La linguistique face à la multiplication des données langagières numériques. Méthodes, risques et enjeux

Jean-Luc Minel

La linguistique face à la multiplication des données langagières numériques. Méthodes, risques et enjeux.

Jean-Luc Minel

MoDyCo, Université Paris Nanterre - CNRS, France,
jean-luc.minel@u-paris10.fr

1 Introduction

Le paysage de la recherche en Sciences du langage a subi d'importantes transformations au cours des dernières décennies. Depuis les années 60, il était institutionnellement dominé par le paradigme computo-représentationnel, qui accorde une place centrale à la notion de système formel. Cependant, depuis quelques années, cette approche se voit questionnée par la montée en puissance de perspectives émergentistes dans lesquelles les dimensions diachronique, énonciative, discursive, textuelle et contextuelle prennent toute leur place. Ces approches renouvelées prennent en compte la réalité de faits linguistiques attestés et mettent au premier plan la diversité des usages attestés. Elles s'appuient donc nécessairement sur des données langagières écrites, orales ou gestuelles rassemblées en corpus. Cette approche nécessite différentes étapes : recueil de données, annotation et transformation des données brutes en observables pour des analyses multifactorielles, mise à disposition et outillage. Par leur diversité, leur hétérogénéité, leur multimodalité et leur évolution continue, ces ressources interrogent les modèles classiques tout en remettant en cause étanchéité des niveaux de description.

L'interrogation des modèles théoriques établis et leur confrontation aux données d'usage rassemblées dans des ressources méthodologiquement construites et scientifiquement éprouvées conduit à la construction de "modèles sur corpus ".

Compte tenu de l'échelle des données à analyser, l'étude de ces données langagières ne peut plus être abordée uniquement par une approche descriptive : elle nécessite l'élaboration d'appareillages mathématiques et statistiques, notamment pour alimenter des modèles d'apprentissage supervisés et/ou non supervisés. Par ailleurs, les possibilités ouvertes par les artefacts (smartphones, tablettes, écrans tactiles, reconnaissance vocale) interrogent les modes de représentation des connaissances et ressources produites par les linguistes (dictionnaires, monographies, etc.) et nécessitent une réflexion sur leur intégration dans l'écosystème du Web sémantique. Le passage quasi obligé par des logiciels d'annotation, d'exploitation et de visualisation, que l'on peut apprhender par le concept de dispositif sociotechnique pose la question de leur portée congitive atuant que socio-politique, [1],[13],[11].

Mon exposé visera à dégager les principaux apports, mais aussi les risques qu'impliquent l'usage des outils d'annotation, d'exploration de corpus annotés,

du Linked Open Data, du "Machine Learning et du Deep Learning". On peut aussi considérer que ce questionnement relève des humanités numériques, terme qui en France constitue un chapeau sous lequel s'abrite, pour des raisons tant scientifiques que sociologiques, l'ensemble de ces approches.

1.1 A propos de l'interdisciplinarité

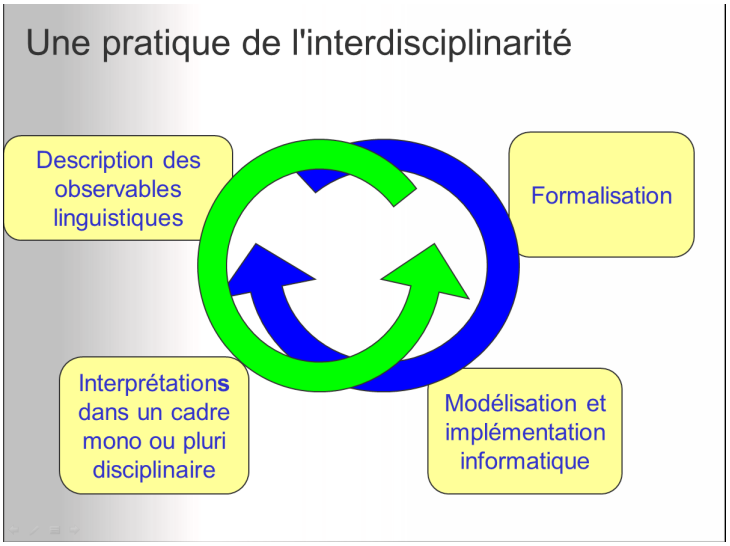

Fig. 1. Interdisciplinarité

Soumettre un projet de recherche sans mentionner qu'il est interdisciplinaire le voue d'emblée à l'échec. En France, depuis plus d'une vingtaine d'années, l'interdisciplinarité est devenu un passage obligé, une sorte de mantra qu'il convient de réciter. Néanmoins, l'exploration et l'exploitation de corpus nécessite indubitablement une approche interdisciplinaire qui convoque des savoirs qui relèvent des Sciences du langage, de l'informatique, des statistiques et quelquefois de la sociologie ainsi que des sciences de la cognition. Je vais donc préciser ce que j'entends par interdisciplinarité (cf. figure 1).

