SDHn 2018 atelier Sciences des Données et Humanités Numériques

Nicolas Turenne

SDHn 2018

atelier

Sciences des Données et Humanités Numériques

23 janvier 2018

Maison des Sciences de l'Homme de Paris Nord

Comité d'organisation

Nicolas	**Turenne**	UMR	LISIS	UPEM-INRA-CNRS	Paris

Comité de lecture

Frédéric	**Amblard**	IRIT	-	UT1	Toulouse
Bruno	**Bachimont**	Costech		UTC	Compiègne
Delphine	**Battistelli**	MODYCO		Paris 10	Paris
Olivier	**Baude**	Université		d'Orléans	Orléans
Serge	**Bauin**	CNRS			Paris
Patrice	**Bellot**	LSIS	Université Aix-Marseille		Marseille
Charles	**Bouveyron**	Université de Nice			Nice
Davide	**Buscaldi**	LIPN	Paris	13	Paris
Vincent	**Claveau**	IRISA			Rennes
Mickaël	**Coustaty**	L3i			La Rochelle
Benoit	**Crabbe**	UFRL	Paris	7	Paris
Béatrice	**Daille**	LS2N			Nantes
Liana	**Ermakova**	Université de Bretagne	Occidentale		Brest
Olivier	**Ferret**	CEA	LIST		Paris
Serge	**Fleury**	Université Sorbonne Nouvelle - Paris		3	Paris
Claire	**François**	INIST			Nancy
Jean-Gabriel	**Ganascia**	LIP6	Paris	6	Paris
Natalia	**Grabar**	STL	Université de Lille		Lille
Serge	**Heiden**	UMR	IHRIM	ENS de Lyon	Lyon
Agata	**Jackiewicz**	praxiling			Montpellier
Marie-Paule	**Jacques**	Lidilem			Grenoble
Jean-Charles	**Lamirel**	LORIA			Nancy
Thomas	**Lebarbé**	UMR	Litt&Arts		Grenoble
Jean-Marc	**Leblanc**	CEDITEC	UPEC		Paris
Jean-Philippe	**Mague**	ENS			Lyon
Denis	**Maurel**	Université Francois Rabelais			Tours
Francesca	**Musiani**	ISCC	CNRS/Paris-Sorbonne	UPMC	Paris
Thierry	**Poibeau**	UMR	LATTICE ENS		Paris
Pascal	**Poncelet**	LIRMM	INRIA		Montpellier
Céline	**Poudat**	UMR	BCL - Université de Nice		Nice
Violaine	**Prince**	LIRMM			Montpellier
Pierre	**Ratinaud**	LERASS,Toulouse 2			Toulouse
Camille	**Roth**	Sciences Po			Paris
Francis	**Rousseaux**	IRCAM			Rheims
Benoît	**Sagot**	ALMAnaCH	INRIA		Paris
Xavier-Laurent	**Salvador**	LDI	Paris	13	Paris
Amalia	**Todirascu**	LILPA,	Université de Strasbourg		Strasbourg
Katerina	**Tzompanaki**	Université de Cergy-Pontoise			Paris
Mathieu	**Valette**	ERTIM	INALCO		Paris
Julien	**Velcin**	ERIC	Lab,	Lyon 2	Lyon
Marie-Luce	**Viaud**	VIM	Institut National de l'Audiovisuel		Paris
Serena	**Villata**	SPARKS-WIMMICS			Nice

Avec le soutien du :

Et du :

Comparaison de deux critères de rupture thématique pour l'extraction d'indices de segmentation et de liage dans une grande collection de textes

Yves Bestgen*

*Chercheur qualifié du FNRS, UCL, 10 Place Mercier, 1348 Louvain-la-Neuve Belgique
yves.bestgen@uclouvain.be

Résumé. Une technique automatique est proposée et évaluée pour extraire de grands corpus de textes des indices de rupture et de continuité thématique, tels que les expressions adverbiales, les connecteurs et les anaphores, qui retiennent l'attention de chercheurs tant en linguistique qu'en psychologie cognitive et en TAL, mais qui sont le plus souvent étudiés par l'analyse qualitative d'un très petit nombre de textes.

1 Introduction et approche proposée

Depuis de nombreuses années, la linguistique textuelle s'intéresse aux indices de rupture et de continuité thématique, tels que les expressions adverbiales, les connecteurs et les anaphores (Adam, 2005 ; Charolles, 1995 ; Bestgen & Piérard, 2014). Ces indices retiennent également l'attention de chercheurs en psychologie cognitive et en traitement automatique du langage parce qu'ils sont susceptibles de faciliter la compréhension d'un texte et d'améliorer les performances d'algorithmes de segmentation automatique. En linguistique textuelle, ces marques sont étudiées par l'analyse qualitative approfondie d'un très petit nombre de textes, ce qui rend difficile toute comparaison entre des genres de textes. De plus, nombre d'indices linguistiques de la structure ont un taux d'occurrence très faible, imposant l'analyse de grands corpus si on cherche à en dresser un tableau représentatif.

L'objectif de la présente étude est d'évaluer une approche basée sur des techniques du TAL qui permet l'extraction de ces indices dans de grandes quantités de textes. L'approche, qui s'inspire des travaux sur la détection de paragraphes (Sporleder & Lapata, 2006), repose sur la construction, par d'une procédure d'apprentissage supervisé, de modèles prédictifs capables de discriminer dans des textes des paires de phrases contigües qui sont ou non séparées par une rupture thématique. Nous avons employé une machine à vecteurs de support (SVM, version linéaire) bien connue pour son efficacité en catégorisation de textes. Les indices potentiels analysés sont composés des unigrammes, bigrammes et trigrammes de lemmes et d'étiquettes morphosyntaxiques présents dans les phrases. Lors de leur extraction, les n-grammes en début de phrase sont distingués des autres parce que les indices de rupture occupent fréquemment cette position (Bestgen & Piérard, 2014). L'hypothèse sous-jacente à cette approche est que les indices les plus efficaces pour discriminer les phrases en situation de rupture de celles en situation de continuité seront des candidats à la fonction d'indices de segmentation et de liage.

Pour identifier les phrases en situation de continuité ou de discontinuité thématiques, deux critères ont été comparés (Piérard & Bestgen, 2006). Le premier est dérivé d'une mesure de cohésion lexicale issue de l'analyse sémantique latente (ASL, Landauer et al., 2007). Il s'agit d'une technique d'analyse statistique de corpus dont l'objectif est d'inférer les similarités sémantiques entre des mots sur la base de leurs occurrences dans des documents. Le sens de chaque mot ou de chaque phrase est représenté par un vecteur dans l'espace sémantique. Pour calculer la similarité sémantique entre deux phrases, on calcule le cosinus entre les vecteurs qui les représentent. Plusieurs autres approches pour extraire des espaces sémantiques ont été proposées (Levy et al., 2015). L'intérêt de l'ASL est que son efficacité pour estimer la cohérence dans des textes est considérée comme bien établie (Landauer et al., 2007, p. 10). Le second critère est plus simple puisqu'il s'appuie directement sur la structure du texte telle qu'elle est rendue visible par les sections (le niveau 1 dans Wikipédia pour la présente étude).

2 Évaluation

L'objectif de cette étude est d'évaluer l'intérêt de l'approche proposée en répondant aux deux questions suivantes : quelle est son efficacité pour discriminer les phrases en situation de rupture de celles en situation de continuité et les indices les plus utiles pour la SVM sont-ils compatibles avec les conclusions des analyses linguistiques plus qualitatives ?

2.1 Méthode

Les analyses ont porté sur des phrases extraites de l'encyclopédie Wikipédia francophone au moyen de *wikiextractor* de Attardi. Les textes ont été lemmatisés et étiquetés morphosyntaxiquement par le TreeTagger (Schmid, 1994). Les phrases utilisées ont été sélectionnées après suppression des documents très courts et très longs par rapport à la longueur moyenne ainsi que de ceux contenant un grand nombre de phrases très brèves ou un nombre disproportionné d'étiquettes morphosyntaxiques des catégories *abréviation*, *nom propre*, *nombre* et *signe de ponctuation*. L'espace sémantique employé pour calculer les cosinus a été extrait de cette collection de textes sur la base d'une segmentation arbitraire en unité de 250 mots.

Pour l'ASL, on a calculé le cosinus entre les deux phrases de toutes les paires de phrases contigües d'au moins 7 mots et d'au plus 43 mots. Ont été considérés en situation de rupture les 2,5% de paires de phrases dont le cosinus était le plus petit. Les paires de phrases en situation de continuité sont celles dont le cosinus fait partie des 2,5% les plus élevés. Les n-grammes servant de traits pour l'apprentissage supervisé ont été extraits de la deuxième phrase de chaque paire. Au total, il y a 35 629 phrases de chaque type.

Pour le critère basé sur les sections, on a analysé les quadruplets de phrases contigües répondant aux mêmes critères de longueur. Pour être considérée en situation de rupture, la troisième phrase de ces quadruplets devait être précédée par une rupture de section principale et il ne pouvait y avoir d'alinéas entre les autres phrases du quadruplet. La troisième phrase d'un quadruplet était considérée en situation de continuité s'il n'y a avait aucune rupture de section, ni aucun alinéa entre les quatre phrases. On a extrait le maximum possible de phrases en situation de rupture (N = 19 216) et un échantillon aléatoire d'une même taille de phrases en situation de continuité. Les n-grammes pour la SVM ont été extraits comme pour l'ASL.

Pour chaque analyse, les phrases cibles ont été divisées en 80% pour l'apprentissage et 20% pour l'évaluation. Deux paramètres ont été optimisés sur 25% des données d'apprentissage : le seuil de fréquence minimal des traits dans le matériel analysé et le paramètre C de régularisation. Les 200 indices les plus utiles pour la SVM (Chang & Lin, 2008) ont été sélectionnés pour être analysés qualitativement.

2.2 Résultats

La première analyse indique que les deux critères de continuité thématique sont liés puisque l'indice de continuité dérivé de l'ASL est plus faible lorsque deux phrases contigües sont séparées par un changement de section (moy. = 0.15) que lorsqu'elles ne le sont pas (moy. = 0.24) (test *t* pour comparaison de moyennes, $p < 0.0001$). Comme le montre le tableau 1, les performances [1] obtenues par la procédure d'apprentissage supervisée sur la base de chacun des deux critères sont relativement élevées. Discriminer les ruptures obtenues par l'ASL est plus simple que discriminer celles basées sur les sections. Le tableau 1 indique aussi les performances obtenues lorsque la SVM ne peut employer que les n-grammes en début de phrase ou que les autres n-grammes. La SVM basée sur l'ASL bénéficie nettement moins des n-grammes en début de phrase que la SVM basée sur les sections alors que les études qualitatives en linguistique soulignent l'importance de la position initiale (Adam, 2005).

Le tableau 2 présente le pourcentage d'indices parmi les 200 les plus utiles qui sont en position initiale, des lemmes et des unigrammes. On observe que les indices sélectionnés sur la base des deux critères sont très différents : l'ASL privilégie nettement des unigrammes de lemmes qui ne commencent pas les phrases. Il s'agit de lemmes comme *album, cheval, langue, Disney, mètre, vin, espèce* et *film* pour signaler une rupture et comme *jeune, peu, politique, siècle* et *aller* pour signaler une continuité. Ces indices ne semblent pas interprétables dans le cadre des travaux linguistiques sur les marques de segmentation et de liage (Adam, 2008), contrairement à ceux mis en évidence lorsqu'il s'agit de prédire la présence ou non d'un changement de section dont les plus utiles sont les pronoms personnels, les déterminants démonstratifs et des connecteurs comme *mais*, *ainsi* et *cependant* pour la continuité et une suite de deux noms propres, *naître*, *après_le* et *depuis_le* pour la discontinuité, à chaque fois en début de phrase.

	Nbr. d'exemplaires	Tout	Début	Suite
ASL	71258	78%	62%	77%
Section	38432	74%	67%	70%

TAB. 1 – *Efficacité (exactitude) pour distinguer les ruptures des continuités*

	Début	Lemme	Unigramme
ASL	10%	93%	86%
Section	40%	38%	31%

TAB. 2 – *Indices les plus utiles selon le type*

1. L'exactitude est employée comme mesure de performance parce que, dans toutes les analyses effectuées, elle est égale au rappel et à la précision et donc aussi à la mesure F1.

3 Conclusion

Les résultats obtenus sur la base des sections sont relativement élevés et les indices sont interprétables. Cette approche semble être une voie prometteuse pour étudier les indices de continuité et de liage dans de grands corpus et il serait intéressant d'appliquer la technique aux sous-sections et aux paragraphes, ainsi qu'à des corpus journalistiques et littéraires. Si le critère structurel dérivé de l'ASL donne lieu à une meilleure performance, les indices les plus utiles sont nettement moins interprétables. Ce résultat questionne l'emploi de l'ASL pour déterminer la cohérence de textes, une procédure largement popularisée par des outils libres d'accès comme Coh-Metrix. Tirer une conclusion définitive de la présente étude serait toutefois prématuré puisqu'on n'a pas procédé à une analyse approfondie des phrases considérées par l'ASL comme en situation de rupture ou de continuité. Il serait aussi intéressant d'évaluer des procédures d'extraction d'espaces sémantiques proposées récemment (Levy et al., 2015).

Références

Adam, J.-M. (2008). *La linguistique textuelle.* Paris : A. Colin.

Bestgen, Y. et S. Piérard (2014). Sentence-initial adverbials and text comprehension. In L. Sarda, S. Carter Thomas, B. Fagard, et M. Charolles (Eds.), *Adverbials in Use : From Predicative to Discourse Functions*, pp. 151–168. PUL.

Chang, Y.-W. et C.-J. Lin (2008). Feature ranking using linear SVM. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, pp. 53–64.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique : Revue Internationale de Linguistique Française 29*, 125–151.

Landauer, T., D. McNamara, D. Simon, et W. Kintsch (2007). *Handbook of Latent Semantic Analysis*. Mahwah: Erlbaum.

Levy, O., Y. Goldberg, et I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *TACL 3*, 211–225.

Piérard, S. et Y. Bestgen (2006). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL 47*(2), 89–110.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sporleder, C. et M. Lapata (2006). Broad coverage paragraph segmentation across languages and domains. *ACM Transactions on Speech and Language Processing 3*, 1–35.

Summary

An automatic technique is proposed and evaluated to extract from large corpora markers of thematic continuity and discontinuity, such as adverbials, connectives and anaphoras, which attract the attention of researchers in linguistics, cognitive psychology and natural language processing, but are often studied by the qualitative analysis of a very small number of texts.

Analyse de données prosopographiques. Application aux officiers angevins (XIIIe-XVe siècles)

Anne Tchounikine*, Maryvonne Miquel*
Thierry Pécout**
Jean-Luc Bonnaud***

*LIRIS-CNRS UMR 5205, INSA-Université de Lyon, Lyon
prenom.nom@insa-lyon.fr
**UMR LEM-CERCOR, Université Jean Monnet, Saint Etienne
thierry.pecout@univ-st-etienne.frl
***Université de Moncton
jean-luc.bonnaud@umoncton.ca

Résumé. Le projet Europange, qui réunit informaticiens et médiévistes, a pour ambition d'étudier la constitution du corps d'administrateurs des territoires placés sous domination angevine aux XIIIe-XVe siècles à partir du parcours individuel de chaque officier, de sa carrière, ses réseaux, sa formation. Dans cet article, nous décrivons les méthodes et outils conçus pour l'analyse de ces données prosopographiques. Ces analyses incluent notamment des analyses OLAP et des analyses de réseaux associées à des outils de visualisation cartographiques et chronologiques.

1 Contexte

Le travail que nous présentons a été effectué dans le cadre du programme EUROPANGE financé par l'Agence Nationale de la Recherche et du contrat quinquennal de l'École Française de Rome. Il émane d'une équipe associant informaticiens et médiévistes s'interrogeant sur les rythmes et les méthodes d'élaboration des communautés politiques [1]. Les XIIIe - XVe siècles constituent un moment privilégié pour observer ces phénomènes, car ils voient la mise en place des organismes, des discours, des méthodes et des corps administratifs qui assurent aux états nationaux et princiers leur premier développement. Dans le cadre de ce projet, nous proposons d'en examiner un aspect significatif, à travers l'émergence d'un milieu et d'une société politique, en questionnant la constitution d'un corps d'administrateurs, les officiers, avec leurs réseaux, leur formation, leurs compétences, dans l'ensemble des espaces politiques des terres angevines : Anjou, Maine, Provence, Lorraine, Italie du Sud et Sicile, Piémont, Lombardie et

1. UMR 8584 LEM-CERCOR de l'Université de Saint-Étienne, École Française de Rome, Università degli Studi della Campania «Luigi Vanvitelli», Università degli Studi di Bergamo, Università del Salento, Università degli Studi di Salerno, Centre de Recherches en Sciences Humaines de l'Académie des Sciences de Hongrie, Université de Moncton (Canada), CERHIO FRE CNRS de l'Université d'Angers, UMR 7303 TELEMME de l'Université d'Aix-Marseille, EA 4583 CEMM de l'Université de Nîmes, UMR 5205 LIRIS (Laboratoire d'Informatique en Image et Systèmes d'Information) de l'Université de Lyon-INSA Lyon).

Toscane, Hongrie, Pologne, Morée, Balkans. Cet espace disjoint dans l'espace et discontinu dans la durée interroge les capacités des autorités politiques à rassembler et à administrer, à susciter un discours politique commun, ce qui revêt un sens particulier face à nos réflexions actuelles sur la construction européenne.

Reconsidérer l'individu comme acteur de l'histoire est une voie d'investigation pour les historiens depuis une vingtaine d'années comme en attestent des projets passés ou en cours (Heloise; Parisiense). Selon Pierre-Marie Delpu (Delpu (2015)), « *Une prosopographie pourrait être définie, a minima, comme une étude collective qui cherche à dégager les caractères communs d'un groupe d'acteurs historiques en se fondant sur l'observation systématique de leurs vies et de leurs parcours*.» Cette démarche, associée à des méthodes quantitatives et statistiques et des outils de visualisation avancée, a montré que la prosopographie est la méthode la plus appropriée pour cerner les contours sociologiques d'un groupe déterminé à travers les trajectoires individuelles de ses membres (Zalc et Lemercier (2008); Messai et Devogele (2014); Bouveyron et al. (2014)). À la lumière de cette technique d'analyse, qui consiste à tenter de dessiner le portrait le plus fin possible d'un groupe d'officiers, nous tentons d'en reconstituer leurs carrières, les réseaux qu'ils partagent et les éventuelles stratégies professionnelles qui ont été les leurs.

2 Contributions

Les informations à recueillir et analyser sont tous les éléments biographiques concernant les officiers et leur entourage issus du dépouillement de sources archivistiques et iconographiques (registres de la chancellerie angevine, actes notariés, sources épigraphiques...). Notre contribution s'articule autour de trois éléments :

— une application collaborative de saisie et de restitution des informations sous la forme de « pages prosopographiques » constituées de l'ensemble consolidé, mis en forme, et enrichi des données saisies (Tchounikine et al. (2016)),
— une application d'analyse pour la constitution et l'étude de populations multicritères,
— une base de données prosopographiques utilisée dans les deux applications.

L'ensemble des ces éléments constituent la suite logicielle « Prosopange », qui est actuellement déployée sur le TGIR Huma-Num. Cette suite logicielle (voir Figure 1) est une application client-serveur 3 tiers. Le tiers de données est constitué par une base de données permettant de structurer l'ensemble des informations caractérisant un officier. Ces informations sont généralement directement saisies par les utilisateurs après dépouillement et analyse de sources bibliographiques, mais on y trouve aussi des informations qui sont calculées automatiquement à partir des données saisies. Le tiers de traitement héberge les algorithmes de création ou de recherche des données dans la base, les différents calculs et contrôles, les algorithmes d'analyse. Enfin, le tiers client comprend l'interface Web utilisateur avec les outils pour la saisie, le requêtage et la visualisation des données.

Dans cet article, nous nous intéresserons plus particulièrement à l'application d'analyse des données prosopographiques. Cette application d'analyse inclut des analyses multidimensionnelles et des analyses de réseaux associées à des représentations cartographiques et temporelles des informations. Ces analyses sont menées sur des populations d'officiers préalablement sélectionnées par un filtre multi-critères portant sur la période historique, les territoires concernés, et/ou des caractéristiques des individus (confession, genre...).

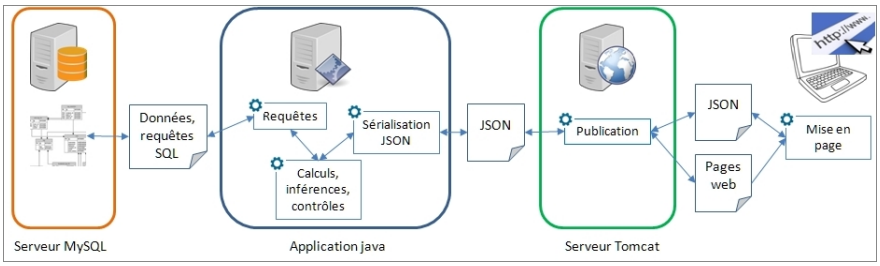


FIG. 1 – *Architecture de la suite Prosopange.*

3 Cube et analyses OLAP

Leonardus Afflicto de Scallis, originaire del Giustizierato di Terra di Lavoro Citra (Royaume de Sicile), titulaire du titre de « professor juris civilis » depuis le 24/09/1372 a été nommé par Johanna I en 1373 en tant que « judex major et secundarum appelationum » (charge de type « judex ») dans les Comtés de Provence et de Forcalquier. Il est dénommé « nobilis » et « vir » dans les sources postérieures à 1378 : voilà le type d'informations dont nous disposons dans la base de données.

Il est apparu très vite que les chercheurs historiens partenaires du projet avaient des attendus très différents en termes d'analyse de ces données. Cette diversité est d'abord le reflet de la diversité des thématiques de recherche : certains s'intéressent à l'étude de la circulation des officiers, leur zone d'action et d'influence, d'autres étudient l'organisation politique des territoires, les jeux de nomination et des lieux d'affectation, ou bien encore la sociologie des corps constitués, les lieux d'origine des officiers, leur entourage et les prédicats ou les titres universitaires qui les qualifient. Les méthodes et le choix des populations analysées sont elles aussi diverses : analyses quantitatives de masse, analyses transverses dans le temps ou dans l'espace, ou analyses d'un groupe identifié d'officiers, un clan, un corps de métier, voire d'un unique individu. Pour satisfaire l'ensemble de ces exigences, et rester ouvert à d'autres à venir, il était indispensable de proposer une interface flexible, offrant au chercheur la possibilité de construire dynamiquement son analyse spécifique. Les méthodes et outils OLAP (On Line Analytical Processing), qui permettent l'analyse interactive de données multidimensionnelles selon différents axes d'analyse et à des grains différents, fournissent des solutions intéressantes. Leur utilisation suppose d'avoir au préalable défini les différents axes susceptibles d'être utilisés dans l'élaboration d'une analyse. Pour les historiens, le défi a été d'identifier, de définir et de créer ces différents axes d'analyse. Pour les informaticiens, la difficulté a été d'adapter les concepts et opérateurs OLAP aux nombreuses singularités des données prosopographiques.

Les analyses OLAP qui nous intéressent sont centrées sur un fait unique : les officiers. Il s'agit de dénombrer, et d'identifier (i.e. de lister) des officiers selon différents axes d'analyse. Ces axes d'analyse peuvent qualifier l'officier, sa ou ses charges, son ou ses titres (titres universitaires, prédicats,...). Ces dimensions présentent de nombreuses irrégularités. L'association entre fait et dimension peut être multi-valuée (e.g. un officier peut exercer plusieurs charges, en même temps ou successivement), il en est de même entre niveaux de la hiérarchie de dimension (e.g. plusieurs lieux d'affectation pour une charge). Les grains renseignés dans la base de données sont variables, l'information recueillie dans les sources manuscrites

étant souvent incomplète ou imprécise. De plus, plusieurs dimensions sont non orthogonales, par exemple la dimension lieu d'affectation dépend de la dimension charge, et non équilibrées (voir Figure 2, un extrait de la dimension « *type de charge* » dans le tableau croisé). Il est donc difficile de construire un schéma multidimensionnel et un cube a priori, comme on le ferait dans une application décisionnelle classique. L'usage d'un moteur OLAP traditionnel est donc exclu.

Dans le modèle multidimensionnel, nous définissons 3 catégories de dimension :

— Les dimensions liées aux individus : lieu d'origine, genre, lieu de séjour, confession, dates...
— Les dimensions liées aux charges exercées par ces individus : type de charge, lieu d'exercice, d'affectation, nomination, dates d'exercice...
— Les dimensions liées aux qualificatifs de ces individus : prédicats, titres universitaires, dates d'obtention des qualificatifs....

Pour chacune de ces dimensions, les hiérarchies sont prédéfinies par les historiens, les membres sont construits au fur et à mesure des insertions dans la base de données. Par exemple, pour les lieux d'origine, lieux de séjour, lieux d'exercice et d'affectation, le schéma hiérarchique défini est : « *localités* » < « *subdivisions* » < « *espaces politiques* » < « *territoires* » < « *all* » ; pour les types de charges, une hiérarchie arborescente non équilibrée est utilisée. Le choix des dimensions et la constitution de ces nomenclatures ont été effectués en fonction des besoins exprimés par les historiens et de la nécessité pour les informaticiens de disposer d'un référentiel pour des analyses comparatives. Ainsi, pour les dimensions spatiales, l'organisation en hiérarchie découle d'un consensus sur un découpage territorial. Cependant, le choix d'intégrer des membres selon les besoins permet de prendre en compte de nouveaux lieux lors de la saisie des officiers et d'enrichir ainsi la hiérarchie.

Pour l'analyse OLAP, l'utilisateur choisit dynamiquement dans ces 3 catégories les dimensions qu'il souhaite faire figurer dans son tableau multidimensionnel, en croisant ou non les catégories : par exemple (« lieu d'origine » × « confession ») ou (« lieu d'origine » × « confession » × « types de charge »). L'hypercube est alors calculé à la volée côté serveur avec les données de la base. L'algorithme (voir Algorithme 1) prend comme entrée chaque officier. Pour chacun, on transforme les informations de cet officier en un point dans l'espace multidimensionnel cible et on met à jour l'ensemble des points agrégats.

La stratégie mise en oeuvre dans l'algorithme 1 nous permet de gérer correctement les données connues ou saisies à des grains différents. Par exemple, un officier dont le lieu d'origine n'est connu qu'au niveau subdivision sera bien comptabilisé dans les agrégats « *subdivisions* » et agrégats de niveaux supérieurs (« *espaces politiques* », « *territoires* », « *all* »). A tous les niveaux plus détaillés de cette dimension, cette imprécision se traduira par l'ajout dynamique d'un membre « non renseigné » afin d'assurer un décompte correct. Les données multivaluées sont elles aussi prises en compte, un officier ayant exercé des charges de type *« judex »* et *« notarius »* apparaitra à la fois dans l'agrégat *« Justice »* et l'agrégat *« Rédaction »*, et dans les agrégats de niveaux supérieurs (« *Ecrit* », « *Charge centrale* », etc.). Contrairement à la démarche classique OLAP, ici l'espace multidimensionnel utilisé est construit dynamiquement au fur et à mesure du balayage des données associées aux officiers. L'hypercube est ainsi entièrement matérialisé, l'ensemble des agrégats est calculé, puis envoyé au client. Ce choix de la matérialisation complète du cube et de balayage de la table officier est rendu possible par le faible volume des données (6000~ officiers) et permet d'assurer de bonnes performances lors

de la navigation OLAP.

Côté client, nous avons implémenté l'ensemble des opérateurs OLAP classiques (Gray et al. (1996); Rafanelli (2003)) :

— *sort* pour trier les membres des dimensions (i.e. les entêtes de lignes ou de colonnes dans le tableau),
— *drill-down* pour passer à un niveau plus fin de la hiérarchie (i.e. déployer les descendants d'un entête de ligne ou de colonne du tableau),
— *roll-up* pour passer à un grain plus grossier de la hiérarchie (i.e. agréger les descendants d'un entête de ligne ou de colonne du tableau),
— *slice* pour supprimer un membre du cube (i.e. supprimer un entête et ses descendants dans les lignes ou les colonnes du tableau),
— *drill-through* pour effecteur un drill-down suivi d'un slice sur un entête de ligne ou de colonne.

Les résultats sont fournis sous forme de tableaux, graphiques, et cartes choroplèthes. Les figures 2 et 3 montrent le résultat de l'analyse portant sur le nombre d'officiers par type de charge et personne ou type de personne ayant procédé à leur nomination. La hiérarchie des charges est ici déployée jusqu'au niveau détaillé pour le membre « *Commandement général* » et la hiérarchie des type de personne est déployée pour le type « *Souverain* ». Le clic sur une cellule du tableau surligne la cellule en vert et permet d'obtenir la liste des officiers concernés et, pour chacun d'eux, leur page prosopographique.

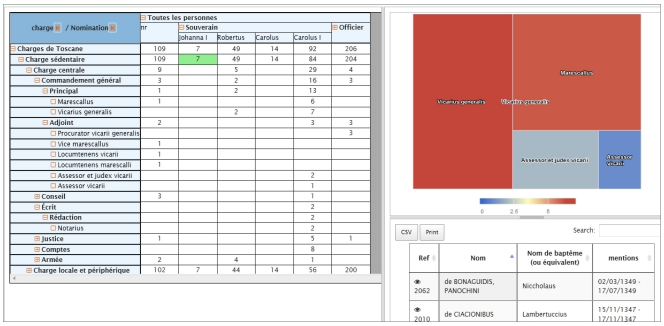


FIG. 2 – *Analyse OLAP du cube « type de charge » × « nomination ».*

4 Analyse de réseaux

Un des objectifs du projet Europange est de mettre en évidence les liens inter-personnels (familiaux, professionnels, amicaux...). Ces informations, décrivant les relations liant un officier à d'autres individus de la base, officiers, souverains ou autres personnes, sont représentées par des n-uplets (officier source du lien, individu cible du lien, type de lien, date du lien) par exemple (*officier Johannes Arlatani*, *souverain Ludovicus III*, *familiaris*, *06/07/1422*). Les liens sont typés et orientés, et éventuellement associés à un lien réciproque. Lors de la saisie, le type du lien est sélectionné au sein d'une hiérarchie arborescente non équilibrée (voir Figure 3, cadre de gauche).

charge / Nomination	Toutes les personnes					
	nr	Souverain				Officier
		Johanna I	Robertus	Carolus	Carolus I	
Charges de Toscane	109	7	49	14	92	206
Charge sédentaire	109	7	49	14	84	204
Charge centrale	9		5		29	4
Commandement général	3		2		16	3
Principal	1		2		13	
Marescallus	1				6	
Vicarius generalis			2		7	
Adjoint	2				3	3
Procurator vicarii generalis						3
Vice marescallus	1					
Locumtenens vicarii	1					
Locumtenens marescalli	1					
Assessor et judex vicarii					2	
Assessor vicarii					1	
Conseil	3				1	
Écrit					2	
Rédaction					2	
Notarius					2	
Justice	1				5	1
Comptes					8	
Armée	2		4		1	
Charge locale et périphérique	102	7	44	14	56	200

FIG. 3 – *Détail du tableau croisé de la figure 2.*

A partir de ces informations, nous construisons 3 types de graphes :

— Un graphe inter-personnel dont les sommets sont les individus, et dont les arcs sont les liens saisis étiquetés par les dates de première et dernière mention du lien et pondérés par la durée connue du lien.

— Un graphe officier-charge dont les sommets sont les individus et les types de charges, et dont les arêtes symbolisent l'exercice de la charge, étiquetées par les dates d'affectation et pondérés par la durée connue de l'exercice.

— Un graphe de coïncidence de séjour ou de coïncidence d'affectation dont les sommets sont les officiers, et dont les arêtes relient les officiers ayant séjourné ou ayant été affectés dans le même lieu aux mêmes dates et sont pondérées par la durée connue de coïncidence (voir Algorithme 2).

Ces différents graphes, calculés côté serveur, peuvent être visualisés côté client sous la forme de réseaux ou de matrice de co-occurrence. Des outils de personnalisation permettent aux historiens de mener des analyses structurales et d'interpréter et de questionner ces graphes. Ainsi, l'utilisateur peut sélectionner les types de lien à faire apparaitre ou colorer dans le graphe, il peut également faire évoluer le graphe dynamiquement sur une période de temps choisie ou rechercher et surligner des sommets en fonction d'un label.

