ; Cornish, 1999), mais aucune qui ne tienne vraiment compte de l'ensemble de leurs spécificités discursives (études longitudinales, explorations des liens avec le genre textuel, études diachroniques). Ce constat découle probablement du manque de corpus annotés en chaînes de référence, le seul corpus – de grande taille – disponible pour la langue française étant le corpus ANCOR, qui propose des annotations d'anaphores sur de l'oral transcrit (https://www.ortolang.fr/market/corpora/ortolang-000903).

2. Au niveau de la méthodologie, c'est-à-dire de la linguistique de corpus outillée, il existe de multiples outils permettant la gestion, l'annotation et l'interrogation de corpus (Fort, 2012 ; Poudat & Landragin, 2017), mais aucun qui permette à l'utilisateur linguiste d'appréhender des chaînes de référence dans leur globalité, ni qui fournisse des outils adaptés (statistiques descriptives, graphiques) pour les interroger. Un premier pas a été fait avec l'outil GLOZZ (Widlöcher & Mathet, 2009), dont le projet Democrat s'est largement inspiré, notamment au niveau de la structuration des annotations – modèle URS (unité, relation, schéma), puis avec l'outil ANALEC (Landragin *et al.*, 2012). Mais un enjeu technique important était d'intégrer ces premiers pas dans une plateforme performante pour la gestion de corpus, et le choix de Democrat s'est porté sur la plateforme TXM (Heiden *et al.*, 2010).

3. Au niveau des applications de traitement automatique des langues (TAL), le français n'est pas représenté dans les grandes campagnes d'évaluation internationales (SemEval, CoNLL), qui se focalisent sur l'anglais (Lassalle, 2015 ; Clark & Manning, 2016 ; Lee *et al*., 2017…), l'espagnol (Recasens, 2010), ou le polonais (Ogrodniczuk *et al*., 2015). De fait, les seuls systèmes de détection automatique de chaînes de coréférence(s) pour le français sont des systèmes assez anciens, à base de règles – et non d'apprentissage artificiel – ce qui les rend très peu susceptibles d'évoluer. Alors qu'il existe pour l'anglais plusieurs systèmes bout-en-bout (c'est-à-dire partant de texte brut, sans aucune annotation) à base d'apprentissage – et même de *deep learning* –, il n'en existait aucun pour le français.

Le projet Democrat s'est donné pour missions de compenser l'ensemble de ces manques. Les analyses linguistiques réalisées tout au long du projet renouvellent l'état de l'art sur la coréférence, l'anaphore et les chaînes de référence ; le corpus Democrat vient combler le manque crucial de données annotées en français écrit ; l'extension URS de la plateforme TXM permet l'annotation et l'exploration conjointe des chaînes de référence sur des corpus écrits ; enfin, les systèmes COFR et DeCOFR sont les premiers à être capables de détecter automatiquement les coréférences dans des textes tout-venant en français contemporain.

La totalité des enjeux initiaux du projet a fait l'objet d'avancées significatives. Aucun élément scientifique n'a été laissé de côté pendant les quatre années du projet Democrat. Mieux : plusieurs domaines de recherche supplémentaires ont été explorés avec succès :

1. La conception d'interfaces graphiques innovantes permettant l'annotation rapide (outil SACR), la visualisation et l'exploration des chaînes de référence (concordanciers et diagrammes à barres adaptés aux chaînes, ainsi que diagrammes de progression de TXM).

2. L'exploration de l'approche du *deep learning* pour la résolution automatique des coréférences, approche qui n'était pas encore répandue lors du lancement du projet, et que les membres spécialistes de TAL de Democrat ont su approfondir et adapter à la résolution des coréférences en français. En plus de deux systèmes, le projet fournit ainsi des résultats significatifs sur les architectures de réseaux neuronaux artificiels.

C.3 APPROCHE SCIENTIFIQUE ET TECHNIQUE

Compte tenu du besoin de données attestées et annotées, les premiers efforts de Democrat ont porté sur la constitution et l'annotation manuelle d'un corpus de grande taille, avec une répartition à peu près équilibrée entre genres textuels (narratifs *versus* non narratifs) et états de langue (français médiéval *versus* français moderne et contemporain). La constitution s'est faite en identifiant des extraits de texte d'environ 10 000 mots, taille choisie selon différents critères : adéquation par rapport à la longueur attendue des chaînes, mais aussi par rapport aux extraits exploités dans d'autres projets, de manière à favoriser les compatibilités.

L'annotation a suivi l'approche désormais indispensable en linguistique de corpus outillée (Fort, 2012) : (a) décisions collectives concernant les phénomènes linguistiques à annoter et les consignes pour le faire ; (b) écriture collective d'un manuel d'annotation, dont la faisabilité a été vérifiée *via* plusieurs expérimentations chronométrées ; (c) test expérimental de l'exploitation des résultats d'un *chunker* en tant que pré-annotateur automatique, de manière à accélérer le processus (le choix d'exploiter ou non ce *chunker* était laissé à l'annotateur) ; (d) annotation proprement dite : soit purement manuelle, soit aidée par l'exécution d'outils spécifiques permettant de repérer les annotations manquantes (repérage des pronoms, par exemple) et d'identifier rapidement des erreurs (doublons, par exemple) ; (e) évaluation de la reproductibilité des annotations réalisées, en procédant à une

double annotation de 10% du corpus et un calcul de l'accord inter-annotateurs ; (f) homogénéisation des annotations avant (g) publication du corpus.

Les analyses linguistiques ont suivi plusieurs approches, notamment celle de la linguistique fondée sur corpus (Bilger, 2000 ; Biber & Conrad, 2009), et se sont appuyées sur un ensemble d'indicateurs quantitatifs communs à tous les participants du projet : nombre de chaînes dans une portion de texte donnée ; nombre moyen de maillons (éléments de chaînes) ; distance moyenne entre deux maillons, etc. Certains de ces indicateurs sont nouveaux et montrent la volonté de Democrat de faire un pas en avant dans le domaine des statistiques textuelles (Lebart & Salem, 1994 ; Poudat & Landragin, 2017), en explorant de nouvelles méthodes d'analyse des chaînes de référence. Au final, les analyses combinent diverses modalités d'exploitation : quantitative, statistique et qualitative.

Les aspects TAL du projet ont suivi deux voies parallèles – raison principale de la livraison de deux systèmes différents – du fait des contraintes temporelles du projet. L'approche technique suivie est celle de l'apprentissage, qui nécessite des données annotées propres. Or le corpus Democrat n'a été disponible que dix mois avant la fin du projet. Comme il n'était pas question de reléguer à la dernière année les recherches relevant du TAL, c'est dans un premier temps le corpus ANCOR (cf. plus haut) qui a été exploité et qui a conduit à des améliorations significatives du processus d'apprentissage. La conception du système DeCOFR – entraîné sur ANCOR – a ainsi permis une multitude d'avancées scientifiques et techniques, aussi bien au niveau du choix des traits utiles à l'apprentissage (syntaxe, par exemple) qu'à celui de la participation du projet aux efforts internationaux de normalisation (spécification du format de représentation des données XML-TEI-URS).

C.4 RESULTATS OBTENUS

Le corpus publié (voir indicateurs quantitatifs en C.1) a une taille comparable à celle du corpus ANCOR. Avec ces deux corpus, la communauté dispose désormais de deux ensembles de données sur la référence en français, l'un pour l'écrit, l'autre pour l'oral – tous les deux compatibles avec des exploitations TAL, ce qui n'était pas le cas des corpus précédemment constitués par la communauté. Le corpus Democrat est un résultat majeur.

Afin que les membres de Democrat échangent autour de l'exploitation de leurs données, se synchronisent sur la méthodologie et déterminent des voies possibles pour des études riches et variées (longitudinales d'un genre donné, incluant des comparaisons avec d'autres langues, etc.), le projet a mis en œuvre l'organisation de trois workshops, avec à chaque fois des présentations de membres du projet, des conférences invitées (de spécialistes extérieurs à Democrat) et une table ronde finale permettant de confronter les points de vue et d'expliciter un programme de recherche. Ces workshops montrent les voies ainsi explorées :

1. Lundi 27 novembre 2017 : « Référence, coréférence et structure textuelle ».
2. Mercredi 14 mars 2018 : « Approches contrastives des chaînes de référence ».
3. Vendredi 14 juin 2019 : « Mesures statistiques et approches quantitatives ».

Chacun de ces workshops a fait l'objet de soumissions d'articles. Notons que, juste avant le lancement du projet, c'était déjà le cas d'un workshop organisé entre les futurs partenaires sur l'étude des chaînes de référence en corpus. Ce tout premier workshop a eu comme résultat principal la publication d'un numéro thématique de la revue *Langue française* (voir publications). Les trois workshops cités ont tous fait l'objet de soumissions d'articles. La parution de certains d'entre eux est imminente (revue *Discours*), d'autres viendront rapidement. Le projet Democrat a déjà produit plus d'une cinquantaine de publications.

Autre résultat majeur, la plateforme TXM comporte désormais une extension permettant l'annotation de corpus. L'intérêt est double : non seulement les fonctionnalités d'annotation sont conçues pour manipuler des objets aussi complexes que des chaînes de référence et, de plus, ces fonctionnalités sont interfacées (*via* des outils spécifiques) avec les innombrables possibilités d'exploration et d'interrogation qu'offre TXM, *via* notamment l'intégration dans celui-ci d'un moteur de recherche (CQP) et d'un module d'analyse statistique (R).

Enfin, les systèmes de TAL développés dans Democrat fournissent de nouvelles possibilités de traitement automatique des textes (voir annexe F.4 pour la qualité des sorties de ces systèmes). Leur publication date de la toute fin du projet, donc aucune exploitation d'envergure n'en a encore été faite. En particulier, il n'a pas été possible de faire participer l'un ou l'autre de ces systèmes à une campagne d'évaluation internationale. Il ne fait aucun doute néanmoins que le français pourra désormais être représenté dans une telle campagne.