Dans une première étape, à parti d'un cadre théorique ou empirique, les chercheurs décident des observables linguistiques ou sémiotiques qui vont constituer leurs terrains de recherches. Par exemple, l'étude d'articles de la presse généraliste et professionnelle en langue française dans le domaine de la protection des données personnelles, ou l'analyse des tweets envoyés pendant un événement culturel comme les Journées Européennes du patrimoine, etc. Les temps forts de cette étape consistent à définir quel est le périmètre (temporel, sectoriel, etc.) des données observées, la granularité des données observées (texte, phrase, syntagme,

forme graphique, etc.). La deuxième étape vise à formaliser les descriptions en employant un langage formel (par exemple un arbre syntaxique ou des réseaux sémantiques). La troisième étape modélise cette formalisation dans des outils informatiques. La quatrième étape combine les savoirs et les expertises des linguistes, des sociologues, des psychologues, etc. pour interpréter des résultats. Je voudrais insister sur plusieurs points. Tour d'abord, ce processus est incrémental, c'est-à-dire que les recherches interdisciplinaires enchaînent plusieurs fois ces différentes étapes. Ensuite, le mot "Interprétations" est au pluriel, pour insister sur la pluralité des interprétations possibles qui dépendent des savoirs des chercheurs et des finalités de leurs recherches.

2 Corpus et annotations

Il y a deux approches dans l'usage des corpus en linguistique: une approche dénommée linguistique fondée sur le corpus (corpus Based) qui relève d'une démarche déductive, et une linguistique guidée par le corpus (corpus driven) qui relève d'une démarche inductive [25]. Mon expérience et mes pratiques actuelles m'inclinent plutôt à parler d'un parcours méthodologique, qui d'une part, hybride les deux approches et d'autre part, s'appuie sur les notions d'annotations, de méthodes et d'outils d'exploration ancrés sur des langages de requêtes. Hybridation, car "les processus d'interprétation et d'analyse sont toujours le produit d'un va et vient entre induction et déduction : le développement - voire la constitution même - d'un corpus est difficilement exempt d'hypothèses, qu'elles soient linguistiques ou non" [22]. Ainsi le travail de Biber [3], souvent cité comme exemple emblématique d'une approche inductive, repose, entre autre, sur l'hypothèse que des traits syntaxiques, permettent de catégoriser des textes. Que l'on privilégie une approche inductive ou déductive, la difficulté vient plutôt de la nécessaire articulation au sein du parcours méthodologique entre outils d'exploration et de traitements. Dans un monde idéal, il conviendrait que le chercheur mette en œuvre son parcours méthodologique indépendamment du choix des outils. Dans le monde réel, académique, où les contraintes de publication et d'évaluation (h-index) se font de plus prégnantes, cette disjonction n'est pas réaliste, notamment parce que la courbe d'apprentissage d'un outil se mesure au minimum en semaines si ce n'est en mois, et que peu de chercheurs disposent du temps nécessaire pour s'approprier plusieurs outils.Or le choix d'un outil va conditionner les formats d'annotations, la palette de traitements, l'interopérabilité et par conséquent la capitalisation des connaissances.

2.1 Annotations

L'usage des annotations est bien antérieur aux traitements des données numériques ainsi que l'illustre la figure 2 avec un manuscrit du XI siècle[1]. L'exploration

[1] Codicis Justiniani libri cum glossa (XIe siècle) Montpellier, Bibliothèque universitaire de médecine, ms. H. 82, fol. 12 (détail de la marge latérale de droite)