5 Analyse égocentrée

L'objectif de cette proposition d'analyse est d'enrichir les outils traditionnels de la méthode prosopographique qui ne donnent le plus souvent qu'une vision descriptive des carrières et des liens sans expliquer véritablement leur formation. L'étude de réseaux égocentrés permet de comparer le capital relationnel de plusieurs individus. Elle nécessite des changements d'échelle

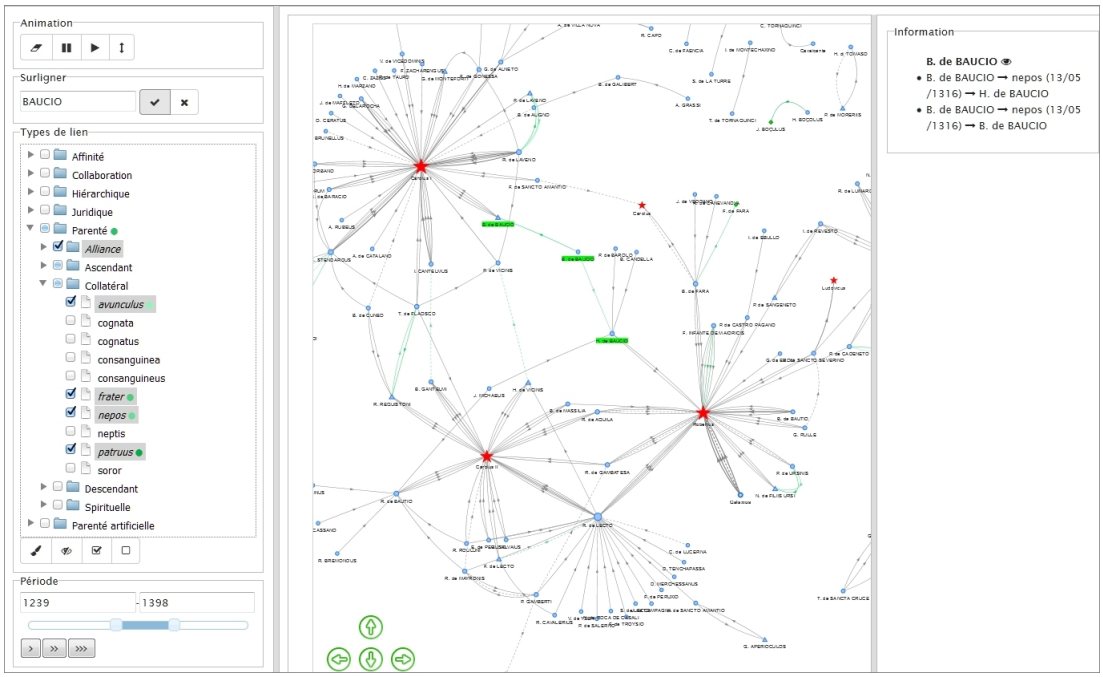


FIG. 4 – *Visualisation d'un réseau inter-personnel avec coloration des liens familiaux et paramétrage de la période.*

(passer de l'étude du groupe en son entier à l'étude de cas spécifiques) et impose une vision dynamique. En effet, une vision statique des liens entre individus est artificielle puisqu'elle ne tient pas compte de ce qui est le matériau de l'historien : le temps. Dans l'optique d'une étude des officiers, un des enjeux principaux est de s'interroger sur la constitution d'un milieu d'officiers homogène et comment ce milieu se construit. Un des angles d'approches possibles est d'étudier le processus d'intégration des nouveaux arrivants à leur nouveau milieu. Parmi les informations à explorer et à mettre en regard, on trouve bien sûr les mouvements migratoires (origine des officiers, lieux de séjour et d'exercice des charges) ; les mariages, qui sont un marqueur de l'intégration à la société d'accueil, et qui influencent parfois directement les attributions de titres nobiliaires et les carrières ; la construction progressive de réseau local professionnel, le jeu des nominations et des promotions ; les politiques individuelles ou familiales d'achats et de vente de biens, les dons aux lieux de cultes locaux et les lieux de sépultures choisis permettent eux aussi de mesurer l'intégration des immigrants.

Pour mener à bien cette analyse, nous construisons récursivement un graphe centré sur l'individu central (appelé « égo ») agrégeant les graphes inter-personnels et les graphes de coïncidences de l'égo et des officiers qui lui sont connectés. Le nombre d'itérations est un paramètre fixé par l'utilisateur. A ce graphe, nous associons une cartographie des officiers (lieux d'origine, séjour, charges) présents dans le graphe, et la chronologie de l'égo composée des différentes dates qui lui sont associées (dates des différentes charges, dates des titres universitaires, dates des liens etc.). Des outils d'animation temporelle synchronisée sur les différents

volets (réseau, cartographique et chronologie) permettent de mettre en évidence la dynamique des constitutions des réseaux et des trajectoires professionnelles (voir Figure 4).

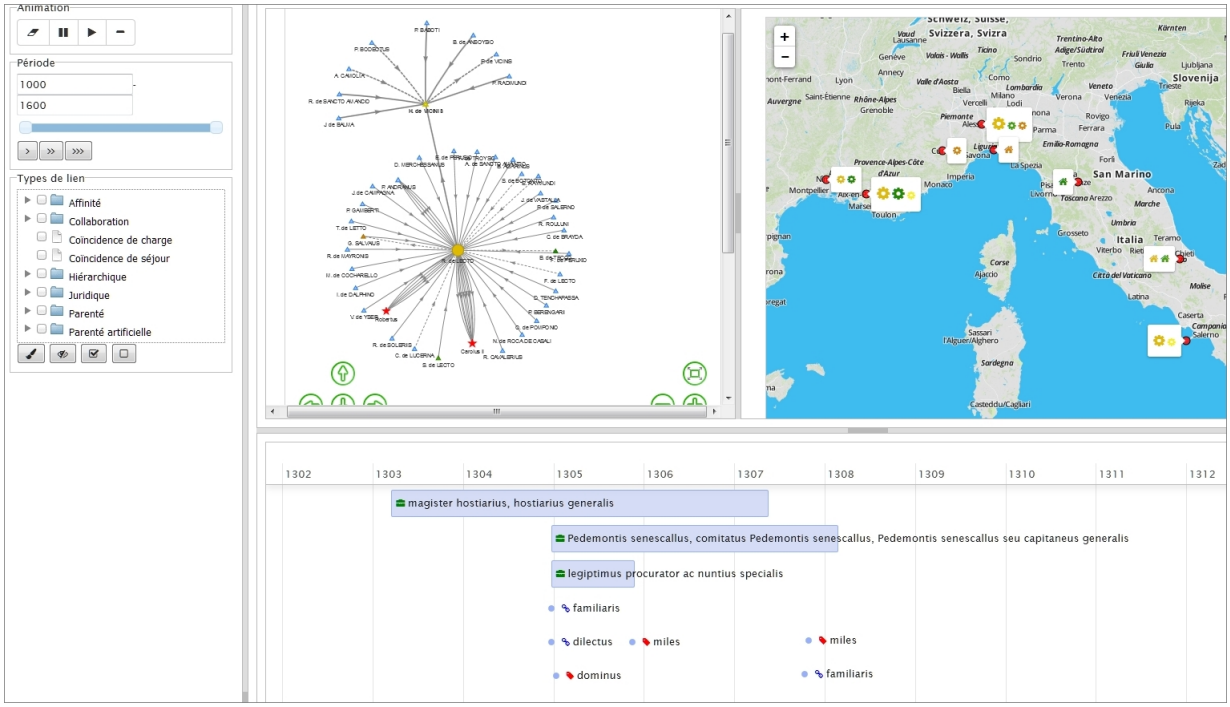


FIG. 5 – *Visualisation d'un réseau égocentré avec cartographie des officiers sélectionnés dans le graphe et frise chronologique de l'égo.*

6 Conclusions et perspectives

La suite Prosopange et sa base de données sont actuellement en production sur le site des humanités numériques Huma-Num. Une cinquantaine de médiévistes (français, italiens, hongrois, polonais, allemands,...) l'alimente quotidiennement en données recueillies grâce au dépouillement de sources archivistiques manuscrites (actuellement plus de 6000 officiers couvrant une période allant de 1210 à 1539). La base de données est déjà aujourd'hui la plus grosse base dans le domaine en nombre d'officiers et en nombre d'informations par officier, elle est utilisée pour les études bibliographiques des chercheurs, doctorants et post-doctorants qui en font la demande. La définition des différentes nomenclatures composant les dimensions des charges, des liens et des statuts, a exigé un effort considérable pour mettre en commun, nommer, synthétiser, et trouver un consensus englobant les termes utilisés, leur catégorisation et leur hiérarchisation. Ce travail est totalement inédit et constitue un résultat en soi. La nomenclature des lieux, quant à elle, pose les bases d'une cartographie administrative. L'utilisation de la base et des outils d'analyse met en évidence l'intérêt d'une approche collaborative : les

historiens enrichissent leur connaissance des carrières individuelles avec la mise en commun de données collectées par des chercheurs spécialistes de différents espaces politiques ; l'analyse fait apparaitre des liens entre individus qui transcendent les territoires et permet la mise en perspective et la comparaison de l'organisation du corps dans les différents espaces et dans le temps.

Prosopange a permis d'accumuler une énorme masse d'outils et de matériaux inédits. L'exploitation de toutes ces ressources n'en est qu'à ses débuts et permet d'ors et déjà d'alimenter différents sujets de recherche portant par exemple sur la typologie des offices et les profils de recrutement, l'analyse des moments où se nouent des partis en périodes de graves tensions politiques, ou encore sur l'étude de la relation entre normes administratives et pratique de l'office. D'un point de vue informatique, le travail se poursuit dans deux directions. Une première perspective est d'étendre les outils d'analyse en particulier vers la dynamique et la navigation dans les réseaux et vers la recherche de régularités ou de singularités dans les trajectoires professionnelles des officiers. Un deuxième défi concerne la prise en compte de la qualité et de la quantité très variables des données, problème particulièrement récurrent lorsqu'il s'agit de traiter de sources documentaires médiévales. En effet, la biographie de certains officiers sera très documentée, très précise, quand les informations connues à ce jour sur d'autres officiers seront parcellaires voire douteuses. Dans ce contexte, toute analyse quantitative est à prendre avec précaution et devrait s'assortir de métriques en mesurant la justesse et la significativité.

Références

Bouveyron, C., L. Jegou, Y. Jernite, S. Lamassé, P. Latouche, et P. Rivera (2014). The Random Subgraph Model for the Analysis of an Ecclesiastical Network in Merovingian Gaul. *Annals Of Applied Statistics 8*(1), 377–405.

Delpu, P.-M. (2015). La prosopographie, une ressource pour l'histoire sociale. *Hypothèses* (18), 263–274.

Gray, J., A. Bosworth, A. Layman, et H. Pirahesh (1996). Data cube : A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In *ICDE*, pp. 152–159. IEEE Computer Society.

Heloise. Projet heloise, european network on digital academic history, http ://heloise.hypotheses.org/.

Messai, N. et T. Devogele (2014). Visualisation de données de prosopographie pour la reconstruction de carrières de personnages et de réseaux socio-professionnels. In *14 ème conférence internationale sur l'extraction et la gestion des connaissances*, Rennes, France, pp. 557–560.

Parisiense. Projet studium parisiense, http ://lamop-vs3.univ-paris1.fr/studium/.

Rafanelli, M. (Ed.) (2003). *Multidimensional Databases : Problems and Solutions, Hershey, PA - USA, 2003*. Idea Group Inc.

Tchounikine, A., M. Miquel, et T. Pécout (2016). Modélisation de données pour une base de données prosopographique. In *Les officiers et la chose publique dans les territoires angevins (XIIIe-XVe siècle) Vers une culture politique ?*, Colloque international de Saint-Étienne, Université Jean Monnet.

Zalc, C. et C. Lemercier (2008). *Méthodes quantitatives pour l'historien*. coll. « Repères ».

Summary

The Europange project, involving both medievalists and computer scientists, aims to study the make-up of the organisation of the corps of administrators of the Angevin controlled territories in the XIII-XV centuries This is based on the study of the biography of each officer, his career, his training and his networks. In this paper, we describe methods and tools designed to analyze this prosopographical data. These include OLAP analyses and network analyses associated with cartographic and chronological visualization tools.

Algorithme 1 Construction d'un cube

Entrée : population d'officiers, liste de dimensions
Sortie : cube
Pour chaque officier *o* de la population
  Pour chaque dimension d_i liée aux individus
    Créer $E_{d_i} = \{\}$ *// ensemble de membres de dimensions*
    Pour chaque information *j* correspondant à d_i renseignée dans *o* au niveau hiérarchique *l*
      $E_{d_i} = E_{d_i} \bigcup \{ j \} \bigcup parents(j, d_i) \bigcup nr(j, d_i, l)$
      // où $parents(j, d_i)$ renvoie les ancêtres de j dans la hiérarchie de d_i
      // et $nr(j, d_i, l)$ renvoie un descendant de j « *non renseigné* » pour chaque niveau inférieur à *l*
    FinPour
  FinPour
  $E_{individus} = E_{d_1} \times E_{d_2} \times ...$
  Pour chaque charge c_i de l'officier o
    Pour chaque dimension d_i liée aux charges
    $E_{d_i} = \{\}$
      Pour chaque *j* correspondant à d_i renseignée dans *o* au niveau *l*
        $E_{d_i} = E_{d_i} \bigcup \{ j \} \bigcup parents(j, d_i) \bigcup nr(j, d_i, l)$
      FinPour
    FinPour
    $E_{c_i} = E_{d_1} \times E_{d_2} \times ...$
  FinPour
  $E_{charges} = \cup E_{c_i}$
  Pour chaque qualificatif q_i de l'officier o
    Pour chaque dimension d_i liée aux qualificatifs
      // traitement similaire aux charges
    FinPour
  FinPour
  Pour chaque n-uplet de $E_{individus} \times E_{charges} \times E_{qualificatifs}$
    créer un point dans l'espace multidimensionnel et y associer la mesure *o*
  FinPour
FinPour

Algorithme 2 Construction de la matrice de co-occurence de charge

Entrée : population d'officiers
Sortie : matrice de co-occurrence de charge
Calculer le cube pour la population en entrée et les dimensions « lieu d'affectation » et « date de la charge »
Exécuter un roll-up du cube sur le niveau « subdivisions » dans la dimension « lieu d'affectation » et « année » dans la dimension « date de la charge »
Pour chaque point (s, a) dans le cube
  Obtenir la mesure qui contient la liste des officiers concernés
  Créer ou incrémenter la cellule correspondante dans la matrice de co-occurrence
FinPour

Measuring and Explaining Political Sophistication Through Textual Complexity*

Kenneth Benoit† Kevin Munger‡ Arthur Spirling§

October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a measure of word rarity derived from term frequencies in the Google books dataset. Our technique not only shows which features are appropriate to the political domain and how, but also provides a measure easily applied and rescaled to political texts in a way that facilitates comparison with reference to a meaningful baseline. We reassess patterns in US and UK political corpora to demonstrate how substantive conclusions differ when using our improved approach.

`Sophistication` software available: `http://github.com/kbenoit/sophistication`.

Word Count: 9,494 (excluding Supporting Information)

*This research was partly supported by the European Research Council grant ERC- 2011-StG283794-QUANTESS.

†Professor of Quantitative Social Research Methods, London School of Economics (`kbenoit@lse.ac.uk`)

‡PhD Candidate, Department of Politics, New York University (`km2713@nyu.edu`)

§Associate Professor of Politics and Data Science, New York University (`arthur.spirling@nyu.edu`)

1 Introduction

A key concern in the study of politics is how the nature of political communication has changed. At the same time that the challenges of governing have grown in complexity, the sophistication of political speech, by many measures, appears to have declined. Thus, within academic studies, typically as part of a broader discussion concerning "dumbing down" (Gatto, 2002), observers have applied measures of textual complexity from educational fields to find that the sophistication of political language has steadily decreased over the past 200 years (e.g. Lim, 2008). Such concerns are echoed in popular presentations too: in 2013, *The Guardian* newspaper[1] used the Flesch-Kincaid grade-level estimates to document a decline in the textual complexity of US Presidential State of Union Addresses.[2]

By contrast, and with more optimistic conclusions, other social science studies have used measures of textual complexity to link linguistic sophistication to outcomes, with a focus on the concrete benefits to clarity. Jansen (2011), for instance, studies the reading level of communications by four central banks, equating lower reading levels of bank communication with greater clarity, which they link to positive effects on the volatility of returns of financial markets. Likewise, Owens and Wedeking (2011) and Spriggs (1996) examine the complexity of Supreme Court decisions, pointing to the importance of clarity in court opinions. In the context of the British parliament, Spirling (2016) applies readability measures to document the democratizing effects of franchise reform on elite speeches. Studying post-war Austrian and German elections, Bischof and Senninger (Forthcoming) find that simpler manifestos make for better informed voters. Finally, as a meta-analysis to defend against charges of elitism and jargon (e.g. Diamond, 2002; Kristof, 2014), Cann, Goelzhauser and Johnson (2014) show that while the reading ease of articles in the top political science journals has declined since 1910, the typical political science article requires less reading ability than the average article in *Time Magazine* or *Reader's Digest*.

These applications share one trait: They equate important substantive characteristics of po-

[1] `http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level`

[2] Although see Benoit, Munger and Spirling (Forthcoming) for a data-driven critique of this claim.

litical, economic, or legal communication such as clarity or sophistication with indexes such as the Flesch Reading Ease (FRE) score (Flesch, 1948) (or something similar to it). These measures, however, were developed decades earlier in entirely different contexts, namely educational research and applied psychology. And it is not clear that they are still relevant for our applications—or indeed if they ever were. As a consequence, we are uncertain as to the true direction of change for specifically *political* communication. More importantly perhaps, we are also unclear about what any such change actually represents in terms of underlying dynamics of language. For example, a trend toward greater verbal simplicity could be a positive development if it improves the clarity of communication, but also might be negative if it represents "dumbing down" in the form of reduced sophistication.

To address such unresolved questions, here we systematically review the properties and statistical performance of current measures of textual difficulty, and develop a new measure of for political language. In what follows, we use the terms "difficulty," "sophistication," and "complexity" interchangeably. Our approach uses experimental data based on human pairwise comparisons of short extracts of political speech (e.g. Lowe and Benoit, 2013; Montgomery and Carlson, Forthcoming), which we then use to scale linguistic sophistication using a simple but well-defined statistical model. In particular, we employ a scaling approach developed by Bradley and Terry (1952) in which clarity of a text is treated as "ability." By moving measurement to a model-based approach, with the statistical mechanics that brings, we allow for sensible statements about uncertainty and inference: thus, one can make claims about the *probability* that a given text is easier or harder than another. This allows us to make meaningful *ratio-level* claims: that, for example, one text is twice as easier (on average) than another (relative to a baseline). This is impossible with all extant techniques of which we are aware. For convenience, and to be consistent with previous efforts, we also provide a continuous version of our measure (designed to be) on the 0–100 interval. Our preferred model is more general than others in the sense that it considers the association of a large collection of features on the difficulty of political texts (rather than just one or two somewhat arbitrarily chosen variables). This includes a comprehensive measure of rarity, extracted from the

Google books corpus. Precisely because it is trained on a relevant domain, this technique yields a measure of textual complexity that is by construction more appropriate for political text than classical measures and with a model fit that is better than more traditional alternatives.[3] Furthermore, we can be precise about each feature's relative contribution to complexity—via the inspection of a $\hat{\beta}$ in a standard generalized linear model arrangement. More generally, our methodological contribution is to provide a work-flow for scholars interested in measuring textual complexity for any substantive area.

To demonstrate how this new measure allows us to gain new insights on old problems, we compare it to the FRE in two related but different applications of elite discourse. In the first—the State of the Union addresses since the founding of the Republic—we show that our measure has considerably more variation than the FRE and, if anything, texts in the modern period are much easier to follow than traditional approaches would suggest. That said, once we introduce uncertainty bounds via a text-based bootstrap, general claims about dumbing down are much more dubious. Second, we apply our approach in its continuous form to three million speeches from the UK's *Hansard* House of Commons records for the period 1935–2013. We show that by our measure, speeches since 1985 have increased in sophistication, mainly because of a rise in the usage rate of unusual terms, which classical measures developed from other domains fail to capture. We relate this to technological changes in how speeches are recorded and broadcast. Furthermore, we show that Labour governments look increasingly like Conservative ones, in terms of the language they use—especially after the 1980s. By setting out clear principles for measuring linguistic sophistication in the political domain, furthermore, we demonstrate the methodological superiority of our approach, and outline a general method for fitting appropriate measures to any context.

[3] Although for reasons we explain, this is a tricky comparison to make.

Table 1: Overview of commonly used reading ease measures in order of citation via Google scholar at the time of writing.

Author	Name of Method	Year	Citations
Flesch	Flesch Reading Ease	1948/49	3793
McLaughlin	SMOG	1969	1402
Dale and Chall	Dale-Chall	1948	1389
Gunning	Gunning Fog Index	1952	1232
Kincaid et al	Flesch-Kincaid Grade Level	1975	1093
Fry	Fry Graph	1968	1007
Spache	Spache Formula	1953	355
Coleman and Liau	Coleman-Liau	1975	261

2 The Challenges of Measuring Linguistic Sophistication

Measuring linguistic complexity is not a new endeavor (see Klare, 1963, for an overview), with early work dating at least to the 19th Century (e.g. Sherman, 1893). The context is typically education, in the sense that the task is matching learning materials to students, based on their age and cognitive ability, with the emphasis being on the easy measurement of the "readability" of a document. While there are a large number of indices for this task—indeed, Michalke (2015) references and implements no fewer than 27 of them—this variety conceals two facts. First, the measures are actually very similar to one another in principle and in practice. And second, a few of the methods completely dominate applied work in terms of use and citation.[4] To see this latter point, in Table 1 we list eight commonly seen metrics—some of which have been adjusted over the years and republished in very similar forms—and their Google Scholar citations at the time of our writing. Inevitably, the number of citations understates the actual use of the methods in practical scenarios, but readers can nonetheless see that the various Flesch-based measures (including the Flesch-Kincaid measure) garner the lion's share of attention, with SMOG and the Dale-Chall measure somewhat behind. Readers will also note that while scholars have continued to be interested in the problem of studying readability after 1975 (e.g. Anderson, 1983), these measures were generally not designed or validated in the modern period.

[4]We ignore metrics for languages other than English here, though there certainly exists a literature dealing with them (e.g. Fucks, 1955; Yuka, Yoshihiko and Hisao, 1988)

In terms of technical details, for a given document, the available measures take into account some combination of: (average) sentence length (e.g. Flesch, 1948, 1949; Gunning, 1952; Fry, 1968; Kincaid et al., 1975); the (average) number of syllables per word (e.g. Flesch, 1948, 1949; Gunning, 1952; Wheeler and Smith, 1954; Fry, 1968; Kincaid et al., 1975); the parts of speech represented in the document (e.g. Coleman and Liau, 1975); and the familiarity of the terms used (e.g. Dale and Chall, 1948; Spache, 1953).

To get a sense of what it means to "take into account" these characteristics, consider the original work of Flesch (1948) (later updated by Kincaid et al. 1975). Flesch studied the reading comprehension of school children. In particular, he was interested in the average grade of students who could correctly answer at least 75% of some multiple choice questions regarding a few select texts. This dependent variable was subsequently transformed to a zero to 100 scale. Fitting a linear regression with a constant and two predictors (average sentence length and average number of syllables per word), ultimately yielded the following formula for scoring documents:

$$206.835 - 1.015\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right) - 84.6\left(\frac{\text{total number of syllables}}{\text{total number of words}}\right).$$

As designed for the original application, this "Flesch Reading Ease" measure had the intended range "for almost all samples taken from ordinary prose" (225 Flesch, 1948).[5] Subsequently, Kincaid et al. (1975) introduced a mechanical conversion of the formula that yields values roughly equivalent to the US grade school level required to understand a text.

Other than indirectly through syllable counts, the Flesch formula does not explicitly take into account the actual familiarity of the words used in a text. An example of an approach that does is Dale-Chall (Dale and Chall, 1948), the formula for which has been adjusted over time but for exposition may be rendered as

$$0.1579\ (\text{percentage of difficult words}) + 0.0496\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right).$$

[5] In practice, the statistic is bounded at an upper "ease" limit of 121.22 for texts consisting of one-syllable, one-word sentences, and bounded from below only by an offset of the average word length.

This yields an (average) grade level at which a reader could be expected to comprehend the document in question. Here, the "percentage of difficult words" refers to any terms not a pre-ordained list of 763 (subsequently around 3000) "'familiar words" in English, deemed to be those known by 80% of fourth grade children (in 1948).

While social scientists have not ignored the measurement of readability *per se* (e.g. Cann, Goelzhauser and Johnson, 2014), there has not been especially great interest in using such methods to produce independent or dependent variables for analysis. None of the studies to which we refer above developed their own measures fit to the domain, but rather adopted some variant of the existing indices, giving rise to an "out-of-domain prediction problem."

2.1 The Out-of-Domain Prediction Problem

Regardless of the specific mechanical details behind current techniques, they were not designed, optimized or tested on the types of social science data to which they are being applied. When political scientists score documents using these methods, they are essentially calculating out-of-domain (and obviously out-of-sample) *predictions*.[6] The problems caused by jumping contexts has frequently been noted in dictionary applications of text analysis (see e.g. Loughran and McDonald, 2014, on using generic sentiment dictionaries on financial documents), but produces more specific problems when designed to measure the sophistication of language.

First, the approaches were designed to match texts to the formal education level of potential readers. They were never intended for the more general task of measuring the "sophistication" of texts in a given domain such as politics, where abstract conceptual appeals to "democracy" or "liberty" might make documents significantly more difficult to follow over and above their sentence structure or average number of syllables. Second, and closely related, the indices were originally for assessing children, rather than adult citizens. Yet this second group will differ not

[6]Technically, the term "out-of-sample" could also be used alone here, but we opt for the stronger "out-of-domain" to draw attention to the fact that the concern is not simply that the estimates are applied to children in the 1940s or 1950s who happened not to be in the original study via random sampling: they are applied to completely different subjects in completely different contexts.

simply in their education level from younger people, but also in their knowledge and understanding of the political process, since presumably they will be exposed to affairs of state on a more regular basis. Third, as the citation dates make clear, these indices were mostly created in the 1940s and 1950s, subsequent to which we can well imagine that language and linguistic style has developed considerably.[7]

Fourth, while the measures are certainly simple—typically consisting of two or three easily calculable text features multiplied by constants—the objective functions they embody are poorly defined when applied to new data. To see this, consider the FRE. This is derived from a linear regression where, as usual for such approaches, the minimization problem (ordinary least squares) is well-defined. In the original context it would yield an R^2 variance explained statistic. However, when taken to State of the Union speeches, it is difficult to know whether the measure—i.e. the model—is performing well or not. That is, the scores of the documents represent out-of-sample predictions, yet there is no readily available metric for assessing the quality of those predictions. An immediate consequence of this issue is that, fifth, it is hard to compare measures (models for the data, essentially) and contend that one is systematically better than another in a given context. Put very simply, if measure A has document i as more difficult than document j, yet measure B implies the opposite, it is not clear which should be preferred, nor on what criteria the predictions ought to be judged. Crudely, once out-of-domain, there is no "ground truth" for comparison.

Precisely because all scores are out-of-domain, a sixth problem emerges: there is no natural way to interpret fine-grained differences in document scores. Consider, for example, documents i and j which score as 70 and 75 respectively on the FRE. In principle, one could claim that were the original sample of children given the speeches, a particular proportion would understand questions relating to the texts in a way that gives rise to the scores. This is a strange counterfactual since, of course, all the texts may have been written after the original study took place. But in any case, the interpretation is extremely awkward. The researcher would like to know the *probability* of understanding one speech over another, or their relative appeal were they in a head-to-head

[7] We return to this idea in some detail below, but as a trivial example to fix ideas, the term *computer* may have been difficult to understand in 1956, but much less so in 2016.

contest for a reader of a given comprehension level. But such information is not forthcoming. The scores are hard to interpret for a related, seventh reason: there are typically no uncertainty estimates around these out-of-sample, out-of-domain, predictions. That is, if document i is scored similarly to document j in terms of point estimates, we would surely be more confident in such a measurement for i if it was 3000 words long relative to j at 30 words.

2.2 Other Problems

Existing measures of readability are *composite indices* whose inputs are weighted. Since those weights are static (i.e. from one point in time), applying them to dynamic data such as time series causes particular inferential problems. To see this, suppose we hypothesize that the State of the Union addresses have gradually adopted less sophisticated language over time. If we use FRE or its close allies to assess this claim, we assume that the only relevant information for the hypothesis test comes from the features of the documents—that is, the Xs. But "dumbing down" could occur (or not) as a consequence of changing *weights* (the $\hat{\beta}$s) too. Traditional approaches cannot speak to such claims directly.

For the reasons we have advanced above, there are compelling reasons to take into account the familiarity of the language used when calculating a document score. For a modern reader, *Indeed, the shoemaker was frightened* would presumably be easier to understand than *Forsooth, the cordwainer was afeared*, yet both would be scored identically by FRE. When such matters are taken into account by current approaches however, it is in a fairly arbitrary manner. For instance, the Dale-Chall method provides a list of 3,000 familiar words, with any word outside this set having a constant weight, regardless of its actual commonality. Such lists are not updated as language changes. Within the Dale-Chall words, we find *locomotive*, a term relatively unknown in this century outside of children's shows; and *telephone*, a term signifying technological advances in 1948, but unknown in 1848 and archaic in 2008. By contrast, *television* is absent from the list.

2.3 Qualities of a Better Approach

Some of the problems we discuss are straightforward to solve. For example, a better approach will study adults in obviously political settings for the contemporary period. This will immediately rectify the central "out-of-domain" issue. Other matters are more subtle. Ultimately, as in the educational literature, humans are the "gold standard" for coding complexity of language. With that broad understanding in mind, an ideal way forward is to either use small numbers of experts or, better yet, large numbers of non-experts who can code texts in a fast, reproducible manner, recruited through a crowd-sourcing platform (Benoit et al., 2016).

Because our interest is more general than education, we want the coders to score the documents directly. At least since the work of Thurstone (1927), we know that having humans perform (large numbers of) pairwise comparisons between texts is likely preferable to other hand-coding systems (see Montgomery and Carlson, Forthcoming, for discussion). In the pairwise case, political scientists (e.g. Loewen, Rubenson and Spirling, 2012; Lowe and Benoit, 2013) have used the Bradley-Terry model (Bradley and Terry, 1952) as a fast and well-grounded way of converting the pairwise binary decisions over items (here, documents) and placing them into continuous score space. This simple approach has a natural interpretation, insofar as its fundamental building block is the probability that 'i beats j'—here, the probability that i is easier to understand than j—when the two documents are compared one to another. This probability is well-defined, and is strictly between zero and one. Finally, because the latent characteristic of the item can be modeled via a linear predictor—that is, $\mathbf{X}\beta$—one can talk meaningfully about the "effects" of certain characteristics, such as document length, syllable number, the familiarity of tokens etc on the linguistic complexity of a document. Notice that such estimates will be *sample specific* and once some domain coding has been undertaken, the researcher is not required to simply apply a rote formula again and again however dubious a given application.

3 Method: Crowdsourcing Complexity

With the above considerations in mind, we aim to discover the textual features that constitute complexity, in the context of specifically political language. At an intuitive level, our procedure is quite simple, and it begins by producing a series of short texts of one or two sentences each—fragments we refer to as "snippets"—which are given to human coders to compare, pairwise.[8] The coders tell us which of the two texts is easier to understand, and they do this multiple times for various pairwise combinations of snippets. We go from these pairwise decisions to a continuous scale of reading ease via the application of an unstructured Bradly-Terry model. Then, given those scores on the scale, we learn features of the snippets best predict their relative difficultly, as rated by the humans.

The human coding is performed on a crowdsourcing platform in batches of ten short comparison tasks, following the general procedures described by Benoit et al. (2016). The precise questions asked of coders, the way in which we ensured consistent quality in their responses, and the exact nature of the comparisons required from them is discussed in some detail in Supporting Information A. In our particular case, the snippets were drawn from the 70 State of the Union Addresses (SOTUs) delivered after 1950.[9] We used these texts because the purpose of the SOTU addresses has remained relatively unchanged in the postwar period, and because of the attention these speeches have received in previous examinations of readability. This gives us a benchmark of interpretation to which to compare our findings below, although our approach may easily be adapted to measure linguistic sophistication in other contexts.

Some preprocessing of the addresses prior to creating snippets was required: in particular, we removed some organizational non-sentence pieces of text (mostly referring to the medium by which the address was delivered). Once cut down for comparison, we disqualified some snippets from consideration: those which were outside the 0–121 range of the FRE; any containing more

[8] We take a candidly "bag of snippets" approach as a document model: we assume all relevant information is within the snippet rather than where or how it occurs in the document.

[9] As we explain below, we subsequently supplement these with some a small amount of earlier pilot comparison data we had.

than two numeric years; any with large numbers; and any beginning with the title of a document section.

