Terminologica

TOTh 2020

Terminologie & Ontologie : Théories et Applications

Presses Universitaires Savoie Mont Blanc

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2020

Université Savoie Mont Blanc

26 & 27 novembre 2020

Les ouvrages TOTh précédents sont disponibles :

- sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr)
- sur le site de l'Université Savoie Mont Blanc (btk.univ-smb.fr/livres)
- ou auprès de : contact@toth.condillac.org

Éditeur : Presses Universitaires Savoie Mont Blanc
27 rue Marcoz
BP 1104
73011 CHAMBÉRY CEDEX
www.univ-smb.fr

Réalisation : C. Brun, C. Roche
Collection « Terminologica »
ISBN : 978-2-37741-065-1
ISSN : 2607-5008
Dépôt légal : mai 2021

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2020

Université Savoie Mont Blanc

26 & 27 novembre 2020

http://toth.condillac.org

avec le soutien de :

- Université Savoie Mont Blanc
- École d'ingénieurs Polytech Annecy Chambéry
- Ministère de la Culture. Ce projet est soutenu financièrement par le Ministère de la Culture - Délégation Générale à la Langue Française et aux Langues de France

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Presses Universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Universidade Nova de Lisboa
Humbley John	Université Paris 7
Kockaert Hendrik	University of Leuven
Christophe Roche	Université Savoie Mont Blanc

Comité de programme 2019

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http://toth.condillac.org/committees

Amparo Alcina	Universitat Jaume I – Spain
Xiaomi An	Renmin University – China
Albina Auksoriute	Institute of the Lithuanian Language – Lithuania
Bruno Bachimont	Université Technologie de Compiègne – France
Jean-Paul Barthès	Université Technologie de Compiègne – France
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Nicolleta Calzolari	Instituto di Linguistica Computazionale, CNR – Italy
Stéphane Chaudiron	Université de Lille 3 – France
Manuel Célio Conceição	Universidade do Algarve – Portugal
Rute Costa	Universidade NOVA de Lisboa – Portugal
Luc Damas	Université Savoie Mont-Blanc – France
Dardo De Vecchi	Kedge Business School – France
Thierry Declerck	DFKI – Germany
Valérie Delavigne	Université Paris 3 – France
Sylvie Desprès	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Cécile Frérot	Université Stendhal Grenoble 3 – France
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
François Gaudin	Université de Rouen – France
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Jean-Yves Gresser	ancien Directeur à la Banque de France – France
Gernot Hebenstreit	University of Graz – Austria
John Humbley	Université Paris 7 – France
Yangli Jia	University of Liaocheng – China
Kyo Kageura	University of Tokyo – Japan
Barbara Karsch	BIK Terminology – USA

Hendrik Kockaert	University of Leuven – Belgium
Heba Lecocq	INALCO – France
Hélène Ledouble	Université de Toulon – France
Georg Löckinger	University of Applied Sciences Upper Austria – Austria
Bénédicte Madinier	Dispositif d'enrichissement de la langue française – France
Rodolfo Maslias	TermCoord, European Parlement – Luxembourg
John McCrae	National University of Ireland – Ireland
Christine Michaux	Université de Mons – Belgium
Fidelma Ní Ghallchobhair	Foras na Gaeilge, Irish-Language Body – Ireland
Donncha Ó Cróinín	Foras na Gaeilge, Irish Terminology Committee – Ireland
António Pareja Lora	Universidad Complutense de Madrid – Spain
Silvia Piccini	Italian National Research Council – Italy
Suzanne Pinson	Université Paris Dauphine - France
Maria Pozzi	el colegio de méxico – Mexico
Michele Prandi	Università degli Studi di Genova – Italy
Bihua Qiu	China National Committee for Terms in Sciences and Technologies – China
Jean Quirion	Université d'Ottawa – Canada
Renato Reinau	Université de Genève – Switzerland
Christophe Roche	Université Savoie Mont-Blanc – France
Mathieu Roche	CIRAD – France
Laurent Romary	INRIA & HUB-ISDL – Germany
Micaela Rossi	Università degli studi di Genova – Italy
Klaus-Dirk Schmitz	Cologne University – Germany
Bernadette Sharp	Staffordshire University – Great Britain
Frieda Steurs	University of Leuven – Belgium
Toma Tasovac	Belgrade Center for Digital Humanities – Serbia
Anne Theissen	Université de Strasbourg – France
Philippe Thoiron	Université Lyon 2 – France
Tanja Wissik	Austrian Academy of Sciences – Austria
Maria Teresa Zanola	Università Cattolica del Sacro Cuore – Italy
Rim Zarrad Zarrad	University of Tunis El Manar – Tunisia