C.5 EXPLOITATION DES RESULTATS

Premier corpus de français annoté en expressions référentielles et chaînes de référence, le corpus Democrat est un corpus unique parmi les innombrables corpus annotés en coréférence(s), du fait de sa dimension diachronique (textes du 12e au 21e siècle), trans-générique (poésie, romans, mémoires, traités, pamphlet) et de son inscription dans des domaines variés (histoire, philosophie, droit, presse, savoirs encyclopédiques). En fonction des domaines d'intérêt des utilisateurs et de leurs objectifs de recherche, le corpus Democrat offre un énorme potentiel et se prête à des exploitations multiples et diversifiées :

– Dans le domaine des sciences du langage, il permet des travaux à la fois quantitatifs et qualitatifs, à caractère diachronique (étude des changements dans l'expression de la cohésion référentielle au fil du temps, des expressions référentielles considérées comme vieillies, étude de la linguistique textuelle diachronique), sémantico-référentiel (de nombreux travaux ou théories sur la coréférence et l'anaphore se sont élaborés « hors sol » sur la base de l'intuition des linguistes, le corpus Democrat permet de vérifier à grande échelle leur plausibilité et de leur apporter les amendements salutaires des approches empiriques), statistique, textométrique, relevant de la linguistique des genres (les chaînes de référence constituent une faisceau d'indices structuré à théoriser et à porter au crédit des approches des genres fondées sur des faisceaux d'indices et/ou les faits de langue) ou de la linguistique contrastive (sur la base des études et des thèses contrastives réalisées dans le projet, et dans la perspective des travaux anglophones existants, il est possible de multiplier les études, à échelle variable, et de tester les hypothèses élaborées par les typologues ou comparatistes).

– Dans le domaine littéraire, dans la mesure où les chaînes de référence sont un indice générique robuste, il devient possible de caractériser des genres à travers le temps, d'étudier l'évolution et le changement des genres ainsi que la notion de *genres comparables* (sachant que certains genres ont disparu au fil du temps quand d'autres sont apparus), de caractériser sur la base d'indices linguistiques structurés des séquences de genres (*incipit*, structuration, type d'entités instanciées), voire les évolutions de style (la désignation des personnages des romans du début du 19e siècle n'a rien à voir avec celle du début du 21e siècle).

– Dans le domaine philosophique, l'étude des genres philosophiques (traités, contes, fables) suivant les principes sus-mentionnés, l'élaboration de concepts philosophiques (entités abstraites) au fil du texte et de la réflexion, l'évolution des notions à travers le temps (par le bais de leur mise en texte), etc.

– Dans le domaine juridique, les textes juridiques annotés (conventions, code de procédure pénale) peuvent servir aux études de droit par divers questionnements : qu'est-ce qu'un texte univoque ? à quoi tient la « complexité » (prétendue ou avérée) des textes juridiques ? quelles sont leurs possibilités de simplification ?

– Dans le domaine de la didactique du français (FLM/FLE) : le corpus Democrat offre une diversité de modèles ou de routines d'écriture permettant d'aider à développer des habiletés langagières en production (modalités d'introduction des entités référentielles suivant leur type ontologique, leur genre d'occurrence, modalités de reprise et leurs effets, micro-synthèses opérées par les anaphores dites résomptives) et en compréhension des textes (repérage des points de blocage ou, au contraire, de facilitation référentiels) ainsi que de construire du matériel didactique de remédiation.

– Dans le domaine de la psychologie cognitive, si les études portant sur la résolution des anaphores pronominales sont pléthoriques, l'impact de la référence discursive n'a pas fait l'objet de travaux conséquents – le corpus Democrat permet de fournir aux psychologues des ressources à même de construire des modèles d'interprétation appuyés sur les procédures d'encodage perceptibles à travers les chaînes de référence.

– Le corpus Democrat peut aussi servir de « modèle » à la constitution de ressources annotées en coréférence(s) dans des domaines où les habiletés langagières normales ou déficientes ont besoin d'être documentées : le domaine médical (détection de pathologies diverses), le domaine du web (recherches et indexation documentaires, textes collaboratifs). Sur la base des études réalisées, il est possible de concevoir des outils logiciels afin d'apporter des aides diverses aux professionnels ou aux usagers.

C.6 Discussion

La totalité des objectifs initiaux du projet a été atteinte. Comme nous l'avons vu, des travaux non envisagés initialement ont même été entrepris, ainsi que des avancées significatives dans le domaine des recherches sur les architectures de réseaux neuronaux artificiels.

Les nombreuses discussions entre les membres du projet ont permis de déterminer un ensemble de perspectives de recherche. Reprenons ainsi la liste des workshops cités en C.4 :

1. Lundi 27 novembre 2017 : « Référence, coréférence et structure textuelle ». En plus des résultats obtenus, soulignons qu'un élargissement possible pourrait consister à annoter automatiquement un certain nombre de structures textuelles sur le corpus Democrat, puis à croiser ces annotations avec celles des chaînes de référence. Ce travail – qui sort largement des objectifs initiaux de Democrat – a commencé à être théorisé et envisagé techniquement. Il devrait se concrétiser par des analyses et des publications.

2. Mercredi 14 mars 2018 : « Approches contrastives des chaînes de référence ». Comme prévu, le corpus Democrat ne comporte que des textes écrits en français. Néanmoins, plus d'une dizaine de membres du projet ont procédé à des études de corpus multilingues, voire ont annoté des textes dans d'autres langues, dans un but de comparaison. Là aussi, cet élargissement pourrait conduire non seulement à des analyses contrastives supplémentaires, mais aussi à la constitution d'un corpus multilingue aligné, fondé sur le corpus Democrat.

3. Vendredi 14 juin 2019 : « Mesures statistiques et approches quantitatives ». Ce workshop a permis de clarifier une méthodologie quantitative et de faire un pas significatif vers des statistiques textuelles adaptées aux chaînes. Mieux : une perspective de recherche consiste à intégrer à la méthodologie de la textométrie une prise en compte efficace des annotations et, qui plus est, d'annotations discursives comme le sont les chaînes de référence.

Concernant l'impact industriel et sociétal des résultats de Democrat, notons que disposer d'un outil capable de détecter automatiquement des chaînes de référence est susceptible d'augmenter l'efficacité des moteurs d'indexation et de recherche, pour ne citer que deux des applications majeures du TAL dans l'industrie et la société. Les grandes sociétés du web travaillent depuis des années sur cette tâche de détection automatique des coréférences, et le projet Democrat n'apporte ni plus ni moins que la prise en compte de la langue française.

C.7 CONCLUSIONS

Le projet Democrat a réuni des chercheurs afin de proposer des avancées en linguistique, en linguistique de corpus outillée et en TAL. La pluridisciplinarité qui caractérise l'approche suivie est à souligner, car les chercheurs de Democrat ont su dialoguer et conjuguer les besoins du TAL avec les analyses linguistiques. Les membres du projet se réjouissent des résultats obtenus, et notamment de la publication du corpus Democrat, cosignée par pas moins de 48 personnes, signe de l'effort collectif réalisé.

C.8 REFERENCES

BIBER D. & CONRAD S. (2009), *Register, Genre, and Style*, Cambridge : Cambridge University Press.

BILGER M. (éd., 2000), *Corpus. Méthodologie et applications linguistiques*, Paris : Honoré Champion.

CHAROLLES M. (2002), *La référence et les expressions référentielles en français*, Paris : Ophrys.

CLARK, K. & MANNING, C. (2016), "Improving Coreference Resolution by Learning Entity-Level Distributed Representations", In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 643-653.

CORNISH F. (1999), *Anaphora, Discourse, and Understanding. Evidence from English and French*, New York: Oxford University Press.

FORT K. (2012), Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. Thèse de doctorat, Université Paris 13.

HABERT B. (2005), *Instruments et ressources électroniques pour le français*, Paris : Ophrys.

HEIDEN S., MAGUE J.-P. & PINCEMIN B. (2010), « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *Actes des 10e Journées Internationales d'Analyse statistique des Données Textuelles (JADT 2010)*, Rome, 1021-1032.

LANDRAGIN F., POIBEAU T. & VICTORRI B. (2012), "ANALEC: a New Tool for the Dynamic Annotation of Textual Data", *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, 357-362.

LASSALLE E. (2015), Structured Learning with Latent Trees: a joint approach to coreference resolution, Thèse de l'Université Paris Diderot.

LEBART L. & SALEM A. (1994), *Statistique textuelle*, Paris : Dunod.

LEE, K., HE, L., LEWIS, M. & ZETTLEMOYER, L. (2017), "End-to-end Neural Coreference Resolution", In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark, pp. 188-197.

NG V. (2007), "Shallow Semantics for Coreference Resolution", *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, 1689-1694.

OGRODNICZUK M., GŁOWIŃSKA K., KOPEĆ M., SAVARY A. & ZAWISŁAWSKA M. (2015), *Coreference in Polish: Annotation, Resolution and Evaluation*. Berlin: Walter De Gruyter.

POUDAT C. & LANDRAGIN F. (2017), *Explorer un corpus textuel. Méthodes, pratiques, outils*, Louvain-la-Neuve : De Boeck Supérieur.

RECASENS M. (2010), Coreference: Theory, Annotation, Resolution and Evaluation, PhD thesis, Barcelona: University of Barcelona.

SCHNEDECKER C. (1997), *Nom propre et chaînes de référence*, Paris : Klincksieck.

SCHNEDECKER C. (2005), Les chaînes de référence dans les portraits journalistiques : éléments de description, *Travaux de linguistique* 51, 2005/2, 85-133.

WIDLÖCHER A., MATHET Y. (2009), « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus », *Actes de la conférence TALN 2009*, Senlis.

D LISTE DES LIVRABLES

Quand le projet en comporte, reproduire ici le tableau des livrables fourni au début du projet. Mentionner l'ensemble des livrables, y compris les éventuels livrables abandonnés, et ceux non prévus dans la liste initiale.