We constrained the snippets drawn for comparison from our texts to three bands of approximately equal lengths, to avoid comparisons where deciding on the "easier" snippet appears easy because one is noticeably shorter than the other. Within each group of snippets of similar lengths, we sorted the snippets once by their FRE scores in ascending order and again in descending order, and combined the two lists to create a set of comparisons that vary from (very) dissimilar to (very) similar FRE scores.[10]

3.1 Incorporating Familiarity: Google n-grams and parts of speech

Corpus linguistics has progressed significantly since the early measures of reading ease were developed, giving us access to a huge amount of detail about word rarity and how it evolves over time. Our test data spans political speech dating to the 1790s, and a major contribution of our measure is that it incorporates a benchmark of how unusual (and hence how difficult to understand) each word from that time span is in contemporary usage. To this end, we downloaded the unigram frequency datasets from the Google Book corpus dataset,[11] which yields token counts on a yearly basis from 1505 until 2008.

To assess the how unusual might be a text for a modern audience, we computed the frequency of each term it contained relative to the frequency of the word *the* today.[12] This allowed us to compare the relative frequencies of terms without being affected by changes in overall word quantities or transcription accuracies (which vary significantly over the time sampled). For instance, *husbandry* (the cultivation and breeding of crops and animals) was used much more often in the 1790s than in current times. Its inclusion in a speech would therefore make that document harder

[10] To be precise, we matched three sets of two-sentence snippet pairs: those with lengths between 345-360, 360-375, and 375-390 characters respectively. We also created an additional 210 randomly selected bridging pairs, to form a fully linked network of pairs to enable pairwise scaling.

[11] `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`

[12] We used *the* since it is the most common word in the English language and because its relative frequency has remained relatively unchanged in several hundred years.

for a contemporary audience (such as our crowd coders). (To smooth out individual differences in the yearly samples, we combined the frequency counts for all years from 2000 through 2008.) We give more details on this process in Supporting Information B.

We also computed the relative frequency of parts of speech in each text, to obtain proportions of nouns, adjectives, verbs, prepositions, and so on. We did the same for some syntactic complexity markers such as the number (subsequently, proportion) of clauses in sentences. This allowed us to include these quantities in the feature set for fitting models below to predict reading ease. Our approach to obtaining these quantities is explained in Supporting Information C.

3.2 Bradley-Terry Regression Analysis

Exposition of the Bradley-Terry model (Bradley and Terry, 1952) can be found in numerous textbooks (e.g. McCullagh and Nelder, 1989), but we follow the presentation found in Turner and Firth (2012) for our work here. The input data is the result of our human coders having declared winners in the large number of "easiness contests" between snippets. For a given contest, crowd workers must decide which of two snippets i and j is easier to comprehend (no ties are allowed). If the easiness of i is α_i, and the easiness of j is α_j, then the odds that snippet i is deemed easier than j may be written as α_i/α_j.

Defining $\lambda_i = \log \alpha_i$, the regression model can be rewritten in logit form:

$$\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j. \tag{1}$$

Subject to specifying a particular snippet as a "reference snippet" (whose easiness is set to zero), this setup allows for maximum likelihood estimation of each snippet's easiness. For current purposes though, we wish to make the easiness of the snippets a product of covariates—that is, the average length of words they contain, the average word's number of syllables, etc. This is achieved

by modeling the easiness of a given snippet as

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}. \tag{2}$$

This is known as the structured Bradley-Terry model: the set of β coefficients then tells us the marginal effect of each x-variable on the perceived (relative) easiness of the snippets. Notice further that, on estimating the β parameters, the covariates pertaining to a given document may be used to obtain the (predicted) easiness of that text (even if it did not appear in sample, or not in that given form).

This is a simple model, and it is worth emphasizing what is being assumed about the data generating process when we interpret its relevant output. First, we assume that the outcomes of the contests are (statistically) independent of one another: that what happens in the kth contest does not affect what happens in the $k+1$th contest. Second, we are making no allowance for variability between snippets which have otherwise identical covariate values. That is, we are not using any kind of random effects for the snippets themselves. This means, equivalently, that the contest results for a given snippet are not modeled as correlated. Third, we make no attempt to include so-called "contest-specific predictors" either in their indirect form—such as effects for (the proclivities of) given human coders—or directly—such as allowing for consequences of the order in which the snippets were presented to the subjects who judged them.

The model is sufficiently flexible to be adapted to address these concerns directly, although here we have kept our formulation deliberately simple. Our primary interest is in estimating the complexity of documents by predicting (that is, scaling up) from the snippet results, for which we need estimates of their relative weights in predicting the human ratings of easiness, not a fully specificed model of coder and sentence effects.

3.2.1 Variable Selection via Machine Learning

For any specific application, it is not obvious which variables should be included in a given model of readability, but with our measures from the *unstructured* Bradley-Terry scaling, we can attempt to predict the variation in this ability scale and use the results to choose the relevant covariates for fitting our own, domain-specific measure. Our scaling returns an estimate of an "ability" λ_i (in this case, relative easiness) for each snippet, but makes no use of covariates.[13] We then use all our various text characteristics as features to predict these (unstructured) abilities using a random forests approach (Breiman, 2001), and then inspect the (relative) variable importance estimates for each covariate. Once those characteristics that matter most are identified, they can be used in the structured model of Equation 2 to obtain the relevant coefficient estimates.

4 Results

We have two main sets of results. First, we can compare the standard measures as applied to specifically political text: the first such attempt that we know of. In Supporting Information D we give more details but one observation is worth noting immediately: the models all perform very similarly, with little to separate them in terms of either model fit (Akaike information criterion) or accuracy (proportion correctly predicted). The best performer on our data was the Spache measure, but the FRE is almost exactly as useful and will be preferred on familiarity grounds. We use it in our running comparison for what follows.

Second, and much more importantly, we provide a new measure of complexity based on our crowdsourced data and the inferences we draw from our machine learning approach.

[13] In practice, it is occasionally the case in our sample that a snippet never wins or never loses. The usual consequence of this kind of data separation would be infinite ability estimates. In one run of the model, we simply deleted those missing values, and in another we used the bias-reduction technique of Firth (1993) to ameliorate this problem. The results, in terms of the variable importance order are essentially identical, either way.

4.1 Augmented Bradley-Terry Approach

In Supporting Information E we report details of the random forest models that we ran on the unstructured abilities, along with variable importance plots for the same. We find that the model favors the rarity measure based on the recording the least commonly occurring term in the snippet (relative to the frequency of *the* in the Google corpus)—denoted as `google_min_2000`. And it also suggests average sentence length measured in characters (`meanSentenceChars`) is about as important. Given our discussion above, the fact that these variables are useful is unsurprising. In principle, of course, we could stop there (especially given the relatively large distance of the top two from the other variables). In experiments, however, we found that the third most important variable, `pr_noun`—the proportion of words from the text that are nouns—helped model fit. We thus include that one too to form a basic machine learning model.

How does this simple model perform? To assess that, we construct a baseline model which uses the Flesch reading ease (FRE) as its (only) covariate content. We do this in two ways. First, we include the FRE of the snippet using the weights from Flesch's (1948)'s original formula. Second, we include the variables Flesch (1948) includes, but allow the model to calculate the optimal weights for our political data. In Table 2 we report the findings from those models, in the leftmost two columns. For the "FRE baseline" model (original weights) we see that the Akaike information criterion (AIC) is 26269, while the proportion (of contests in the data) correctly predicted (PCP) is 0.568. When we allow the weights on the relevant variables to adjust to local conditions (column 2) we see a commensurately better model fit: the AIC falls to 25912.69, and the proportion correctly predicted rises to 0.583. This is in line with our thinking above: in particular, that models work best when fit to relevant data. Column 3 represents our basic three variable model as discussed above. Clearly, it does better than the Flesch model with the original weights, but—perhaps surprisingly—not as well as the re-weighted version (AIC is higher, PCP is lower).

Our model does not include a measure of word length, despite this feature being one of the two core components of the Flesch index. Looking down the variable importance plots, the first measure of word length to be recommended (i.e. the one highest up in importance terms) is the average

Table 2: Model comparison, post feature-selection. Note that the last column represents our "optimal" model. "PCP" is proportion (of contests) correctly predicted by the model.

	FRE Baseline	FRE re-weight	Basic RF model	Best Model
FRE	0.02			
	(0.00)			
mean Sentence Length		−0.06		
		(0.00)		
mean Word Syllables		−1.78		
		(0.07)		
Minimum Google books rarity			1310.41	1332.49
			(153.27)	(155.85)
mean Sentence Chars			−0.01	−0.01
			(0.00)	(0.00)
noun proportion			0.61	0.63
			(0.19)	(0.19)
mean Word Chars				−0.31
				(0.02)
N	19430	19430	19430	19430
AIC	26269.20	25912.69	25917.49	25739.93
PCP	0.568	0.583	0.580	0.587

Standard errors in parentheses

All coefficients are statistically significant at the $p \leq .05$ level.

number of characters per word (`MeanWordChars`). As an experiment, we added this variable to our machine learning model and re-ran the analysis. The results of that process are in the fourth column of Table 2 titled "Best Model," which outperforms every other version, with the lowest AIC (25739.93) and the highest PCP (0.587). In an effort to ascertain the robustness of this model, we dropped the parts-of-speech variable (`pr_noun`) and added the next highest rated one (`pr_verb`), but in both cases the fit got worse. This is our preferred model for the analysis that follows. Note, in passing, that all the variable effects are as expected (and are statistically significant at conventional levels): in particular, *ceteris paribus* texts that contain words which have low (minimum) rarities are easier to understand ("Minimum Google books rarity" is positive), texts that contain longer sentences ("mean Sentence Chars") are harder, and texts with longer words ("mean Word Chars") are also more difficult to comprehend. More nouns ("noun proportion"), on average, also adds to simplicity. This is, in fact, in keeping with earlier work by Flesch (1948) who proposed

a "human interest" index in which a text with more (pro)nouns was generally found to be more compelling than one with fewer.

On what types of data, exactly, does our model do better? Unsurprisingly, given they share core terms, it performs best when two documents are similar other than the proportion of nouns they contain, or the rarity of their words. In the contests for which our model outperforms the Flesch version to the greatest extent, it is the word rarity input that matters most. To get a sense of this, compare these two snippets. The first is from Obama's 2009 address, and has an FRE of around 50:

> I speak to you not just as a President, but as a father, when I say that responsibility for our children's education must begin at home.

The second is from Cleveland's 1889 effort,[14] which has an FRE of approximately 67:

> The first cession was made by the State of New York, and the largest, which in area exceeded all the others, by the State of Virginia.

Thus the FRE model predicts this to be a relatively straightforward win for Cleveland's speech. Our model, of course, penalizes the estimate of its simplicity due to the presence of the relatively rare term *cession* (along with there being slightly fewer nouns in the second document). Indeed, the frequency of the least common term in Obama's speech is over three orders of magnitude larger than that of Cleveland's speech. Put crudely, if researchers think the commonality of terms matters for measuring complexity, our approach is preferred.

It is helpful to be candid about several issues pertaining to our results. First, clearly, while we are outperforming the most widely-used measure of readability, our gains are not huge in an *absolute* sense. The largest gains in predictive accuracy come from refitting the Flesch model appropriately to the data rather than using its usual "off-the-shelf" mode. Second, these gains are, however, large in a *relative* sense. Our task was intentionally designed to be difficult. The baseline Flesch predictive accuracy was 56.8%–a mere 6.8% better than chance. Our final model is 8.7% better than chance, a relative increase of 28%. Third, whether or not one uses our *specification*, the

[14]This snippet appears per discussion in Supporting Information A about including some older texts from an earlier pilot study.

Table 3: Examples of covariates from two snippets in the data.

snippet	Min Google rarity	Mean Sent Chars	noun proportion	mean Word Chars
Eisenhower	3.501e-07	158.5	0.23	5.37
Bush	1.40e-08	153.5	0.31	4.72

general *approach*—of training on relevant data and providing model-based estimates—is preferable for the reasons we gave above. Even if one wanted simply to use the Flesch set up (in terms of its component variables) based on Table 2 we would recommend local data for that purpose.

5 Applications to political text

We can apply the results of our model in various ways. We outline two obvious approaches before demonstrating how they might be used in practice. First, given Equations 1 and 2, we can obtain a (point) estimate of the probability that any given text i is easier (or conversely, more difficult) than any other text j by calculating

$$\Pr(i \text{ easier than } j) = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)}. \tag{3}$$

To see how this works, consider two snippets, one from Eisenhower,

> Here in the District of Columbia, serious attention should be given to the proposal to develop and authorize, through legislation, a system to provide an effective voice in local self-government. While consideration of this proceeds, I recommend an immediate increase of two in the number of District Commissioners to broaden representation of all elements of our local population.

and one from George W. Bush

> And the victory of freedom in Iraq will strengthen a new ally in the war on terror, inspire democratic reformers from Damascus to Tehran, bring more hope and progress to a troubled region, and thereby lift a terrible threat from the lives of our children and grandchildren. We will succeed because the Iraqi people value their own liberty - as they showed the world last Sunday.

For each of these snippets, Table 3 gives the relevant covariate values for our best model above.

Using the coefficients from Table 2, it is a simple matter of matrix multiplication to form

$$\lambda_{\text{Eisenhower}} = (1332.49 \times 3.501e-07) + (-0.01 \times 158.5) + (0.63 \times 0.23) + (-0.31 \times 5.37) = -3.10$$

and

$$\lambda_{\text{Bush}} = (1332.49 \times 1.40e-08) + (-0.01 \times 153.5) + (0.63 \times 0.31) + (-0.31 \times 4.72) = -2.80.$$

Following the algebra above, we have

$$\Pr(\text{Eisenhower snippet easier than Bush snippet}) = \frac{\exp(-3.10)}{\exp(-3.10) + \exp(-2.80)} = 0.425.$$

Of course, these comparisons can be made between *any* two documents—so long as we have covariate values for them—including fifth grade texts, as in Flesch's (1948) original work. In our case, we obtained a set of fifth grade texts from a university education department,[15] and estimated the relevant λ to be -2.175897. Thus, the probability that the Eisenhower text is easier than a fifth grade text is estimated to be 0.284; and the probability that the Bush text is easier to follow than the fifth grade works is 0.324. We can place confidence intervals around the point prediction by resampling the sentences in the texts (in the sense of Lowe and Benoit, 2013). Note that the differences between texts mean something extremely well-defined here: we can make concrete statements about *how much* easier one document is relative to another, and the quantity refers back to a sensible model. This is quite unlike FRE, where as we noted, a difference of 5 points on the scale has no natural, cardinal interpretation.

Along with model-based estimates, researchers may also want a quantity analogous to the continuous 0–100 scores from the Flesch (1948) (regression) formula. Our proposal is to simply rescale all the λs (that is, the $\mathbf{X}\beta$s, without applying the exponential function) themselves to be on the relevant interval.[16] For a given data set, a sensible way to proceed is to include a text(s) at the

[15] `https://projects.ncsu.edu/project/lancet/fifth.htm`

[16] See Supporting Information F for an alternative approach.

fifth grade level (designated a score of 100), and one at the post-college level (designated a score of 0)—or whatever minimum and maximum is preferred)—and to then (linearly) scale all resulting λs based on those two end points.[17]

Experimenting with the continuous measure on the SOTU snippet corpus performs well in the sense that it returns point estimates on a 0–100 scale commensurate (but not identical) to the FRE equivalents. This works because it replaces a logit-style calculation that is not linear in the predictors with a linear sum (i.e $\sum_{r=1}^{p} \beta_r x_{ir}$), exactly like the regression-based formula for FRE. In Figure 1 we provide a scatterplot of our measure for the snippets (y-axis) relative to the FRE for the same data (x-axis). Clearly the correlation over the full range of points (~ 0.7) is reasonably large and positive. The internal box allows for a more direct comparison of our measure to the (theoretical) minimum and maximum of the FRE: in general, our measure performs similarly. This implies that for the great majority of documents for which FRE is used, our measure—preferred on theoretical grounds—is a good choice that will behave as expected. Outside the box, particularly to the bottom left of the plot, our measure tends to score the points differently. Indeed, we assign a considerably lower ("harder") rating for the hardest texts.

5.1 Reanalyzing the *State of the Union* addresses

Recall that our snippets came from the SOTU time-series, a dataset of considerable interest to academics and journalists. Using our model-based probability measure—here, with a fifth grade text as a baseline for comparison—Figure 2 plots the relevant point estimates and 95% (simulated) confidence intervals (y-axis) plotted against the date of the relevant text. The probability estimates are drifting upwards over time, but generally stay below 0.50. But because we are using a well-defined statistical model, we can say more about the data. In particular, the confidence intervals allow us to make comments about sampling uncertainty. Note that there is considerable overlap between the intervals for the post-war period (for example, some of the speeches in the early

[17] We used the collection of fifth grade texts we mentioned above for the easy end of the scale, and the most difficult snippet (which had an FRE of around 3) for the "hard" end.

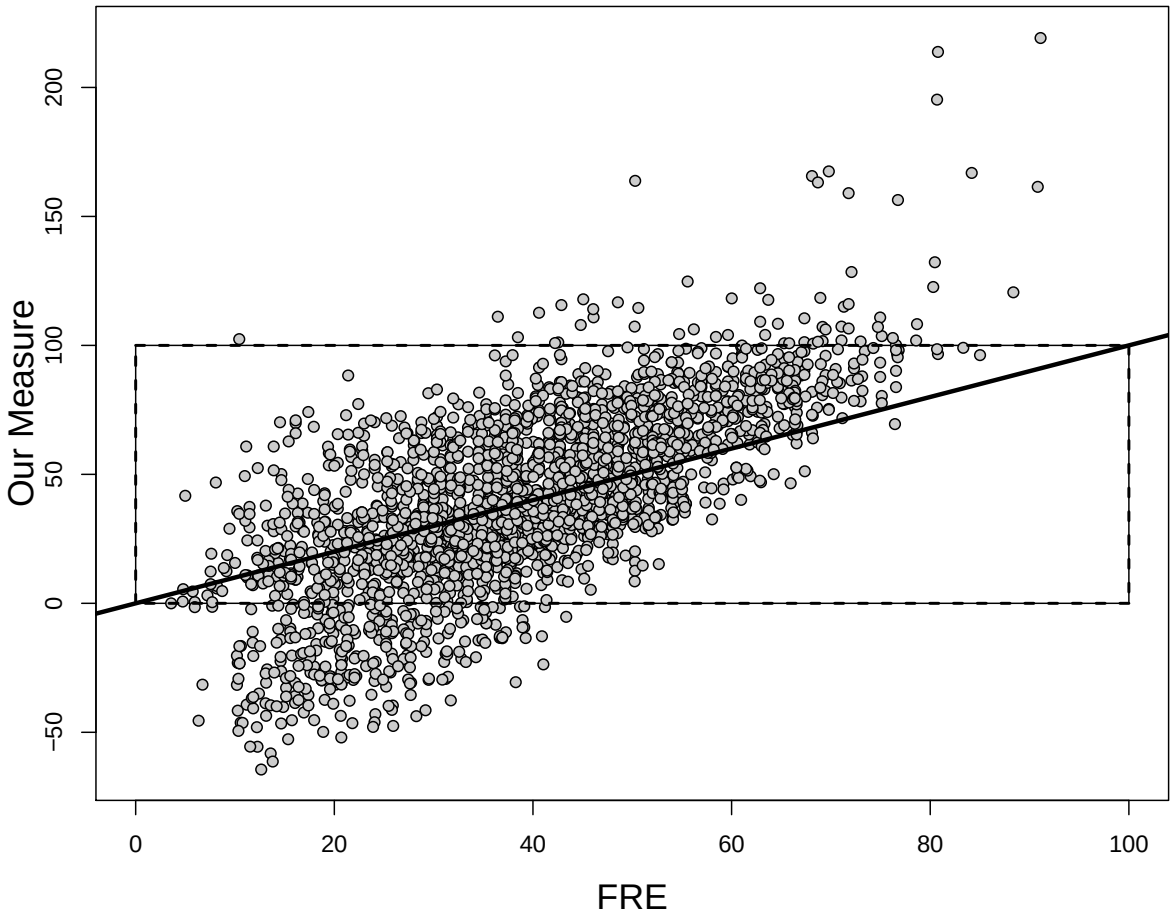


Figure 1: Comparing the "linear" version of our measure to FRE of the snippets. Correlation is generally high, especially for the theoretical range of the FRE (inner box).

2000s are not so different to those in the early 1950s). This implies that statements about the simplification of language may be correct in some aggregate sense if we consider the entire period since the founding of the Republic, but less clear for modern times specifically.

Of course, since other measures in the literature are not based directly on a statistical model, it is hard to compare our output here with more traditional approaches. Fortunately, the continuous version of our measure does allow a direct comparison, and in Figure 3 (where we label it "'MBE" for [m]odel [b]ased [e]stimate(s)) we show it plotted against the FRE (which has been smoothed and given a 95% confidence band calculate by sentence-level bootstrap). Clearly, the conclusions

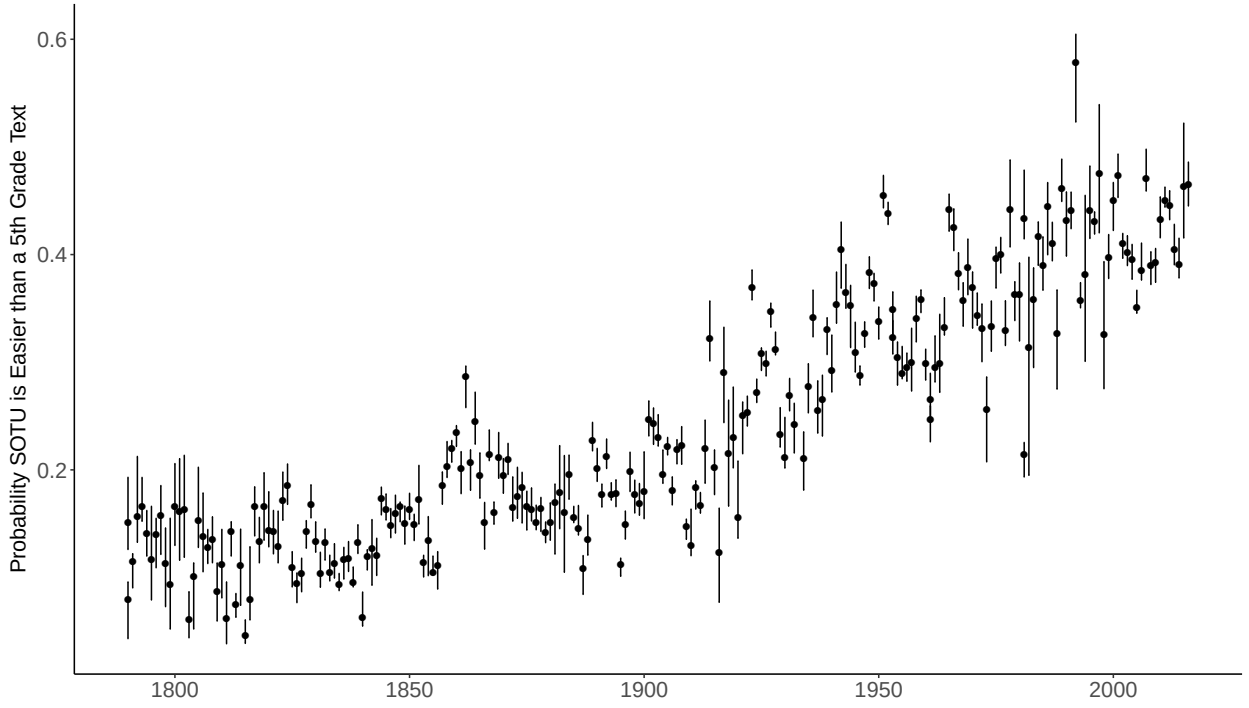


Figure 2: The probability that a State of the Union address is easier to understand than a fifth grade text baseline.

from the measures agree in terms of general direction: addresses become easier over time. But conclusions differ in terms of magnitude. In particular, our measure has the speeches prior to around 1910 being considerably more difficult to understand than FRE claims they were. And then, post 1910, our measure tends to have the estimated ease of understanding the passages as higher than FRE. To the extent that one believes that new technology, such as the radio and the television, lead to speeches that are easier to follow after the first decade of the 20th Century, this makes sense. And, to reiterate, our model is actually trained on appropriate, political data. Why do we estimate the earlier speeches as being so much more difficult than FRE has them? Mostly, this is because of our rarity variable. Recall that it uses the relative commonality of a word in 2000 as a baseline. Of course, as one moves back into history words that are rare and archaic today become more common. Thus, our measure allows us to more accurately judge how difficult texts are from the *perspective of a modern reader*. Notice that if this is undesirable, e.g. one may want difficulty estimated for contemporaneous audiences in 1800, 1810, 1820 etc, our framework allows one to

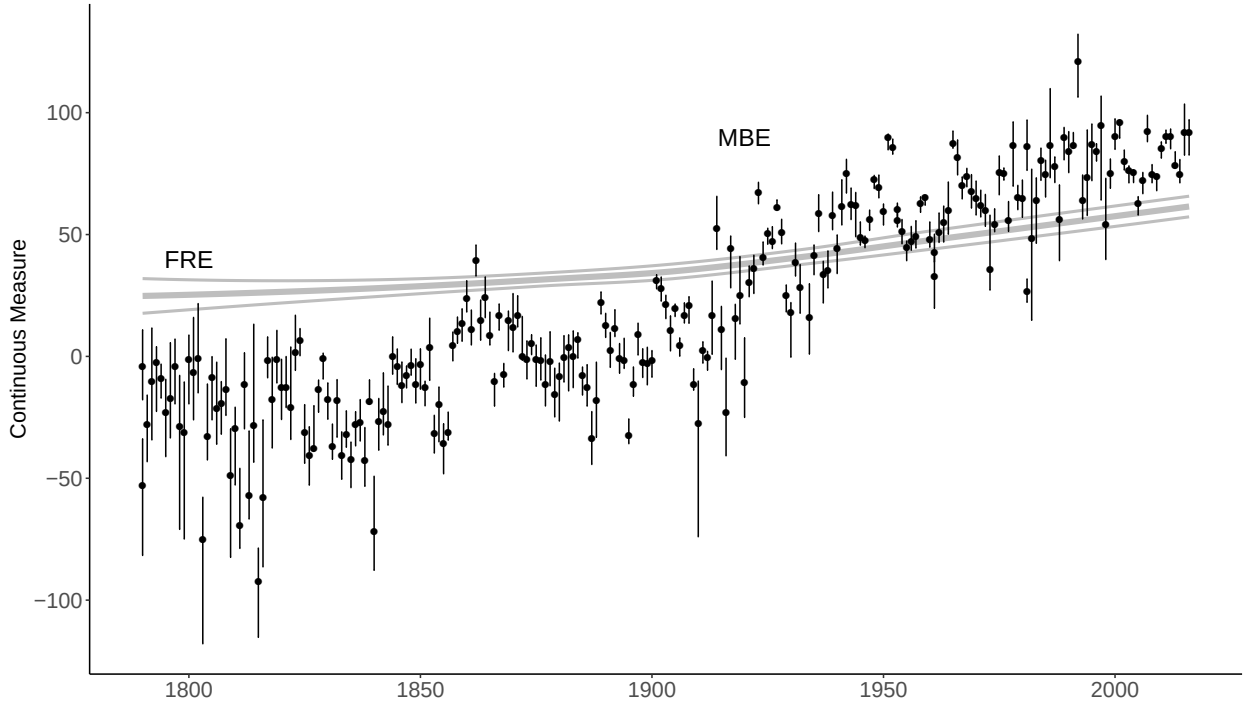


Figure 3: Comparing the linear, continuous version of our model based estimates (points plus 95% confidence intervals, denoted MBE) to FRE (smooth lines, with outer edges representing 95% confidence intervals) of the State of Union addresses. Confidence intervals estimated by sentence-level bootstrap.

do that. It would simply require using the relevant Google books corpus for the decade in which the text originated: that is, this rarity would become a dynamic variable in the modeling set-up, rather than fixed to its levels in 2000.[18]

5.2 *Hansard*, 1935–2013

As our final application, and to demonstrate the different types of conclusions one might reach using our measure versus FRE, we analyzed 78 years of House of Commons debates. This *Hansard* corpus includes essentially all speeches (some 3 million in number) by all Members of Parliament (MPs) for the period under study (See Rheault et al., 2016, for description). To keep our analysis simple, we focus solely on Labour and Conservative legislators, who represent around 90% of all

[18]Of course, we do not have coders from any other period, so one would need to make simplifying assumptions about the relevant coefficients.

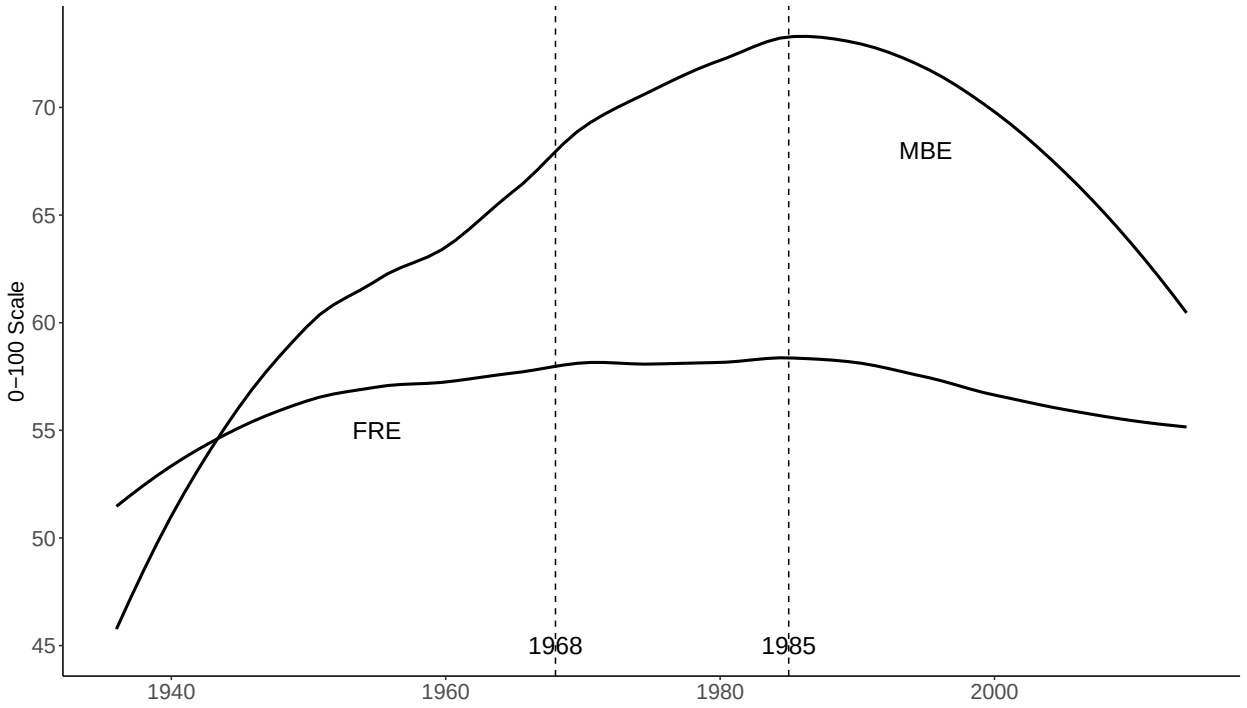


Figure 4: Comparing our mean model based estimate (MBE) with FRE estimates for 3 million speeches delivered by Members of Parliament. Note the break point for our measure is 1985, while for the FRE it is 1968.

MPs in the corpus. The data is compiled in "sessions" of parliamentary time, which last around a year a piece. We begin with by looking at the behavior of our continuous measure (relative to FRE) and then study the model based probabilities.

5.2.1 Speeches and Technology Changes

To begin, for each of the parliamentary sessions, we estimated the mean of the FRE and our continuous measure, for all MPs. The results of those calculations can be seen in Figure 4.

Although the lines start in approximately the same place region of complexity (the rescaled measure on the *y*-axis), the speeches quickly become easier according to the MBE measure, before the trend reverses in 1985. The FRE, by contrast, is almost constant at around 55 on the 0–100 scale, after 1968. To identify the different inflection points, we conducted a generalized version of the Chow (1960) test. For each session in the data, we segmented the time series into two parts (before and after the session in question). We then looked for evidence of structural instability

between the two segments, using standard defaults as described by Zeileis et al. (2002). For the FRE series, the optimal break is in session 33, or around 1968. For our preferred approach, the optimal break is in session 50, or around 1985. Interestingly, both of these change points correspond approximately to technological shifts in terms of recording and broadcasting House of Commons proceedings.[19] In particular, in the spring of 1968, the House of Commons experimented with sound broadcasting. Ultimately, parliament would install permanent means of recording in 1978. By contrast, it wasn't until the late 1980s that television recording was approved—and it began in November 1989.