Avant-propos

2020 aura été une année bien sombre. La pandémie de COVID-19 n'aura rien épargné. Comme de nombreux événements, la conférence TOTh 2020 a été fortement impactée. Reportée de la traditionnelle première semaine de juin à fin novembre dans l'espoir, vite déçu, d'une tenue en présentiel, nous nous sommes résolus à organiser à distance la formation et la conférence TOTh. Elles se sont déroulées du 24 au 27 novembre 2020, et malgré quelques appréhensions, cela a été une réussite. La visioconférence s'est déroulée sans accroc ni incident technique, à un point tel que ce mode de participation, même si rien ne remplace la présence physique et la chaleur des rencontres indispensables à la vie d'une communauté comme TOTh, devra maintenant être considérée. Une réussite aussi avec, comme depuis plusieurs années, une participation internationale importante, 27 nationalités étaient représentées.

La Conférence TOTh s'est ouverte avec la conférence invitée de notre collègue Nuno Fonseca de l'Université Nouvelle de Lisbonne portant sur « La logique de Port Royal : une logique des idées et une sémantique des termes ». L'ouvrage demeure une référence pour la terminologie en tant que science et sa lecture a des vertus prophylactiques en particulier pour les notions de concept et de définition.

Les 14 communications retenues pour publication ont permis d'aborder de nombreux sujets tant théoriques que pratiques portant sur des domaines aussi variés que la sécurité alimentaire, la finance ou la karstologie, autant d'exemples de la diversité et de la richesse de notre discipline. A noter que deux de ces communications sont directement liées au projet Européen ELEXIS. Je vous invite à les découvrir à travers ces actes publiés aux Presses Universitaires Savoie Mont Blanc.

Avant de vous souhaiter bonne lecture, j'aimerais terminer en remerciant le Ministère de la Culture et plus précisément la Délégation Générale à la Langue Française et aux Langues de France, l'Université Savoie Mont-Blanc et l'école Polytech Annecy-Chambéry pour leur support et leur aide financière à l'organisation de la conférence et à la publication des actes.

Christophe Roche
Président du comité scientifique

Sommaire

Conférence d'ouverture 13

La Logique de Port Royal :
Une logique des *idées* et une sémantique des *termes*
Nuno Fonseca 15

Articles 39

Eugen Wüster's Semiotic Quadrangles
Marija Ivanović 41

Difficulties of unifying international anatomical nomenclature
Éva Katalin Varga, Andrea Barta, Ákos Zimonyi 63

Première approche de la modélisation ontoterminologique
de la balance des paiements
Stéphane Carsenty 81

Construction d'un corpus sur la problématique
de la sécurité alimentaire guidée par un lexique
et des approches de fouille de textes
Camille Schaeffer, Roberto Interdonato, Mathieu Roche 101

Extraire des *patterns* pour améliorer l'idiomaticité de résumés
semi-automatiques en finances : le cas du lexique support
Abdelghani Laifa, Laurent Gautier, Christophe Cruz 113

Exploitation des données orales pour
l'enrichissement de ressources lexicales
Hélène Flamein et Iris Eshkol-Taravella 137

Construction collaborative d'une ressource termino-ontologique
pour le domaine de l'orthophonie fondée sur la temporalité
Frédérique Brin-Henry, Rute Costa, Sylvie Desprès 153

Terminologie et Ontologie pour l'Héritage Culturel : Le projet Tao Ci
Wei Tong, Christophe Roche, Yangli Jia, Maria Papadopoulou 175

Une ontologie pour les systèmes multi-agents ambiants dans les villes intelligentes
Nathan Aky, Denis Payet, Sylvain Giroux, Rémy Courdier 199

Automated Formalisation of unstructured Terminologies
Thierry Declerck and Tanja Wissik 217

A bilingual approach to specialised adjectives through word embeddings in the karstology domain
Larisa Grčić Simeunović, Matej Martinc, Špela Vintar 233

Formal Representation of Linguistic Restrictions expressed in Terminologies
Thierry Declerck, Stefania Racioppa, Tanja Wissik 251

Unravelling the confusions: defining concepts to record archaeological and historical evidence for knitting
Jane Malcolm-Davies 267

Constitution d'un corpus de contextes définitoires dans le domaine de la propriété intellectuelle : vers la définition de structures linguistiques dans les brevets
Cécile Frérot et Cristina Valentini 283

Construction d'un corpus sur la problématique de la sécurité alimentaire guidée par un lexique et des approches de fouille de textes