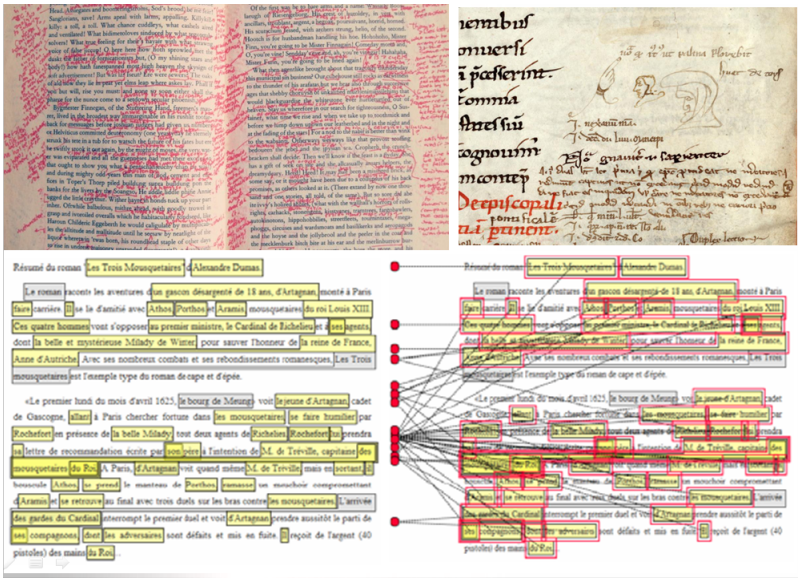

Fig. 2. Différents types d'annotations

de corpus ou de données langagières en nombre nécessite que ces corpus soient annotés automatiquement ou manuellement.Les outils d'exploration de corpus associe généralement l'annotation à la notion de marquable (markable), ce qui ne va pas sans poser problème. De fait ,un marquable est une chaîne de caractères à laquelle on va associer une structure de traits, c'est à dire un ensemble de couple attribut-valeurs qui relèvent de différents niveaux linguistiques (morphologique, syntaxique, sémantique, discursif, textuel). Se pose alors la question du format (encodage et structure) des annotations. Il existe des recommandations et des standards dont le plus connu est la TEI [2] qui conceptuellement est une DTD du langage XML.La TEI propose un modèle générique qui s'avère peu adapté aux annotations linguistiques, et ce sont plutôt des formats proposés par différents projets (Penn Treebank, CoNLL, etc.) qui sont utilisés. L'absence de standards d'annotations, au niveau structurel comme au niveau des valeurs (par exemple le jeu d'étiquettes morpho-syntaxiques) est d'un des obstacles majeurs que rencontre un chercheur qui souhaite s'investir dans le traitement des données langagières numérisées.A cette étape le choix de l'outil d'annotation s'avère crucial et je vais l'illustrer par un exemple.

Soit l'énoncé suivant emprunté à [22] : *Jean a rencontré Michel hier soir; cela faisait dix ans qu'ils ne s'étaient pas vus.* On souhaite annoter l'antécédent anaphorique de "ils". Cet antécédent ne correspond pas à un marquable continu mais aux deux expressions référentielles discontinues "Jean" et "Michel". Il faut donc que l'outil d'annotation propose la création d'unité abstraite. C'est par

[2] www.tei-c.org

exemple ce que propose des outils comme MMAX2 [17] ou Glozz [26] mais ce n'est pas le cas d'outils comme TXM [10] ou UAM [19]. Le chercheur doit faire face à un dilemme. Choisir un outil qui offre une puissance d'expression complète peut complexifier l'appréhension cognitive tout en limitant l'interopérabilité ce qui risque de compromettre l'achèvement du travail, ou de par le moins de rallonger la durée de l'étude. Le choix d'un outil plus limité dans sa puissance d'expression réduit ainsi les risques d'échec, tout en sachant que certains traitements ne seront pas possibles ou qu'il sera nécessaire de construire des artifices pour les achever.

2.2 Premiers jalons sur corpus et annotation

L'exploration de corpus annoté nécessite des choix cruciaux qui conditionnent très fortement les types de traitement que le chercheur pourra conduire. Premièrement, le choix du périmètre du corpus, sachant que tout corpus est le choix d'une construction sous tendue par des hypothèses que celles-ci soient explicites ou implicites. Deuxièmement, le choix du modèle d'annotation et de la structure de traits typés choisie, simple (ensemble de couples attributs valeurs) ou complexe (schéma récursif). Troisièmement, le choix de l'outil d'annotation, sachant que les deux dernières décisions sont fortement corrélées. Au final, le chercheur disposera alors d'un ou de plusieurs outils qui offrent un ou plusieurs langages de requête avec lequel il pourra entreprendre des traitements exploratoires.

2.3 Quels traitements ?

La numérisation des données langagières ouvre la voie à de nombreux traitements et l'une des difficultés à laquelle doit faire face le chercheur est de choisir, parmi l'éventail des outils proposés, celui ou ceux qui vont répondre à ses besoins. Mon point de vue est d'adopter, à cette étape, un parcours méthodologique qui va enchaîner approche inductive et approche déductive. Dans une première étape, l'observation des données au moyen de statistiques descriptives (indicateurs de tendance centrale et de dispersion, histogramme, écart type) permet d'identifier des régularités, des hapax, etc. Je n'insisterai pas sur cette étape qui est largement connue.

Analyses de données Dans une seconde étape, je proposerai d'appliquer des méthodes qui permettent d'approcher la structure d'un corpus "par les frontières" [22]. En effet, la structure d'un corpus s'explore en débusquant les oppositions les plus significatives. On cherchera ainsi à identifier ce que les statisticiens appellent des "outliers" (des déviants, ou des points aberrants) en regard de la norme endogène que constitue le corpus. L'analyse factorielle (analyse en composantes principales)[2] constitue un bon outil pour cette étape. Je vais illustrer son usage à partir d'un exemple emprunté à[22].