Obviously, it is very difficult to make causal claims from such aggregated, observational data. Still, the effects seem to be similar: with new technology, and new visibility, speeches become (on average) more complex. Why might this be? One argument made in the press[20] is that television, in particular, encourages members to make longer opening speeches in debates. The idea here is that they do this to ensure their presence is noted by cameras, and that they can be quoted—possibly at length—on news programs. In general, making longer, more structured reports as speeches will tend to depress readability indices, especially if they substitute for shorter, punchier statements. In the Canadian context,[21] there is some belief that television broadcasting encourages MPs to read their speeches, rather than speaking off-the-cuff. If so, this formalism will tend to drive the average statement to be more complex as measured by any approach. To get a sense of the plausibility of this argument, in Figure 5 we disaggregate our measure into its four component parts, and study their (mean) behavior over time. We add a lowess curve in each case, and vertical lines for estimated breakpoints (see Bai and Perron, 2003) in the data (as implemented by Zeileis et al., 2002). The patterns are clear: the proportion of nouns per speech is rising over time (top left); the average length of words is rising (bottom left); speeches contain words that are rarer (bottom right); sentence lengths got shorter and then longer again (top right). This latter point is

[19] See House of Commons briefing on "Broadcasting Proceedings of the House"': https://www.parliament.uk/documents/commons-information-office/g05.pdf

[20] See e.g. "Have TV cameras in Parliament made political debate coarser?" http://www.telegraph.co.uk/news/politics/11244147/Have-TV-cameras-in-Parliament-made-political-debate-coarser.html

[21] See "Television and the House of Commons", https://lop.parl.ca/content/lop/ResearchPublications/bp242-e.htm

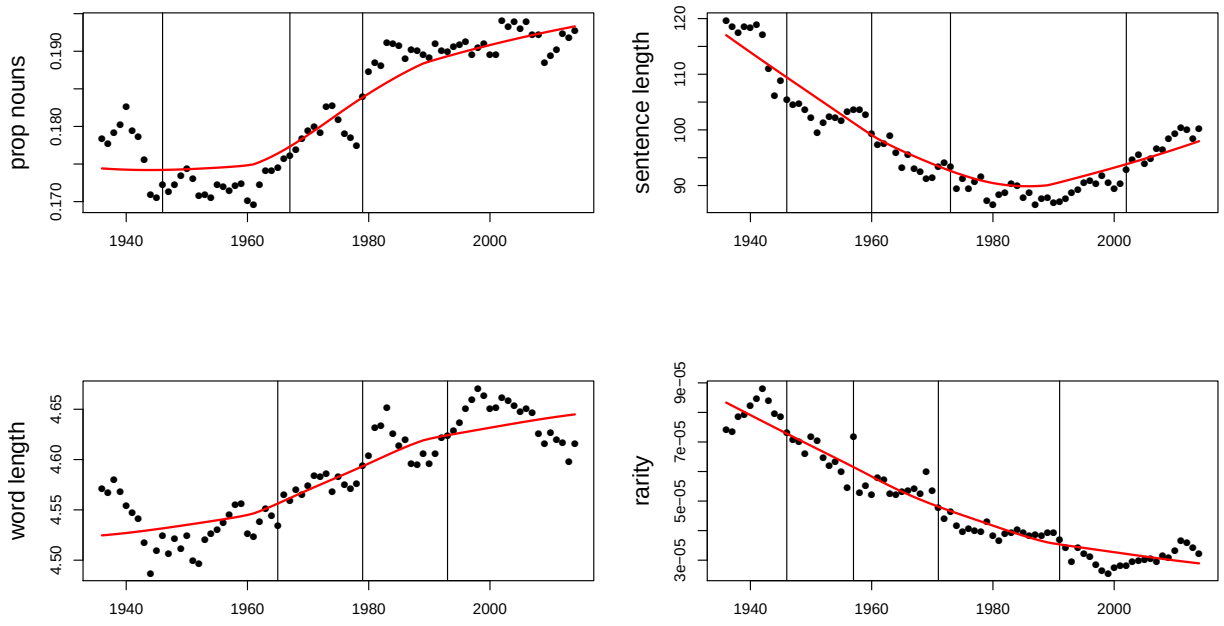


Figure 5: *Hansard* time series disaggregated by covariate in our measure, each point representing the average value for that session (with lowess line smoother added). Horizontal axis is the date of the session. Vertical lines represent estimated change points.

the key for our inference here: that is, only sentence length shows a pattern consistent with Figure 4. In particular, it seems that the most contemporary speeches involve longer sentences, which corroborates our earlier claims about the effects of television: somewhere between 1980 and 2000, something—we would argue the introduction of television—altered the data generating process.

5.2.2 Sociological Change in the House of Commons

The idea that descriptive representation might be an important characteristic of elected officials is not new (Pitkin, 1967). In recent times, however, scholars of British politics have specifically addressed its effects in the context of social class in the House of Commons (e.g. Heath, 2016). Empirically, they note that fewer and fewer Labour MPs in the post-war period come from (objectively) working class backgrounds, with an especially steep decline during and after the 1980s (Heath, 2015). Other scholars note that this is also true of subjective measures, wherein MPs are

asked to self-identify in class terms (Norris and Lovenduski, 1995). While the typical focus is on voter perceptions of politics, our measure allows to investigate how such changes affect discourse in parliament. Recall from Equation 3 that it is trivial for us to produce a probability that one text is easier than another. For the entirety of the *Hansard* data, we do just that for the mean value of λ (the "easiness" of a speech) for all Conservative and Labour MPs. That is, we calculate, for every session, the probability that the mean Conservative speech is easier to comprehend than the mean Labour speech. Note the contrasting strength of our approach with the weakness of traditional efforts. In particular, such a ratio is not directly interpretable in the Flesch context: e.g. the fact that text *A* has an FRE of 100, and text *B* has one of 50, does not mean *A* is "twice as easy" as *B*. For us though, the probabilities can be interpreted directly in these terms.

The results of this calculation are shown in Figure 6 as the plotted points. Those points are (blue) circles when the Conservatives are in government, and (red) squares when it is Labour. The means are equal at the 0.5 point on the *y*-axis, as noted by the broken line. We see immediately that when parties are in government, their (average) speech is easier to follow: this must be the case, because all the Conservative sessions in power are above the $\Pr(\text{Conservative easier}) = 0.5$ line, while all the Labour sessions are below it. But more interestingly, the trend of the data is towards a probability of 0.5, and we see this from the solid lowess line we imposed on the plot.[22] Put otherwise, Labour and Conservative speeches increasingly resemble one another in terms of difficulty. Why might this be? One possibility is that, with the general decline of working-class Labour MPs, both Conservative and Labour members are more similar in education, class and background than before. If we think social background matters for communication styles, then the convergence may be simply a consequence of sociological change in the House of Commons. In Supporting Information G we show that one possible mechanism is via changing word rarity: in particular, especially from around the 1980s onwards, Labour speeches (on average) use words that are rarer than in the past, and indeed rarer than those used by contemporaneous Conservatives. That is, one possibility here is that in line with their changing social and educational position,

[22] A simple linear regression with a dependent variable equal to the absolute deviation of the relevant probability from 0.5, with session number and party of government as regressors, corroborates the trend claim.

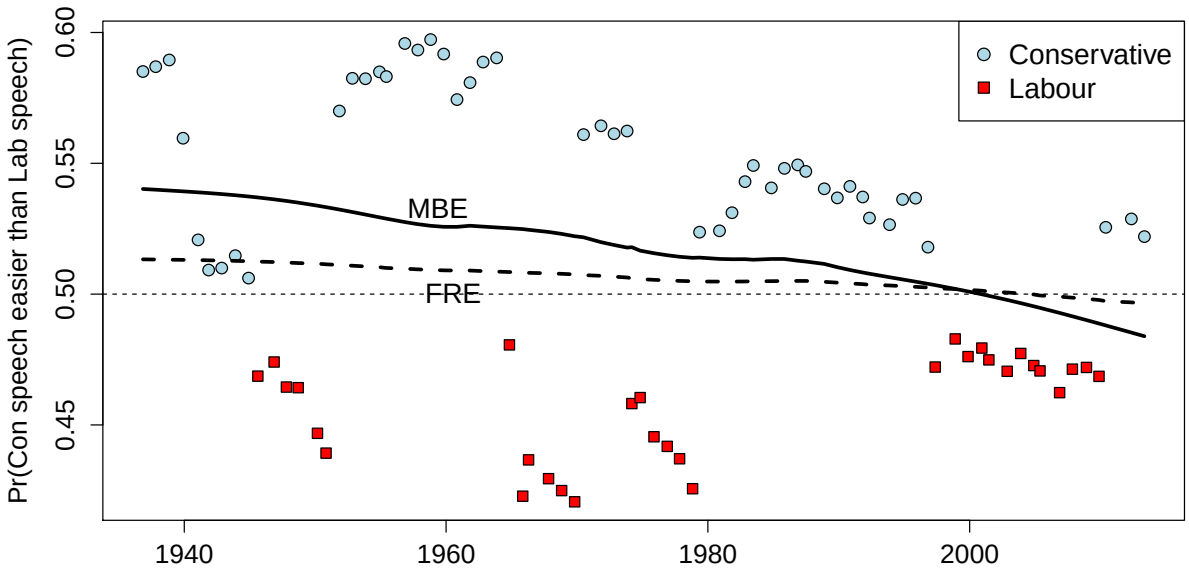


Figure 6: Estimated probability that the mean Conservative speech is easier than the mean Labour speech, over time. Point colors and shapes represent which party was in government at the time: (red) squares are Labour, (blue) circles are Conservative. Solid black line is lowess of probability over time ("MBE"). Broken line is lowess of the FRE *ratio* measure.

Labour members are departing from more basic vocabularies in favor of (relatively) more abstruse terms.

To show how our measure here improves over standard approaches, we include the lowess for an FRE *ratio*: the mean FRE for Conservative speech in a given year divided by the sum of the means for the Conservatives and Labour. While this is not a well-defined probability, its interpretation is more directly comparable to our model-based probabilistic estimate. One observations is immediate: the FRE ratio is considerably less variable than our measure, showing more stability over time. The shallower angle indicates that it fails to capture the full effect of the changing pattern in the political sophistication of language shown by our technique. (In Supporting Information G, we provide regression-based details, based on detrending the time series, that brings this difference in magnitude changes into starker relief.) In sum, our model is more sensitive to changing patterns in linguistic complexity in the Hansard example, because it was fit to the specific context

required.

6 Summary and Discussion

The nature of the messages that political actors send one another are of key interest to political science, whether it be in American politics, international relations or from a comparative perspective. Yet a curious gulf has emerged in our studies. On the one hand, we have plenty of theory and empirical evidence that such communication matters: whether it be "dog whistle" in nature (Albertson, 2015), rhetorical (Riker, 1996), vague (Lo, Proksch and Slapin, 2016), or more explicitly designed to appeal to certain types of agents. On the other hand, the discipline has been slow to adopt textual complexity measures in any context, preferring instead to code documents using pre-existing dictionaries. This is despite the fact that the various readability measures are easy to use and scale in a straightforward way—which is important, given the sheer amount of textual data now available to scholars. Presumably, part of this reticence is lack of familiarity with such approaches. But part of it is likely a very reasonable skepticism about the merits of these educational measures—a concern echoed in other fields of social science (e.g. Sirico, 2007; Loughran and McDonald, 2014) and indeed, increasingly in education itself (Ardoin et al., 2005).

Rather than attempt to rehabilitate the indices, here we focused on producing something better: Table 4 summarizes our contribution with respect to the problems we raised in Section 2.

Table 4: Summary of our approach as a solution to a series of problems with traditional approaches.

Problem with traditional approach	**Solution via our approach**
1. Designed for education	1. Designed for *politics*
2. Tested/validated on children	2. Tested/validated on *adults*
3. Designed for readers in 1940/50s, not easily updated	3. Designed for *contemporary* readers, easy to update (via crowdsourcing).
4. Cannot assess quality/fit of predictions for documents	4. Straightforward to assess *absolute model fit* (in training set) via usual metrics like percent correctly predicted
5. Cannot compare models of different forms	5. Straightforward to assess *relative model fit* (in training set) via usual metrics like AIC, BIC.
6. Cannot interpret fine-grained differences in document scores	6. Natural *model-based interpretation* of document estimates (via Bradley-Terry model).
7. No uncertainty around estimates.	7. Uncertainty *estimates available* both for variables in model, and on document scores (via bootstrap).
8. Composite indices/aggregate form hides changes in variables "under the hood."	8. Straightforward to examine all *changes to component parts*.
9. Rarity of terms accounted for in *ad hoc* inflexible way, if at all.	9. Rarity of terms *systematically derived* from large corpus, and available for any period of interest in past 200 years.

In particular, we used human coders (via the crowd) to provide relative assessments of short texts, and from there we built a well-defined statistical model. That model uses variables that differ from standard approaches, including word rarity and parts-of-speech information. The final version performs better in fit terms too, although precisely because the approach is on much firmer probabilistic grounds it is hard to compare directly to previous approaches. Fundamentally then, we believe we have improved practice here: the approach is transparent, sensible and model-based and trained on relevant domain data. It is also flexible, in the sense that the workflow and software we have designed allows end-users to calibrate the method to their specific problems.

While our contribution is helpful for those interested in communication in politics, it is hardly the last word on the matter. We have provided a statistical machinery, and variables, for thinking more carefully about the measurement of sophistication or clarity in texts. What we have not done is produced a straightforward way to distinguish between more subtle understandings of such con-

cepts. For example, one can imagine a politician—a president of the United States even—who uses relatively common terms in simple sentence constructions, but is not especially clear. By contrast, great academic writers might be able to describe extremely complicated ideas in straightforward ways for popular audiences. Our approach would generally be better than previous ones, but is still unlikely to place these two extremes correctly on the same scale. This is, of course, because a sophisticated idea (like democracy, or inclusivity or conservatism) need not be complicated in expression, and vice versa. More attempts should be made—not least at the coding/crowdsourcing level—to iron out these differences, possibly by introducing different dimensions of complexity at the point of testing or modeling. We leave such efforts for future work.

References

Albertson, Bethany. 2015. "Dog-Whistle Politics: Multivocal Communication and Religious Appeals." *Political Behavior* 37(1):3–26.

Anderson, Jonathan. 1983. "Lix and Rix: Variations on a Little-known Readability Index." *Journal of Reading* 26(6):490–496.

Ardoin, Scott P, Shannon M Suldo, Joseph Witt, Seth Aldrich and Erin McDonald. 2005. "Accuracy of Readability Estimates' Predictions of CBM Performance." *School Psychology Quarterly* 20(1):1.

Bai, Jushan and Pierre Perron. 2003. "Computation and analysis of multiple structural change models." *Journal of Applied Econometrics* 18(1):1–22.

Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2).

Benoit, Kenneth, Kevin Munger and Arthur Spirling. Forthcoming. "Dumbing Down? Trends in the Complexity of Political Communication." `http://kmunger.github.io/pdfs/BenoitMungerSpirling_SSRCchapter.pdf`. Prepared for 'Anxieties of Democracy' volume (editors Frances Lee and Nolan McCarty).

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Bischof, Daniel and Roman Senninger. Forthcoming. "Simple politics for the people? Complexity in campaign messages and political knowledge." *European Journal of Political Research* .

Bradley, Ralph and Milton Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39(3/4):324–345.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Cann, Damon, Greg Goelzhauser and Kaylee Johnson. 2014. "Analyzing Text Complexity in Political Science Research." *PS: Political Science & Politics* 47:663–666.

Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28(3):591–605.

Coleman, M and T Liau. 1975. "A computer readability formula designed for machine scoring." *Journal of Applied Psychology* 60(2):283–284.

Dale, Edgar and Jeanne Chall. 1948. "A Formula for Predicting Readability." *Educational Research Bulletin* 27(1):11–20.

Diamond, Larry. 2002. "What Political Science Owes the World." *PS: Political Science & Politics Online Forum* pp. 113–27.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

Flesch, Rudolf. 1949. *The Art of Readable Writing*. New York: Harper.

Flesch, Rudolph. 1948. "A new readability yardstick." *Journal of Applied Psychology* 32(3):221–233.

Fry, Edward. 1968. "A Readability Formula That Saves Time." *Journal of Reading* 11(7):513–578.

Fucks, Wilhelm. 1955. *Unterschied des Prosastils von Dichtern und anderen Schriftstellern: ein Beispiel mathematischer Stilanalyse*. Bouvier.

Gatto, John Taylor. 2002. *Dumbing us down: The hidden curriculum of compulsory schooling*. Vancouver: New Society Publishers.

Gunning, Robert. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.

Heath, Oliver. 2015. "Policy Representation, Social Representation and Class Voting in Britain." *British Journal of Political Science* 45(1):173193.

Heath, Oliver. 2016. A growing class divide: MPs and voters. In *Sexier Lies and the Ballot Box*, ed. Philip Cowley and Robert Ford. Biteback.

Jansen, David-Jan. 2011. "Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies." *Contemporary Economic Policy* 29(4).

Kincaid, J Peter, Robert Fishburne, Richard Rogers and Brad Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel*. Vol. Research Branch Report 8-75 Naval Air Station Memphis: Chief of Naval Technical Training.

Klare, George. 1963. *The measurement of readability*. Ames, Iowa: University of Iowa Press.

Kristof, Nicholas. 2014. "Professors, We Need You!" https://www.nytimes.com/2014/02/16/opinion/sunday/kristof-professors-we-need-you.html. New York Times, Online, February 15, 2014.

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.

Lim, Elvin. 2008. *The Anti-Intellectual Presidency*. New York: Oxford University Press.

Lo, James, Sven-Oliver Proksch and Jonathan B Slapin. 2016. "Ideological clarity in multiparty competition: A new measure and test using election manifestos." *British Journal of Political Science* 46(3):591–610.

Loewen, Peter, Daniel Rubenson and Arthur Spirling. 2012. "Testing the power of arguments in referendums: A Bradley–Terry approach." *Electoral Studies* 31(1).

Loughran, Tim and Bill McDonald. 2014. "Measuring Readability in Financial Disclosures." *The Journal of Finance* 69(4):1643–1671.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

McCullagh, Peter and John Nelder. 1989. *Generalized linear models*. New York: CRC press.

Michalke, Meik. 2015. *koRpus: An R Package for Text Analysis, V0.05-6*. https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.pdf.

Montgomery, Jacob and David Carlson. Forthcoming. "Human computation scaling for measuring meaningful latent traits in political texts." *American Political Science Review* . Accessed October 30, 2017: http://pages.wustl.edu/montgomery/sentimentit.

Norris, Pippa and Joni Lovenduski. 1995. *Political Recruitment*. Cambridge: Cambridge University Press.

Owens, Ryan and Justin Wedeking. 2011. "Justices and Legal Clarity: Analyzing the Complexity of Supreme Court Opinions." *Law & Society Review* 45(4):1027–1061.

Pitkin, Hanna. 1967. *The Concept of Representation*. Berkeley, CA: University of California Press.

Rheault, L, Beelen K, Cochrane C and Hirst G. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLOS ONE* 11(12).

Riker, William H. 1996. *The strategy of rhetoric: Campaigning for the American Constitution*. New Haven: Yale University Press.

Sherman, Lucius. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.

Sirico, Louis J. 2007. "Readability Studies: how technocentrism can compromise research and legal determinations." *QLR* 26:147.

Spache, George. 1953. "A new readability formula for primary-grade reading materials." *The Elementary School Journal* 53(7):410–413.

Spirling, Arthur. 2016. "Democratization of Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *Journal of Politics* 78(1):120–136.

Spriggs, James F. II. 1996. "The Supreme Court and Federal Administrative Agencies: A Resource-Based Theory and Analysis of Judicial Impact." *American Journal of Political Science* 40:1122–1151.

Thurstone, L. L. 1927. "A law of comparative judgment." *Psychological Review* 34(4):273–286.

Tränkle, U. and H. Bailer. 1984. "Kreuzvalidierung und Neuberechnung von Lesbarkeitsformeln für die deutsche Sprache." *Zeitschrift f˙ür Entwicklungspsychologie und Pädagogische Psychologie* 16(3):231–244.

Turner, Heather and David Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package." *Journal of Statistical Software* 48(1):1–21.

Wheeler, Lester and Edwin Smith. 1954. "A practical readability formula for the classroom teacher in the primary grades." *Elementary English* 31:397–399.

Yuka, Tateisi, Ono Yoshihiko and Yamada Hisao. 1988. A Computer Readability Formula of Japanese Texts for Machine Scoring. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*. COLING '88 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 649–654.

Zeileis, Achim, Friedrich Leisch, Kurt Hornik and Christian Kleiber. 2002. "strucchange: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software* 7(2):1–38.

Supporting Information

A Details on crowd-sourcing, "gold questions" and snippet construction

We labeled the task as "Identify Which of Two Text Segments Contains Easier Language." Upon accepting the task, we provide the workers with a number of example comparisons, with one option correctly labeled as more complex. The specific instructions provided to each worker were:

> Your task is to read two short passages of text, and to judge which you think would be easier for a native English speaker to read and understand. An easier text is one that takes a reader less time to comprehend fully, requires less re-reading, and can be more easily understood by someone with a lower level of education and language ability.

A crucial aspect of crowdsourcing any coding operation is ensuring that workers provide high quality responses. To that end, we employ "gold standard" tasks: tests in which one snippet is unambiguously easier than the other, interspersed with normal rating tasks at a rate of one in ten. To create the gold standard test questions, we select the snippet pairs with the largest disparity in FRE scores, verified through inspection. Prior to being accepted for the task, a crowd worker had to pass a qualification test consistently entirely of test questions, answering at least 7 of 10 correctly. Following successful qualification, a coder performed job lots of ten pairwise comparisons, where one of these was a test question. Workers who did not maintain an overall accuracy rate of 30% correct on the test questions were removed from the pool of workers and their answers dropped from the dataset.[23]

To create the snippets, we formed two-sentence segments from the State of the Union corpus, with three levels of ranges of the total number of characters: between 345–360, 360–375, and 375–390 in length, from which we randomly selected 2000 pairs of snippets for direct comparison, in a way that guaranteed the connectivity of pairs for comparison to enable Bradley-Terry scaling.[24]

[23] Following Berinsky, Margolis and Sances (2014), we also included some "screener" questions, which appear to be the same as normal comparisons but include at some point the phrase "Disregard the content and code this sentence as EASIER." Of the test questions, approximately 10% were screeners.

[24] To increase the range of data and to use results from an pilot study of coding, we also combined the post-1950

Finally, we added another 15% of gold questions plus 5% of special gold "screener" questions. After removing duplicates, our dataset of snippets to be compared consisted of 7,236 total pairings for comparison, including 836 "gold" questions, of which 310 were screeners. We crowd-sourced the comparisons using a minimum of three coders per pair, yielding 19,810 total comparisons, of which 13,430 did not involve screeners or test questions. To aid the automation of this process and to provide both reproducibility and transparency, we implemented all of the functions to sample snippets, create pairs and test questions, prepare the data for Crowdflower, and to process the crowd-coded data in an R package sophistication, which also includes the cleaned version of the SOTU corpus.

B Details on using the Google-books corpus

After filtering out tokens that occurred fewer than five times or that did not match a dictionary of 133,000 English words and word forms, we ended up a table of frequencies for 82,558 unique word types from the total corpus.[25]

To see how this works, consider the following two snippets:

> Numerous are the providential blessings which demand our grateful acknowledgments. . . too important to escape recollection. (George Washington, 1791)

> Now, we have to build a fence. And it's got to be a beauty. (Donald Trump, 2015)

These are 15 and 14 tokens in length, but the mean frequency relative to *the* in the 2000s for the first was 0.11, and 0.14 for the second, indicating that the mean word in Washington's speech was relatively much less frequently used than in Trump's. The word that is used least commonly (relative to *the* in the 2000s) in the two snippets induces a large difference in the measurements of the texts: for Washington, it is *providential* which has a ratio of 0.00002085 relative to *the*

texts with some with one- and two-sentence snippets from an earlier set of crowd work. This earlier set used a range of 180–300 characters and 180–400 characters respectively, but our dataset included just nine unique snippets, used in 99 different comparisons with post-1950 snippets, and in all of the 36 pairwise comparisons against one another.

[25] This was a fairly massive reduction from the over 615 billion term counts in the original term-year dataset. One reason for the massive drop in the number of word types is that many appear to be artifacts of errors introduced in optical character recognition.

(implying *the* is used about 48,000 times as often). For Trump, the relevant word is *fence*, for which the ratio is an order of magnitude higher, at 0.00025 (meaning *the* is used about 4000 times as often). (We note also that the Flesch Reading Ease for the Washington text is 5.5, compared to 105.1 for the Trump snippet.)

C Details on obtaining part-of-speech information

We began by tagging the snippets using the Google Universal tagset[26] using the spacyr package built on the spaCy NLP library for Python.[27] This follows some readability indexes, such as Tränkle and Bailer (1984), that consider conjunctions and prepositions, and Coleman's "C3" and "C4" indexes (Coleman and Liau, 1975) that take into account the frequency of pronouns and propositions. Converting these to relative frequencies for each snippet gave us the information required.

D Comparing the standard measures

In Table 5 we consider two natural ways to compare the fit of the standard approaches in the literature. For each of the traditional measures, we fit a Bradley-Terry model which has one predictor: the score for the snippets on a given measure. Thus, the first row refers to a model in which the only covariate is the (difference in the) snippets' Flesch scores (a model we return to below), the second row refers to a model in which the only covariate is the (difference in the) snippets' Dale-Chall scores, and so on. We report the Akaike information criterion for each of these models, along with the proportion of contests correctly predicted by the model. This latter statistic is calculated by generating the relevant λ_is from the linear predictor, using the $\hat{\beta}$ from the model, multiplied by inputs for a given snippet. We then calculate the probability that the snippet which actually won a contest would be expected to do so given the estimated parameters—in the sense of Equation 1. If

[26] See `https://github.com/slavpetrov/universal-pos-tags`.
[27] See `http://spacy.io`.

this probability is greater than 0.5, then we declare that a success for the model.

Table 5: Model performance of the standard measures. The overall fit of the Bradley-Terry model using the scores for a given measure is reported in two ways: the Akaike information criterion (AIC) and the Proportion of contest results correctly predicted (where a correctly predicted contest is one in which there is > 0.5 probability that the actual winner would win).

	AIC	Proportion Correct
FRE	26269.2	0.568
Dale-Chall	26227.9	0.573
FOG	26084.8	0.573
SMOG	26188.2	0.526
Spache	26025.6	0.577
Coleman-Liau	26574.4	0.550

E Random forest variable importance plots

As noted in text, we ran our random forest model (1000 trees, otherwise standard defaults in the sense of Liaw and Wiener (2002)) for both sets of unstructured estimates—that is, with and without bias-reduction. The results of that process, in terms of the variable importance plots, are given in Figure 7. As usual, variables (on the *y*-axis) with points further right are deemed "more important" for predicting the outcome (here, the snippet's ability). Notice that the ordering of the variables is similar, regardless of which approach we take (i.e. with or without bias reduction).

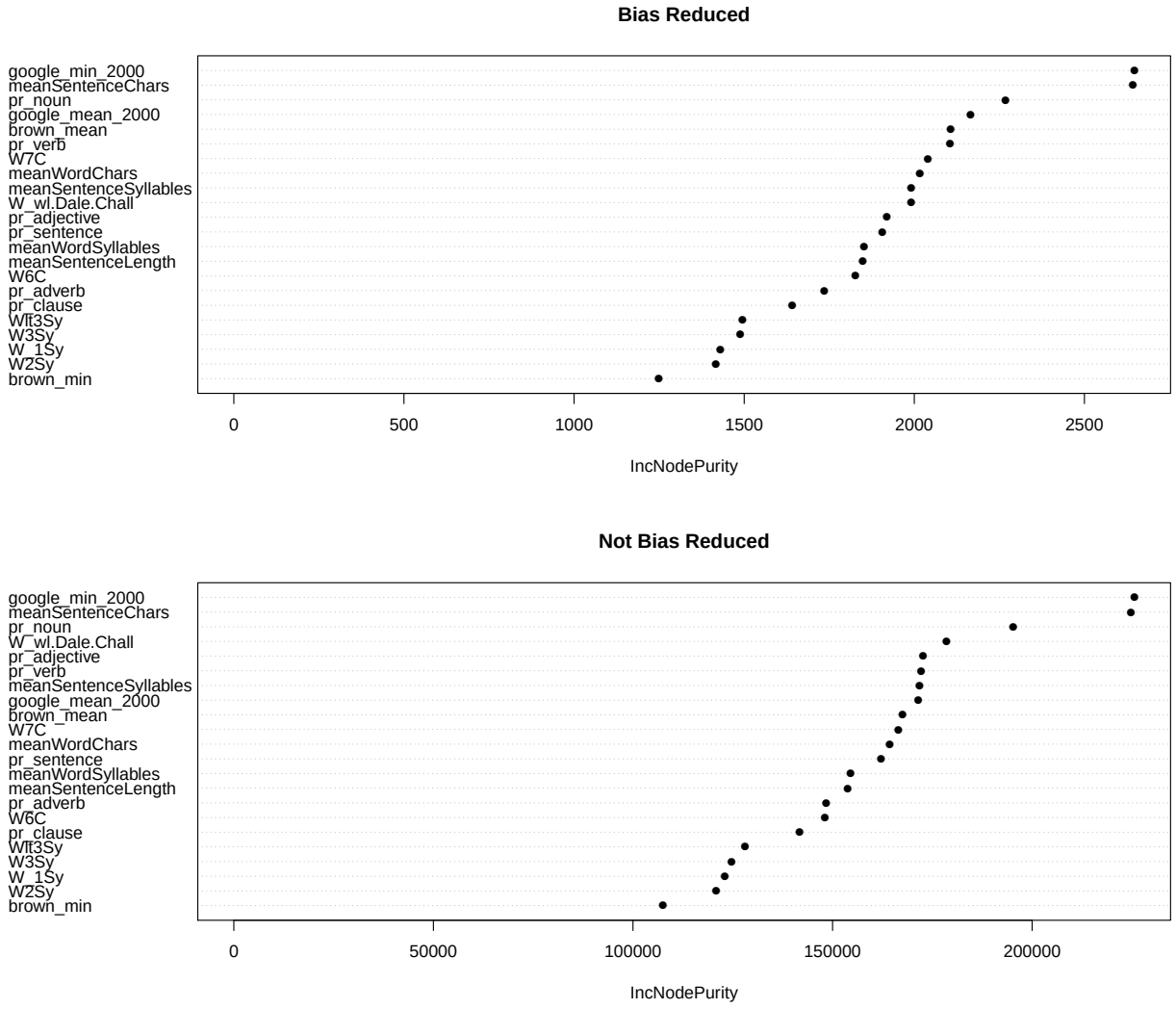


Figure 7: Variable Importance Plots for (unstructured) readability estimates. Note that points further to the right imply more important variables. Top panel is for bias-reduced estimates; bottom panel is for non-bias reduced estimates.

F An alternative continuous measure

There are ways to rescale the λ estimates that may be of greater theoretical appeal. To see this, using Equation 3 denote the $\Pr(i \text{ easier than } j)$ term as p. Then, supposing that we have an appropriate example of a (set of) fifth grade text(s), we can substitute $\exp(\lambda_i)$ for 100 (or, indeed, any number preferred) and then rescale $\exp(\lambda_j)$ as $100 \times (\frac{1}{p} - 1)$. Though this preserves the model-based interpretation of the quantity of interest, in practice it tends to return quite low numbers once

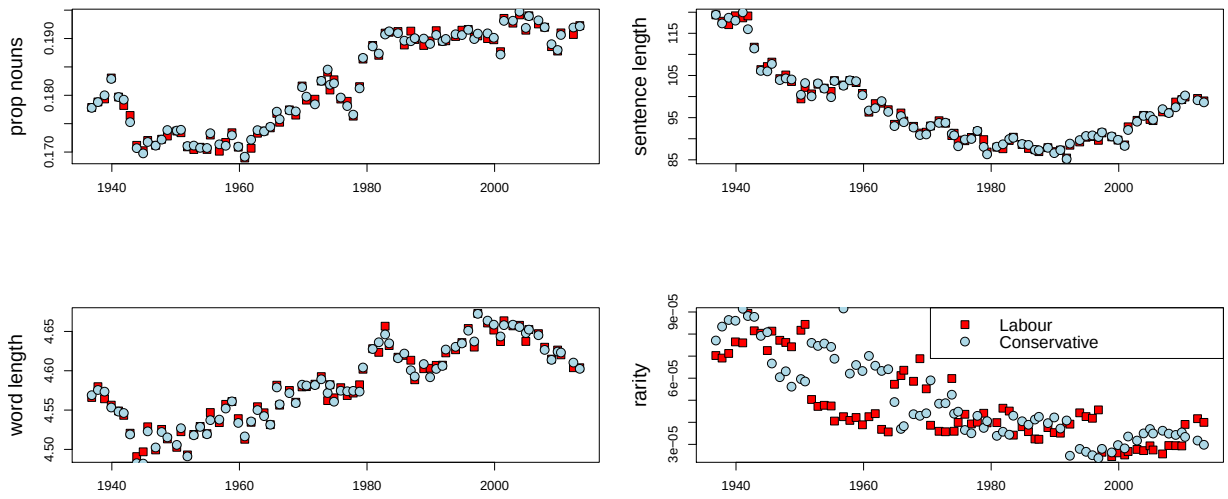


Figure 8: Disaggregation of speech difficulty by party over time (Conservative vs Labour). Note that the parties are essentially similar on all components, except rarity of speech.

one is even slightly removed from a fifth grade text. For example, a spotcheck on a document with an FRE of around 84 implies a rescaled score of 35, which seems very low. Again, this is not wrong—it is simply rescaling in a way that preserves the probability structure inherent in the model. But it may well be confusing for end-users, who expect a number approximately commensurate with the original interpretation given by Flesch.