Camille Schaeffer*, Roberto Interdonato* **, Mathieu Roche* **

* TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
Prenom.Nom@teledetection.fr
** CIRAD - Centre de coopération internationale en recherche agronomique pour le développement, Montpellier, France
Prenom.Nom@cirad.fr

Résumé. Des systèmes d'alerte précoce sur la gestion des risques liés à la sécurité alimentaire sont mis en place pour informer les acteurs, afin d'adapter et améliorer le développement de l'agriculture. Ces systèmes analysent principalement des images satellitaires (par exemple, des images de champs agricoles) et des données quantitatives (par exemple, le prix du marché alimentaire). Cependant, peu de données textuelles sont intégrées dans leur traitement, or elles sont de plus en plus abondantes et accessibles sur Internet. Elles ne décrivent pas parfaitement la situation géographique, mais elles peuvent apporter une information ou une observation complémentaire aux images satellitaires. Dans ce contexte, cet article propose d'acquérir un corpus d'articles de sites d'actualités en lien avec le thème de la sécurité alimentaire au Burkina Faso et d'extraire les informations spatio-temporelles présentes dans les articles.

1. Introduction

Avec la révolution du numérique, les données de plus en plus abondantes et accessibles peuvent être utilisées pour rechercher des informations spécifiques et apporter certaines observations sur des événements. Dans ce contexte, des systèmes d'alerte précoce sur la gestion des risques liés à la sécurité alimentaire sont mis en place pour informer les acteurs afin d'adapter et améliorer le développement de l'agriculture. Les systèmes suivent et ana-

lysent notamment la situation agricole et la situation pluviométrique grâce aux images satellitaires, ainsi que la situation économique (par exemple, le prix du marché alimentaire) grâce aux données quantitatives. Cependant, ces systèmes d'alerte précoce intègrent peu de données textuelles dans leur traitement.

Le travail présenté dans cet article consiste à appliquer des traitements de fouille de textes consacrés à la problématique de la sécurité alimentaire au Burkina Faso en analysant des articles extraits de sites d'actualités de ce pays. Les informations issues de ces journaux peuvent apporter des connaissances complémentaires par rapport aux données présentes dans les systèmes d'alerte précoce, afin de mieux anticiper et informer en temps quasi-réel les potentiels problèmes de sécurité alimentaire au Burkina Faso.

De manière plus précise, cet article décrit un processus automatique de collecte et de sélection d'articles d'actualité liés à la sécurité alimentaire guidé par des approches de fouille de textes et des lexiques dédiés. À partir de ces textes identifiés, un processus d'extraction d'entités spatio-temporelles est mis en place.

La section suivante décrit les sources de données identifiées. Ensuite, les sections 3 et 4 détaillent, respectivement, le processus mis en place pour sélectionner les textes liés à la thématique de la sécurité alimentaire puis les méthodes d'extraction d'entités spatiales et temporelles à partir de ces données textuelles. Ces processus sont évalués en section 5. Enfin, quelques perspectives à ce travail sont résumées en sections 6 et 7.

2. Corpus

Il existe sur le web plusieurs sites d'actualité du Burkina Faso, dont les principaux sont *l'Observateur Paalga*, *Burkina24*, *Lefaso.net* qui sont des quotidiens d'informations générales (politique, société, économie, coopération, culture, portraits, multimédia, sport, international). Par un système automatique de « crawling », nous avons collecté un corpus en français décrit dans le Tableau 1. Ce corpus sera analysé dans la section suivante en mobilisant des approches de fouille de textes.

Nom journal	Date parution articles	Nombre d'articles
Observateur paalga	Depuis 2012	2 019
Burkina24	Depuis 2011	6 365
Lefaso.net	Depuis 2003	23 229
Total		31 613

Tab. 1 – *Articles collectés sur la base de journaux du Burkina Faso.*

3. Processus d'identification des documents pertinents

Après avoir recueilli les articles, nous devons identifier ceux qui traitent de la problématique de la sécurité alimentaire. N'ayant pas de données étiquetées pour l'analyse thématique des sujets, nous avons appliqué un algorithme non supervisé en utilisant le modèle LDA (Latent Dirichlet Allocation) (Jelodar *et al.* 2019) qui permet une catégorisation plus précise des articles. LDA réalise une modélisation du processus de génération des données à savoir des paires documents-termes à l'aide de facteurs latents. Une telle approche a été utilisée dans des problématiques similaires comme les catastrophes naturelles (Nazer *et al.* 2017). Après avoir appliqué LDA, nous obtenons des thématiques (*topic*s) composées de termes avec une probabilité d'occurrence pour la thématique donnée et les articles ayant une probabilité d'appartenance à une ou plusieurs thématique(s).