Date de livraison	N°	Titre	Nature (rapport, logiciel, prototype, données, …)	Partenaires (souligner le responsable)	Commentaires
03/18	1	Manuel d'annotation du corpus et organisation de formations sur l'annotation (« L2 »)	rapport, formations	tous, Lattice (Frédéric Landragin)	Rendu public lors de la livraison du livrable ci-dessous
06/19	2	Mise à disposition du **corpus annoté** sur le site WEB du projet, sur une plateforme pérenne prévue dans ce but, et information *via* les listes de diffusion de la communauté (« L1 »)	données	tous, Lattice (Frédéric Landragin)	Données rendues publiques sur la plateforme Ortolang
06/19	3	Mise à disposition de l'**outil d'annotation manuelle**, information *via* les listes de diffusion de la communauté) ; actions de formation pour des utilisateurs potentiels (« L3.1 »)	logiciel, formations	tous, IHRIM (Serge Heiden et Céline Guillot-Barbance)	Outil rendu public sur le site web de la plateforme TXM
02/20	4	Mise à disposition d'**outils de détection automatique de chaînes de référence** (« L3.2 »)	logiciel	tous, Lattice (Frédéric Landragin)	Les deux logiciels livrés sont rendus publics sur les pages github de leurs auteurs
Tout au long du projet	5	Séries de publications et de communications dans des colloques dédiés, nationaux et internationaux (« L4 » et « L5 »)	publications, workshops, formations	tous, LILPA (Catherine Schnedecker)	Toutes les publications ont été déposées sur HAL avec le code ANR-15-CE38-0008

E IMPACT DU PROJET

Ce rapport rassemble des éléments nécessaires au bilan du projet et plus globalement permettant d'apprécier l'impact du programme à différents niveaux.

E.1 INDICATEURS D'IMPACT

Nombre de publications et de communications (à détailler en E.2)

Comptabiliser séparément les actions monopartenaires, impliquant un seul partenaire, et les actions multipartenaires résultant d'un travail en commun.

***Attention** : éviter une inflation artificielle des publications, mentionner uniquement celles qui résultent directement du projet (postérieures à son démarrage, et qui citent le soutien de l'ANR et la référence du projet).*

		Publications multipartenaires	Publications monopartenaires
International	**Revues à comité de lecture**	0	0
	Ouvrages ou chapitres d'ouvrage	1 : publication n° 1	0

	Communications (conférence)	7 : publications 2 à 8	9 : publications 9 à 19
France	**Revues à comité de lecture**	3 : publications 20 à 22	11 : publications 23 à 33
	Ouvrages ou chapitres d'ouvrage	0	0
	Communications (conférence)	1 : publication 34	19 : publications 35 à 53
Actions de diffusion	**Articles vulgarisation**	0	1 : publication 54
	Conférences vulgarisation	0	2 : publications 55 à 56
	Autres	0	7 : publications 57 à 63

Autres valorisations scientifiques (à détailler en E.3)

Ce tableau dénombre et liste les brevets nationaux et internationaux, licences, et autres éléments de propriété intellectuelle consécutifs au projet, du savoir faire, des retombées diverses en précisant les partenariats éventuels. Voir en particulier celles annoncées dans l'annexe technique).

	Nombre, années et commentaires (valorisations avérées ou probables)
Brevets internationaux obtenus	0
Brevet internationaux en cours d'obtention	0
Brevets nationaux obtenus	0
Brevet nationaux en cours d'obtention	0
Licences d'exploitation (obtention / cession)	0
Créations d'entreprises ou essaimage	0
Nouveaux projets collaboratifs	1 : 2019, pour l'AAP « Instituts Thématiques Interdisciplinaires », proposition intitulée « ILC, Institut du Langage et de la Communication », portée par Catherine Schnedecker et Amalia Todirascu, LiLPa – et Democrat – avec Nicolas Amadio (DynamE), Elisabeth Demont (LPC) et Philippe Viallon (Lisec)
Colloques scientifiques	3 : 2017, 2018 et 2019. Democrat a organisé 3 workshops scientifiques, regroupant des membres du projet et des invités extérieurs (voir annexe F.1) : 1. « Référence, coréférence et structure textuelle » (Lyon, 27 novembre 2017). 2. « Approches contrastives des chaînes de référence » (Paris, 14 mars 2018). 3. « Mesures statistiques et approches quantitatives » (Strasbourg, 14 juin 2019).
Autres (préciser)	Actions de formations aux fonctionnalités logicielles développées dans le projet Democrat (cf. E.3 pour le détail) : 1. formations à ANALEC. 2. formations à l'extension URS de TXM. Mise en place de collaborations : 1. Collaboration avec le consortium CORLI (CORpus, Langues et Interactions) du TGIR Huma-Num (https://www.huma-num.fr/), pour intégrer à son programme de formation des séances dédiées aux avancées Democrat de TXM. 2. Collaboration ponctuelle avec le projet ANR Alector (« Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques »), ANR-16-CE28-0005, voir l'annexe F.4. 3. Collaboration avec le Labex EFL (« Empirical Foundations of Linguistics »), ANR-10-LABX-0083, voir l'annexe F.4.

E.2 LISTE DES PUBLICATIONS ET COMMUNICATIONS

Répertorier les publications résultant des travaux effectués dans le cadre du projet. On suivra les catégories du premier tableau de la section E.1 en suivant les normes éditoriales habituelles. En ce qui concerne les conférences, on spécifiera les conférences invitées.

Notes préliminaires :

- Les communications réalisées lors des workshops organisés par le projet Democrat n'ont pas été incluses ici ; leur liste est visible *via* les programmes des workshops dans l'annexe F.1.
- Dans chaque catégorie, les publications multipartenaires apparaissent en premier (dans l'ordre chronologique inverse), puis sont suivies par les publications impliquant un seul partenaire (également dans l'ordre chronologique inverse).
- Toutes les publications ont fait l'objet d'un dépôt sur HAL. Une « collection » HAL a été mise en place : https://hal.archives-ouvertes.fr/DEMOCRAT/.

E.2.1 International

Ouvrages ou chapitres d'ouvrage

1. Baumer, E., Dias, D., Gardelle, L. & Prak-Derrington, E. (2020, sous presse) « Peut-on parler aujourd'hui de rédaction de presse globalisée ? Étude comparée d'un corpus trilingue (français / anglais / allemand) ». A paraître dans un ouvrage, au titre non encore communiqué, aux éditions Peter Lang, https://hal.archives-ouvertes.fr/hal-02465871

Communications (conférences)

2. Wilkens, R., Oberle, B., Landragin, F. & Todirascu, A. (2020, sous presse) « French coreference for spoken and written language ». In : *Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, https://hal.archives-ouvertes.fr/hal-02476902

3. Baumer, E., Dias, D. & Schnedecker C. (2019) « Les chaînes de référence dans un corpus contrastif (allemand-anglais-français) de romans réalistes du XIXe siècle : analyse quantitative et qualitative ». In : *Colloque Phraséologie et stylistique de la langue littéraire (PhraseoRom)*, Friedrich Alexander-Universität, Erlangen-Nürnberg, Germany, https://hal.archives-ouvertes.fr/hal-01999089

4. Grobol, L., Landragin, F. & Heiden, S. (2018) « XML-TEI-URS: Using a TEI Format for Annotated Linguistic Resources ». In: *CLARIN Annual Conference 2018*, Pise, Italie, https://hal.archives-ouvertes.fr/hal-01827563v1

5. Quignard, M., Heiden, S., Landragin, F. & Decorde, M. (2018) « Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes ». In: *Fourteenth International Conference on the Statistical Analysis of Textual Data (JADT 2018)*, Rome, Italie, https://hal.archives-ouvertes.fr/hal-01814858

6. Baumer, E. & Dias, D. (2018) « Chaînes de référence et genres discursifs : étude exploratoire d'un corpus de presse trilingue (français/anglais/allemand) ». In: *La globalisation communicationnelle 10 ans après : les défis de l'interculturel*, Université de Gdańsk.

7. Grobol, L., Landragin, F. & Heiden, S. (2017) « Interoperable annotation of (co)references in the Democrat project ». In: *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*, Montpellier, https://hal.archives-ouvertes.fr/hal-01583527

8. 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

9. Baumer, E., (2019) « Reference chains across languages : a cross-linguistic study ». In : *Premier Colloque International de linguistique contrastive*, Université des Langues et Cultures de Beijing, Chine, https://hal.archives-ouvertes.fr/hal-02486510

10. 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

11. Guillot-Barbance, C. (2019) « Les chaînes de référence en français : analyse d'un corpus diachronique de textes narratifs (début 12e – fin 15e s.) ». In : *Le français en diachronie (DIACHRO IX)*, Universidad de Salamanca, Salamanque, Espagne, https://halshs.archives-ouvertes.fr/halshs-02472741

12. 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

13. Grobol, L., Tellier, I., De La Clergerie, É., Dinarelli, M. & Landragin, F. (2018) « ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations ». In: *11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, https://hal.inria.fr/hal-01744572

14. Oberle, B. (2018) « SACR: A Drag-and-Drop Based Tool for Coreference Annotation ». In: *11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, https://halshs.archives-ouvertes.fr/halshs-01793477

15. Heiden, S. (2018) « Annotation-based Digital Text Corpora Analysis within the TXM Platform ». In: *Fourteenth International Conference on the Statistical Analysis of Textual*

Data (JADT 2018), Rome, Italie, pp. 367-374, https://halshs.archives-ouvertes.fr/hal-02015898

16. Delaborde, M. & Landragin, F. (2018) « Traitement "good-enough" du pronom "on" : vers une modélisation de la coréférence floue ». In: *Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2018)*, Paris, https://halshs.archives-ouvertes.fr/halshs-01795228

17. 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

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

19. Landragin, F. (2018) « Referring Expressions and Coreference Chains in French: Annotation Strategies, Annotating Tool, and Annotated Resources ». Conférence invitée à l'Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, le 12/11/2018.