G Disaggregation of Conservative vs Labour patterns in government

Above, we noted that there is apparent periodicity in the time series of Conservative vs Labour (mean) speech difficulty. In particular, we noted that when a party is in government, its speeches tend to be harder to understand. To see why this might be, in Figure 8 we disaggregate our measure (for the mean speech) into its component parts, and divide out the data into Conservative and Labour means.

Clearly, the time series overlap: the [blue] Conservative circles overlap with the [red] Labour

squares everywhere with the exception of the bottom right—which is our measure of rarity. Looking at that subplot, we see the following pattern: prior to around 1945, when the Conservatives are in government, their (average) word rarity is larger, meaning they use terms that are more common than those used by Labour. The next five years (when Labour are in government) sees Labour using less rare words. Then when the Conservatives are in government in the 1950s and early 1960s, they use more common words. Labour switches to being the party that uses more common words after that (with the exception of the early 1970s when the Tories are in power for four years). By the 1980s—a period in which the Conservatives are completely dominant—the parties are more similar and almost overlap in word rarity terms; meanwhile, in aggregate, the rarest words used become more rare (the level shifts down over time). This pattern continues until the end of the data, although we note that Labour is generally below the Conservatives everywhere after around 1985 (the very end of the data being an exception).

Analysis of Detrended Data

After detrending the two time series (our estimates and those from the FRE ratio), we fit two linear regressions of the form $Y = \beta_0 + \beta_1 X_{\text{after 1997}}$. Here Y is the relevant measure, and $X_{\text{after 1997}}$ is dummy taking the value 1 if the session occurs after the Labour landslide of 1997, and 0 otherwise. If we think the 1980s was the key period of modernization for Labour, and was also a time of changing recruitment, then it makes sense to investigate the extra effect of time once Labour came to power after a break of 18 years.

Unsurprisingly, given the theory that Labour elites were now more similar to their Conservative peers, for both measures there is a negative effect of Labour gaining power in 1997. However, the coefficient for our (detrended) measure ($\hat{\beta_1} = -0.0167$) is about four times as large (in addition the model fit is better, and the p-value smaller) as that for the FRE (-0.0043). In that sense, then, our measure is more sensitive to the advent of new Labour elites than the most common extant approach.

Modélisation Conceptuelle des Bases Prosopographiques - Représentation de l'Information Incertaine

Jacky Akoka[*,**], Isabelle Comyn-Wattiau[***], Stéphane Lamassé[****], Cédric du Mouza[*]

[*]Laboratoire CEDRIC, CNAM
prenom.nom@lecnam.net
[**] Télécom École de Management, Institut Mines-Télécom
Jacky.akoka@telecom-em.eu
[***]ESSEC Business School
wattiau@essec.edu
[****]Laboratoire PIREH, Univ. Paris I
stephane.lamasse@univ-paris1.fr

1 Introduction

La prosopographie est une méthode permettant d'étudier un groupe social en comparant les itinéraires biographiques particuliers de chacun de ses membres. Il s'agit d'approcher un groupe, d'en comprendre les fonctionnements, sans en négliger les trajectoires singulières. Elle repose sur une enquête précise, documentée, de chaque individu de la population déterminée.

En histoire, c'est grâce à une méthodologie et une érudition pointue que l'on collecte l'ensemble des traces qui vont constituer la fiche de chaque personne. Toutes les périodes historiques utilisent cette méthode d'investigation. Si la méthode est ancienne, le mot "prosopographia" apparaît au XVIe siècle, l'analyse quantitative et l'ordinateur ont profondément transformé sa méthodologie au XXe siecle. De nombreux périodiques se sont intéressés à cet aspect en proposant des articles sur ce thème. Plusieurs historiens ont même proposé des développements de logiciels dédiés. On retiendra surtout l'apport des bases de données qui ont permis, par exemple, d'aborder très concrètement le "sourçage" de l'information. Sur les fiches papier, il est difficile d'indicer chaque fait constituant la carrière d'une personne avec le document qui a permis de l'établir. De la même façon, il n'est pas simple de gérer une information contradictoire. Or il est, en effet, possible que deux documents différents apportent des informations contradictoires sur un individu.

Dans cet article nous proposons un modèle conceptuel permettant de décrire de manière générale et enrichie l'information contenue dans une base de données prosopographique. Nous étudions ensuite comment ce modèle peut être instancié avec les bases de données du projet PASE (Bradley et Short, 2005) et STUDIUM (Genet et al., 2016).

État de l'art

Passer d'une collection de fiches à des bases de données implique d'abord une réflexion sur le modèle de données. Les premières propositions de bases de données prosopographiques

se sont appuyées sur le modèle relationnel comme par exemple (Keats-Rohan, 1998). Des travaux récents (Bol, 2012) proposent l'utilisation de systèmes d'information géographiques, utilisant une base de données relationnelle, afin de détecter par exemple des motifs spatiaux. Cette représentation structurée permet d'effectuer des interrogations efficaces, croisant un nombre restreint de tables. La représentation semi-structurée, en plus d'un apport sémantique, permet de limiter les jointures en exploitant la structure arborescente. Elle autorise ainsi les attributs multivalués et l'intégration d'objets (semi-) structurés au sein d'un objet (semi-)structuré. Elle est donc adaptée aux bases prosopographiques ou un élément "personne" peut ainsi être composé d'éléments "production", "diplôme", etc. eux-mêmes éléments structurés. Les projets STUDIUM (Genet et al., 2016) et PROSO (Barabucci et Zingoni, 2013) sont deux exemples d'un tel choix de représentation. Si le modèle semi-structuré permet de représenter des liens entre personnes/objets/lieux/faits, il permet difficilement d'interroger des liens plus complexes entre éléments. Pour cette raison, d'autres travaux récents appliquent des représentations du type "réseaux sociaux" , par exemple (Graham et Ruffini, 2007; Verbruggen, 2007). Cette approche permet notamment la fouille de données pour découvrir des liens entre personnes/objets/lieux/faits, ou des motifs récurrents.

Une des problématiques importantes des bases de données en général, et des bases prosopographiques en particulier, est la qualité des informations stockées. La qualité des données est un domaine de recherche en soi. De nombreuses contributions ont permis de dresser une catégorisation des problèmes de qualité, des métriques permettant de mesurer l'étendue de ces problèmes et des méthodes et outils pour l'améliorer (Berti-Equille, 2012). Pour des raisons d'espace, nous mentionnons ici uniquement quelques aspects qui nous semblent pertinents dans le contexte des sciences humaines. Les attributs d'une entité peuvent avoir des valeurs floues. (Urrutia et al., 2002) classent par exemple ces attributs en quatre types. (Matousek et al., 2007) proposent une catégorisation des assertions temporelles imprécises. (Plewe, 2002) propose un modèle sur la nature de l'incertitude, spécifiquement pour la représentation thématique, spatiale et temporelle des phénomènes géo-historiques. Le but est de fournir un cadre pour la modélisation de données spatio-temporelles dans un cadre historique. A notre connaissance, il n'existe pas de base de données prosopographique intégrant la représentation de l'information incertaine au niveau du modèle. Certaines, telle STUDIUM, insère des marques (principalement le point d'interrogation, ou la langue naturelle) pour alerter l'utilisateur sur le caractère incertain de l'information. Cependant, cette représentation artisanale ne permet pas l'évaluation de la certitude associée à l'information correspondante.

2 Modélisation conceptuelle de bases prosopographiques

Tous les modèles de bases prosopographiques sont bâtis autour de quatre concepts principaux : les personnes, les événements, les lieux et les sources. PASE utilise le modèle factoïde pour représenter les faits et les événements et incorpore des informations très détaillées sur les sources. STUDIUM, qui cible la vie des universitaires du Moyen Age, accorde une importance particulière à la description des cursus et diplômes des personnes.

Le modèle proposé dans cet article représente de façon générique et enrichie les quatre concepts de base (Fig. 1). Il intègre de plus une représentation de l'incertitude associée. La notion d'événement est prise au sens large. Elle inclut les factoïdes, mais aussi l'ensemble des faits qui caractérisent les individus, par exemple une publication est aussi un événement.

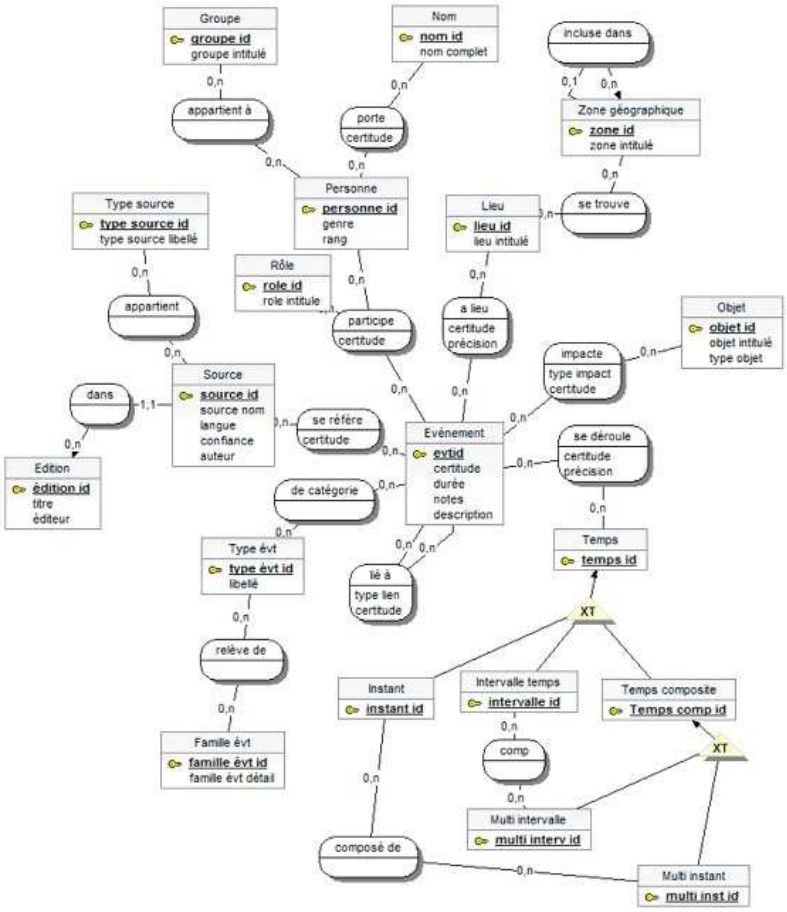


FIG. 1 – *Modèle conceptuel d'une base prosopographique*

Le choix de généraliser l'événement permet un modèle compact sans perdre pour autant la richesse de l'information que l'on peut ainsi représenter. Toutefois, il nous a conduits à définir l'événement avec un nombre plus important de dimensions. Par exemple, nous avons introduit une entité *Objet* qui peut représenter une publication, un bien, etc. Ainsi, le fait qu'un événement impacte un objet nous permet de couvrir : la publication écrite par un auteur, l'achat d'un bien, la dot lors d'un mariage, etc. Les dimensions de l'événement et des autres concepts prosopographiques sont associées à des référentiels hiérarchiques. Par exemple, les lieux, les sources, les personnes et les événements sont généralisés à un ou plusieurs niveaux. Les événements sont regroupés en types d'événements puis en familles, à l'image de PASE ou la confession est un événement de piété chrétienne (type), lui-même un acte religieux (famille). Le critère de regroupement est défini par le concepteur de la base prosopographique en fonction de l'objectif recherché. Selon le domaine ciblé par la base prosopographique, un individu peut

être connu sous plusieurs noms. Aussi, notre modèle représente le nom comme une entité en relation N-N avec la personne. Chaque nom potentiel connu est associé à la personne avec, si elle est disponible, une mesure de la certitude. La plupart des relations entre les concepts sont typées, dans le sens ou un attribut *Type* les décrit. Par exemple, le type d'impact entre l'événement et l'objet permet de préciser que, lors d'un événement de troc, un objet est cédé et un objet est accordé en échange. Le type d'impact peut prendre la valeur "dot" lors d'un mariage. Les événements peuvent être liés entre eux : l'information "type de lien" permet de caractériser ce lien et peut prendre une valeur comme précède, provoque, etc. Le rôle d'une personne dans un événement est aussi un type qui a été représenté sous la forme d'une entité dans la mesure ou une même personne peut parfois jouer plusieurs rôles dans le même événement. La représentation du temps intègre le temps discret (une date), le temps continu (un intervalle) et leur composition (plusieurs dates potentielles, ou plusieurs intervalles possibles, ou plusieurs intervalles cumulés, par exemple il a été présent de 1492 à 1500 puis de 1503 à 1508). Elle est reprise du modele d'AROM-ST Page et al. (2001).

Notre modèle intègre également la gestion de l'information incertaine sous trois formes : une précision, une confiance et une certitude. La précision permet d'attribuer une mesure approximative à l'information. Dans les bases prosopographiques, elle permet de représenter certaines informations spatio-temporelles. Par exemple, la précision peut être relative à l'endroit ou à lieu un événement. Les valeurs qu'elle peut prendre dans ce cas sont : près de, aux alentours de, non loin de, à quelques kilomètres de, etc. Quand elle caractérise le moment ou se déroule un événement, elle peut prendre les valeurs de : aux environs de, avant, bien avant, peu de temps après, etc. La confiance est une caractéristique partagée d'une information, représentée par un degré entre 0 et 1. Dans ce modèle, nous avons restreint son utilisation à la caractérisation des sources d'information, dans la mesure ou c'est la principale information disponible. Enfin, la certitude est une représentation du degré de fiabilité de l'information à laquelle elle est attachée. Généralement, elle prend sa valeur dans l'intervalle [0,1]. Ce peut être le cas, par exemple, lorsque deux documents donnent une information différente et une datation relative au document terminus ante quem ou post quem. L'ensemble des documents concernant Johannes Vitalis nous permettent de dire que son activité se situe entre 1380 et 1395. Il est connu comme franciscain, un ordre mendiant. On sait qu'il a été bachelier, licencié en théologie. Il est cité comme docteur en théologie lors d'une demande de pardon entre le 8 et 11 septembre 1390 d'un autre frère dominicain Johannes Nicolai. On peut donc penser qu'il a obtenu son grade avant ce moment. Puis on le retrouve au procès de Jean Blanchard et dans la convocation des étudiants en théologie pour le procès ou il est cité comme dominicain, ce qui est probablement une erreur. Ainsi, on pourra par exemple associer une certitude à l'information selon laquelle Johannes Vitalis est franciscain et une certitude moindre à celle selon laquelle il serait dominicain. C'est dans le mode d'interrogation de cette base que l'on exploitera cette apparente contradiction. Par exemple, on pourra définir un seuil de certitude en-dessous duquel une information est considérée comme ne devant pas être fournie ou, au contraire, fournir toutes les réponses possibles avec les certitudes associées.

3 Instanciation avec la base PASE et STUDIUM

Le Tableau 1 est un résultat de la comparaison de notre modèle avec PASE (extrait du mapping). Les deux premières colonnes désignent l'entité ou la relation dans notre modèle et la

propriété associée. Les deux dernières désignent la table et la colonne correspondante dans PASE. A titre d'exemple, les groupes de personnes dans notre modèle correspondent aux types représentés dans la table *alfactoidpersontype*. Cet effort de mise en correspondance des deux modèles nous a permis de vérifier que notre modèle intègre toutes les informations de PASE. De plus, l'ajout de certaines dimensions aux événements améliore la représentation de l'information. Par exemple, l'entité OBJET qui permet de structurer la description de certains événements (diplômation, mariage, etc.) évite la description en langue naturelle dans des champs peu structurés, plus difficiles à exploiter par les requêtes.

Objet	Propriété	PASE objet	PASE propriété
Groupe	groupe id	alfactoidpersontype	alfactoidpersontypekey
Groupe	groupe intitulé	alfactoidpersontype	alfactoidpersontype
Nom	nom id	Person	headname
Nom	nom complet	Person	descriptionname
Zone géographique	zone id	allocation	allocationkey
Zone géographique	zone intitulé	allocation	allocation
Type source	type source id	alsourcetype	alsourcetypekey
Type source	type source libellé	alsourcetype	alsourcetype
Personne	personne id	Person	personkey
Personne	genre	AlGender	AlGenderAbrv
Personne	rang	alfactoidpersonrank	alfactoidpersonrank
Lieu	lieu id	factoidlocation	factoidlocationkey
Lieu	lieu intitulé	factoidlocation	alplace
Objet	objet id	Possession	possessionkey
Objet	objet intitulé	Possession	description
Objet	type objet	alpossessiontype	alpossessiontype
Source	source id	Source	sourcekey
Source	source nom	Source	sourcetitle
Source	description	Source	description
Source	auteur	Source	author
Source	langue	alLanguage	allanguage
Source	confiance	Archivequality	archivequalityname
Edition	édition id	Editioninfo	editioninfokey
Edition	titre	Editioninfo	articletitle
Edition	éditeur	Editioninfo	editor
Evénement	evt id	Factoid	factoidkey
Evénement	description	Factoid	shortdesc

TAB. 1 – *Extrait du mapping entre notre modèle et* PASE

De la même façon, nous avons effectué la comparaison entre notre modèle et celui de STUDIUM. Ainsi, les variantes du nom que permet STUDIUM sont représentées, dans notre modèle, par la relation PORTE entre les personnes et les noms. La période d'activité de STUDIUM est représentée par l'événement ACTIVITÉ avec une date de début et une date de fin. La médiane d'activité est une information calculée à partir de ces dates. Le statut d'une personne dans STUDIUM correspond à son rang dans notre modèle. L'information *Bachelier es arts (Paris) 1460* dans STUDIUM correspond à un événement de diplômation ayant lieu a Paris en 1460.

Finalement, notre approche présente l'avantage d'offrir un modèle générique pour toutes ces bases, ce qui permet de mutualiser les efforts de développement et de maintenance. Ainsi, les différentes communautés d'historiens disposeraient chacune de leur base spécifique (PASE, STUDIUM, PBW, etc.), laquelle résulterait de l'adaptation de ce modèle générique à leurs besoins de recherche. De plus, la gestion de l'information incertaine permet une interrogation de meilleure qualité, associant à chaque réponse une certitude.

4 Conclusion

Cet article propose un modèle conceptuel générique couvrant les concepts et relations entre concepts présents dans les différents modèles de données prosopographiques (nous avons vu que ce modèle généralise et enrichit ceux de PASE et de STUDIUM par exemple), mais se distinguant par ailleurs par sa représentation de la qualité des données, comme par exemple l'incertitude, la fiabilité ou la précision. Nos recherches futures vont consister à valider le modèle en le confrontant à d'autres références dans le domaine des bases prosopographiques, à vérifier son applicabilité en le transformant en modèle logique et physique (relationnel, graphe ou document par exemple). Cet article a mis en avant la représentation de l'incertitude, enrichissant les possibilités offertes par les bases prosopographiques. Il reste à intégrer cet aspect dans l'interrogation en définissant les modes d'agrégation de ces représentations de l'incertain.

Références

Barabucci, G. et J. Zingoni (2013). PROSO : prosopographic records. In *Proc. Intl Work. on Collaborative Annotations in Shared Environment, DH-CASE@DocEng*, pp. 3 :1–3 :7.

Berti-Equille, L. (2012). *La Qualité et la Gouvernance des Données : Au Service de la Performance des Entreprises*. Informatique et SI, Recherche, technologie, applications. Lavoisier.

Bol, P. K. (2012). GIS, Prosopography and History. *Annals of GIS 18*(1), 3–15.

Bradley, J. et H. Short (2005). Texts into Databases : The Evolving Field of New-style Prosopography. *Literary and Linguistic Computing 20*(Suppl 1), 3–24.

Genet, J.-P., H. Idabal, T. Kouamé, S. Lamassé, C. Priol, et A. Tournieroux (2016). General Introduction to the Studium Project. *Medieval Prosopography* (31), 156–172.

Graham, S. et G. Ruffini (2007). Network Analysis and Greco-Roman Prosopography. In *Prosopography Approaches and Applications. A Handbook.*, pp. 325–336. K.S.B. Keats-Rohan, (ed.).

Keats-Rohan, K. (1998). Historical Text Archives and Prosopography : the COEL Database System. *History & Computing 10*(1-2-3), 57–72.

Matousek, K., M. Falc, et Z. Kouba (2007). Extending Temporal Ontology with Uncertain Historical Time. *Computing and Informatics 26*(3), 239–254.

Page, M., J. Gensel, C. Capponi, C. Bruley, P. Genoud, D. Ziébelin, D. Bardou, et V. Dupierris (2001). A New Approach in Object-Based Knowledge Representation : The AROM System. In *Intl. Conf. on Indus. and Eng. Applcations of Artificial Intelligence and Expert Systems*, pp. 113–118.

Plewe, B. (2002). The Nature of Uncertainty in Historical Geographic Information. *Trans. GIS 6*(4), 431–456.

Urrutia, A., J. Galindo, et M. Piattini (2002). Modeling Data Using Fuzzy Attributes. In *Intl Conf. of the Chilean Computer Science Society (SCCC)*, pp. 117–123.

Verbruggen, C. (2007). Combining Social Network Analysis and Prosopography. In *Prosopography Approaches and Applications. A Handbook*, pp. 579–601. Linacre College.

Raisonnement Causal et Relations Symboliques dans les Enluminures Médiévales

Djibril DIARRA*, Christophe NICOLLE*,
Martine CLOUZOT**

*Laboratoire Électronique Informatique et Image (LE2I)
dl4djibril@gmail.com,
cnicolle@u-bourgogne.fr,
http://le2i.cnrs.fr/
**Laboratoire Archéologie Terre et Histoire (ArTeHis)
martine.clouzot@u-bourgogne.fr

Résumé. Ce travail concerne les domaines de l'ingénierie des connaissances et des enluminures médiévales. Dans cet article nous considérons une enluminure comme un graphe de connaissances. Ce graphe était utilisé par les élites au Moyen Âge pour se représenter comme groupe social et représenter les évènements de leur vie. Pour ce faire, des combinaisons d'éléments symboliques sont utilisées pour encoder des messages d'influence, plus ou moins implicites. Notre travail est d'identifier la signification de ces éléments au travers d'une approche de modélisation logique en utilisant des ontologies. L'idée à terme est d'identifier des règles de raisonnements logiques et de les simuler à l'aide de mécanismes d'intelligence artificielle pour, d'une part, faciliter l'interprétation des enluminures au regard du contexte et d'autre part fournir une formalisation logique de nouveaux services d'encodages et de transmission de l'information dans les futures évolutions des réseaux sociaux actuels.

1 Introduction

L'ingénierie des connaissances vise à formaliser de manière logique les connaissances humaines afin qu'elles soient manipulables par des systèmes informatisés. Dans cet article, cette science est utilisée pour caractériser et formaliser des relations symboliques entre les concepts dans un contexte en constante évolution. Les enluminures médiévales sont des images qui, au Moyen Âge, étaient conçues et utilisées par les élites pour représenter les évènements de leur vie, mais aussi pour se représenter en tant que groupe social. Elles constituent un système d'informations fondé sur des relations symboliques dont les significations et les messages sont déterminés dans un contexte particulier et évolutif. Elles peuvent être représentées par des graphes de connaissances. Les fortes corrélations de structures entre les enluminures médiévales et les réseaux sociaux demandent d'abord d'expliciter par quels procédés les enluminures servaient de support visuel à la communication sociale de l'époque. Même si le pouvoir des images est explicitement de représenter des scènes, implicitement elles visent surtout à agir sur les perceptions cognitives des utilisateurs et donc sur leurs comportements. Les enluminures

médiévales traitées dans nos travaux désignent uniquement celles relatives à la cour du Duc de Bourgogne (cf. 3.1), Philippe Le Bon [1]. Elles étaient réalisées, pour le Duc, dans le but implicite d'exercer une influence sur ses différents réseaux sociaux : familles, chevaliers, alliés, ennemis, autres princes européens, concurrents, etc.

Cette influence se constate dans la circulation, la ré-appropriation, par l'imitation ou le détournement des enluminures : des messages implicites circulent dans les réseaux, qui les réutilisent et en produisent de nouveaux. À partir des relations symboliques que les enluminures médiévales contiennent, elles peuvent donc être considérées comme les premiers réseaux sociaux et aident à enrichir les structures des réseaux sociaux actuels et futurs.

1.1 Objectif de recherche

Notre travail vise à identifier la signification des éléments symboliques et des relations sémantiques dans les enluminures médiévales à travers une approche de modélisation logique en utilisant des ontologies. Des combinaisons d'éléments symboliques sont utilisées pour encoder des messages d'influence, plus ou moins implicites. Ces combinaisons permettent de décrire et d'analyser des éléments symboliques figurés dans les images. Pour ce faire, nous développons des outils informatiques et des ontologies visant à saisir et modéliser la signification des éléments symboliques et leurs relations. Au-delà d'une simple modélisation taxonomique, nous pouvons contraindre l'ontologie des enluminures que nous avons constituée pour la rendre plus expressive. Le niveau d'expressivité atteint est équivalent au langage SHOIN(D) (cf. 4.1) en logiques de description. Cela nous permet de construire des règles de raisonnement et d'utiliser le moteur d'inférence d'un triplestore [2]. Notre système informatique est donc capable de raisonner sur des éléments du graphe représentant les enluminures pour découvrir de nouvelles connaissances implicites.

1.2 Organisation du document

Dans la suite de ce document, nous présenterons tout d'abord quelques travaux auxquels les nôtres s'apparentent, notre vision sur les réseaux sociaux et montrerons les corrélations entre les enluminures et ces réseaux. Ensuite, nous présenterons une enluminure et détaillerons le contexte historique associé. Après quoi, nous proposerons une ontologie de ces enluminures. Nous donnerons quelques exemples de relations symboliques formalisées. Et nous présenterons une interface en cours d'implémentation pour l'annotation d'enluminures. Enfin nous conclurons et donnerons les perspectives de quelques travaux futurs.

2 Quelques Travaux Liés

L'utilisation des ontologies pour la représentation des graphes de connaissances, de manière générale, n'est pas nouvelle et beaucoup de travaux la traitent. Plus spécifiquement, pour

1. Philippe Le Bon (1419-1467) est le plus célèbre des Ducs de Bourgogne et compte parmi les princes européens les plus puissants à l'époque de la Guerre de Cent Ans (1337-1453).

2. Un triplestore est une base de données spécialement conçue pour le stockage et la récupération de données RDF (Resource Description Framework, un format de représentation de connaissances sous forme de triplet (sujet, prédicat, objet)). Tout comme une base de données relationnelles, un triplestore stocke des données et il les récupère via un langage de requête.

les réseaux sociaux qui sont des graphes de connaissances composées d'entités et de relations sociales, une ontologie célèbre a été modélisée. C'est l'ontologie FOAF[3] dont la spécification est détaillée dans les travaux de (Brickley et Miller, 2007). Elle nous sert de base pour montrer la similitude entre les enluminures et un réseau social. En outre, notre ontologie des enluminures l'utilise, l'étend pour être plus précise dans notre description.

Concernant la description des images médiévales, du patrimoine culturel en général, l'auteur de (Dörr, 2002) propose une ontologie (CIDOC object-oriented Conceptual Reference Model, CRM) qui modélise des informations sur le patrimoine culturel. Elle décrit formellement les concepts et les relations qui sous-tendent les structures documentaires utilisées dans ce domaine. Cette ontologie est très générale car elle essaye de couvrir tout type de document utilisé dans le domaine du patrimoine culturel. Celle que nous proposons contribue également à la numérisation du patrimoine culturel mais elle est spécifique aux enluminures du Duc de Bourgogne. Les auteurs de (Doerr et al., 2006) élargissent la CRM en la combinant avec une autre qui décrit le domaine de la librairie digitale. Ce qui favorise l'intégration de connaissances. Notre ontologie peut, de même, être étendue à d'autres modèles (FOAF, par exemple). Les données du patrimoine culturel sont considérées comme syntaxiquement et sémantiquement très hétérogènes, multi-langues, sémantiquement très riches et hautement reliées car elles sont produites par différentes entités (musées, archives, fouilles archéologiques, etc). Ainsi l'auteur de (Hyvönen, 2012) donne un aperçu sur quand, pourquoi et comment utiliser dans la pratique les technologies du Web Sémantique pour publier des connaissances du patrimoine culturel sur le Web. Il évoque la plupart des formalismes que nous utilisons dans nos travaux. Les principales raisons qui motivent nos travaux sont : le partage de connaissances explicites ou implicites extraites des enluminures, l'interopérabilité et l'intégration avec d'autres connaissances similaires et la mise à disposition d'un modèle valide qui aide au développement de systèmes informatiques utilisés dans la gestion des connaissances du patrimoine culturel.

3 Réseaux Sociaux et Enluminures Médiévales

Les réseaux sociaux sont le sujet de nombreux travaux de recherche académiques et industriels. Ces travaux traitent différents aspects des réseaux sociaux : l'analyse des relations sociales (Raad, 2011), l'analyse des sentiments sur les réseaux sociaux (Martínez-Cámara et al., 2014), la découverte des communautés implicites dans les réseaux sociaux (Leprovost et al., 2012), l'extraction des profils dans un réseau social (Ramiandrisoa et Mothe, 2017), la diffusion d'informations sur les réseaux sociaux (Bakshy et al., 2012), etc.

Tous ces travaux nécessitent une masse d'informations importante et variable dans des contextes différents où la notion de réseau social désigne toute relation entrainant des interactions sociales régulières entre des individus, des organisations, des entreprises, des régions, des pays, etc. Ces relations sont fondées sur, entre autres, la connaissance, la collaboration, la collégialité, l'amitié. Elles peuvent être directes (une entité A a une relation directe avec l'entité B, $A \rightarrow B$) ou indirectes ($A \rightarrow B \rightarrow ... \rightarrow X$, donc $A \rightarrow X$), symétriques ($A \rightarrow B$ implique $B \rightarrow A$) ou asymétriques ($A \rightarrow B$ n'implique pas $B \rightarrow A$).

Tout réseau social est maintenu et entretenu par des opérations de partage de ressources qui peuvent être matérielles (argent, bétails, alimentation, équipements, armes) ou non matérielles

3. FOAF, Friend Of A Friend est une ontologie populaire qui décrit les relations sociales entre des entités et leurs centres d'intérêt dans un réseau social.

(informations, stratégies, décisions, état humeur, activités). Plus les réactions des membres vis-à-vis de ces ressources et leur partage multipartite sont grands dans le réseau, plus celui-ci devient dense et important. Ces opérations constituent l'une des caractéristiques fondamentales d'un réseau social, qu'il soit représenté par une communauté virtuelle : site web sur internet (Facebook [4], LinkedIn [5], Viadeo [6], Twitter [7], etc.) ou communauté réelle de la vie.

Pour résumer, un réseau social est un ensemble de représentations (avatars) de personnes réelles ou morales ; contributrices dans une plateforme informatique ou non, par des messages ou des documents ; et qui sont incitées à interagir explicitement ou implicitement à ces informations diffusées. La majorité des réseaux sociaux actuels est structurée autour de la valorisation du "moi" des utilisateurs réelles à travers leur avatar.

3.1 Enluminure : définition et principes de conception

Cette représentation du "moi" et de "mon" environnement est une constante dans l'histoire de l'humanité. Des peintures rupestres illustraient une vision d'une activité qui se transmettait sur un support horizontal, l'écriture et le dessin (par opposition à la transmission orale, dite verticale). Au Moyen Âge, pour s'adresser aux lettrés et catégories sociales éduquées, les enluminures ont été développées dans des livres. Une enluminure est une peinture fixe, faite sur les feuilles de parchemin (habituellement du cuir d'animal tanné) d'un manuscrit. C'est une représentation codifiée pour valoriser le "moi" du commanditaire. Cette codification peut facilement être représentée sous la forme d'un graphe décrivant des relations sémantiques et symboliques entre des objets, des idéaux, des personnages exprimant des concepts et transmettant des messages explicites et implicites. Elle est structurée selon des relations de positionnement (topologie), de hiérarchie, des relations sémantiques (par exemple la méronymie) et des relations métaphoriques (par exemple une représentation animale pour exprimer une valeur morale humaine).