Pour réaliser nos expérimentations avec l'algorithme LDA, la librairie Python Scikit-learn a été utilisée. Les résultats obtenus n'ont pas nécessairement mis en évidence le thème de la sécurité alimentaire car les sujets abordés par ces journaux sont très généralistes. Seules trois thématiques (*topics*) mettent en évidence un vocabulaire lié à la sécurité alimentaire (*céréales, plantes, aliments, repas, nutrition, alimentation, eau*) qui sont associées, en général, à des thématiques de santé/médecine. Dans ce contexte, nous avons combiné l'approche LDA à une méthode *Word2vec* guidée par un lexique dédié afin d'identifier des articles spécifiques à notre étude (Interdonato *et al. 2018*).

Word2vec (plongement lexical) (Mikolov *et al.* 2013) est un ensemble de modèles de réseau neuronal qui permet de transformer les mots en vecteurs,

et de comparer ces vecteurs entre eux. Cette comparaison permet d'obtenir une proximité sémantique des mots, par exemple, les mots *aliment* et *repas* sont modélisés par des vecteurs proches. *Word2vec* a besoin d'avoir un modèle de vecteurs pré-entraîné sur des données volumineuses. Nous utilisons un modèle pré-entraîné sur Wikipédia français. Wikipédia étant publié dans un langage standard (orthographe et grammaire stables), le modèle semble adapté à nos données qui utilisent également un langage standard.

Afin de trouver les thématiques (générées par LDA) les plus proches du thème de la sécurité alimentaire, nous utilisons un lexique ayant 90 entrées de noms communs, comportant des termes de sécurité alimentaire comme *alimentaire*, *agriculture*, *élevage*, *légumineuse*, *céréales*, *malnutrition*, etc. Le lexique provient du site web *Cultivo*[1] , contenant des informations historiques et scientifiques. Notre approche classe les thématiques en fonction de leur similarité sémantique globale avec le lexique de la sécurité alimentaire (en considérant tous les articles de la thématique comme un seul texte) avec le modèle *Word2vec.* Dans le contexte de notre étude, dans un premier temps, les thématiques les mieux classées (au-dessus de 0.7) ont été sélectionnées puis les articles présents dans ces thématiques sont classés selon leur similarité avec le lexique. L'évaluation de ce classement est discuté en section 5.

4. Extraction d'information dans les articles

Une fois les articles pertinents identifiés, c'est-à-dire en lien avec la thématique de la sécurité alimentaire, les entités nommées et plus spécifiquement les informations spatio-temporelles peuvent être extraites à partir de ces textes.

4.1. Reconnaissance d'entités nommées

Les entités nommées (EN) ont été définies dans le contexte de MUC (Message Understanding Conferences) dans les années 90 comme "tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique".

La reconnaissance des entités nommées consiste à identifier l'entité et déterminer la catégorie à laquelle elle fait référence, à savoir un *nom de*

1 https://www.cultivoo.com/index.php/developpement-durable/agriculture/2590-vocabulaire-sur-la-securite-alimentaire

personne, un *lieu*, une *organisation*, etc. (Schmitt *et al*. 2019). Il existe deux familles d'approches pour la reconnaissance des entités nommées :

(1) Les méthodes fondées sur des règles *ou rule-based*. Ce type d'approche implique de définir des listes de noms (par exemple, gazetiers) et des motifs correspondant aux entités nommées à extraire. En général, ces méthodes sont précises mais longues et fastidieuses à mettre en place. Elles rencontrent également des limites quant à leur exhaustivité.

(2) Les méthodes fondées sur des approches d'apprentissage automatique (*machine learning*). Dans ce contexte, des modèles sont entraînés à partir de données textuelles annotées (c'est-à-dire, entités nommées annotées), afin de reconnaître les entités à partir de nouvelles données textuelles. L'apprentissage est alimenté grâce à des indices présents dans le texte tels que la forme des mots (majuscules, nombres, etc.), le contexte des mots, les étiquettes morpho-syntaxiques, etc.

4.2. Reconnaissance d'entités spatiales

Afin de reconnaître les entités spatiales, le logiciel SpaCy[2] a été utilisé. SpaCy reconnaît différents types d'entités nommées, par exemple :

- PER (nom d'une personne, nom de famille),
- LOC (nom d'une localisation, c'est-à-dire villes, provinces, pays, régions, fleuve, lac, montagne, etc.),
- ORG (nom d'une organisation, etc.),
- MISC (nom d'entités diverses comme les événements, produits, etc.).