E.2.2 France

Revues à comité de lecture

20. Oberle, B., Schnedecker, S., Baumer, E., Capin, D., Glikman, J., Guo, C., Revol, T., Todirascu, A. & Tushkova, J. (2018) « Les chaînes de référence dans les textes encyclopédiques du 12e au 21e siècle : étude longitudinale ». *Travaux de linguistique*, 2018/2, 77, pp. 67-141, https://hal.archives-ouvertes.fr/hal-02145155

21. Schnedecker, C., Glikman, J. & Landragin, F. (2017) « Les chaînes de référence : annotation, application et questions théoriques ». *Langue française*, 195, pp. 5-15, https://halshs.archives-ouvertes.fr/halshs-01580785

22. Obry, V., Glikman, J., Guillot-Barbance, C. & Pincemin, B. (2017) « Les chaînes de référence dans les récits brefs en français : étude diachronique (XIIIe – XVIe siècles) ». *Langue française*, 195, pp. 91-110, https://hal.archives-ouvertes.fr/hal-01598911

23. Oberle, B. (2020) « Types de chaînes de référence dans les articles de recherche de format IMRaD ». *Discours*, 25, accepté, sous presse.

24. Rousier-Vercruyssen, L. & Landragin, F. (2020) « Interdistance et instabilité au sein des chaînes de référence : indices textuels ? ». *Discours*, 25, accepté, sous presse.

25. Guillot-Barbance, C. & Quignard, M. (2020) « Chaînes de référence et structure textuelle dans les *Essais sur la peinture* de Diderot ». *Discours*, 25, accepté, sous presse, https://halshs.archives-ouvertes.fr/halshs-02484981

26. Delaborde, M. & Landragin, F. (2019) En quoi le pronom « on » a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de « on ». *Cahiers de Praxématique*, 72, en ligne, pp. 1-18, https://hal.archives-ouvertes.fr/hal-02161902

27. Schnedecker, C. (2019) De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de Praxématique*, 72, en ligne, pp. 1-19, https://hal.archives-ouvertes.fr/hal-02317889

28. Baumer, E. (2019) Expressions référentielles et point de vue dans les nouvelles en anglais et en français. *Les Cahiers du Forellis* (*Formes et Représentations en Linguistique, Littérature et dans les arts de l'Image et de la Scène)*, https://cahiersforell.edel.univ-poitiers.fr/index.php?id=693, https://hal.archives-ouvertes.fr/hal-02457169

29. Landragin, F. (2018) Étude de la référence et de la coréférence : rôles des petits corpus et observations à partir du corpus MC4, *Corpus*, 18, https://hal.archives-ouvertes.fr/hal-01834692

30. 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

31. Landragin, F. (2017) Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ? *Langue française*, 195, pp. 17-34, https://halshs.archives-ouvertes.fr/halshs-01580784

32. Schnedecker, C. (2017) Les chaînes de référence : Une configuration d'indices pour distinguer et identifier les genres textuels, *Langue française*, 195, pp. 53-72, https://hal.archives-ouvertes.fr/hal-01591017

33. Baumer, E. (2017) Chaînes de référence et point de vue dans la fiction littéraire : le cas des nouvelles courtes, *Langue française*, 195, pp. 73-90, https://hal.archives-ouvertes.fr/hal-01585659

Communications (conférences)

34. 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

35. Delaborde, M. & Landragin, F. (2019) « De la coréférence exacte à la coréférence complexe : une typologie et sa mise en œuvre en corpus ». In : *10èmes Journées*

internationales de Linguistique de Corpus (JLC 2019), Grenoble, France, https://hal.archives-ouvertes.fr/hal-02286100

36. 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

37. 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

38. Landragin, F. & Delaborde, M. (2018) Faut-il compter ou ignorer les occurrences de "ce" dans les chaînes de coréférences ? In : Ce *disant, que fait-on ? Aspects grammaticaux et discursifs de* ce *en français*, Strasbourg, https://halshs.archives-ouvertes.fr/halshs-01836380

39. Delaborde, M. & Landragin, F. (2018) En quoi le pronom "on" a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de "on". In : *Gérer L'Anaphore en Discours : vers une approche interdisciplinaire (GLAD 2018)*, Grenoble, https://halshs.archives-ouvertes.fr/halshs-01795213

40. Schnedecker, C. (2018) De l'intérêt de la notion de chaîne de référence. In : *Gérer L'Anaphore en Discours : vers une approche interdisciplinaire (GLAD 2018)*, Grenoble.

41. 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

42. 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://halshs.archives-ouvertes.fr/hal-02002111

43. 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

44. Oberle, B. (2017) Annotation de la coréférence avec SACR, un nouvel outil reposant sur le drag-and-drop. In: Workshop Eclavit, Marne-La-Vallée, https://halshs.archives-ouvertes.fr/halshs-01715467

45. Landragin, F., Potier, J. & Bothua, M. (2017) Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus, In: *Neuvièmes Journées Internationales de la Linguistique de Corpus (JLC 2017)*, Grenoble, pp. 43-46, https://halshs.archives-ouvertes.fr/halshs-01513810v1

46. Landragin, F. (2016) Conception d'un outil de visualisation et d'exploration de chaînes de coréférences, In: *Thirteen International Conference on Statistical Analysis of Textual Data (JADT 2016)*, Nice, pp. 109-120, https://halshs.archives-ouvertes.fr/halshs-01329414v1

47. Landragin, F. (2020) Le projet Democrat : ses objets d'étude, son corpus, ses applications. Conférence invitée pour le séminaire CLLE-ERSS, Toulouse, le 19/03/2020.

48. Landragin, F. (2019) Étude des chaînes de référence en français : liens entre modélisation linguistique et analyse quantitative. Conférence invitée pour la Journée scientifique « Corpus, analyses quantitatives et modèles linguistiques » de la Société Linguistique de Paris, le 26/01/2019.

49. Landragin, F. (2018) La référence et les chaînes de coréférences : un défi pour la modélisation linguistique, l'annotation manuelle de corpus et le traitement automatique des langues. Conférence invitée pour le séminaire du laboratoire Praxiling, Montpellier, le 08/10/2018.

50. Landragin, F. (2018) Description et modélisation des chaînes de référence : le projet Democrat. Conférence invitée pour la journée d'étude franco-tchèque « Linguistique textuelle, linguistique de corpus » organisée par Guy Achard-Bayle, Université de Lorraine, Metz, le 10/04/2018.

51. Landragin, F. (2018) Les avancées du projet Democrat, description et modélisation des chaînes de référence : outils pour l'annotation de corpus et le traitement automatique. Conférence invitée pour le séminaire du laboratoire Lattice, Montrouge, le 03/04/2018.

52. Landragin, F. (2017) Le projet Democrat : annoter et explorer des chaînes de référence avec TXM. Conférence invitée pour la journée organisée par le consortium CORLI (Corpus, Langues, Interactions) intitulée « Explorer un corpus annoté », Université Paris Diderot, Grands Moulins, Paris, le 25/10/2017.

53. Landragin, F. (2017) Annotation, analyse et identification automatique de chaînes de coréférences : des questions interdépendantes ? Conférence invitée pour le séminaire doctoral du laboratoire MODYCO, Nanterre, le 28/02/2017.

Articles vulgarisation

54. Landragin, F. (2016) Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT), *Bulletin de l'Association Française pour l'Intelligence Artificielle (AFIA)* 92, pp. 11-15, https://halshs.archives-ouvertes.fr/hal-01347949v1

Conférences vulgarisation

55. Landragin, F., Delaborde, M., Dupont, Y. & Grobol, L. (2018) Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours. In : *Cinquième édition du Salon de l'Innovation en TAL (Traitement Automatique des Langues) et RI (Recherche d'Informations), Vingt-cinquième conférence sur le traitement automatique des langues naturelles (TALN 2018)*, Rennes, https://hal.archives-ouvertes.fr/hal-01797982 (poster).

56. Landragin, F., Tellier, I. & Dupont, Y. (2016) DEMOCRAT : description et modélisation des chaînes de référence. Outils pour l'annotation de corpus et le traitement automatique. In : *Salon Partenariats Recherche et Industries de la Langue (PAREIL), Vingt-troisième conférence sur le traitement automatique des langues naturelles (TALN 2016)*, Paris, https://hal.archives-ouvertes.fr/hal-01384485 (poster).

Autres

57. Grobol, L. *et al*. (2020) Logiciel DeCOFR, « Detecting Coreference for Oral French », dépôt à venir sur HAL, cf. annexe F.4.

58. Wilkens, R., Oberle, B., Landragin, F. & Todirascu, A. (2020) Logiciel COFR, « COreference for FRench », https://hal.archives-ouvertes.fr/hal-02486764, annexe F.4.

59. Désoyer, A., Tellier, I. & Landragin, F. (2018) Logiciel CROC, « Coreference Resolver for Oral Corpus », https://hal.inria.fr/hal-01836189

60. Oberle, B. (2018) Logiciel SACR, « Script d'Annotation des Chaînes de Référence », https://hal.inria.fr/hal-01836266

61. Oberle, B. (2018) Logiciel ODACR, « Outil de Détection Automatique des Chaînes de Référence », https://hal.inria.fr/hal-01837101

62. Victorri, B., Landragin, F. & Poibeau, T. (2018) Logiciel ANALEC, « Analyse de l'écrit », https://hal.inria.fr/hal-01836169

63. Heiden, S. & Decorde, M. (2017) Logiciel TXM : mise à jour du 10 avril 2017, https://halshs.archives-ouvertes.fr/halshs-00377694

E.3 LISTE DES ELEMENTS DE VALORISATION

La liste des éléments de valorisation inventorie les retombées (autres que les publications) décomptées dans le deuxième tableau de la section E.1. On détaillera notamment :

- *brevets nationaux et internationaux, licences, et autres éléments de propriété intellectuelle consécutifs au projet.*
- *logiciels et tout autre prototype*
- *actions de normalisation*
- *lancement de produit ou service, nouveau projet, contrat,…*

- *le développement d'un nouveau partenariat,*
- *la création d'une plate-forme à la disposition d'une communauté*
- *création d'entreprise, essaimage, levées de fonds*
- *autres (ouverture internationale,..)*

Elle en précise les partenariats éventuels. Dans le cas où des livrables ont été spécifiés dans l'annexe technique, on présentera ici un bilan de leur fourniture.