La métaphore est un procédé rhétorique important depuis l'Antiquité. Une définition médiévale de la métaphore vient *du grec metaphora et du latin translatio, elle signifie littéralement le remplacement du terme propre par un terme imagé, répondant à une comparaison implicite* (Pernot, 1993). Elle se fait selon quatre modes : de l'animé à l'inanimé, de l'inanimé à l'inanimé, de l'animé à l'animé, de l'inanimé à l'animé ".

Les relations métaphoriques sont très abondantes dans les enluminures que nous traitons, c'est-à-dire les enluminures de la cour ducale.

La cour ducale est une micro-société aristocratique, composée d'un ensemble d'entités dont le Duc lui-même, le Comte (son fils ainé), les ecclésiastiques (évêque, clercs), les chevaliers. Son territoire est une vaste principauté allant du sud de la Bourgogne actuelle jusqu'à Amsterdam. Ses possessions territoriales sont en partie rurales, mais aussi très urbanisées (Bruxelles, Bruges, Gand, Dijon, Lille, etc.). Différents groupes sociaux de son réseau (banquiers, commerçants, universitaires, bourgeois des villes, etc.) fonctionnent aussi en réseaux plus ou moins larges. L'enluminure de la figure 1 illustre un réseau composé du Duc, de son fils le Comte, les Conseillers. Ensemble, ils constituent le réseau social fermé symbolisé par le collier de la Toison d'Or.

4. www.facebook.com
5. www.linkedin.com
6. www.viadeo.com
7. https://twitter.com

Dans les enluminures on considère aussi certains objets, personnages ou l'image dans son entier comme un signe (parfois plus fort que la métaphore ou fonctionnant au moyen de la métaphore). Une enluminure est réalisée dans un contexte toujours savant, religieux ou profane. Elle fait partie d'une chaîne de communication allant du commanditaire (ici, le Duc de Bourgogne), l'auteur (enlumineur) ou le copiste du livre jusqu'au destinataire (le Duc, un(e) noble de sa cour, un homme politique, un roi/une reine ou un personnage important d'une cour européenne). Elle représente des thématiques correspondant au commanditaire/destinataire, c'est-à-dire à son niveau social, à son niveau de culture et elle vise à représenter une image idéale de lui-même, de sa famille, de sa société dans le passé, le présent, le futur (après la mort). Implicitement, elle sert à transmettre des idées, des relations idéalisées, des valeurs chevaleresques et religieuses. Explicitement, elle représente des scènes de vie de la cour comme les banquets, les noces, les bals, les tournois, etc. Les scènes qu'elle décrit et leurs significations ne sont pas fixes, elles varient selon les contextes. Ces scènes d'évènements sont sélectionnées par un écrivain (l'enlumineur, la personne qui peint les images) et dessinées sous forme d'image. Ces images servent de véhicules d'informations sur les activités du Duc et de la cour ducale et sont présentées au public lors d'événements politiques comme le don d'un manuscrit luxueux au Duc (scène décrite dans l'enluminure du Duc illustrée par la figure 1), un banquet ou une grande assemblée de chevaliers.

FIG. 1 – *Enluminure présentant une scène de remise de livre au Duc de Bourgogne. Bruxelles, Bibliothèque royale de Belgique, ms. 9243, folio 185 verso, Chroniques de Hainaut de Jean Wauquelin, 1446*

3.2 Correspondance entre enluminure et réseaux social

De part ces vertus et ces usages, un livre enluminé constitue un réseau social à l'instar de Twitter ou LinkedIn, en termes d'outil de communication et de partage de ressources. Cependant une enluminure en soit peut être assimilée à une ressource d'animation, d'entretien d'un réseau social (le réseau social du Duc) dans la mesure où elle raconte et partage avec les

membres du réseau une histoire. Son but est ainsi de montrer l'influence du Duc, de l'exercer et de la véhiculer. On déduit ces influences à travers certaines relations dans l'enluminure. Quelques-unes de ces relations d'influences sont :

- la soumission de la cour au Duc ; les conseillers, l'enlumineur, le Compte, les chevaliers (à part celui en vert), le lévrier sont fidèles au Duc. De même, les conseillers se tiennent derrière le Duc et l'écrivain est agenouillé devant le Duc ;
- la richesse du Duc ; au travers des bourses qu'il porte et de sa tenue ;
- le commandement du Duc ; il tient un bâton de commandement.

4 Formalisation Sémantique des Enluminures

Cette partie est consacrée à notre formalisation sémantique des enluminures. Une succincte description des notions et termes utilisés est faite au fur et à mesure de la présentation des différents résultats.

4.1 Modélisation Ontologique d'une Enluminure et Fondamentaux sur les Termes et Formalismes utilisés

Étymologiquement lié à la théorie de l'existant, le terme "ontologie" admet beaucoup de définitions dans la littérature, car elle est applicable à de nombreux domaines (la philosophie, les sciences de l'information, la linguistique, l'ingénierie des connaissances, l'intelligence artificielle, etc.). Dans notre projet, nous retiendrons la définition de (Studer et al., 1998) : *"An ontology is an explicit and formal specification of a shared conceptualization"*.

Une ontologie représente une conceptualisation formelle d'un domaine (Gruber, 1993). Ici, un domaine désigne l'environnement que l'on souhaite décrire. Une ontologie inclut une organisation hiérarchique des concepts pertinents du domaine, des relations qui existent entre ces concepts ainsi que des règles et axiomes qui les contraignent dans leur fonctionnement. Les connaissances d'un domaine sont formalisées, dans une ontologie, en utilisant principalement cinq types de composants à savoir :

- les concepts, aussi appelés classes, correspondent aux abstractions pertinentes, du domaine, retenues en fonction des objectifs que l'on se donne et de l'application envisagée pour l'ontologie ;
- les relations traduisent les associations pertinentes existant entre les concepts. Ces relations sont de type hiérarchique (généralisation/spécialisation, agrégation/composition ; instance de), associatif, équivalent (synonymie, homonymie, antonymie, etc.)
- les axiomes constituent des assertions, acceptées comme vraies, à propos des abstractions du domaine ;
- les instances constituant la définition extensionnelle de l'ontologie. Ces objets véhiculent les connaissances statiques ou factuelles du domaine.

Pour l'enluminure de la figure 1, des exemples de composantes ontologiques sont : les concepts (Duc, Livre, Lévrier, Chevalier, Activité, Personne, Animal) ; les relations de généralisation (Duc est une Personne, Lévrier est un Animal), d'agrégation (un Projet est composé d'Activités), de synonymie (Prince est synonyme de LeComte), associative (l'Ecrivain offre le Livre, le Livre est offert au Duc) ; les instances (banquet, jeu, chasse, fauconnerie sont des instances d'activité).

La représentation ontologique d'un univers de discours doit bannir toute ambigüité significative. Cela permet de disposer d'un support de connaissances uniforme pour la communauté utilisatrice, d'avoir une base de connaissances réutilisable, d'assurer le partage de connaissances et une communication efficace.

Cette contrainte est assurée par l'utilisation d'un langage formel : les logiques de description (DL) dans le cadre de nos travaux, à travers sa variante SHOIN(D). Cette variante des DL est très utilisée dans les représentations ontologiques à cause de son expressivité, sa décidabilité et sa complexité maitrisée. Ses constructeurs (l'ensemble des symboles lexicaux et opérateurs utilisés dans les DL) lui confèrent une expressivité suffisante pour la description ontologique. Ces constructeurs sont S(ALC et R+), H, O,I, N (Baader et al., 2005), (Baader, 2011). Leur signification est :

- **S(ALC et R+)** : est la désignation donnée à une sous-variante des DL qui regroupe les constructeurs ALC (sous-variante basique des DL composée des opérations de - définition du concept global (Top, représenté par $\top$), du concept néant (bottom, $\bot$), d'un concept quelconque (Duc par exemple) ; de conjonction de concepts (Duc ET $Chevalier$, $Duc \sqcap Chevalier$), de quantification universelle (Tous les enfants du Duc, $\forall enfant.Duc$) et existentielle(le père du Comte, $\exists pere.LeComte$), et de négation d'un concept(non Animal, $\neg Animal$)) et du constructeur R+ (qui permet la composition de rôles ou relations. *Exemple :* $pere(pere(X, Y), Z)$ *pour dire que X est le grand-père de Z où* X, Y, Z *sont des concepts et* $pere$*, une relation,* $pere.pere.Z$) ;
- **H** : désigne le constructeur de la hiérarchie entre concepts. *Exemple :* Duc *est une* $Personne$, $Duc \subseteq Personne$;
- **O** : désigne le constructeur pour les instances. *Exemple : jeux, banquet sont des instance d'activité,* $Activite\{jeux, banquet\}$;
- **I** : pour designer l'inverse d'un rôle (une relation). *Exemple : la relation* $enfant$ *est l'inverse de* $pere$, $enfant.\top \equiv \exists pere^{-1}.\top$;
- **N** : pour la restriction du nombre. *Exemple : Les chevaliers sont au maximum 8,* $nombreChevalier._{<=}8Chevalier$.

En outre, ces constructeurs peuvent être combinés, par des opérateurs de hiérarchisation (subsomption, exprimée par $\subseteq$) ou d'équivalence ($\equiv$), pour définir d'autres concepts ou pour les organiser à travers des règles (ou axiomes).

Le tableau 1 présente quelques concepts, relations et individus de l'enluminure de la figure 1. La figure 2 décrit une vue des concepts de l'enluminure, modélisés dans l'outil logiciel protégé [8].

Cette modélisation, une fois finie, peut être récupérée sous la forme d'une syntaxe formelle dans un langage [9] tel que RDF/XML, Turtule ou OWL, générée par l'outil. Elle est extensible et peut être combinée à d'autres ontologies, pour compléter la modélisation initiale, telle que FOAF avec laquelle elle a beaucoup de termes en commun comme *foaf :Person, foaf :member, foaf :interest, foaf :Group, foaf :depict, foaf :Image*.

En outre, nous avons développé une plateforme web permettant de sélectionner une enluminure et d'indexer sémantiquement ses éléments constitutifs. Sur cette plateforme accessible via un navigateur web, il est possible de télécharger une version numérique d'une enluminure

8. http ://protege.stanford.edu/

9. RDF/XML, Turtle, OWL sont des langages informatiques utilisés dans la création des ontologies. Afin de respecter le nombre de pages autorisé, nous ne n'avons pas développé la spécification de ces langages

sous forme d'image (au format jpeg ou png). Ensuite, des outils de sélection graphique sont utilisés pour encadrer des éléments importants de l'image en tant que concepts. Une fois ces concepts indexés, il est possible de construire des relations sémantiques entre eux. Cela permet d'étendre l'ontologie. Pour l'instant, l'ajout des contraintes pour la vérification des relations ajoutées et la définition d'inférences n'est pas encore traité et fera l'objet de prochains travaux.

Néanmoins l'annotation des concepts et leurs relations sont faites manuellement suivant une liste de ces concepts et relations fournie par les experts métiers. Ce qui garantit une certaine cohérence dans les concepts et relations indexés. La figure 3 illustre une interface de cette plateforme. Elle nous montre quelques annotations effectuées et mises en relation dans l'enluminure de la figure 1. Les annotations et relations créées peuvent être récupérées et sauvegardées sous format JSON [10]. Ce fichier est construit pour pouvoir étendre l'ontologie à la fois sur sa Tbox (l'ensemble des concepts d'une ontologie et leurs relations) et sur sa Abox(l'ensemble des instances, des faits ou individus dans le modèle).

La construction de cette ontologie est réalisée à partir du savoir faire métier des médiévistes. Les connaissances du domaine permettent de former un ensemble formel et cohérent de description des concepts et des relations formant le graphe sémantique de l'enluminure. L'ontologie que nous avons obtenue peut être interrogée par des requêtes SPARQL [11]. L'intérêt de l'ontologie est sa capacité à découvrir des connaissances nouvelles à partir des connaissances initiales. Cette découverte est réalisée par la construction de règles logiques et leur application au sein d'un moteur d'inférence. Au-delà de notre description terminologique des enluminures, certains éléments d'interprétations symboliques restent à définir par le biais des règles logiques. Nous travaillons actuellement sur la construction de ces règles en langage SWRL [12] (Semantic Web Rule Language). Ainsi nous souhaitons dans un futur proche construire des règles pour faire du raisonnement. Comme par exemple tel personnage est le Duc de Bourgogne en fonction de la description de ses habits, de ses bijoux, de sa posture, du contexte ou des relations topologiques.

4.2 Exemple de formalisation de relations symboliques : les métaphores

La construction de règles logiques sous la forme de clause de Horn nous permettra d'interpréter des métaphores, nombreuses dans les enluminures. La métaphore est une figure de rhétorique qui effectue une comparaison non explicite mais intuitivement perceptible entre deux concepts dissemblables. Bien que les deux concepts reliés par la comparaison appartiennent à des champs sémantiques différents, ils partagent tout de même une caractéristique commune qui permet d'établir l'analogie entre eux. Par exemple, on peut louer la bravoure d'un homme en le désignant par un lion mais homme et lion restent des concepts qui sont totalement distincts (homme est un humain alors que lion est un animal). Bien qu'il n'y ait

10. JSON (JavaScript Object Notation - Notation Objet issue de JavaScript) est un format léger d'échange de données. Il est facile à lire ou écrire, aisément analysable.

11. SPARQL est le langage utilisé pour interroger une base de données RDF (ou triplestore défini plus haut). Il est similaire au langage SQL utilisé pour interroger les base de données relationnelles

12. SWRL - Langage de Règles pour le Web Sémantique, est le langage qui permet de construire des clauses de Horn (règles logiques d'inférences). Il contribue à étendre l'expressivité de certaines variantes du langage OWL (Ontologie Web Language) en leur permettant de créer des règles complexes. OWL est un autre langage dédié à la création des ontologies pour le Web. Il admet beaucoup de variantes : OWL-DL (correspond au SHOIN(D)), OWL-Lite (utilise moins de constructeurs), OWL-Full (OWL complet, qui est indécidable).

pas eu de consensus mutuel de la part des linguistes sur une typologie universelle des métaphores (Perrenoud, 2002), on peut cependant énumérer deux principaux types : métaphore in praesentia et métaphore in absentia.

Pour **la métaphore in praesentia**, les deux concepts (comparé et comparant) sont présents et malgré l'absence de l'outil de comparaison, il est possible de percevoir assez aisément le lien qui les unit. Cela rend la comparaison moins allusive et atténue relativement la force d'expression de la métaphore. *Exemple :" Le papillon, fleur sans tige " (Nerval).* Le papillon est comparé à une fleur pour mettre en valeur sa splendeur. Cette métaphore peut être exprimée en logiques de description par :

$$\textit{Fleur} \equiv \textit{Papillon}$$
$$\textit{Fleur} \sqcap \textit{Papillon} \equiv \exists estSplendide.\top$$

De même qu'en langage OWL comme indiqué dans la figure 4

La **métaphore in absentia** est caractérisée par la présence unique du concept comparant et l'absence du concept comparé dont l'existence est insinuée par le contexte. *Exemple : " Mon esprit amer, d'une aile inquiète et folle vole sur la mer " (Verlaine).* Le comparant "esprit" est clairement exprimé alors que le comparé "oiseau" est devinable grâce aux mots aile et vole qui font partie de son champs lexical. Il s'agit d'une métaphore qui n'est pas exprimée ouvertement, le lien entre le comparant et le comparé est établi par inférence logique. Celle de l'exemple peut être exprimée en logiques de description comme comme indiqué ci-après :

$$\textit{Oiseau} \subseteq \exists \textit{aAile}.\top \sqcap \exists \textit{peutVoler}.\top$$
$$\exists \textit{aAile}.\top \sqcap \exists \textit{peutVoler}.\top \equiv \textit{Esprit} \sqcap \textit{Légèreté}.$$

Par transitivité, on pourra déduire que $\textit{Oiseau} \subseteq \textit{Esprit} \sqcap \textit{Légèreté}$.
Cette même métaphore peut être exprimée en langage OWL à l'instar de l'exemple ci-dessus.

Il existe d'autres types de métaphores, comme la métaphore filée qui ajoute de nouveaux termes empruntés au lexique d'une autre métaphore afin de compléter le sens de cette dernière et intensifier son effet, ou la métaphore lexicalisée (ou catachrèse). C'est une métaphore qui au fil de son usage est entièrement intégrée dans le langage courant (ex : les dents d'une scie, le bras d'un fauteuil...). Des exemples de ces métaphores et leur formalisation en logiques de description sont :

Métaphore filée, un axiome d'équivalence relie le comparant à l'intersection du comparé avec le concept commun. Quant à l'union des nouveaux termes métaphoriques ajoutés, elle est subsumée par l'ensemble des choses qui ont au moins une instance de la propriété " champs-lexicalDe " vers le concept comparant.

Exemple :" Cette femme est une fleur, la corolle de son visage m'obsède, les pétales de ses joues m'enivrent " L'expression de cette métaphore en logique de description est :

$$\textit{Fleur} \equiv\subseteq \textit{Femme} \sqcap \textit{Beauté}$$
$$\textit{Corolle} \sqcap \textit{Pétales} \subseteq \textit{champlexicalDe.Fleur}.$$

Métaphore lexicalisée (catachrèse), l'emploi fréquent de cette métaphore finit par lui faire perdre sa puissance poétique, conduisant alors à son assimilation totale au langage courant. En conséquence, des relations méronymiques ont été créées entre comparants et comparés. Par exemple :

$$\textit{Bras} \subseteq lmentsDe.Fauteuil, \textit{Coucher} \subseteq \textit{phaseDe.soleil}.$$

Hormis le texte, l'expression d'une métaphore peut être visuelle. Ces métaphores sont des images qui recèlent des métaphores sous-jacentes, visibles grâce à des formes ou des symboles qui incarnent des phénomènes, des évènements, des personnages ou autres. Le travail que nous avons réalisé permet de formaliser l'expression de la métaphore visuelle dans le langage OWL pour les enluminures médiévales.

5 Conclusion et Travaux Futurs

Ce papier présente une recherche en cours associant des savoirs et techniques issus du domaine de l'ingénierie des connaissances et du domaine de l'analyse historique de documents médiévaux. Ce travail nous a permis tout d'abord d'identifier le processus de conception et de diffusion des enluminures comme une expression médiévale d'un réseau social. Ce réseau social obéit aux mêmes motivations et aux mêmes codes que les réseaux sociaux numériques actuels. Néanmoins, ce système d'expression de la connaissance utilise des types de relations plus complexes, comme les métaphores. Pour faciliter l'expression du savoir-faire d'interprétation des médiévistes et développer un système automatique de compréhension des métaphores, nous avons proposé une approche de formalisation en logiques de description d'une ontologie du domaine. Cette ontologie décrite en SHOIN(D) permet aux médiévistes de décrire à l'aide d'une interface web les composantes des enluminures et d'encoder sous la forme de règles logiques le raisonnement associé aux métaphores. Dans les travaux futurs, nous cherchons à combiner notre ontologie avec des systèmes de représentation sémantique des réseaux sociaux actuels. L'objectif est de proposer à terme une extension fonctionnelle des réseaux sociaux visant a améliorer par une analyse qualitative le calcul de l'influence de ces membres.

6 Remerciements

Ces travaux de recherches sont financées par le CNRS dans le cadre d'un projet PEPS, l'Ambassade de France à Bamako et le gouvernement Malien pour le co-financement d'une thèse, la Région Bourgogne Franche-Comté et l'université de Bourgogne. Nous tenons à remercier l'ensemble des ces institutions ainsi que Monsieur Rafik Zebidi et Monsieur Florian Lacroix pour leur aide sur la partie formalisation et implémentation.

Références

Baader, F. (2011). What ?s new in description logics. *Informatik-Spektrum 34*(5), 434–442.

Baader, F., I. Horrocks, et U. Sattler (2005). Description logics as ontology languages for the semantic web. In *Mechanizing Mathematical Reasoning*, pp. 228–248. Springer.

Bakshy, E., I. Rosenn, C. Marlow, et L. Adamic (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528. ACM.

Brickley, D. et L. Miller (2007). Foaf vocabulary specification 0.91.

Doerr, M., J. Hunter, et C. Lagoze (2006). Towards a core ontology for information integration. *Journal of Digital information 4*(1).

Dörr, M. (2002). The cidoc crm-an ontological approach to semantic interoperability of metadata, 2001. *AI Magazine, Special Issue on Ontologies, Nov*.

Gruber, T. (1993). What is an ontology. *WWW Site http ://www-ksl. stanford. edu/kst/whatis-an-ontology. html (accessed on 07-09-2004)*.

Hyvönen, E. (2012). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web : Theory and Technology 2*(1), 1–159.

Leprovost, D., L. Abrouk, et D. Gross-Amblard (2012). Discovering implicit communities in web forums through ontologies. *Web Intelligence and Agent Systems : An International Journal 10*(1), 93–103.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Urena-López, et A. R. Montejo-Ráez (2014). Sentiment analysis in twitter. *Natural Language Engineering 20*(1), 1–28.

Pernot, L. (1993). *La rhétorique de l'éloge dans le monde gréco-romain : Histoire et technique*, Volume 137. Institut d'études augustiniennes.

Perrenoud, P. (2002). D'une métaphore à l'autre : transférer ou mobiliser ses connaissances ? In *L'énigme de la compétence en éducation*, pp. 45–60. De Boeck Supérieur.

Raad, E. (2011). *Découverte des relations dans les réseaux sociaux*. Ph. D. thesis, Dijon.

Ramiandrisoa, F. et J. Mothe (2017). Profil utilisateur dans les réseaux sociaux : Etat de l'art. In *CORIA*, pp. 395–404.

Studer, R., V. R. Benjamins, et D. Fensel (1998). Knowledge engineering : principles and methods. *Data & knowledge engineering 25*(1-2), 161–197.

Concept	**Rôle ou relation**	**Individu**
Animal	avoirEnfant(Personne, Personne)	Lieu(PALAIS, SALLE)
Duc	êtrePère(Personne, Personne)	ActivitéExtraProfessionnelle (LECTURE, FAUCONNERIE,DANSE, MUSIQUE, BAIN)
Palais	entourer (Personne, Duc)	
CourDucale	êtreReunie(Personne, Duc)	
Chevalier	êtrePositionné(Personne, Position)	
Conseiller	êtreAgenouilléDevant(Enlumineur, Duc)	
LeComte	êtreOffert (Livre, Duc)	
Ecrivain	parler(Duc, Personne)	
Livre	peindre(Livre, Ecrivain)	
Enluminure	tenirSous(Duc, Dais)	ActivitéProfessionnelle (DECISION, POLITIQUE, JUSTICE, DIPLOMATIE, FINANCES, COMMANDEMENT MILITAIRE)
Enlumineur	regarder(ChevalierEnVert, Duc)	
Collier	êtreLier(Personne, Personne)	
GoupeToisonDor	habiller(Personne, Vêtement)	
PierrePrecieuse	êtreHors(Ecrivain, Cadre)	
Lévrier	nommer(Duc, Personne)	
Personne	interesser(Personne, Activité)	
Chaussure	êtreMembre(Personne, GroupeToisonDor)	
ActivitéProfessionnelle	composer(Assemblé, Personne)	
ActivitéExtraProfessionnelle	avoirAge(Personne, entier)	
Positionnement		
BatonCommandement		
Assemblé		
Activité		

TAB. 1 – *Les composantes ontologiques contenues dans l'enluminure de figure 1. Les listes de concepts, relations et individus présentes dans ce tableau ne sont pas exhaustives. Elles ne présentent qu'une portion des composantes de cette enluminure*

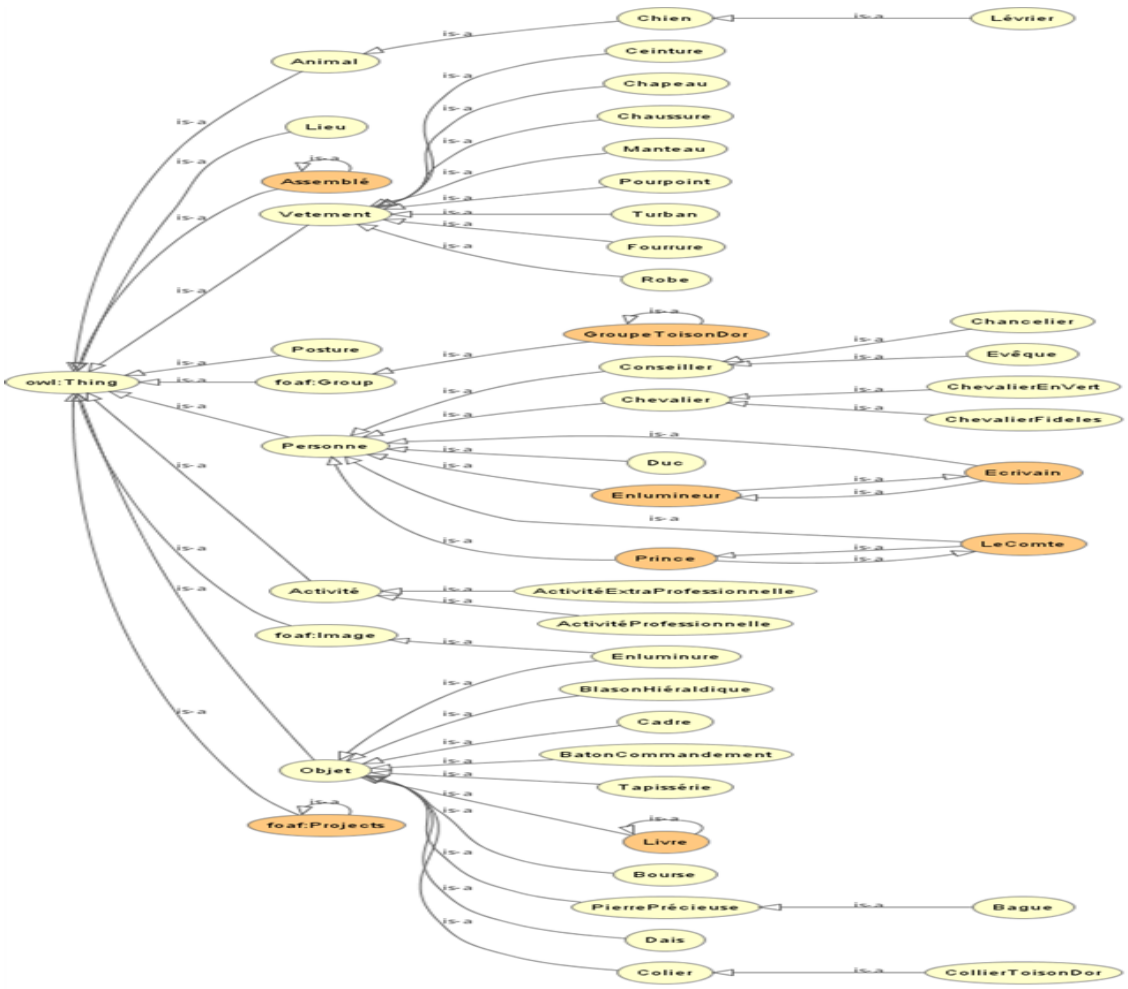


FIG. 2 – *Les concepts de l'enluminure, modélisés dans Protégé 2000.*

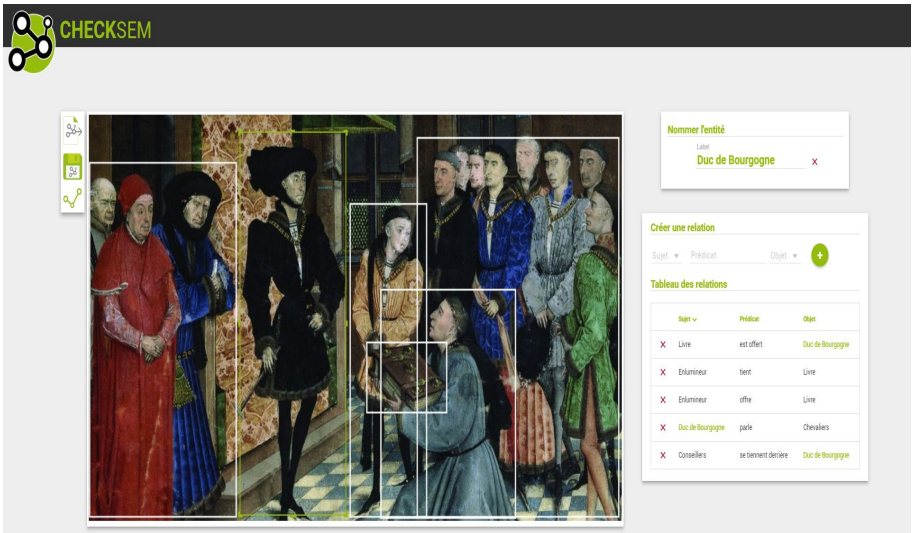


FIG. 3 – *Une interface de notre outil de tag des enluminures.*

```
<owl:ObjectProperty rdf: ID ="estSplendide"/>
<owl:Class rdf: ID ="Papillon"/>
<owl:Class rdf:ID ="Fleur">
    <owl:equivalentClass rdf:resource="#Papillon"/>
</owl:Class>
<owl:Class>
  <owl:intersectionOf rdf=parseType="Collection">
        <owl:Class rdf:about ="#Fleur"/>
        <owl:Class rdf :bout="#Papillon"/>
  </owl:intersectionOf>
  <owl:equivalentClass>
        <owl:Restriction>
          <owl:onProperty rdf:resource="#estSplendide"/>
          <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
        </owl:Restriction>
  </owl:equivalentClass>
</owl:Class>
```

FIG. 4 – *Formalisation owl de la métaphore "Papillon, fleur sans tige"*

Utilisation des relations sémantiques des mots-clés pour la catégorisation d'articles scientifiques

Bastien Latard*,** Jonathan Weber*
Germain Forestier*, Michel Hassenforder*

*MIPS, Université de Haute-Alsace, Mulhouse, France
prenom.nom@uha.fr,
**MDPI AG, Bâle, Suisse

Résumé. Soumises au déluge de données, la recherche bibliographique scientifique est de plus en plus chronophage. Par conséquent, maintenir ses connaissances à l'état de l'art dans un domaine spécifique est une tâche complexe et pénible. L'objectif de notre travail est de créer un moteur de recherche intelligent prenant en compte le contenu des articles scientifiques. Les travaux préliminaires présentés dans cet article sont la première brique d'un tel système. Les relations sémantiques entre les différents mots-clés des articles sont extraites dans le but de catégoriser les articles et de les rapprocher d'autres articles similaires en se basant sur la sémantique des mots-clés. En exploitant la corrélation entre les catégories et domaines des mots-clés, issus de BabelNet, notre méthode sera capable de regrouper des articles proches en termes de contenu ayant des mots-clés différents.

1 Introduction

La recherche bibliographique est une étape cruciale pour tout chercheur. En effet, la connaissance des travaux existant peut faire gagner un temps précieux tant pour le choix de la méthode à adopter que pour être à jour des dernières avancées. Néanmoins, trouver des articles similaires reste une tâche compliquée et pénible autant pour les domaines étendus que réduits. Les chercheurs passent un temps considérable à chercher des travaux proches de leurs intérêts de recherche. Cette étape est cependant incontournable dans tout projet de recherche afin de confronter de nouvelles idées à des solutions existantes, ainsi que pour l'acquisition de connaissance à propos d'un domaine spécifique. C'est pourquoi l'amélioration de la recherche bibliographique pourrait avoir un impact très positif pour la communauté scientifique.

Un éditorial de *Nature* (Editorial) (2012) exprime explicitement la frustration constante de la communauté scientifique au regard du potentiel incroyable que représenterait la fouille de données de l'ensemble de la littérature scientifique. Cependant, les *text miners* doivent faire face aux restrictions légales des maisons d'éditions (accès fermé). La croissance de la littérature scientifique est estimée à environ 3 millions d'articles de journaux ou conférence par an sur les 4 dernières années, d'après Scilit [1]. Collecter et analyser manuellement cette masse

1. http://www.scilit.net

de donnée est très long et presque impossible, surtout lorsqu'elle est disséminée dans 47000 revues scientifiques appartenant à quelques 6000 éditeurs différents, toujours d'après Scilit. Pour contrer l'éparpillement de ces articles à travers de multiples plates-formes décentralisées et isolées, les scientifiques doivent s'appuyer sur de grandes bases de données ou des entreprises d'indexation qui fournissent soit un corpus incomplet (critères de sélection), soit affichent seulement des articles de leurs propres plates-formes. De plus, ces moteurs de recherche proposent souvent des fonctionnalités de recherche très limitées. Une solution à cette limitation pourrait être de rendre les systèmes recommandés plus intelligents en intégrant les avancées récentes sur la fouille de données (*text mining*) et l'analyse sémantique. Cela a récemment engendré un intérêt et une attention considérable, et plusieurs approches ont déjà été implémentées (voir Section 2).