Dans nos travaux, nous utilisons un modèle SpaCy (cf. Fig 1) qui a été préalablement entraîné sur deux corpus en français :

- Corpus Sequoia qui repose sur des données textuelles provenant d'Europarl (Parlement européen), du journal Est Républicain (journal quotidien), de Wikipédia français (encyclopédie collective en ligne), et des documents de l'Agence Européenne du Médicament.
- Le corpus WikiNER pour la reconnaissance d'entités nommées (données textuelles en français provenant de Wikipédia français).

Notons que, dans ce processus, seuls les toponymes sont reconnus. Par exemple, avec le segment « *marché de Bamako* », seul le toponyme "*Bamako*" est finalement identifié.

Pour pallier certaines erreurs issues de SpaCy, le système de reconnaissance d'entités nommées a été enrichi d'un lexique dédié et spécifique

2 https://spacy.io/api/doc

à notre cas d'étude. De manière concrète, une liste de noms de lieux (provinces, régions, communes, villes et villages) du Burkina Faso a été ajoutée au lexique de base utilisé par SpaCy. Cette liste a été produite et utilisée par l'État Burkinabé pour des enquêtes permanente agricole (EPA). Ceci nous permet de réaliser des annotations en meilleure adéquation avec les villes et villages du Burkina Faso.

Par ailleurs, deux autres lexiques ont été ajoutés, à savoir un dictionnaire d'adverbes et un lexique de mots génériques qui étaient initialement reconnus comme des lieux par le système de reconnaissance d'entités spatiales.

```
Projet «Un ménage vulnérable une vache» : Un bol… de lait pour des familles burkinabè.
25 Jui 2014.
Dans les mois à venir, des familles burkinabè connaîtront une amélioration de leur
alimentation à travers le projet «un ménage vulnérable, une vache» du ministère des
Ressource animales et halieutiques.  Le lancement de ce projet, qui veut augmenter la
consommation  de lait au ['Burkina', 'LOC'] , a eu lieu le  lundi 23 juin 2014 à
['Kaya', 'LOC'] dans la région du ['Centre-Nord', 'LOC'] , concomitamment avec celui de
la campagne de production fourragère et la pose de la première pierre de l'abattoir
frigorifique de la ville.    5 000 vaches à la disposition de ménages vulnérables,
augmentation de la production laitière de 840, 1080 ou 1440 litres par an, accroissement
du revenu par ménage, amélioration de la situation nutritionnelle des bénéficiaires… le
projet «un ménage vulnérable, une vache» va donc changer les habitudes des populations
après sa mise en œuvre. Raison pour laquelle son initiateur, le ministère des Ressources
animales et halieutiques, a pris toutes les dispositions, notamment en mobilisant 22 208
018 161 F CFA pour son exécution. «Celui-ci procurera à la famille du lait pour
améliorer son bol alimentaire quotidien, mais aussi contribuera à améliorer ses revenus
à partir de la vente du surplus de lait», affirme d'ailleurs le chef du département,
```

Fig. 1 – *Exemple d'article annoté par SpaCy.*

4.3. Reconnaissance d'entités temporelles

Pour reconnaître les entités temporelles, nous utilisons HeidelTime, un programme, développé par le Database Systems Research Group de l'Université de Heidelberg, permettant d'annoter les entités temporelles multilingues de documents selon la norme d'annotation TIMEX3. HeidelTime nécessite, en entrée, un texte avec des phrases décomposées en tokens (mots) et avec un étiquetage morpho-syntaxique (nom, verbe, etc.). Ces traitements ont été réalisés avec TreeTagger[3]. HeidelTime extrait et normalise alors les entités temporelles.