1. Les logiciels développés dans le cadre du projet Democrat ont tous été déposés sur HAL en tant que logiciels (voire liste à la fin de la section E.2) :

- Grobol, L. *et al*. (2020) Logiciel DeCOFR, « Detecting Coreference for Oral French », dépôt à venir sur HAL. Voir le descriptif complet dans l'annexe F.4.

- Wilkens, R., Oberle, B., Landragin, F. & Todirascu, A. (2020) Logiciel COFR, « COreference for FRench », https://hal.archives-ouvertes.fr/hal-02486764. Voir le descriptif complet dans l'annexe F.4.

- Désoyer, A., Tellier, I. & Landragin, F. (2018) Logiciel CROC, « Coreference Resolver for Oral Corpus », https://hal.inria.fr/hal-01836189
  We present CROC (Coreference Resolution for Oral Corpus), the first machine learning system for coreference resolution in French. One specific aspect of the system is that it has been trained on data that are exclusively oral, namely ANCOR (ANaphora and Coreference in ORal corpus), the first corpus in oral French with anaphorical relations annotations. In its current state, the CROC system requires pre-annotated mentions. We detail the features that we chose to be used by the learning algorithms, and we present a set of experiments with these features. The scores we obtain are close to those of state-of-the-art systems for written English. Then we give future works on the design of an end-to-end system for oral and written French.

- Oberle, B. (2018) Logiciel SACR, « Script d'Annotation des Chaînes de Référence », https://hal.inria.fr/hal-01836266
  SACR is an easy to use annotation tool, specifically designed to annotate coreference chain. It is written in HTML, CSS and Javascript, and consists in one web page. You can use it freely at http://boberle.com/projects/sacr or download it from there. It is distributed under the terms of the Mozilla Public Licence, version 2.0 (see de LICENSE file).

- Oberle, B. (2018) Logiciel ODACR, « Outil de Détection Automatique des Chaînes de Référence », https://hal.inria.fr/hal-01837101
  ODACR est un détection automatique des chaînes de référence. Son implémentation suit les principes des systèmes à base de règles. ODACR exploite de nombreuses ressources linguistiques disponibles sur le web.

- Victorri, B., Landragin, F. & Poibeau, T. (2018) Logiciel ANALEC, « Analyse de l'écrit », https://hal.inria.fr/hal-01836169
  ANALEC is a tool which aim is to bring together corpus annotation, visualization and query management. The main idea is to provide a unified and dynamic way of

annotating textual data. ANALEC allows researchers to dynamically build their own annotation scheme and use the possibilities of scheme revision, data querying and graphical visualization during the annotation process.

- Heiden, S. & Decorde, M. (2017) Logiciel TXM : mise à jour du 10 avril 2017, https://halshs.archives-ouvertes.fr/halshs-00377694
  Logiciel open-source d'analyse et de mise en ligne de corpus textuels compatible Unicode, XML & TEI, basé sur le moteur de recherche plein texte CQP et l'environnement statistique R :
  Version pour poste (Windows, Mac OS X & Linux) :
  - TXM 0.7.8 (avril 2017), http://textometrie.ens-lyon.fr/files/software/TXM/0.7.8
  - TXM 0.7.9 (janvier 2018), http://textometrie.ens-lyon.fr/files/software/TXM/0.7.9/
  - TXM 0.8.0 (juin 2019), http://textometrie.ens-lyon.fr/files/software/TXM/0.8.0/
  Version portail web :
  - TXM portal 0.6.1 (9 décembre 2014),
  http://sourceforge.net/projects/txm/files/software/TXM%20portal
  L'extension URS (Unité, Relation, Schéma) provient des développements effectués dans le projet Democrat.

2. Des actions de formation à ces logiciels ont eu lieu tout au long du projet, notamment pour les logiciels dotés de fonctionnalités d'annotation et d'interrogation des données annotées (plus complexes à utiliser que les autres logiciels produits par le projet). Ces actions ont eu lieu dans un premier temps dans le périmètre des partenaires de Democrat : plusieurs journées dites « Marathon Annotation Democrat » – au minimum une dans chacun des trois laboratoires partenaires – se sont réparties fin 2016 ; 6 mars 2017 : formation de 3h « TXM et annotation URS » à l'ENS de Lyon pour les partenaires de Strasbourg, animée par Serge Heiden (IHRIM) ; 12-13 février 2019 : formation de deux demi-journées à Strasbourg animée par Matthieu Quignard (ICAR) ; etc. La liste suivante ne mentionne que les actions de formation destinée à un public plus large :

- Le 30 mars 2016 : dans le cadre de la formation « e-philologie » de l'ENS Paris, l'EPHE, l'ENC et l'EHESS, cours de 3h de Frédéric Landragin (Lattice) sur l'étude des expressions référentielles et des chaînes de référence (illustrations tirées d'ANALEC).

- Le 6 avril 2016 : dans le cadre de la formation « e-philologie » de l'ENS Paris, l'EPHE, l'ENC et l'EHESS (suite de l'item précédent), cours de 3h de Frédérique Mélanie-Becquet (Lattice) sur l'utilisation d'ANALEC pour annoter et interroger les données annotées, en prenant comme exemples plusieurs façons d'annoter des chaînes de référence.

- Le 9 mars 2017 : dans le cadre de la formation « e-philologie » de l'ENS Paris, l'EPHE, l'ENC et l'EHESS, cours de 3h de Frédéric Landragin (Lattice) sur l'étude des chaînes de référence, avec illustrations (ANALEC) issues du projet Democrat.

- Le 23 mars 2017 : dans le cadre de la formation « e-philologie » de l'ENS Paris, l'EPHE, l'ENC et l'EHESS, cours de 3h de Frédérique Mélanie-Becquet (Lattice) sur

l'utilisation d'ANALEC pour annoter et interroger des chaînes de référence, avec illustration du projet Democrat.

- Le 26 octobre 2017 : formation de 3h « TXM pour annoter » dans le cadre du consortium CORLI (CORpus, Langues, Interactions) de la TGIR Huma-Num à l'Université Paris-Diderot, animée par Serge Heiden (IHRIM).

- Le 8 février 2018 : dans le cadre du master « Humanités numériques et computationnelles » de l'Université Paris Sciences et Lettres, cours de 3h de Frédéric Landragin (Lattice) sur la méthodologie d'annotation et l'annotation de chaînes de référence (illustrations tirées de Democrat, avec ANALEC).

- Le 30 octobre 2019 : formation de 3h « TXM : annotation en plein texte (application à la co-référence) » dans le cadre du consortium CORLI de la TGIR Huma-Num, à l'Université Paris-Diderot, animée par Matthieu Quignard (ICAR), http://drehu.linguist.univ-paris-diderot.fr/corli-2019/?fichier=programme

- Le 19 mars 2020 (appel lancé en février 2020) : dans le cadre du consortium CORLI, formation à l'extension URS de Democrat, qui se déroulera à l'Université Toulouse Jean Jaurès et sera animée par Matthieu Quignard (ICAR), https://corli.huma-num.fr/node/216

3. Des actions de normalisation sont en cours, à la suite notamment d'une proposition de format de fichier XML-TEI-URS (pour le codage de corpus annotés selon le format URS – unité, relation, schéma), voir publications suivantes :

- Grobol, L., Landragin, F. & Heiden, S. (2017) Interoperable annotation of (co)references in the Democrat project. In: *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*, Montpellier, https://hal.archives-ouvertes.fr/hal-01583527
  This paper proposes XML-TEI-URS, a generic TEI-based format for the annotation of coreferences in arbitrary corpora. This proposal is made in the context of Democrat, a French Agence Nationale de la Recherche project that aims to produce a large corpus of written French with coreference annotations, in an attempt to design a corpus that is usable both by humans and automated tools and as compatible as possible with future concurrent annotations.

- Grobol, L., Landragin, F. & Heiden, S. (2018) XML-TEI-URS: Using a TEI Format for Annotated Linguistic Resources. In: *CLARIN Annual Conference 2018*, Pise, Italie, https://hal.archives-ouvertes.fr/hal-01827563v1
  This paper discusses XML-TEI-URS, a recently introduced TEI-compliant XML format for the annotation of referential phenomenons in arbitrary corpora. We describe our experiments on using this format in different contexts, assess its perceived strengths and weaknesses, compare it with other similar efforts and suggest improvements to ease its use as standard for the distribution of interoperable annotated linguistic resources.

4. **Un nouveau partenariat** a bénéficié des travaux effectués dans le projet Democrat avec, sous la coordination de deux membres de Democrat, Catherine Schnedecker et Amalia Todirascu, une collaboration avec Nicolas Amadio (DynamE), Elisabeth Demont (LPC) et Philippe Viallon (Lisec) pour la mise en place d'un projet intitulé « ILC, Institut du Langage et de la Communication », suite à l'appel à projets « Instituts Thématiques Interdisciplinaires » à l'Université de Strasbourg.

Notons également que le projet Democrat a initié deux collaborations avec d'autres projets ANR, toutes les deux pour les travaux d'expérimentations et de développement relevant du volet TAL du projet :

- Collaboration ponctuelle avec le projet ANR Alector (« Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques »), ANR-16-CE28-0005, voir l'annexe F.4.
- Collaboration avec le Labex EFL (« Empirical Foundations of Linguistics »), ANR-10-LABX-0083, voir l'annexe F.4.

5. Autres éléments de valorisation

Il est encore trop tôt pour recueillir d'autres indicateurs d'impact du projet Democrat, à la date de finalisation de celui-ci, mais notons que beaucoup parmi les publications listées ci-dessus font déjà l'objet de citations dans la communauté scientifique, comme le montre la figure 1 – qui indique par ailleurs le h-index du projet en date de février 2020, à savoir : h-index $_{\text{Democrat}}$ = 5.