Dans cet article, une nouvelle méthode d'extraction de connexions entre les catégories et les domaines des mots-clés des articles scientifiques est proposée. Les limites de notre approche naïve héritée de la recherche exacte ont été soulignées dans Latard et al. (2017) , et cet article fournit une amélioration qui s'attaque à ce problème. Notre recherche a pour but d'intégrer les relations sémantiques dans les moteurs de recherche scientifiques. En effet, la connexion sémantique des mots-clés par leurs catégories / domaines peut être utile pour valider la catégorie principale d'un article. Zhang et al. (2008) ont suggéré que l'utilisation de relations sémantiques entre les mots-clés des articles pourrait améliorer la précision et le rappel de leur méthode d'extraction de mots-clés. De plus, Effendy et Yap (2016) ont avancé que l'utilisation d'outils d'extraction sémantique pour extraire la catégorie d'une conférence pourrait être prometteuse. Par conséquent, l'extraction de relations sémantiques est la première étape vers notre objectif final qui est de rendre les moteurs de recherche plus intelligents. Effectivement, en fonction du nombre de résultats renvoyés (et des catégories / domaines correspondants), un résultat plus raffiné / étendu pourrait être proposé à l'utilisateur.

2 Travaux Similaires

En raison de la croissance exponentielle des données numériques, les méthodes de catégorisation, classification ou plus généralement de regroupement (*clustering*) ou d'extraction de texte ont été largement étudiées dans la littérature scientifique. Dans cet article, nous distinguons les mots clés (mots-clés de l'article) et sujets principaux d'un article. Menaka et Radha (2013) ont développé une méthode basée sur *Term Frequency Inverse Document Frequency (tf-idf)* utilisant WordNet comme base de connaissances pour extraire les sujets principaux des articles scientifiques et ensuite appliquer un algorithme de *Machine Learning — k-Nearest Neighbor (kNN), Decision Trees (DT), Naive Bayes (NB) —* pour les classifier. Zhang et al. (2008) ont développé une méthode probabilistique de champs conditionnels aléatoires (*Conditional Random Fields – CRF*), pour extraire les sujets principaux d'articles scientifiques. Ces méthodes sont les méthodes classiques de classification (tf-idf, extraction des sujets principaux, ...) appliquées à la littérature scientifique. Des études complètes sur les méthodes de classification ont été écrites par Fernández-Delgado et al. (2014) et Berry et Castellanos (2008).

L'analyse de la littérature scientifique n'est pas une tâche simple et a suscité une attention particulière au cours des dernières décennies. Gil-Leiva et Alonso-Arroyo (2007) ont convenu, après plusieurs études de la littérature scientifique, que les mots-clés fournis par les auteurs sont une source d'information très significative. Shah et al. (2003) ont cherché à savoir s'il

était légitime de se concentrer sur le résumé pour extraire les sujets principaux des articles scientifiques (en biologie) et ont conclu que les résumés contiennent le meilleur ratio des sujets principaux par total de mots. De plus, une étude récente et complète de 200 articles sur les systèmes de recommandation dans la littérature scientifique est donnée par Beel et al. (2016). Il en résulte que les approches utilisant les citations ont été largement développées. Effectivement, Reyhani Hamedani et al. (2016) ont appliqué à la littérature scientifique une approche générale utilisant les citations pour calculer la similitude entre les objets. González-Pereira et al. (2010) ont inventé un indicateur de prestige de journaux (indicateur SJR) en combinant les citations et le prestige de la revue citante, à l'aide d'une version modifiée de l'algorithme PageRank. Cependant, les citations peuvent être très générales, voir hors sujet (Ex., des citations philosophiques en science exacte). De plus, les citations sont disponibles seulement dans une petite partie de nos données, ce type d'approche n'est donc pas fiable dans notre contexte.

Pour contrer les inconvénients liés aux systèmes de recommandation basés sur la co-occurrence, certaines approches utilisent le filtrage collaboratif dans diverses applications telles que celles proposées par McNee et al. (2002) et Pennock et al. (2000). L'objectif principal de ces systèmes de recommandation est de proposer du contenu proche de ce que les lecteurs (ou lecteurs similaires) lisent ou aiment, en analysant les interactions des utilisateurs. Cette méthode est particulièrement intéressante car aucune analyse de texte n'est nécessaire. Malheureusement, les approches de filtrage collaboratif font face au problème du démarrage à froid (*cold start*). En effet, les articles ne peuvent être proposés que lorsqu'ils ont déjà été évalués ou accédés par d'autres utilisateurs, et la motivation à participer est souvent très faible.

En raison des limites de ces deux autres approches (filtrage collaboratif et co-occurrence), le filtrage basé sur le contenu est l'approche la plus appropriée à notre contexte. C'est aussi la plus utilisé dans les systèmes de recommandation scientifique, selon Beel. Par conséquent, utiliser le contenu pour proposer les articles les plus pertinents ou pour calculer la similitude entre les articles semble être un choix naturel, et le meilleur. Jiang et al. (2012) ont proposé une méthode qui extrait les problèmes / solutions des résumés et qui calcule des modèles de similarité à l'aide de modèle td-idf et de modèle de thème / concept. Bien que leur approche repose toujours sur des citations pour sélectionner les articles pertinents, le classement des articles se fait en comparant la similitude entre vecteurs. Nascimento et al. (2011) ont implémenté une approche exclusivement basée sur le contenu qui prend un article en entrée et extrait des articles pertinents depuis trois bibliothèques scientifiques. Les candidats potentiels sont ensuite classés en calculant la combinaison linéaire de matrices de similarité cosinus à partir des résumés et des titres.

3 Matériels et Méthodes

3.1 Approche Proposée

Notre approche utilise BabelNet, une base de données lexicographique et encyclopédique multilingue développée par Navigli et Ponzetto (2012) et basée sur la superposition intelligente de lexiques sémantiques (WordNet, VerbNet) et d'autres bases de données collaboratives (Wikipedia et autres données Wiki). Une requête pour un terme à travers BabelNet renvoie des "entrées de dictionnaire", des synonymes, des catégories ou des domaines. C'est donc la base de connaissances que notre approche va exploiter. BabelNet a déjà été largement utilisé pour

la fouille de texte et de données, et la pertinence de ses données a été prouvée. En effet, son utilisation a aidé, entre autre, Gábor et al. (2016) à extraire les données les plus significatives des articles scientifiques, Rashidghalam et al. (2016) pour la génération de résumé à partir de documents, ou encore Romeo et al. (2015) à réaliser la désambiguïsation du sens des mots (*word sense disambiguation – WSD*) pour la classification multilingue des documents.

Cette base de connaissance est intégrée afin de rechercher tous les mots-clés de la base de données de littérature scientifique, Scilit [2]. Développée par MDPI [3], Scilit contient à ce jour les métadonnées de plus de 97 millions d'articles. Pour l'instant, notre approche est évaluée avec un sous-ensemble de 595 articles afin de pouvoir annoter et analyser les résultats (voir Section 4.1 pour plus de détails). Comme exprimé dans la Section 2, les mots-clés fournis par les auteurs et générés à partir du résumé sont les plus significatifs. Étant donné que les mots clés proviennent essentiellement des auteurs, nous les considérons comme légitimes et les exploitons dans cette approche.

Le concept de synset (combinaison de *synonym* et *set*), hérité de WordNet (et inclus dans BabelNet) est défini comme un ensemble de mots partageant la même signification par Navigli et Ponzetto (2012). En d'autres termes, un synset (S) renvoyé par BabelNet peut être considéré comme une entrée de dictionnaire — *ou un mot dans un concept spécifique* — à partir duquel on peut obtenir ses catégories (C), domaines (D), synonymes (syn) et autres données intéressantes. On définit donc $S = \{C, D, syn\}$, et la fonction $F(K)$ retournant un ensemble de synsets ($\{S_1, \ldots, S_n\}$) pour un mot-clé K. BabelNet est composé de 34 domaines globaux (Ex., 'health and medicine' ou 'physics and astronomy') et beaucoup de catégories spécifiques (Ex., 'peripheral nervous system disorders' ou 'exact solutions in general relativity') héritées principalement de Wikipédia.
Notre méthode vise à extraire les connexions entre les catégories (et les domaines) des différents synsets résultant de tous les mots clés de l'article. Une catégorie est dite connectée lorsqu'elle est partagée par au moins deux mots-clés. En faisant cela, nous filtrons le bruit provenant de faux amis d'autres disciplines. La meilleure catégorie $\mathcal{C}_\mathcal{A}$ de l'article $\mathcal{A}$ est calculée par l'équation suivante :

$$\begin{aligned} C_A \in A_C \text{ tel que } \forall\, c' \in A_C,\ count_{C_A} \geq count_{c'} \\ \text{où } count_c = \sum_{i=1}^{n} 1_{A_{S_i}}(c) \\ \text{et } 1_{A_{S_i}}(c) := \begin{cases} 1 \text{ si } c \in\ A_{S_i} \\ 0 \text{ si } c \notin\ A_{S_i} \end{cases} \end{aligned} \tag{1}$$

A_C et A_S sont respectivement toutes les catégories et synsets de l'article A. La même logique appliquée aux domaines fournit plus de connexions car ils sont moins nombreux (34) et plus généraux que les catégories. La Figure 1 illustre la logique principale de notre framework.

3.2 Recherche Exacte

La recherche exacte est l'approche naïve de notre framework qui prend des mots-clés sans préformatage et tente de faire une recherche exacte sur BabelNet. La Figure 2 montre les limites de cette approche. En effet, aucun résultat n'est retourné pour les mots-clés composés

2. http://www.scilit.net
3. http://www.mdpi.com

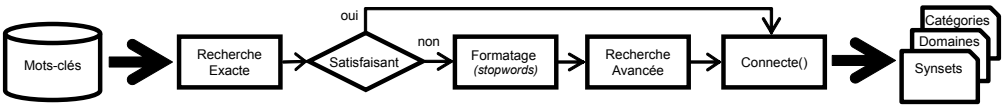


FIG. 1 – *Illustration de la logique générale de notre approche*

(mots-clés composés de plusieurs mots) "flapping flight" et "normalized lift". Ceci est problématique étant donné que de nombreux mots-clés sont en réalité composés — *76% des mots-clés de Scilit*. De plus, plus le mot-clé contient de mots, moins il y a de chances d'obtenir une réponse de BabelNet. Dans cet exemple, "lift coefficient" procure un seul synset, mais il peut y avoir beaucoup de synset potentiels pour un seul mot.

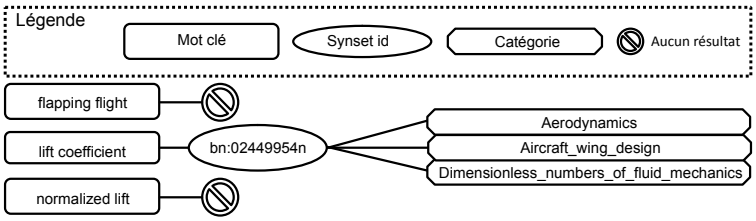


FIG. 2 – *Limites de la recherche exacte : un seul mot-clé sur trois retourne des résultats.*

L'approche par recherche exacte fournit une bonne précision (de 0,95 à 1 — Table 3) mais ne recouvre qu'entre 4% et 22% du total des articles en entrée, en fonction de la valeur du paramètre de seuil α (décrit dans la Section 4). Afin d'améliorer le rappel et de recouvrir plus d'entrées, les mots-clés composés ne fournissant aucun résultat sont séparés par les espaces. Ceci est notre deuxième phase appelée recherche avancée, développée dans la section suivante.

3.3 Recherche Avancée

Lorsqu'une recherche exacte ne renvoie aucun résultat pour un mot-clé composé, les mots vides (*stopwords*) sont supprimés (car ils n'ont aucun sens dans la détection de catégorie / domaine) et le mot-clé est divisé sur les espaces. Cette étape s'appelle la recherche avancée dans le reste de cet article. La division sur espaces offre plus de chances d'obtenir des résultats, mais aussi le risque d'obtenir des synsets dans le mauvais contexte (moins il y a de mot, moins il devient spécifique). Afin de diminuer le risque engendré par cette division, deux modes légèrement différents ont été créés : Split et Multi. Ils tendent tous deux à chercher les différentes combinaisons syntaxiques possibles à partir de ces sous mots-clés. Nous appelons une *window* la fenêtre de sélection des mots et *window size (WS)* sa taille, qui n'est autre que le nombre de mots sélectionnés. En partant de la plus grande fenêtre possible après suppression des *stopwords*, la taille de la fenêtre est réduite jusqu'à ce que des données soient trouvées.

Lorsque des synsets sont retournés pour une recherche contenant plusieurs mots (WS > 1), ils sont considérés comme le résultat du mot-clé initial, même s'il n'y a pas de connexion catégorie / domaine. Cependant, à partir de la recherche de mots uniques (WS = 1), seuls les synsets partageant au moins deux catégories / domaines entre sous mots-clés sont renvoyés, car ils sont moins spécifiques (et donc procurent plus de bruit). La seule exception concerne

les mots-clés à deux mots (Ex., flapping flight). Pour ceux-ci, même les synsets non connectés sont renvoyés car ils sont presque aussi significatifs que le mot-clé initial. Nous perdons en précision avec cette règle (environ 6%) mais gagnons considérablement en rappel (environ 11%).

Split. Le mode Split favorise la proximité entre les mots en exploitant les n-grammes les plus larges possible. En d'autres termes, pour un mot-clé composé "A B C", le mode Split essaye d'abord "A B" et "B C" (mais pas "A C") et s'il n'y a pas de résultat, cherche des résultats pour les mots uniques "A", "B" et "C". ". Un exemple réel est donné dans la Figure 3. Un *round* est le déplacement complet de la fenêtre sur les mots clés. Le processus de réduction de la fenêtre s'arrête à la fin d'un *round*, dès lors que des données sont trouvées. Ainsi, les résultats provenant des étapes moins spécifiques ne polluent pas ceux extraits de recherche plus précises (car plus de mots). Dans l'exemple de la Figure 3, des résultats sont retournés dans le round 2 pour "flight control" et "control systems", donc le round 3 n'est pas exécuté.

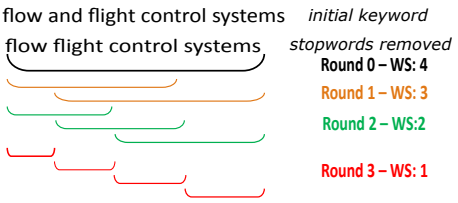


FIG. 3 – *Le mode Split*

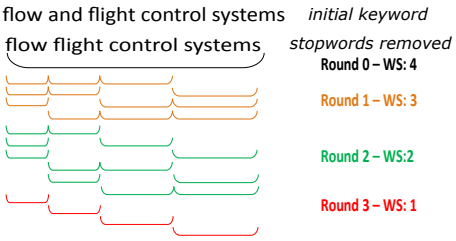


FIG. 4 – *Le mode Multi*

Multi. Le mode Multi suit la même logique que le Split en termes de processus de réduction de la fenêtre. Il démarre également à partir de la plus grande fenêtre possible, mais tente toutes les combinaisons linéaires au lieu de seulement les adjacentes. Pour un mot-clé "A B C", le mode Multi recherche "A B" et "B C" comme le mode Split, mais aussi "A C". Cependant, les éléments ne sont pas permutés ("C A", "C B", "B A") car cela procure plus de bruit. La Figure 4 illustre la logique de propagation de ce mode sur le même exemple que celui donné pour le mode Split (Figure 3). En comparant la Figure 3 et la Figure 4, nous voyons l'avantage principal du mode Multi. Effectivement, "flow control" est plus significatif que "flow flight" pour le mot-clé "flow and flight control systems". Le mode Multi peut cependant également fournir des résultats inattendus en cherchant pour des combinaisons indésirables, en particulier pour de longs mots-clés. Par exemple, la requête "particle Vaidya" est effectuée pour le mot-clé "particle detectors in Vaidya". Dans notre jeu de données, le mode Multi fournit une meilleure précision que le mode Split car il extrait certaines données pour lesquelles les recherches adjacentes ne renvoient aucun résultat. Les résultats seront décrits plus en détail dans la Section 4.

Valeur Ajoutée de la Recherche Avancée. Les modes Multi et Split fournissent les mêmes résultats pour les mots-clés à deux mots, car ils ne diffèrent qu'à partir des mots-clés à trois mots. En regardant notre exemple initial de la recherche exacte (Figure 2), aucune catégorie n'est renvoyée pour le mot-clé composé "flapping flight". En séparant sur les espaces, notre méthode est capable d'extraire "Aerodynamics" comme la catégorie principale de "flapping

flight" à partir de la connexion entre les synsets de ces deux sous mots-clés, comme le montre la Figure 5. Le mot-clé "normalized lift" ne retourne également aucun résultat. Cependant, les 43 synsets de "lift" sont utilisés malgré le manque de connexion, car il s'agit d'un mot-clé à seulement deux mots.

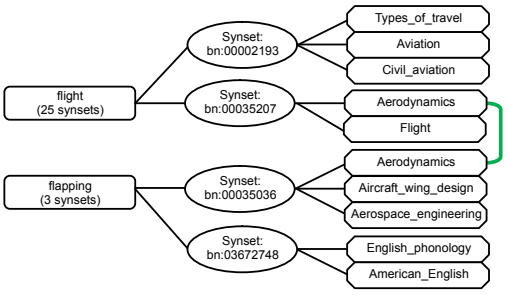


FIG. 5 – *La catégorie "Aerodynamics" est renvoyée comme catégorie principale de "flapping flight". Les autres catégories sont naturellement filtrées par cette connexion.*

Finalement, notre approche est capable de fournir "Aerodynamics" comme catégorie principale de "lift coefficient ; normalized lift ; flapping flight", par l'intermédiaire des modes Split ou Multi. La Figure 6, représentant toutes les différentes manières d'extraire les catégories principales, est un bon résumé de notre framework. En effet, "lift coefficient" résulte de la recherche exacte et une catégorie est retournée pour le "flapping flight" grâce à la recherche avancée. Les 43 synsets hérités de "lift" sont conservés dans "normalized lift" en raison de la règle des mots-clés de deux mots — *les synsets sont conservés même si aucun chevauchement n'est observé à partir de recherches provenant de mots seuls (WS = 1) lors de l'exploitation de mots-clés de deux mots*. Enfin, les synsets de "lift" n'ayant aucun rapport sont naturellement filtrés et la catégorie principale de l'article est extraite avec succès.

3.4 Filtrage

Notre méthode filtre le bruit en supprimant les domaines et catégories non connectés (et sans rapport) dès lors que des domaines / catégories d'autres synsets sont connectés (voir Éq. 1). La Figure 5 montre que cette méthode aide à retourner la catégorie "Aerodynamics" pour "flapping" et "flight" qui renvoient initialement trois et 25 synsets. Cependant l'approche de recherche exacte renvoie encore "Living people ; English-language films ; Celestial mechanics ; American films" comme catégories principales pour les mots-clés "nonlocal gravity ; celestial mechanics ; dark matter" après filtrage. Un bruit constant (**_singer*, **_album*, etc.), n'ayant aucun sens dans notre contexte scientifique a été identifié. Un paramètre peut maintenant être réglé pour filtrer automatiquement du bruit identifié. La plupart du bruit restant est finalement filtré, et "Celestial mechanics" est retournée comme catégorie principale.

La connexion de synsets par leurs domaines génère généralement plus de résultats car ils sont beaucoup plus généraux que les catégories — *il y a seulement 34 domaines dans BabelNet*. De ce fait, les connexions connues sont utilisées pour filtrer les éléments non connectés (catégories / domaines). Cette méthode de filtrage diminue considérablement le nombre de faux positifs. Par conséquent, la précision est améliorée, mais le rappel est réduit en raison de

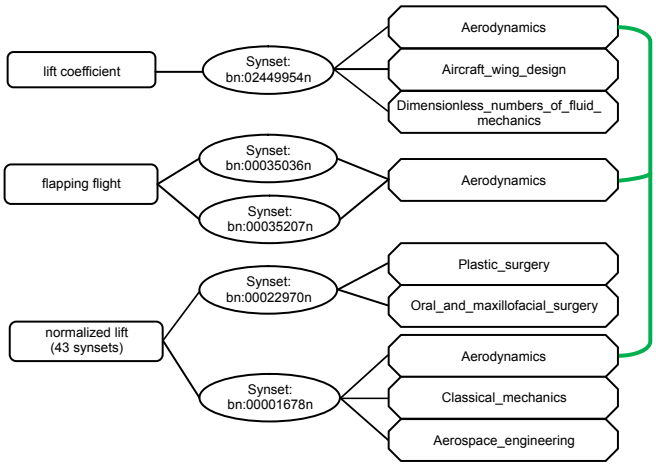


FIG. 6 – *La recherche avancée extrait avec succès la catégorie partagée par l'ensemble des trois mots-clés.*

notre politique de ne proposer que des données connectées en sortie. En effet, les données non connectées ne sont jamais marquées comme valides.

Critère de sélection. La connexion des catégories / domaines à partir des différents synsets peut être réalisée de manière plus ou moins restrictive en modifiant la valeur du critère de sélection minimum (seuil), également appelé paramètre α. Sa valeur s'étend de 1 (plus restrictive) à 4 (plus flexible). Le tableau 1 fournit des explications détaillées sur ses différentes valeurs.

TAB. 1 – *Le paramètre (α) définit le seuil de restriction pour le critère de sélection*

Valeur	Contrainte
1	minimum trois mots-clés ont le même élément en commun
2	deux mots-clés ont un seul et même élément en commun
3	deux mots-clés ont une à trois catégories en commun et le domaine est validé (avec $\alpha = 1$)
4	minimum deux mots-clés ont un, deux ou trois éléments en commun

- un élément peut être une catégorie ou un domaine
- les contraintes d'un α plus petit sont appliquées aux valeurs supérieures
(c'est à dire, 1 est appliqué à 1/2/3/4, 2 à 2/3/4, et ainsi de suite)

4 Résultats

Afin de développer et d'évaluer notre framework, sept revues de deux éditeurs sont utilisées. Tous les journaux sont dans le domaine des sciences physiques sauf un (*Children*), ce qui sert à valider l'application de notre approche à d'autres domaines (ici pédiatrie). En effet, la précision pour *Children* atteint 0.92 pour les catégories, ce qui correspond à la moyenne glo-

bale. Les résultats détaillés sont disponibles depuis ce lien [4]. Notre jeu de données contient tous les articles de quatre journaux (*Galaxies*, *Aerospace*, *Universe*, *Children*) de MDPI, tous les articles de *Preprints* [5] dans le domaine des sciences physiques et cent articles de deux journaux (*Classical and Quantum Gravity*, *The Astrophysical Journal*) de IOP Publishing [6].

4.1 Jeu de données

TAB. 2 – *Échantillon du jeu de données – Des mots-clés du journal Galaxies.*

Keywords	classical general relativity ; cosmology
doi	10.3390/galaxies2010013
Title	Two-Body Orbit Expansion Due to Time-Dependent Relative Acceleration Rate [...]
Keywords	dark energy ; cosmological principle ; inhomogeneous anisotropic universe
doi	10.3390/galaxies2010022
Title	Large Scale Cosmological Anomalies and Inhomogeneous Dark Energy
Keywords	dark energy ; analogue spacetime ; hyperbolic metamaterial
doi	10.3390/galaxies2010072
Title	Metamaterial Model of Tachyonic Dark Energy

Notre jeu de données est composé de 595 articles qui ont été collectés en décembre 2016. Il contient les métadonnées des articles, les résultats des différentes méthodes et les données correctes / erronées pour chaque article (utilisables pour une évaluation automatique). Le dossier compressé, téléchargeable à ce lien[4], contient des résultats détaillés (par journal, par article et par mode). Le Tableau 2 montre un exemple de données en entrée.

4.2 Analyse

Afin d'évaluer l'exactitude de notre approche, les catégories / domaines proposés ont été vérifiés manuellement et annotés comme étant correctes ou incorrectes. La Précision (P) est donnée dans le Tableau 3. Le Tableau 4 est une métrique appelée taux de recouvrement (C), représentant le ratio entre le nombre de catégories correctement proposées et le nombre total d'articles, ce qui permet de visualiser combien de réponses correctes sont retournées pour un échantillon donné. Le Rappel (R) est lui affiché dans le Tableau 5, et $R0$ (Tableau 6) représente le rappel pour les domaines ou catégories identifiables, qui sont ceux pour lesquels il existe au moins une connexion dans les résultats de mots-clés — *NB : un domaine pour un article donné peut être identifiables (c.-à-d. partagé par plusieurs mots-clés), alors que la catégorie pour le même article ne l'est pas (car il n'y a pas de connexion)*. Notre approche ne propose aucun résultat pour les catégories non liées étant donné que nous ne pouvons définir aucun degré de confiance pour celles-ci. Par conséquent, il est logique de différencier le rappel général de tous les résultats renvoyés et celui uniquement pour les identifiables (c'est-à-dire utilisables). Le Tableau 7 représente $F1$, un indicateur unique permettant d'identifier le paramètre α qui fournit le meilleur ratio P/R. $F1(R0)$ (Tableau 8) est une variante utilisant $R0$ au lieu de R.

4. `http://img.mdpi.org/data/latard_egc2018.zip`
5. `http://www.preprints.org/`
6. `http://ioppublishing.org`

TAB. 3 – *Catégories – Précision.*

Recherche	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exacte	**1.00**	0.96	0.96	0.96
Split	0.90	**0.96**	**0.96**	0.95
Multi	0.88	**0.93**	0.92	0.92

TAB. 4 – *Catégories - Recouvrement.*

Recherche	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exacte	0.04	0.18	0.21	**0.22**
Split	0.08	0.27	0.35	**0.38**
Multi	0.08	0.27	0.35	**0.38**

TAB. 5 – *Catégories - Rappel.*

Recherche	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exacte	0.09	0.38	0.44	**0.47**
Split	0.14	0.50	0.65	**0.70**
Multi	0.17	0.55	0.72	**0.77**

TAB. 6 – *Catégories - Rappel ($R0$).*

Recherche	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exacte	0.18	0.76	0.89	**0.93**
Split	0.18	0.66	0.85	**0.92**
Multi	0.20	0.63	0.83	**0.89**

Grâce à ces tableaux, nous pouvons étudier les métriques en fonction de la valeur de α. Multi avec $\alpha = 4$ est le meilleur compromis en ce qui concerne les quatre métriques (P, R, C et $F1$). Effectivement, il fournit une bonne précision (0.92), un rappel acceptable pour les éléments identifiables (0.89) et recouvre correctement 38% des 595 articles. Une tendance similaire est observée pour le mode Split, qui fournit même des résultats légèrement meilleurs si l'on observe uniquement les résultats identifiables ($R0$). Si la précision importe plus, α peut être diminué, mais le rappel et le recouvrement diminuent de manière significative . Si acceptable, il peut être plus intéressant de diminuer légèrement la précision afin de gagner considérablement en rappel et recouvrement. Par soucis de cohérence, nous n'avons donné que les résultats pour les catégories dans cet article. Pour ce qui est des domaines, le mode multi avec $\alpha = 3$ fournit les meilleurs résultats, car la précision diminue de trop lorsque $\alpha = 4$ (0.88). Globalement, la même tendance est observée avec une bonne précision ($P = 0,93$), mais un rappel ($R0 = 0,96$, $R = 0,92$) et un recouvrement ($C = 0,74$) beaucoup plus élevé . Le ratio précision / rappel / recouvrement est meilleur pour l'extraction des domaines que pour celle des catégories. Ce n'est pas surprenant car les domaines sont bien plus généraux que les catégories et par conséquent, se chevauchent plus souvent.

5 Conclusion

Le développement des systèmes de recommandation scientifique plus intelligents est crucial pour aider les scientifiques dans leur phase de recherche bibliographique, étape obligatoire et fastidieuse. L'approche présentée ici est la première étape pour créer un tel système intelligent. Nous prévoyons d'améliorer la précision de notre framework en extrayant le part-of-speech (POS) des mots-clés composés à l'aide d'un analyseur syntaxique lors de la recherche avancée. Ainsi, nous pourrions rapprocher les articles en termes de triplets (sujet, verbe, objet), à la manière de Amir et al. (2016). De plus, nous planifions d'effectuer une évaluation plus poussée, utilisant des jeux de données standards ou bien une évaluation en ligne.

TAB. 7 – *Catégories - F1.*

Recherche	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exacte	0.16	0.54	0.61	**0.63**
Split	0.24	0.65	0.77	**0.80**
Multi	0.29	0.69	0.81	**0.84**

TAB. 8 – *Catégories - F1 ($R0$).*

Recherche	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
Exacte	0.30	0.85	0.92	**0.94**
Split	0.31	0.78	0.90	**0.94**
Multi	0.32	0.75	0.87	**0.90**

D'autre part, en validant les entrées dans le dictionnaire à partir des catégories principales, le sens des mots-clés (et de l'ensemble de leurs synsets) est également vérifié. Dans de futurs travaux, nous pourrions utiliser les catégories extraites et leurs données BabelNet correspondantes pour construire un graphique représentant un arbre de connaissance scientifique afin de regrouper dynamiquement les articles similaires.

En effet, les résultats (Section 4) confirment que l'utilisation de relations sémantiques entre mots-clés fournit un bon moyen de classifier les articles scientifiques. Il fournit effectivement une bonne précision (environ 0,92 pour les catégories et les domaines) et un bon rappel $R0$ (0,89 pour les catégories, 0,96 pour les domaines). Enfin, les catégories correctes sont trouvées pour 38% des articles, et 74% d'articles ont obtenu des domaines corrects. Notre travail ouvre donc de nombreuses possibilités de recherches futures que nous envisageons d'explorer à l'avenir.

Références

Amir, S., A. Tanasescu, et D. A. Zighed (2016). Une mesure de similarité entre phrases basée sur des noyaux sémantiques. In *EGC*, pp. 141–146.

Beel, J., B. Gipp, S. Langer, et C. Breitinger (2016). Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries 17*(4), 305–338.

Berry, M. W. et M. Castellanos (2008). *Survey of text mining II*. Springer.

(Editorial), N. (2012). Gold in the text ? *Nature 483*(7388), 124–124.

Effendy, S. et R. H. C. Yap (2016). The Problem of Categorizing Conferences in Computer Science. In *TPDL*, pp. 447–450. Springer.

Fernández-Delgado, M., E. Cernadas, S. Barro, et D. Amorim (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research 15*(1), 3133–3181.

Gábor, K., H. Zargayouna, D. Buscaldi, I. Tellier, et T. Charnois (2016). Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *LREC*, pp. 3694–3701.

Gil-Leiva, I. et A. Alonso-Arroyo (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology 58*(8), 1175–1187.

González-Pereira, B., V. P. Guerrero-Bote, et F. Moya-Anegón (2010). A new approach to the metric of journals' scientific prestige : The SJR indicator. *Journal of Informetrics 4*(3),

379–391.

Jiang, Y., A. Jia, Y. Feng, et D. Zhao (2012). Recommending academic papers via users' reading purposes. In *RecSys*, pp. 241–244. ACM.

Latard, B., J. Weber, G. Forestier, et M. Hassenforder (2017). Towards a Semantic Search Engine for Scientific Articles. In *TPDL*, pp. 608–611. Springer.

McNee, S. M., I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, et J. Riedl (2002). On the recommending of citations for research papers. In *CSCW*, pp. 116–125. ACM.

Menaka, S. et N. Radha (2013). Text classification using keyword extraction technique. *International Journal of Advanced Research in Computer Science and Software Engineering 3*(12).

Nascimento, C., A. H. Laender, A. S. da Silva, et M. A. Gonçalves (2011). A source independent framework for research paper recommendation. In *JCDL*, pp. 297. ACM Press.

Navigli, R. et S. P. Ponzetto (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193*, 217–250.

Pennock, D. M., E. Horvitz, S. Lawrence, et C. L. Giles (2000). Collaborative filtering by personality diagnosis : A hybrid memory-and model-based approach. In *UAI*, pp. 473–480. Morgan Kaufmann Publishers Inc.

Rashidghalam, H., M. Taherkhani, et F. Mahmoudi (2016). Text summarization using concept graph and BabelNet knowledge base. In *AIR*, pp. 115–119. IEEE.

Reyhani Hamedani, M., S.-W. Kim, et D.-J. Kim (2016). SimCC : A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences 334-335*, 273–292.

Romeo, S., D. Ienco, et A. Tagarelli (2015). Knowledge-based representation for transductive multilingual document classification. In *ECIR*, pp. 92–103. Springer.

Shah, P. K., C. Perez-Iratxeta, P. Bork, et M. A. Andrade (2003). Information extraction from full text scientific articles : Where are the keywords ? *BMC bioinformatics 4*(1), 20.

Zhang, C., H. Wang, Y. Liu, D. Wu, Y. Liao, et B. Wang (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems 4*(3), 1169–1180.