3 https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Il existe deux principaux types d'expressions temporelles : les expressions temporelles absolues (par exemple, *mercredi 12 juin 2019*, *2011*, etc.) et relatives (par exemple, *cette semaine*, *3 heures*, etc.). Les entités temporelles sont annotées selon la typologie de HeidelTime. Ainsi, la balise <TIMEX3> permet de marquer les heures (type=TIME), les dates (type=DATE), les intervalles et les durées (cf. Fig 2).

```
<?xml version="1.0"?>
<!DOCTYPE TimeML SYSTEM "TimeML.dtd">
<TimeML>
['Campagne', 'LOC'] agricole : Entre espoir d'une bonne pluviométrie et lutte contre la
menace chenille.
Publié <TIMEX3 tid="t2" type="DATE" value="2018-07-20">le 20 juillet 2018</TIMEX3> at
<TIMEX3 tid="t9" type="TIME" value="2018-07-20T14:11">14 h 11</TIMEX3> min.
La campagne agricole s'est effectivement installée dans toutes les régions du
['<Burkina>', 'LOC'] . C'est la principale information donnée <TIMEX3 tid="t13"
type="DATE" value="2018-07-20">ce vendredi 20 juillet 2018</TIMEX3> par le ministre de
l'agriculture  et des aménagements hydrauliques Jacob Ouédraogo au cours d'une
conférence de presse à ['<Ouagadougou>', 'LOC'] . Selon les relevés, le cumul de
pluviométrie au <TIMEX3 tid="t22" type="DATE" value="2018-07-10">10 juillet
2018</TIMEX3> comparativement à la même période de la campagne précédente indique une
situation déficitaire dans la majeure partie du territoire, exception faite aux régions
du <'<Sahel>', 'LOC'> , du ['<Centre-Ouest>', 'LOC'] et des ['<Cascades>', 'LOC'] qui,
elles, ont enregistré un cumul pluviométrique excédentaire. Cependant, rien
d'inquiétant selon le ministre Jacob Ouédraogo qui espère même une campagne agricole
fructueuse. D'après l'Agence Nationale de la Météorologie (ANAM), une pluviométrie
excédentaire à normale est annoncée sur la quasi-totalité du territoire national pour
cette campagne agricole. Pour autant, une pluviométrie excédentaire ne signifie pas
forcément une sécurité alimentaire. Et le ministre Jacob dit en être conscient. C'est
d'ailleurs pour cela qu'il encourage « les producteurs à l'utilisation des semences
améliorées, des engrais de qualité et des pesticides homologuées ». En plus, dans le
but d'une meilleure sécurisation des exploitations agricoles face à d'éventuelles
poches de sécheresse, le ministre en charge de l'agriculture invite les producteurs à
la pratique de l'irrigation de complément à partir des bassins de collecte d'eau de
ruissèlement et de toute retenue d'eau. La chenille légionnaire d'automne réapparait,
le gouvernement riposte… Mais il convient de noter que la campagne agricole n'est pas
seulement menacée par la pluviométrie capricieuse. <TIMEX3 tid="t27" type="DATE"
value="PRESENT_REF">Désormais</TIMEX3> et cela depuis <TIMEX3 tid="t26" type="DATE"
value="2017">2017</TIMEX3>, la chenille légionnaire d'automne fait partie des
```

Fig. 2 – *Exemple d'article (Burkina24) annoté par SpaCy et HeidelTime.*

5. Évaluation

5.1. Évaluation qualitative du processus d'identification des documents pertinents

Pour évaluer le processus d'identification des articles pertinents, les premiers articles retournés par notre processus ont été annotés. Pour chacun de

ces articles, les questions suivantes ont été posées à des utilisateurs : *l'article porte-t-il sur la sécurité alimentaire au Burkina Faso ? (Oui tout à fait, Oui partiellement, Non).* À noter que nous définissons un article traitant de la problématique de la sécurité alimentaire les textes mentionnant ou ayant des liens avec les termes de la *famine*, du *climat*, des *prix alimentaires*, de l'*agriculture*. Deux types d'évaluation décrites dans le paragraphe ci-dessous ont été réalisées sur les corpus l'*Observatoire Paalga* et *Burkina24*.

Dans un premier temps, nous avons réalisé une étude par un utilisateur non expert sur les 10 premiers textes (bloc 1), puis une dizaine de textes extraits entre les rangs 11 et 40 (bloc 2) en appliquant notre processus qui repose sur l'application de LDA suivie de *Word2vec* présentés en section 3.

Avec l'*Observatoire Paalga*, sur les 10 premiers textes, 5 sont donnés comme pertinents et 1 comme partiellement pertinent. Notons que 3 des textes non pertinents sont placés aux rangs 8, 9 et 10. Dans le deuxième bloc (11e au 40e document), 8 textes sont non pertinents et 2 partialement pertinents. Ce type de résultat est confirmé avec le corpus *Burkina24* avec 7 articles pertinents (les premiers) et 2 articles partiellement pertinents dans le premier bloc alors qu'il n'y a que 4 articles pertinents ou partiellement pertinents dans le deuxième bloc.

Ces résultats mettent en évidence que les premiers articles retournés par notre approche sont les plus pertinents.