	Cites	Per year	Rank	Authors	Title	Year
☑ h	13	4.33	16	Y Dupont, M Dinarelli, I Tellier	Label-dependencies aware recurrent neural networks	2017
☑ h	12	4.00	4	M Dinarelli, V Vukotić, C Ray...	Label-dependency coding in simple recurrent networks for spoken language understanding	2017
☑ h	9	2.25	17	F Landragin	Conception d'un outil de visualisation et d'exploration de chaînes de coréférences	2016
☑ h	7	2.33	21	C Schnedecker, J Glikman, F ...	Les chaînes de référence: annotation, application et questions théoriques	2017
☑ h	6	2.00	23	C Schnedecker	Les chaînes de référence: une configuration d'indices pour distinguer et identifier les genre...	2017
☑	3	0.75	8	A Désoyer, F Landragin, I Tell...	Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR	2016
☑	3	1.00	12	F Landragin, J Potier, M Both...	Annotation manuelle d'expressions référentielles: expérimentations pour simplifier les pris...	2017
☑	3	1.00	27	F Landragin	Analyse, visualisation et identification automatique des chaînes de coréférences: des quest...	2017
☑	3	1.00	30	V Obry, J Glikman, C Guillot-...	Les chaînes de référence dans les récits brefs en français: étude diachronique (xiiie-xvie s.)	2017
☑	2	1.00	1	S Heiden	Annotation-based digital text corpora analysis within the TXM platform	2018
☑	2	0.67	2	B Oberle	Coreference annotation with SACR, a new drag-and-drop based tool	2017
☑	2	1.00	5	L Grobol, I Tellier, ÉV de La C...	ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations	2018
☑	2	1.00	11	B Oberle	SACR: A drag-and-drop based tool for coreference annotation	2018
☑	2	2.00	18	M Dinarelli, L Grobol	Seq2biseq: Bidirectional output-wise recurrent neural networks for sequence modelling	2019
☑	2	0.67	28	E Baumer	Chaînes de référence et point de vue dans la fiction littéraire: le cas des nouvelles courtes	2017
☑	1	0.33	15	L Grobol, F Landragin, S Hei...	Interoperable annotation of (co) references in the Democrat project	2017
☑	1	0.33	22	L Grobol, I Tellier, ÉV de La C...	Apports des analyses syntaxiques pour la détection automatique de mentions dans un cor...	2017
☑	1	0.50	24	C Schnedecker	Reference Chains and Genre Identification	2018
☑	1	0.50	31	F Landragin	Étude de la référence et de la coréférence: rôle des petits corpus et observations à partir d...	2018

Figure 1 : Citations des publications de Democrat – copie d'écran du logiciel « Publish or Perish », requête du 26 février 2020.

Notons également que le corpus Democrat a déjà été téléchargé de nombreuses fois sur la plateforme Ortolang (https://hdl.handle.net/11403/democrat/v1.1), comme le montre la figure 2. Là encore, la parution du corpus est trop récente pour que ces statistiques de consultations et de téléchargements aient vraiment un sens. Notons également qu'Ortolang, qui propose aux propriétaires de corpus un suivi de ces statistiques, n'indique la provenance ni des consultations, ni des téléchargements. On ne peut donc pas savoir si ce sont des chercheurs individuels, des laboratoires et surtout, on ne peut pas savoir pour quel usage le corpus est téléchargé. Depuis février 2020, des pointeurs vers le corpus Democrat ont été ajoutés sur divers sites web, notamment celui de la Base de Français Médiéval (BFM) – http://txm.bfm-corpus.org/ – et celui de la Société Internationale de Diachronie du Français (SIDF) – https://diachronie.org/2020/02/25/corpus-democrat.

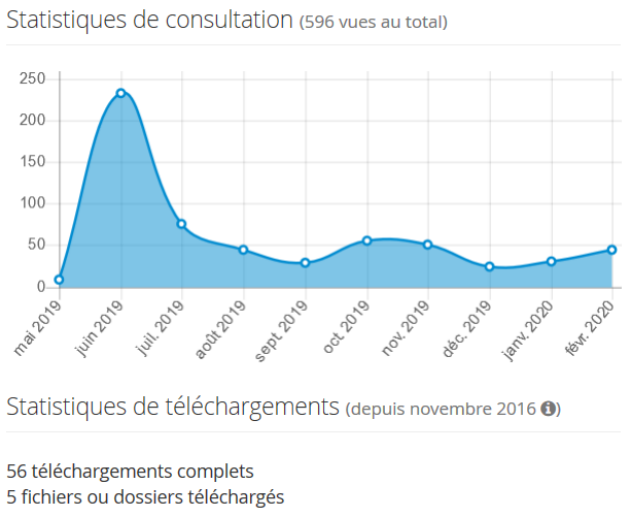


Figure 2 : Statistiques de téléchargement du corpus Democrat – copie d'écran du site Ortolang, requête du 26 février 2020.

E.4 BILAN ET SUIVI DES PERSONNELS RECRUTES EN CDD (HORS STAGIAIRES)

Ce tableau dresse le bilan du projet en termes de recrutement de personnels non permanents sur CDD ou assimilé. Renseigner une ligne par personne embauchée sur le projet quand l'embauche a été financée partiellement ou en totalité par l'aide de l'ANR et quand la contribution au projet a été d'une durée au moins égale à 3 mois, tous contrats confondus, l'aide de l'ANR pouvant ne représenter qu'une partie de la rémunération de la personne sur la durée de sa participation au projet.
Les stagiaires bénéficiant d'une convention de stage avec un établissement d'enseignement ne doivent pas être mentionnés.

Les données recueillies pourront faire l'objet d'une demande de mise à jour par l'ANR jusqu'à 5 ans après la fin du projet.

Identification				Avant le recrutement sur le projet			Recrutement sur le projet				Après le projet				
Nom et prénom	Sexe H/F	Adresse email (1)	Date des dernières nouvelles	Dernier diplôme obtenu au moment du recrutement	Lieu d'études (France, UE, hors UE)	Expérience prof. Antérieure, y compris post-docs (ans)	Partenaire ayant embauché la personne	Poste dans le projet (2)	Durée missions (mois) (3)	Date de fin de mission sur le projet	Devenir professionnel (4)	Type d'employeur (5)	Type d'emploi (6)	Lien au projet ANR (7)	Valorisation expérience (8)
Delaborde Marine	F	marine.delaborde@gmail.com	février 2020	Master	France	CDD ingénieur (2 ans)	Lattice (ENS Paris)	Doctorant	40 dont 36 financés par Democrat	31 octobre 2019	encore sur son doctorat	-	-	-	-
Decorde Matthieu	H	matthieu.decorde@ens-lyon.fr	février 2020	Master	France	CDD ingénieur d'étude (7 ans)	ENS Lyon	Ingénieur	22,5	30 avril 2018	CDD	enseignement et recherche publique	ingénieur	même partenaire	oui
Le Mené Marine	F	lemeneguigoures@unistra.fr	février 2020	Doctorat	France, Canada	CDD post-doc (1 an)	LiLPa (Université de Strasbourg)	Post-doc	14	8 mars 2019	CDI	enseignement et recherche publique	enseignante-chercheure	même partenaire	oui
Rousier-Vercruysse n Lucie	F	vercruyssenlucie@gmail.com	février 2020	Doctorat	France, Suisse, Belgique	CDD post-doc (1 an)	Lattice (ENS Paris)	Post-doc	12	31 décembre 2019	recherche d'emploi	-	-	-	-

Aide pour le remplissage

(1) Adresse email : indiquer une adresse email la plus pérenne possible
(2) Poste dans le projet : post-doc, doctorant, ingénieur ou niveau ingénieur, technicien, vacataire, autre (préciser)
(3) Durée missions : indiquer en mois la durée totale des missions (y compris celles non financées par l'ANR) effectuées sur le projet
(4) Devenir professionnel : CDI, CDD, chef d'entreprise, encore sur le projet, post-doc France, post-doc étranger, étudiant, recherche d'emploi, sans nouvelles
(5) Type d'employeur : enseignement et recherche publique, EPIC de recherche, grande entreprise, PME/TPE, création d'entreprise, autre public, autre privé, libéral, autre (préciser)
(6) Type d'emploi : ingénieur, chercheur, enseignant-chercheur, cadre, technicien, autre (préciser)
(7) Lien au projet ANR : préciser si l'employeur est ou non un partenaire du projet

(8) Valorisation expérience : préciser si le poste occupé valorise l'expérience acquise pendant le projet.

Les informations personnelles recueillies feront l'objet d'un traitement de données informatisées pour les seuls besoins de l'étude anonymisée sur le devenir professionnel des personnes recrutées sur les projets ANR. Elles ne feront l'objet d'aucune cession et seront conservées par l'ANR pendant une durée maximale de 5 ans après la fin du projet concerné. Conformément à la loi n° 78-17 du 6 janvier 1978 modifiée, relative à l'Informatique, aux Fichiers et aux Libertés, les personnes concernées disposent d'un droit d'accès, de rectification et de suppression des données personnelles les concernant. Les personnes concernées seront informées directement de ce droit lorsque leurs coordonnées sont renseignées. Elles peuvent exercer ce droit en s'adressant l'ANR (http://www.agence-nationale-recherche.fr/Contact).

F ANNEXES (PARTIES NON CONFIDENTIELLES)

Cette partie propose des informations complémentaires qui reprennent (en partie) les contenus des livrables et illustrent à ce titre les éléments de contenu des parties C, D et E.