Summary

Given that the data deluge in scientific bibliographic research is increasingly time-consuming, staying up-to-date in any specific area is a tedious and complex task. Our final goal is to create an intelligent search engine that also takes into account the content of scientific articles. The preliminary work presented in this article is the starting point of such a system. Semantic relations between the different keywords of the articles are extracted in order to categorize the articles and bring them closer to other similar articles . By exploiting the correlations between categories and domains of keywords, inherited from BabelNet, our method will be able to group related articles in terms of keywords' similarity.

An automatic extraction method of static and dynamic spatial contexts from texts

Ludovic Moncla* Mauro Gaio**
Ekaterina Egorova*** Christophe Claramunt*

*Naval Academy Research Institute, Brest, France
ludovic.moncla,christophe.claramunt@ecole-navale.fr
**LIUPPA, Université de Pau et des Pays de l'Adour, Pau, France
mauro.gaio@univ-pau.fr
***Department of Geography, University of Zurich, Zurich, Switzerland
ekaterina.egorova@geo.uzh.ch

Abstract. Spatial descriptions, with or without motion, are the main issues addressed by this paper. We describe construction grammars implemented in the PERDIDO platform with cascaded finite-state transducers which aims at marking and formalizing relations between extended named entities, geographical terms, spatial relations and motion verbs. These grammars can be seen as a computational synthesis of the work on the expression of space and motion in natural language. The proposed method for geographical information extraction has been tested for three different projects within the digital humanities using specific corpora. The first task deals with the extraction of place names from French novels, the second task deals with the extraction of motion events from hiking descriptions written in Romance languages (French, Spanish and Italian) and the third task aims at identifying fictive motion expressions in English alpine journals.

1 Introduction

It is established that one of the best ways to approach spatial semantics is through its representations in language. In all the viable representations, two subsets may be distinguished. Representations where the spatial situation described is motionless and representations where the spatial relation between entities changes over time. In other words, the spatial context may be static or dynamic. Whatever the context is, spatial descriptions involve three main components: a located entity called "target" (Vandeloise, 1991); a reference entity called "landmark" (Langacker, 1987; Vandeloise, 1991) and a spatial relation between these two entities.

For static descriptions, the relation is often carried by at least one adpositional element applied to the noun denoting the landmark (Levinson and Wilkins, 2006). For dynamic descriptions, motions events or displacements are introduced by one or more verbal and adpositional elements also applied to the noun representing landmark. Although these patterns are not unique (see e.g. Levinson and Wilkins (2006)), it has been observed that landmarks are

larger, more salient and stable than targets, since the main purpose of this kind of descriptions is to locate one entity with respect to another. Regarding dynamic scenes, a wide range of publications have specifically addressed the expression of motion in language (Talmy, 1985; Tenny and Pustejowsky, 1999; Talmy, 2000; Hickmann, 2006). In particular, lexicon-grammar approaches have significantly contributed to the represenation of dynamic spaces (Asher and Sablayrolles, 1995; Borillo and Sablayrolles, 1993; Laur, 1993; Muller and Sarda, 1998; Aurnague, 2011). Other studies oriented to an integration of additional spatial semantics have integrated the ontological nature of the landmark entities denoted by the nominal elements that propositions and verbal units select. Among the many phenomena tackled in this literature, an interesting pattern that appears is that some motion verbs and constructions are likely generate some static interpretations. This phenomenon is often called "fictive motion" (Talmy, 2000) or "non-actual motion" (Blomberg and Zlatev, 2014).

Spatial descriptions, with or without motion, are the main issues addressed by this paper. A first one mainly concerns the expression of the located entity called target. Thus, the task of "Named Entity Recognition and Classification" (i.e., NERC algorithms) is considered to play an important part in the processing of spatial descriptions. When considering such motionless spatial descriptions, a first experiment has been done to automatically recognize and extract the places mentioned in the context of French novels. This process also known as "geoparsing" can be non-straightforward for fictional texts because a novelist has often multiple ways of evoking a given place: either directly (by giving an explicit name) or more elusively (by using relative references, e.g., near, behind, or two blocks further, relative to other places mentioned before). Some places can be even deliberately disguised and others can be completely imaginary. These different cases, among many others, can be found in a same novel. Automatizing the recognition of all these kinds of places is even more difficult when referring to ancient texts (Matei-Chesnoiu, 2015). Additionally, the way real vs. fictive motion occurs in discourse is a quite unexplored question.

A second set of issues addressed in this paper concerns the semantic and syntactic relationships between the verb and the possible adverbial elements appearing in motion descriptions. More precisely with respect to fictive or non-actual motions, our objective is oriented to the identification of the whole range of verbs that are likely to occur in fictive motion descriptions. The ontological nature of the target entities appearing in this kind of interpretation of motion verbs have been another tackled issue for which an in-depth analysis of French texts have provided valuable insights.

Two different experiments have been conducted to illustrate the potential of our method oriented towards an automatic information extraction of real and fictive motion expressions in texts. The first experiment focuses on motion and geo-spatial information extraction for itinerary reconstruction from texts written in Romance languages. The second experiment implements a similar method for automatic extraction and classification of fictive motion expressions in an English corpus.

2 Geoparsing places

2.1 Construction grammar for Extended Named Entities extraction

It is generally accepted that proper name is the most frequent component of a Named Entity (NE). As proper name is a linguistic issue still under discussion, we have adopted a definition according to various criteria. Words (one or several following each other) starting with an uppercase letter is the most commonly cited criterion (Fourour et al., 2002), especially in Romance languages. But this criterion alone is not discriminatory because it suffers of many exceptions (vallée de la mort or vallée de la Mort, gave de Pau, massif armoricain or Massif armoricain). Again in Romance languages the second criterion is morphosyntactic, it shows that most of the time a proper name does not involve determiners and could not be inflected, but this criterion also suffers of some exceptions on at least one of the two rules (Le Havre, La Rochelle, La Pierre Saint Martin, Gaves Réunis) or both (Les Deux Alpes). However, let us not confuse with expressions like l'Aquitaine, les Pyrénées, le Rhône, where the determiner is not really part of the proper name even if the rule could be considered as waived. A third criterion is the non-significance of a proper name, but here again this criteria suffers of exceptions (Archipel des Sanguinaires, Petit Mont Blanc, gare du Nord). A last criterion is the uniqueness of the proper name reference but as the others criteria some exceptions can be found and moreover some common names can have a single reference (the sun, the house on the left after the crossroad). As regards the syntactic shape we selected Jonasson's (Jonasson, 1994) categories of proper names. A proper name can be categorised as pure or descriptive. Pure proper names are simple (i.e., composed of a single lexeme) or complex (i.e., composed of several lexemes) and are composed of proper names only. Descriptive proper names refer to a composition of proper names and common names (i.e., descriptive expansion). In other words, a descriptive proper name overlaps a pure proper name and refers to NE built with a pure proper name and a descriptive expansion. This expansion can change the implicit type (e.g., location, person, etc.) of the initial pure proper name and then of the NE. However, the presence of a proper name in NE is not mandatory. In some specific contexts, la Ville Lumière, le pays du Soleil-Levant ou le sommet du Monde, could be considered as NE. As mentioned in (Kleiber, 1981) these expressions are part of what the author calls "la description définie" (the defined description) namely expressions having the ability to make reference to an entity identifiable as such in a given context.

Finally, according to these concepts of proper name and Named Entity, we introduced in a previous work the concept of Extended Named Entity (ENE) (Gaio and Moncla, 2017). The notion of "extended" is pretty near to the one named "mixed" proposed in (Fourour et al., 2002).

We defined several levels of overlapping (0, 1, 2, etc.) for the representation of ENE. Each level is encapsulated in the previous one. For instance, level 0 refers to pure proper names and can be seen as the core component of an ENE. Thus, we consider NE as a special kind of ENE. Then, level >0 refers to descriptive proper names composed of another descriptive proper name or of a pure proper name (i.e., an entity of level 0) and a common noun. Descriptive expansions may or may not change the implicit or default nature of the object described by the proper name. Indeed, when the associated term has not the same type of the intrinsic or default feature type of the pure proper name, it defines a new entity that overlaps the pure proper name one.

(1) a. *sindaco di* [*Genova*]
'mayor of Genoa'

b. *club de pelota vasca de la* [*ciudad de* [*San Sebastián*]]
'Basque pelota club of San Sebastian city'

c. *dépression karstique dans le* [territoire aride au sud de la [*région* d' [Aragon]]]
'karstic depression in the arid land south of the Aragon Region'

Examples (1a-1c) show that an entity may contain the name of another entity, and that the new entity may have a different feature type. For instance, 'Genoa' refers to a location whereas 'mayor of Genoa' refers to a person or a function (see example (1a)). Additionally, there is not really a limit to the overlapping. However, it is quite uncommon to find an ENE of a level greater than 3 (see example (1c)). We have considered the annotation of ENE as a shallow parsing and the grammar to be used as a specific construction. The core of the grammar [1] is set as follow:

$S \rightarrow ENE$
$ENE \rightarrow ENEA \quad | \quad (Term)\ ENER$
$ENER \rightarrow Offset\ ENEA \quad | \quad Offset\ ENER$
$ENEA \rightarrow (Term)\ ProperNoun \quad | \quad Term\ ENEA$
$Term \rightarrow Nominal\ Det$

This grammar integrated in the PERDIDO platform is implemented as a cascade of finite-state transducers using the CasSys program available in the Unitex/GramLab platform [2]. We developed an hybrid solution combining a preprocessing step for the disambiguation of grammatical categories (using part-of-speech taggers) and the cascade of transducers. The proposed PERDIDO NERC tool is based on a bottom-up strategy where each level of the ENE is marked, from the pure proper name to the whole ENE. It can distinguish between two types of ENE, 'absolute' referring to standard spatial ENE and 'relative' referring to spatial ENE associated with spatial relations (i.e., 'offset' and 'measure'). The cascaded finite-state transducers produce a generic annotation of ENE (i.e., ENE boundaries are identified but not classified). Therefore, for geocoding, PERDIDO implements a gazetteer lookup method to classify them and uses the local linguistic context (i.e., feature type within ENE), when available, to identify subtypes associated with ENE (e.g., city, street, church) to classify them and, more specifically to identify the spatial ones.

With respect to the specific problem of the NERC category of place names, one might move beyond reducing a place to a name and then geocoded with a single set of coordinates, a model that is still predominant in Geographic Information Science (Purves and Derungs, 2015). For instance, taking the example (1c), using the PERDIDO NERC tool, this produces the result represented in a feature structure form in Fig. 1. We argue that for a fine-grained task, especially in digital humanities, such as marking, classifying and disambiguating named entities, it is essential to consider ENE (1c) as a composition of entities. In a such case, standard NER

1. $Offset$ can be seen as an adverbial clause
2. http://unitexgramlab.org/

tools such as OpenCalais [3], OpenNLP [4] and Stanford-NER [5] consider only the entity 'Aragon Region', and therefore lead to inaccuracies in classification and/or disambiguation.

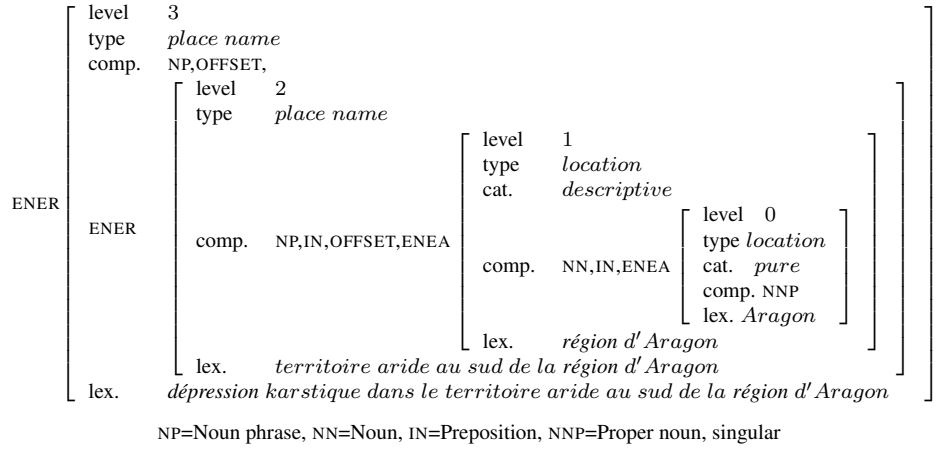


FIG. 1: Feature structure representation of ENE (1c)

2.2 Automatic extraction of place names from French novels

The method developed so far for automatically retrieving place names has been applied to French novels of the 19th century (Moncla et al., 2017). This work has been conducted in the context of a research project whose aims are to provide a method for the cartographic analysis of Paris street names in French novels. The corpus used for this experiment is composed of 31 French novels covering different periods of the 19th century centered on Paris.

As described in the previous section, the proposed construction grammar is applied to the extraction of "complex" place names such as example (2) [6]. This allows us to extract relative place names composed of spatial relations (*offset*) and ENE. This is particularly useful for the extraction of the static spatial context and not just standard place names. Moreover, this method can also be applied for the extraction of place names at different scales, such as places included inside other places (e.g., buildings or streets located inside a city or another administrative entity).

(2) *[...] se trouvait au coin de la* ***rue des Poissonniers*** *et du* ***boulevard de Rochechouard***
'[...] is located at the corner of ***Poissonniers Street*** *and* ***Rochechouard Boulevard***'

Figure 2 shows an excerpt of the XML/TEI annotation [7] of the place name shown in example (2). This figure shows the construction of the relative ENE (identified by the TEI element:

3. http://www.opencalais.com/
4. http://opennlp.apache.org/
5. http://nlp.stanford.edu/ner/
6. This sentence is extracted from the novel *L'Assommoir* written by Emile Zola in 1877.
7. The values of attributes *type* and *subtype* of the *geogName* element refer to GeoNames feature codes: http://www.geonames.org/export/codes.html

placeName) using an offset (*at the corner*) and two absolute ENE (*Poissonniers Street* and *Rochechouard Boulevard*).

```
<placeName type="relative" subtype="compound">
  <offset type="inclusion">
    <w lemma="au" type="PREPDET">au</w>
    <w lemma="coin" type="PREP">coin</w>
    <w lemma="de" type="PREP">de</w>
  </offset>
  <placeName n="1" type="absolute">
    <geogName type="R" subtype="ST">
      <w lemma="le" type="DET">la</w>
      <geogFeat>
        <w lemma="rue" type="N">rue</w>
      </geogFeat>
      <w lemma="de" type="PREP">des</w>
      <name>
        <w lemma="Poissonniers" type="NPr">Poissonniers</w>
      </name>
    </geogName>
  </placeName>
  [...]
</placeName>
```

FIG. 2: XML/TEI annotation of a place name using the PERDIDO NERC tool

According to the results provided by the PERDIDO NERC tool, 112 descriptive expansions of ENE referring to geographical feature types were found in the corpus. In particular, street is the most used geographical feature, which confirms the great interest in these novels for the cartographic analysis of Paris (Moncla et al., 2017) using street names. The proposed method can be used to generate diagrams (showing indicators such as the distribution of the number of occurrences of street names compared to the number of distinct streets mentioned) or maps built using geohistorical gazetteers (see Fig 3). Furthermore, the preliminary results described by Moncla et al. (2017), highlight the great interest for digital humanities in combining the PERDIDO NERC tool with a textometric analysis tool to provide automated analysis of novels based on spatial named entities. Indeed, the direct access to a corpus of texts through the use of place names significantly transforms the ways in which space and fictional landscapes can be explored. It becomes possible to interactively and simultaneously browse through geographical and literary space.

3 Retrieving the dynamic space context from texts

3.1 Construction grammar for motion expressions

For a better understanding of the spatial context, linguists have highlighted the importance of the use of motion verbs and spatial relations, especially in Romance languages (Aurnague, 2011). This leads us to take into account movement verbs and spatial offsets in the parsing process. The core of the 'VT' grammar proposed hereafter can be seen both as a specialisation and as an extension of the ENE construction grammar and it aims to be a computational attempt to provide a synthesis of previous works on how language expresses displacement (Talmy, 1983; Vandeloise, 1986), and on how movement verbs are used in some sentences (Pourcel

FIG. 3: Occurrences of street names represented using proportional lines (Moncla et al., 2017)

and Kopecka, 2005), and on how these verbs are combined with different prepositions (Boons, 1987). The core of the grammar is as follows:

$S \rightarrow V\ \ T$
$V \rightarrow Verb \mid Verb\ \ SO$
$C \rightarrow Conjonction \mid ,$
$LT \rightarrow ENE\ \ C\ \ T$
$T \rightarrow (SO)\ (det)\ ENE \mid (SO \mid ENE)\ T \mid (SO)\ LT$

The symbol V represents a set of movement verbs and the symbol T a set of n-tuples, i.e., a composition of elements belonging respectively to three sets: SO a set of spatial offsets (that can be seen as a spatial adverbial clause), TG a set of geographical noun phrases and E a set of ENE.

(3) *Descendre sur le territoire aride au sud de la région d'Aragon*
'Go down onto the arid land south of the Aragon region.'

Example (3) has the following VT structure= (v, t), with: v = *descendre*, t = *sur le territoire aride au sud de la région d'Aragon*. With t respectively composed of: $tg_3 = \emptyset$, so_3 = *sur*, ENE_2 = *territoire aride au sud de la région d'Aragon*, tg_2 = *territoire aride*, so_2 = *au sud de*, ENE_1 = *région d'Aragon*, tg_1 =*région*, $so_1 = \emptyset$, ENE_0 = *Aragon*.

The set SO of spatial offsets is composed of locative phrases in which, at least in verb-framed languages such as French, the role of prepositions is central. A large number of studies have shown that prepositions are involved in the operation of spatial tracking, or location. With respect to the location concept, following Talmy's work Talmy (1983) and Vandeloise's Vandeloise (1986) proposals, prepositions contribute significantly to reconciliate two entities: a locator and a localised entity (i.e., a landmark and a target in Vandeloise's terms). The part of the phrase used as locator must have spatial properties that facilitate its identification and the explanation of the spatial relationship in which it is involved. Boons (1987) proposed to classify motion verbs according to the aspectual properties of movement called 'aspectual polarity'. The three aspectual polarities are initial (e.g., to leave), median (e.g., to cross) and final

(e.g., to arrive). Without changing the intrinsic aspectual polarity of the verb, the preposition can change what could be called the focus of the displacement. More specifically, the association of a motion verb with a spatial preposition can change the focus of the displacement and take on the aspectual polarity of the preposition instead of that of the verb (Laur, 1993). Undeniably, 'leaving from Paris' and 'leaving for Paris' are two expressions with radically opposite focus of the displacement.

The bottom-up parser, based on the VT grammar and implemented with a cascade of transducers within the PERDIDO platform, can be viewed as searching through the space of possible parse trees to find the correct parse tree for a given 'VT' phrase.

3.2 Reconstruction of itineraries from texts

For experiment purposes, this construction grammar has been applied to the extraction of the dynamic space context from texts. More specifically, we try to use the information of motion expressed in texts to automatically reconstruct trajectories of displacements.

(4) a. *[Emprunter] successivement* ***rue des Capucins*** *et* ***rue de Compostelle***.
'Walk down Capucins Street and then Compostelle Street.'

b. *[Prendre] à gauche après l'entrée de* ***l'usine de Fontanille***.
'Turn left after the entry to the Fontanille factory.'

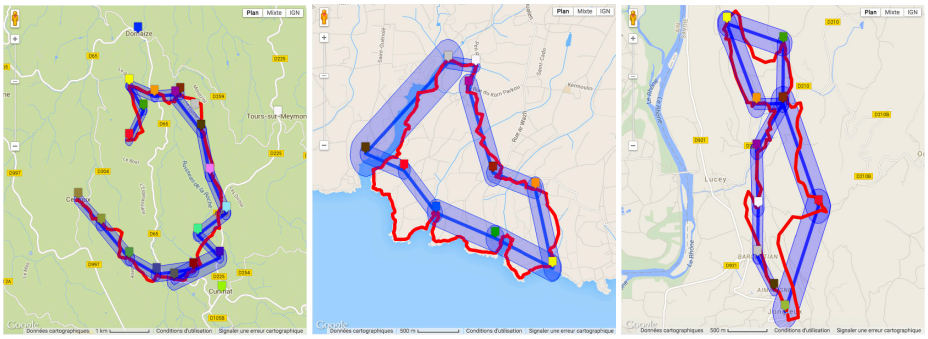
c. *[Suivre] la route depuis le* ***hameau Lic*** *jusqu'à la* ***Chapelle Saint-Roche***.
'Follow the road from the hamlet Lic to the Chapelle Saint-Roche.'

For the automatic reconstruction of itineraries from texts, we proposed a multi-criteria approach combining quantitative and qualitative criteria based on knowledge extracted from the text and geographic databases (Moncla et al., 2016). The proposed method builds a weighted complete graph using the multi-criteria approach where edges represent route segments and vertices represent locations. Then, in order to identify the sequence of waypoints (excluding landmarks) and build an approximation of a plausible footprint of the itinerary described, the graph is transformed into a directed acyclic graph using a minimum spanning tree and spatio-temporal information extracted from the text (see Fig. 4).

For the evaluation of our approach, we used a multilingual corpus (French, Spanish and Italian) of 90 hiking descriptions manually annotated. Each document in the corpus describes one trail and it is associated with the real trajectory (GPS) of the route (used as a comparison basis). Hiking descriptions are a specific type of document describing displacements using geographical information, such as toponyms, spatial and motion relations, and natural features or landscapes, such as shown on example (4a)-(4c).

	# of ENE	*Recall*	*Precision*	*SER*
French	660	95%	96%	17%
Spanish	421	97%	99%	15%
Italian	475	84%	98%	32%
total	1556	92%	97%	21%

TAB. 1: Evaluation of the NERC task with Perdido.



(a) (b) (c)

FIG. 4: Automatic reconstruction of itineraries (Moncla et al., 2016)

Table 1 shows recall, precision and SER scores for the NERC task according to the reference number of ENE in the Perdido gold-standard corpus. The SER (Slot Error Rate) (Makhoul et al., 1999) is the ratio of the total number of slot errors, i.e., insertions, deletions and substitutions (wrong classification and wrong boundaries), divided by the total number of relevant results in the reference. We compared the results obtained with Perdido and with the CasEN system (Friburger and Maurel, 2004) for the automatic annotation and classification of named entities. Although, CasEN obtains good results on a corpus of French newspapers, it obtains an SER score of 51% using the Perdido corpus of French hiking descriptions. This score is mainly explained by the fact that CasEN uses dictionaries of proper names whereas Perdido uses linked data resources. More details about the evaluation of the NERC task for the three languages on the Perdido corpus are given in (Moncla, 2015).

Additionally, the corpus analysis shows that only 2% of ENE are not spatial entities. Furthermore, 810 occurrences of spatial ENE are contained within a VT structure (i.e., 53%) and 47% are associated with feature types (i.e., 53% of spatial ENE belong to the level 0) and a very few number of spatial ENE (3%) are built with more than one expansion (level >1). Additionally, about 59% of verbs are motion verbs (i.e., 1985 occurrences). Median and final motion verbs are the most frequent ones and only 3% of verbs belonging to a VT structure refer to verbs of perception (i.e., 113 occurrences).

4 Fictive motion: static and dynamic scenes

Fictive motion (FM) is an example of the metaphoric nature of human thought and language (Lakoff and Johnson, 2008) that can represent a challenge in the task of automatic identification of motion events in text. Essentially, this linguistic structure represents a static spatial entity as moving, as in (5a) and cannot be interpreted literally. Moreover, there are two types of FM (Matsumoto, 1996). Type I is a static description of a spatial entity and it location in space, as in (5a). Type II is based on "the actual motion of a particular moving entity at a particular time" Matsumoto (1996, p. 361), as exemplified in 5b – imagine this sentences being uttered by a driver, who is actually moving (in the car).

(5) a. The mountain [range goes] from Canada to Mexico.
b. This [highway will enter] California soon.

Egorova et al. (2016) examined the use of FM in a corpus of alpine texts "Text+Berg" (Bubenhofer et al., 2015). They queried the corpus for a spatial entity (based on a list of terms) followed by a verb and manually created a corpus of FM out of candidate phrases. Both types of FM were found in alpine narratives, alongside with two distinct subcategories of Type I: description of a vista along the way (6a) and description of general spatial knowledge about a larger geographic area 6b). Type II, as stated in the definition, encodes the actual motion of the mountaineer, as in (6c). All the three uses of FM (Type I, vista; Type I, spatial knowledge; Type II) were subsequently annotated.

(6) a. Beyond, the desert [hills rose] to the Russian-China border. (Type I, vista)
b. This wonderful [chain runs] NE to SW for some 13 km. (Type I, spatial knowledge)
c. The [ridge went] on forever, but after what seemed an age... (Type II)

We use the "Text+Berg" corpus (Bubenhofer et al., 2015) and the subcorpus of annotated FM (Egorova et al., 2016) to automatically reproduce the half-automated extraction and manual annotation of FM into the three types performed by (Egorova et al., 2016).

For the extraction of FM, we use the lists of geographic entities and motion verbs from (Egorova et al., 2016). Although a list-based extraction is straightforward, resulting in high recall, dealing with various types of false positives – e.g. factive motion (as in 7a) or the use of motion verbs in metaphoric sense s (as in 7b) – represents an interesting task, requiring a set of additional rules that we develop within the PERDIDO platform.

(7) a. [Half the peak fell] in prehistoric times.
b. Out of sheer jealousy the mighty [mountain went on war] against Carihuairazo...

To classify the identified FM into the three types, we identify concepts that can be used for their differentiation. For example, (8) is a vista because of the explicit inclusion of the observer into the frame of reference. We further operationalize these concepts through linguistic structures, based on the literature and thesauri (e.g. expanding potential linguistic encodings of a concept through synonyms or words in the same semantic field).

(8) Down below us the [glacier snaked] away.

5 Conclusion

In this paper, we described a method for the automatic extraction of static and dynamic spatial context from texts. We have proposed construction grammars implemented in the PERDIDO platform with cascaded finite-state transducers which aims at marking and formalizing relations between ENE, geographical terms, spatial relations and motion verbs. These grammars can be seen as a computational synthesis of the work on the expression of space and motion in natural language.

The proposed geoparser tool has been tested for three different tasks (i.e., retrieving place names, motion events and fictive motion expressions) using several corpora (i.e., French novels, French, Spanish and Italian hiking descriptions and English Alpine journal) related with

the digital humanities. We first described the great interest of the concept of ENE for retrieving "complex" place names describing a static spatial context. Then, we also described how local geo-spatial information (referring to space and motion) extracted from the texts can be used for the construction and the representation of more complex geographical objects: here an itinerary. Finally, we have adapted our approach for the extraction of fictive motion expressions which can refer to both static and dynamic spatial descriptions.

The growth of digital corpora opens new perspectives regarding future work in digital humanities and more specifically developing natural language processing and data mining solutions. These few examples show the diversity of needs in the field of digital humanities but also a certain uniqueness in the way people refers to space in a static or dynamic context. This observation strengthen the idea of proposing to the community a set of tools, such as the Web services implemented in the PERDIDO platform, in order to build processing chains adapted to different tasks and to an important variety of needs.

References

Asher, N. and P. Sablayrolles (1995). A typology and discourse semantics for motion verbs and spatial PPs in french. *Journal of Semantics 12*(2), 163–209.

Aurnague, M. (2011). How motion verbs are spatial: The spatial foundations of intransitive motion verbs in French. *Lingvisticae Investigationes 34*(1), 1–34.

Blomberg, J. and J. Zlatev (2014). Actual and non-actual motion: Why experientialist semantics needs phenomenology. *Phenomenology and the cognitive sciences 13*(3), 395–418.

Boons, J.-P. (1987). La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. *Langue Française* (76), 5–40.

Borillo, M. and P. Sablayrolles (1993). The semantics of motion verbs in french. In *Proceedings of the 13th International Conference on Natural Language Processing of Avignon*, pp. 24–28.

Bubenhofer, N., M. Volk, F. Leuenberger, and D. Wüest (2015). Text+Berg-korpus (release 151_v01). XML-Format. The Alpine Journal 1969-2008.

Egorova, E., G. Boo, and R. S. Purves (2016). "the ridge went north": Did the observer go as well? corpus-driven investigation of fictive motion. In *International Conference on GIScience Short Paper Proceedings*.

Fourour, N., E. Morin, and B. Daille (2002). Incremental Recognition and Referential Categorization of French Proper Names. In *LREC 2002 Third International Conference on Language Ressources and Evaluation*, Las Palmas, Canary Islands.

Friburger, N. and D. Maurel (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science 313*(1), 93–104.

Gaio, M. and L. Moncla (2017). Extended named entity recognition using finite-state transducers: An application to place names. In *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017)*.

Hickmann, M. (2006). The relativity of motion in first language acquisition. *Space across Languages: Linguistic Systems and Cognitive Categories. Amsterdam: Iohn Benjamins*,

281–308.

Jonasson, K. (1994). *Le nom propre*. Duculot, Belgique, Louvain-la-Neuve.

Kleiber, G. (1981). Problèmes de référence: descriptions définies et noms propres. *Centre d'Analyse Syntaxique de l'Université de Metz* (6), 538.

Lakoff, G. and M. Johnson (2008). *Metaphors we live by*. University of Chicago press.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, Volume 1. Stanford university press.

Laur, D. (1993). La relation entre le verbe et la préposition dans la sémantique du déplacement. *Langages 27*(110), 47–67.

Levinson, S. C. and D. P. Wilkins (2006). *Grammars of space: Explorations in cognitive diversity*, Volume 6. Cambridge University Press.

Makhoul, J., F. Kubala, R. Schwartz, and R. Weischedel (1999). Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pp. 249–252.

Matei-Chesnoiu, M. (2015). *Geoparsing early modern English drama*. Springer.

Matsumoto, Y. (1996). How abstract is subjective motion?: a comparison of coverage path expressions and access path expressions. *Conceptual structure, discourse, and language*.

Moncla, L. (2015). *Automatic Reconstruction of Itineraries from Descriptive Texts*. Ph. D. thesis, Université de Pau et des Pays de l'Adour, France.

Moncla, L., M. Gaio, T. Joliveau, and Y.-F. Le Lay (2017). Automated geoparsing of paris street names in 19th century novels. In *1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, Los Angeles Area, CA, USA. ACM.

Moncla, L., M. Gaio, J. Nogueras-Iso, and S. Mustière (2016). Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science (IJGIS) 30*(6), 1137–1160.

Muller, P. and L. Sarda (1998). Représentation de la sémantique des verbes de déplacement transitif du français. *TAL. Traitement automatique des langues 39*(2), 127–147.

Pourcel, S. and A. Kopecka (2005). Motion expression in French: typological diversity. *Durham & Newcastle working papers in linguistics 11*, 139–153.

Purves, R. S. and C. Derungs (2015). From Space to Place: Place-Based Explorations of Text. *International Journal of Humanities and Arts Computing 9*(1), 74–94.
anglais

Talmy, L. (1983). *How language structures space*. Number 4 in Berkeley cognitive science report. Berkeley, CA, Etats-Unis: Cognitive Science Program, Institute of Cognitive Studies, University of California at Berkeley.

Talmy, L. (1985). *Lexicalization patterns: Semantic structure in lexical forms. Language typology and syntactic description, vol. 3, Grammatical categories and the lexicon, ed. by Timothy Shopen, 57-149*. Cambridge: Cambridge University Press.

Talmy, L. (2000). *Toward a Cognitive Semantics*. The MIT Press.

Tenny, C. and J. Pustejowsky (Eds.) (1999). *Events as grammatical objects: the converging perspectives of lexical semantics and syntax*. Stanford, CA: CSLI.

Vandeloise, C. (1986). *L'Espace en français. Sémantique des prépositions spatiales*. Editions du Seuil.

Vandeloise, C. (1991). *Spatial prepositions: A case study from French*. University of Chicago Press.

Résumé

L'espace statique par rapport à l'espace dynamique et plus spécifiquement les descriptions spatiales, avec ou sans mouvement, sont les principales questions abordées dans cet article. Nous présentons des grammaires de construction implémentées dans la plateforme PERDIDO à l'aide de cascades de transducteurs à états finis qui visent à marquer et formaliser les relations entre entités nommées étendues, termes géographiques, relations spatiales et verbes de mouvement. Ces grammaires peuvent être considérées comme une synthèse computationnelle du travail sur l'expression de l'espace et du mouvement en langage naturel. La méthode proposée pour l'extraction d'information géographique a été appliquée pour trois projets différents au sein des humanités numériques utilisant des corpus spécifiques. La première tâche concerne l'extraction des noms de lieux à partir de romans français du XIXe siècle, la deuxième tâche traite de l'extraction de déplacements et de trajectoires à partir des descriptions de randonnées écrites en langues romanes (français, espagnol et italien) et la troisième tâche vise à identifier les expressions du mouvement fictif en anglais dans des revues alpines.