Ensuite une experte en économie sur les problématiques de sécurité alimentaire a évalué de manière approfondie les articles du premier bloc. Les résultats de la première expertise sont confirmés et mettent également en avant la pertinence de notre approche avec respectivement 7 et 10 articles globalement pertinents (c'est-à-dire, *le texte contient des informations complémentaires et d'intérêt sur la sécurité alimentaire*) sur l'*Observatoire Paalga* et *Burkina24*. Cependant en analysant de manière plus approfondie ces articles, seuls 2 (*Observatoire Paalga*) et 6 (*Burkina24*) ont été évalués comme parfaitement adaptés. Ceci peut s'expliquer par le fait que beaucoup d'articles publiés peuvent être des publicités « déguisées » et non des articles de presse liés à l'analyse d'une situation particulière. En effet, les institutions (Ministères, ONG, bailleurs) peuvent organiser des évènements dans le cadre de projets ou de journées thématiques et promouvoir leurs activités dans la presse locale ou nationale. Ceci peut créer une certaine déconnexion avec la situation réelle de sécurité ou insécurité alimentaire même si la thématique de l'événement s'en rapproche.

5.2. Évaluation du processus d'identification des entités spatiales

Pour réaliser une évaluation du processus de reconnaissance des entités spatiales avec *SpaCy* et les ajouts de lexiques résumés en Section 4.2, nous avons annoté manuellement les entités spatiales à partir de 25 articles choisis aléatoirement à partir du site *l'Observateur paalga*. Nous avons ensuite comparé les résultats de l'annotation de *SpaCy* avec les résultats de l'annotation manuelle. Une amélioration de la reconnaissance des entités spatiales en termes de précision, rappel et F-mesure est à relever (cf. Tableau 2). La précision permet de déterminer la proportion d'entités extraites qui sont pertinentes. Le rappel identifie l'exhaustivité de cette extraction, à savoir si l'ensemble des entités attendues a été extrait. Enfin, la F-mesure calcule la moyenne harmonique de la précision et du rappel.

	Précision	**Rappel**	**F-mesure**
Reconnaissance des entités spatiales par SpaCy	0,389	0,663	0,475
Reconnaissance des entités spatiales par SpaCy + ajout dictionnaire + post-traitements	0,461	0,674	0,519

Tab. 2 – *Évaluation du processus d'extraction d'entités spatiales.*

Les 20 entités spatiales les plus fréquentes issues des 50 premiers articles de *l'Observateur paalga*, de *Burkina24* et de *Lefaso.net* sont données dans le Tableau 3.

Observateur Paalga	**Burkina24**	**Lefaso.net**
Ouagadougou 88	Ouagadougou 119	Ouagadougou 91
Ouaga 30	Centre 33	Centre 68
Centre 18	Ouaga 19	Sahel 26
Pouytenga 13	Bobo 17	Bobo-Dioulasso 22
Bobo 12	Nouna 15	Diallo 11
Sahel 12	Bobo-Dioulasso 10	Gaoua 11
Bobo-Dioulasso 10	Sahel 10	Baskuy 10
Banfora 10	Tanghuin 9	Bobo 9
Logobou 9	Centre-Est 9	Koubri 9
Cascades 9	Nord 9	Nord 9
Tanzéongo 8	Bassins 8	Zorgho 7

Observateur Paalga	**Burkina24**	**Lefaso.net**
Koudougou 8	Dédougou 8	Central 7
Diébougou 6	Bagré 7	Toma 6
Nord 6	Bombissiri 6	Dolo 6
Gaoua 4	Centre-Ouest 5	Pô 6
Gourcy 4	Djibo 5	Koudougou 5
Mané 3	Arbinda 5	Ouahigouya 5
Sanmatenga 3	Larlé 5	Yatenga 5
Centre-Nord 3	Bagassi 5	Plateau 5
Diébou 3	Sara 4	Centre-Nord 5

Tab. 3 – *Exemples d'entités spatiales extraites avec leur nombre d'occurrences.*

6. Perspectives de mise en lien

Le corpus étant constitué d'articles annotés en entités spatiales et traitant du thème de la sécurité alimentaire au Burkina Faso, nous pourrons les mettre en relation avec d'autres données, en particulier les données satellitaires mobilisées dans les systèmes d'alerte précoce (cf. Fig. 3). Cette mise en lien constitue une problématique de recherche importante abordée par la communauté des Science des Données (Roche *et al*. 2014) dans un contexte de Recherche d'Information multimodal (Bellot *et al*. 2018).