F.1 PROGRAMMES DES WORKSHOPS ORGANISES PAR LE PROJET DEMOCRAT

Journée d'étude

« Référence, coréférence et structure textuelle »

Lundi 27 novembre 2017
ENS Lyon, Bâtiment Buisson salle de réunion 2

9h00 : Accueil des participants

9h10 : Frédéric Landragin (CNRS) et Céline Guillot-Barbance (ENS Lyon) : Introduction et objectifs

9h30 : Jean-Michel Adam (Université de Lausanne) : Paragraphes et chaînes de référence aux paliers micro- et méso-textuel d'analyse

10h20 : Pause

10h30 : Agnès Tutin (Université Grenoble Alpes) : « *Shell nouns* » et anaphore dans l'écrit scientifique

11h20 : Catherine Schnedecker et l'équipe de Strasbourg : Analyse exploratoire longitudinale des chaînes de référence dans des textes de type « encyclopédique » du Moyen Français au Français contemporain

12h00 : Repas

13h30 : Yves Bestgen (Université Catholique de Louvain) : Recherche d'indices de (dis)continuité thématique par une analyse automatique de corpus

14h20 : Céline Guillot-Barbance et l'équipe de Lyon : Chaînes de référence, structuration textuelle et genres textuels en diachronie : premières explorations du corpus Democrat

15h00 : Bruno Oberlé (Université de Strasbourg) : Étude des chaînes de référence dans les articles de recherche de format IMRaD

15h40 : Pause

15h50 : Table ronde (modérateur : Frédéric Landragin) : Pertinence et faisabilité d'une annotation spécifique du corpus Democrat pour étudier les liens entre structure textuelle et chaînes de référence

17h20 : Conclusion et perspectives de la journée

17h30 : Clôture de la journée

Journée d'étude

« Approches contrastives des chaînes de référence »

Mercredi 14 mars 2018
Lattice-ENS, 1 rue Maurice Arnoux, Montrouge, salle 302

9h00 : Accueil des participants

9h10 : Frédéric Landragin (CNRS, Lattice) : « Introduction et objectifs »

9h30 : Xiuli Wang (Beijing Language and Culture University) : « Deux solutions de la chaîne de référence en chinois par rapport au français » (conférence invitée)

10h30 : Pause

10h40 : Chang Guo (Université de Strasbourg) : « Approche contrastive de la coréférence en français-chinois : application dans un corpus de textes encyclopédiques »

11h20 : Emmanuel Baumer (Université Nice Sophia Antipolis, BCL), Laure Gardelle (Université Grenoble Alpes), Dominique Dias (Université Grenoble Alpes) et Emmanuelle Prak-Derrington (ENS de Lyon) :

« À quel point les chaînes de référence peuvent-elles être homogènes au sein d'un même genre discursif ? Exploration contrastive allemand/anglais/français »

12h00 : Repas

13h30 : Shirley Carter-Thomas (Institut Mines-Télécom, Lattice) et Laure Sarda (CNRS, Lattice) : « Anaphores (in)fidèles : analyse contrastive en langue et en genre » (conférence invitée)

14h30 : Zsuzsanna Gécseg (Université de Szeged, Hongrie), Frédéric Landragin (CNRS) et Benjamin Fagard (CNRS) : « Maillons forts et maillons faibles d'une chaîne de référence : une étude contrastive français-hongrois »

15h10 : Jan Dvorak (ENS Lyon) : « Le démonstratif et les chaînes de référence dans les langues parlée et littéraire : une brève comparaison entre le tchèque et le français »

15h50 : Pause

16h00 : Emmanuel Schang (Université d'Orléans), Anaïs Lefeuvre-Halftermeyer (Université d'Orléans) et Jean-Yves Antoine (Université François Rabelais Tours) : « Les chaînes coréférentielles en créole de la Guadeloupe »

16h40 : Bilan de la journée, tour de table, conclusions et perspectives.

17h30 : Clôture de la journée

Journée d'étude

« Mesures statistiques et approches quantitatives pour étudier les chaînes de référence »

14 juin 2019
Strasbourg, Bâtiment le Portique, salle 409

9h15 - 9h45	Accueil des participant.e.s
9h45 - 10h15	Chaînes de référence et « mesures » : le cas du découpage en paragraphes **Catherine Schnedecker** *(LiLPa)*
10h15 - 10h45	Prototypes référentiels **Lucie Vercruyssen** (*Lattice*)
10h45 - 11h15	L'effet des facteurs de "distance" et de "fréquence" sur la saillance des entités **Jiaqi Hou** (*Lattice*)
11h15 - 11h30	*Pause*
11h30 - 12h45	Exploitation de l'annotation en chaînes de référence : point de vue formel, repères techniques et retour d'expérience **Céline Guillot & Bénédicte Pincemin** (*IHRIM*)
12h45 - 14h	*Pause déjeuner*
14h - 14h45	Segmentation textuelle et chaînes de référence : étude de quelques indicateurs **Bruno Oberlé** (*LiLPa*)
14h45 - 15h15	Chaînes de référence et paragraphes : étude d'un corpus en diachronie **Daniéla Capin, Julie Glikman, Marine Le Mené, Catherine Schnedecker, Amalia Todirascu** (*LiLPa*)
15h15 - 16h30	Discussions

F.2 CORPUS DEMOCRAT

Accès au corpus

Conformément aux objectifs du projet Democrat, le corpus annoté dans le cadre du projet est diffusé gratuitement sous licence ouverte CC BY-SA. L'ensemble des fichiers, des métadonnées et des informations de la documentation du corpus a ainsi été déposé sur la plateforme ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue), plateforme vouée à être pérenne.
Lien vers le corpus annoté Democrat : https://hdl.handle.net/11403/democrat

Descriptif succinct

Le corpus Democrat est un corpus textuel annoté en références. Les expressions référentielles sont repérées (sous la forme d'unités « MENTION ») et annotées avec l'identifiant du référent (champ « REF »). Les chaînes de référence sont construites dans des schémas « CHAINE ».
La composition du corpus est établie dans l'objectif d'étudier la variation des chaînes de référence en fonction des genres discursifs et des époques. La composition est établie selon trois critères : époque, type de texte (narratif ou non-narratif), genre textuel. La taille du corpus a été envisagée pour permettre des applications de traitement automatique des langues.
Les textes constituant le corpus sont diffusés sous deux formats différents :

- Fichiers .TXM (un par texte du corpus) : il s'agit de corpus binaires utilisables directement avec le logiciel TXM version 0.8.0 (et ultérieures). C'est le point d'entrée recommandé pour explorer les annotations et effectuer des mesures : calculs de fréquences, etc. Un mode d'emploi est fourni ci-dessous.
- Fichiers .XML (deux par texte du corpus) : il s'agit de fichiers XML suivant les recommandations du consortium TEI (version P5) et les principes de codage d'annotations URS (unités, relations, schémas). Un premier fichier contient les métadonnées (titre, auteur, date, etc.) et le texte (les mots et leurs propriétés : forme graphique, partie du discours, lemme ; et éventuellement la structure logique du texte : chapitres, paragraphes, etc.) ; un second fichier contient les annotations URS (qui sont donc déportées : les annotations de ce fichier pointent vers les mots du premier fichier). C'est le point d'entrée recommandé pour mettre en œuvre des applications de traitement automatique de la langue (apprentissage artificiel).

Textes constituant le corpus

fichier	auteur	titre	source	siècle_composition	type_textuel	genre_textuel
ROLAND	Anonyme	Chanson de Roland	BFM	environ 1100	narratif	poème épique
ENEAS	Anonyme	Eneas	BFM	12e	narratif	roman
SBATH1	Anonyme	Vie de Sainte Bathilde	BFM	13e	narratif	hagiographie
CHARTES-HAIN13	Anonyme	Chartes de Hainaut	NaN	13e	non narratif	chartes
REGCRIM1	NaN	Registre criminel du Châtelet	BFM / DMF	14e	non narratif	registre
DAUDIN	Jean Daudin	De la érudition	BFM / DMF	14e	non narratif	didactique
MOREE	Anonyme	Chronique de Morée	BFM	14e	narratif	chronique historique

fichier	auteur	titre	source	siècle_composition	type_textuel	genre_textuel
COMMYNES	Philippe de Commynes	Mémoires	BFM	15e	mixte (narratif et non narratif)	mémoires
JEHANDEPARIS	Anonyme	Jean de Paris	BFM	15e	narratif	roman
LIVREDESTROISVERTUS	Christine de Pizan	Le livre des trois vertus à l'enseignement des...	BFM	15e	non narratif	manuel d'éducation
DUBELLAYDEFFENSE	Joachim du Bellay	La défense et illustration de la langue française	BVH	16e	non narratif	traité argumentatif
RABELAISPANTAGRUEL-V2	François Rabelais	Pantagruel	BVH	16e	narratif	roman
SERVITUDEVOLONTAIRE	Étienne de la Boétie	Le discours de la servitude volontaire ou le c...	http://classiques.uqac.ca/	16e	non narratif	pamphlet
LERYBRESIL	Jean de Léry	Histoire d'un voyage fait en la terre du Brésil	BVH	16e	mixte (narratif et non narratif)	récit de voyage
DESPERRIERSRECRE	Bonaventure Des Periers	Les nouvelles récréations et joyeux devis	BVH	16e	NaN	NaN
LAFAYETTECLEVES	Marie-Madeleine de la Fayette	Princesse de Clèves	Wikisource/Frantext	17e	narratif	roman
COEFFETEAUHISTOIRE	Nicolas Coeffeteau	Histoire romaine	Frantext	17e	narratif	traité historique
BOSSUETDISCOURS	Jacques-Bénigne Bossuet	Discours sur l'histoire universelle	Frantext	17e	mixte (narratif et non narratif)	traité historique
PERRAULTCONTES	Charles Perrault	Contes (vers)	Frantext	17e	narratif	récit bref
PERRAULTCONTES2	Charles Perrault	Contes (prose)	Frantext	17e	narratif	récit bref
SERRESAGRICULTURE	Olivier de Serres	Le théâtre d'agriculture et ménage des champs	Frantext	17e	non narratif	traité (agriculture ?)
DESCARTESDISCOURS	René Descartes	Discours de la méthode	Frantext	17e	non narratif	texte philosophique
PAULETVIRGINIE	Jacques-Henri Bernardin de Saint-Pierre	Paul et Virginie	Wikisource	18e	narratif	roman
RAMSAYCYRUS	André Michel Ramsay	Les voyages de Cyrus	Frantext	18e	narratif	roman
MONTESQUIEULOIS	Montesquieu	L'Esprit des lois	Frantext	18e	non narratif	traité argumentatif
VOLTAIREESSAI	Voltaire	Essai sur l'histoire générale et sur les moeur...	Frantext	18e	non narratif	traité argumentatif
DIDEROTESSAIS	Denis Diderot	Essais sur la peinture	Frantext	18e	non narratif	traité didactique
FABLES	François de Salignac de la Mothe-Fénelon	Fables et opuscules pédagogiques	Wikisource	18e	narratif	fables
BOUVARDETPECUCHET	Gustave Flaubert	Bouvard et Pécuchet	Wikisource	19e	narratif	roman
CAPITAINEFRACASSE	Théophile Gautier	Le capitaine Fracasse	Wikisource	19e	narratif	roman
VENTREDEPARIS	Émile Zola	Le ventre de Paris	Wikisource	19e	narratif	roman
MORTEAMOUREUSE	Théophile Gautier	La morte amoureuse	Wikisource	19e	narratif	nouvelle