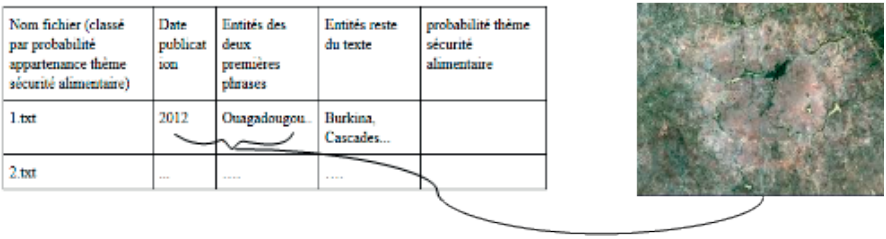

Fig. 3 – *Exemple de mise en lien.*

7. Conclusion

Dans ce travail, nous avons constitué un corpus d'articles pertinents au regard de la problématique de la sécurité alimentaire au Burkina Faso en étant guidé par des ressources lexicales et une approche de fouille de textes. Par ailleurs, un processus de reconnaissance d'informations spatio-temporelles à partir de données textuelles a été utilisé et adapté à notre cas d'étude.

Dans nos futurs travaux, nous souhaitons analyser d'autres sources de données textuelles comme des données issues de réseaux sociaux (Twitter), et les vidéos journalistiques (par exemple, transcriptions issues de YouTube). Enfin, nous souhaitons étudier l'impact des différents éléments des lexiques sur l'approche de fouille de textes (selon leur nombre, leur catégorie morpho-syntaxique, etc.).

Remerciements

Les auteurs remercient Élodie MAITRE D'HOTEL, chercheuse CIRAD, et Maguelonne TEISSEIRE, Directrice de Recherche INRAE pour l'annotation des articles et Agnès BEGUE (co-responsable du DP ISA - Information pour la sécurité alimentaire en Afrique de l'Ouest - du CIRAD) pour ses conseils et recommandations ainsi qu'Hugo DELEGLISE (doctorant CIRAD / #DigitAg) pour son accompagnement sur certains aspects liés à cette étude. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'Avenir portant la référence ANR-16-CONV-0004. Ce travail est également financé par la Région Occitanie et par les Fonds Européens de Développement Régional (FEDER) dans le cadre du projet SONGES - Science des dONnées hétéroGènES.

References

Bellot Patrice, Cappellato Linda, Ferro Nicola, Mothe Josiane, Murtagh Fionn, Nie Jian-Yun, SanJuan Eric, Soulier Laure, Trabelsi Chiraz. 2018. Report on CLEF 2018 : Experimental IR Meets Multilinguality, Multimodality, and Interaction. *SIGIR Forum* 52(2): 72-82.

Interdonato Roberto, Doucet Antoine, Guillaume Jean-Loup. 2018. Unsupervised Crisis Information Extraction from Twitter Data. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM*, 579-580.

Jelodar Hamed, Wang Yongli, Yuan Chi, Feng Xia, Jiang Xiahui, Li Yanchao, Zhao Liang. 2019. Latent Dirichlet allocation (LDA) and topic modeling : models, applications, a survey. *Multimed Tools Appl* 78, 15169–15211.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, Dean Jeffrey. 2013. Distributed Representations of Words and Phrases and Their Compositionality. *In Proceedings of the 26th International Conference on Neural Information Processing Systems* - Volume 2, 3111–3119.

Roche Mathieu, Teisseire Maguelonne, Crémilleux Bruno, Gançarski Pierre, Sallaberry Christian. 2014. ANIMITEX. Analyse d'images fondée sur des informations textuelles. *Ingénierie des Systèmes d'Inf.* 19(3): 163-167

Schmitt Xavier, Kubler Sylvain, Robert Jérémy, Papadakis Mike, Le Traon Yves. 2019. A Replicable Comparison Study of NER Software : StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *Proc. of Sixth International Conference on Social Networks Analysis, Management and Security - SNAMS* : 338-343

Tahora H. Nazer, Guoliang Xue, Yusheng Ji, Huan Liu. 2017. Intelligent Disaster Response via Social Media Analysis A Survey. *SIGKDD Explorations* 19(1): 46-59.

Abstract

Food security early warning systems are set up to inform stakeholders in order to adapt and improve agricultural development. These systems mainly analyze satellite images (e.g. images of agricultural fields) and quantitative data (e.g. food market prices). However, few textual data are integrated into their processing, yet they are increasingly abundant and accessible on the Internet. They do not perfectly describe the geographical situation, but they could provide additional information or observations to satellite images. This paper proposes to acquire a corpus of articles from newspapers sites related to the theme of food security in Burkina Faso and to extract the spatio-temporal information contained in the corpus.