fichier	auteur	titre	source	siècle_composition	type_textuel	genre_textuel
SARRASINE	Honoré de Balzac	Sarrasine	Wikisource	19e	narratif	nouvelle
CHATEAUBRIANDGENIE	François-René de Chateaubriand	Génie du christianisme	Frantext	19e	non narratif	traité argumentatif
CODECIVILFRANCAIS-1	NaN	Code civil des français (Code Napoléon)	Wikisource	19e	non narratif	texte juridique
CODECIVILFRANCAIS-2	NaN	Code civil des français (Code napoléon)	Wikisource	19e	non narratif	texte juridique
MADEMOISELLEFIFI-1	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (1)	Projet Gutenberg	19e	narratif	nouvelle
MADEMOISELLEFIFI-2	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (2)	Projet Gutenberg	19e	narratif	nouvelle
MADEMOISELLEFIFI-3	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (3)	Projet Gutenberg	19e	narratif	nouvelle
MADAMEDEHAUTEFORT	Victor Cousin	Madame de Hautefort	Wikisource	19e	narratif	biographie
PAULINE	Alexandre Dumas	Pauline	Wikisource	19e	narratif	roman
NEMOVILLE	Adèle Bourgeois	Némoville	Wikisource	20e	narratif	roman
DELAVILLEAUMOULIN	Marguerite Audoux	De la ville au moulin	Wikisource	20e	narratif	roman
ESTREPUBLICAIN	NaN	Est Républicain_1 (collection d'articles)	Ortolang	20e	non narratif	articles de presse
CONVTRANSPORTAERIEN	NaN	Convention pour l'unification de certaines règ...	http://eur-lex.europa.eu	20e	non narratif	texte juridique
CONVENTIONMILIEUMARIN	NaN	Convention pour la protection du milieu marin ...	http://eur-lex.europa.eu	20e	non narratif	texte juridique
CODEPROCEDUREPENALE	NaN	Code de procédure pénale	Legifrance	20e	non narratif	texte juridique
CONVENTIONINSTUNIV	NaN	Convention portant création d'un institut univ...	http://eur-lex.europa.eu	20e	non narratif	texte juridique
ADEN	Paul Nizan	Aden Arabie	http://www.ebooksgratuits.com/	20e	non narratif	pamphlet
ELISABETHSETON	Laure CoN/A	Élisabeth Seton	Wikisource	20e	narratif	biographie
ROSALIEDECONSTANT	Lucie Achard	Rosalie de Constant, sa famille et ses amis	Wikisource	20e	narratif	biographie
JEANCHRISTOPHE-1	Romain Rolland	Jean-Christophe (1)	Wikisource	20e	narratif	roman
JEANCHRISTOPHE-2	Romain Rolland	Jean-Christophe (2)	Wikisource	20e	narratif	roman
ESTREPUBLICAIN-2	NaN	Est Républicain_2 (collection d'articles)	Ortolang	20e	non narratif	articles de presse
DOUCELUMIERE	Marguerite Audoux	Douce Lumière	Wikisource	20e	narratif	roman
DIABLEAUCORPS	Raymond Radiguet	Le diable au corps	Wikisource	20e	narratif	roman
CONVENTIONTHONTROPIC	NaN	Convention relative au renforcement de la Comm...	http://eur-lex.europa.eu	21e	non narratif	texte juridique
ARTICLESWIKI	NaN	Articles encyclopédiques "zèbre", "girafe", "s...	Wikipedia	21e	non narratif	articles encyclopédiques

Mode d'emploi pour les fichiers .TXM

1. si nécessaire (à faire une seule fois) :
    - installer TXM 0.8.0, à télécharger depuis http://textometrie.ens-lyon.fr ;
    - lancer une première fois TXM pour finaliser l'installation, puis quitter ;
    - relancer TXM ;
    - installer l'extension « Annotation URS (Unité, Relation, Schéma) » avec la commande « Fichier > Ajouter une extension ». Valider les étapes et relancer TXM pour finaliser l'installation de l'extension.
2. depuis TXM :
    - charger le fichier « .txm » avec la commande « Fichier > Charger > un corpus binaire (.txm)... » ;
    - vous pouvez désormais utiliser tous les outils de TXM sur le corpus : Lexique, Concordance, Progression, Lecture de l'édition, etc. ;
    - outils spécifiques à l'annotation URS :
        - les outils d'annotation et d'exploitation URS sont documentés dans la section « Annotation avec un modèle Unité-Relation-Schéma (URS) au fil du texte » du manuel de TXM 0.8 – en ligne à l'adresse suivante : https://zenodo.org/record/3267345.
        - sont couverts les aspects suivants :
            - annotation plein texte par l'interface ;
            - annotation par scripts ;
            - vérification de la cohérence des annotations ;
            - exploitation (lister, compter, visualiser...) ;
            - export.

Mode d'emploi pour importer dans TXM les fichiers .XML

- séparer les fichiers « *-urs.xml » dans un répertoire « urs » ;
- déplacer les fichiers de textes « .xml » dans un répertoire « democrat » ;
- lancer TXM ;
    - lancer la commande d'import « Fichier > Importer > XML-TEI TXM » sur le répertoire « democrat » ;
    - sélectionner le corpus DEMOCRAT ;
    - lancer la commande « URS > Importer des annotations XML-TEI URS... » sur le répertoire « urs » ;
    - vous pouvez désormais utiliser tous les outils de TXM sur ce corpus contenant l'ensemble des textes annotés.
      Remarque : la lecture des « éditions » de textes peut être moins complète que dans la version « .txm ».

F.3 Extension « Annotation URS » de l'outil d'analyse de corpus TXM

Remarque préalable

L'essentiel du travail réalisé autour des aspects « linguistique de corpus outillée » du projet consiste en une extension appelée « Annotation URS (Unité-Relation-Schéma) » du logiciel TXM d'analyse textométrique de corpus textuels <http://textometrie.ens-lyon.fr>. Cette extension intègre effectivement du code source du logiciel Analec <http://www.lattice.cnrs.fr/ressources/logiciels/analec> pour rendre les fonctionnalités d'annotation dynamique du modèle URS (Unité-Relation-Schéma) compatibles avec l'environnement de la plateforme TXM (architecture des corpus textuels, outils d'exploitation, interface utilisateur intégrée, outils d'import / export de textes, d'annotations et de résultats, etc.) tout en développant de nouvelles fonctionnalités basées sur ce modèle.

Accès à l'outil d'annotation

Conformément aux objectifs du projet Democrat, l'extension « Annotation URS (Unité-Relation-Schéma) » de TXM développée dans le cadre du projet est diffusée gratuitement sous licence ouverte GNU GPL v3. Elle prend la forme d'un composant logiciel (plugin) au standard OSGi pour Eclipse RCP. Ce composant est hébergé dans le site de mise à jour (update site) de la plateforme TXM à l'adresse <http://textometrie.ens-lyon.fr/dist/0.8.0/ext/stable/site.xml> (feature « org.txm.annotation.urs.feature »).

L'extension est téléchargeable automatiquement par le biais du logiciel TXM : pour y accéder l'utilisateur doit d'abord installer le logiciel TXM en version 0.8.0 ou supérieur depuis son site de diffusion <http://textometrie.ens-lyon.fr>, puis ajouter l'extension « Annotation URS (Unité-Relation-Schéma) » depuis TXM par le biais de la commande « Fichier > Ajouter une extension ». L'installation de l'extension ajoute à TXM de nouvelles commandes et interfaces utilisateur pour le travail avec les annotations « URS ».

Documentation

La documentation de l'extension est accessible en ligne à l'adresse https://zenodo.org/record/3267345. Il s'agit d'un extrait du manuel de TXM 0.8.

Elle est composée des sections suivantes :

A) Installation de l'extension
B) Modèle d'annotation par défaut
C) Importer un corpus déjà annoté
D) Importer des annotations
E) Annoter Interactivement
F) Enregistrer les annotations
G) Annoter, Vérifier, Exploiter et Exporter par commandes
H) Outils spécifiques Democrat
I) Exporter des annotations

Utilisation

Les commandes de l'extension « Annotation URS (Unité-Relation-Schéma) » sont compatibles avec les annotations se trouvant dans les fichiers « .txm » livrés dans le livrable L1 : « Corpus annoté » du projet DEMOCRAT (voir le mode d'emploi de ce livrable), et permettent par ailleurs d'annoter n'importe quel nouveau texte ou corpus de textes importé dans TXM.

F.4 OUTILS DE DETECTION AUTOMATIQUE DES CHAINES DE REFERENCE

Remarque préalable

Cette partie concerne 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.

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 des travaux de thèse de Bruno Oberlé et 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

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/decofre

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 des annotations des singletons (expressions référentielles non reprises), 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 conduits 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. Voir https://github.com/boberle/corefconversion et https://github.com/boberle/jsonlines2tei.

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 complique 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 équivalentes (même forme de surface) mais 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 telles, 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.

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] .