Un corpus de 224 articles extraits de 32 numéros de revues de linguistique a été constitué par C. Poudat en 2006 [21]. L'objectif était de réaliser une étude centrée sur le genre de l'article de recherche en linguistique. Un ensemble de 145

variables morphosyntaxiques a été défini et utilisé pour réaliser une analyse en composantes principales (ACP). La figure 3 montre le résultat obtenu : chaque point numéroté correspond à un article scientifique. On constate que deux textes, le numéro 46 (en haut à gauche) et le numéro 2 (en haut à droite) sont positionnés en périphérie sur la carte factorielle. Ces deux "outliers" attirent l'attention du chercheur qui doit identifier les raisons de cet isolement. Dans le cas présent, l'explication est la suivante : Le texte 46 est un texte d'histoire de la linguistique (épistémologie) dans lequel les temps du passé simple de l'imparfait et du plus que parfait prédominent; le texte 2 est un article rédigé sous la forme d'un dialogue socratique. Il appartient au chercheur de décider de conserver ces deux textes dans le corpus pour appliquer les traitements prévus, ou de considérer que leur présence peut perturber les traitements à venir[3]

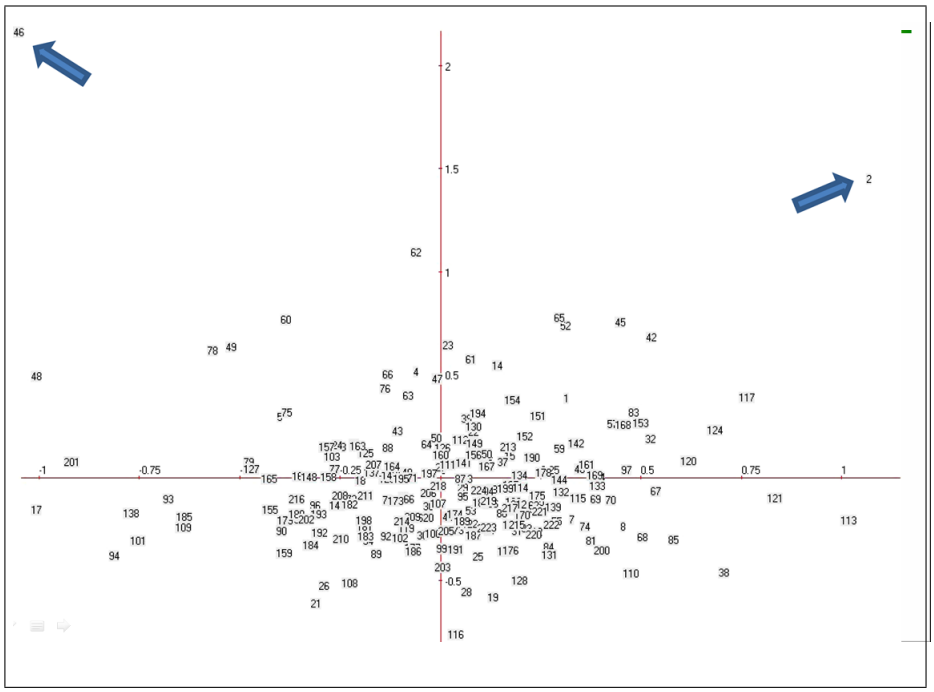

Fig. 3. Analyse en composantes principales [22]

Il convient d'insister sur deux points. Une étude uniquement manuelle n'aurait pas nécessairement permis de détecter ces deux singularités. C'est là un atout majeur des outils d'exploration lorsque la volumétrie des données devient importante et donc difficile à gérer manuellement. La décision d'exclure ou de conserver

[3] Dans le cas présent, C. Poudat a exclu les deux textes après les avoir analysé et à exécuter une nouvelle ACP pour obtenir une carte factorielle moins décentrée.

les singularités est une décision motivée par la finalité de la recherche; elle doit s'appuyer sur une analyse des singularités.

Plus généralement, les méthodes factorielles visent à représenter de manière synthétique des ensembles de données au moyen d'un nombre plus restreint de variables calculées qui sont appelées des facteurs. La carte factorielle offre une synthèse visuelle qu'il est plus facile d'interpréter que des tableaux de de données.

Prenons un exemple : si l'on cherche à étudier une cinquantaine de textes en choisissant les 100 mots les plus fréquents dans ces textes, une analyse qui se fonderait sur les indicateurs de la statistique descriptive serait très difficile à interpréter. Une carte factorielle, qui effectue une réduction des dimensions résout ce problème. Un des attraits de l'ACP est qu'il est possible d'interpréter les axes calculés (les facteurs) et de projeter sur la carte factorielle les variables comme le montre la figure 4.

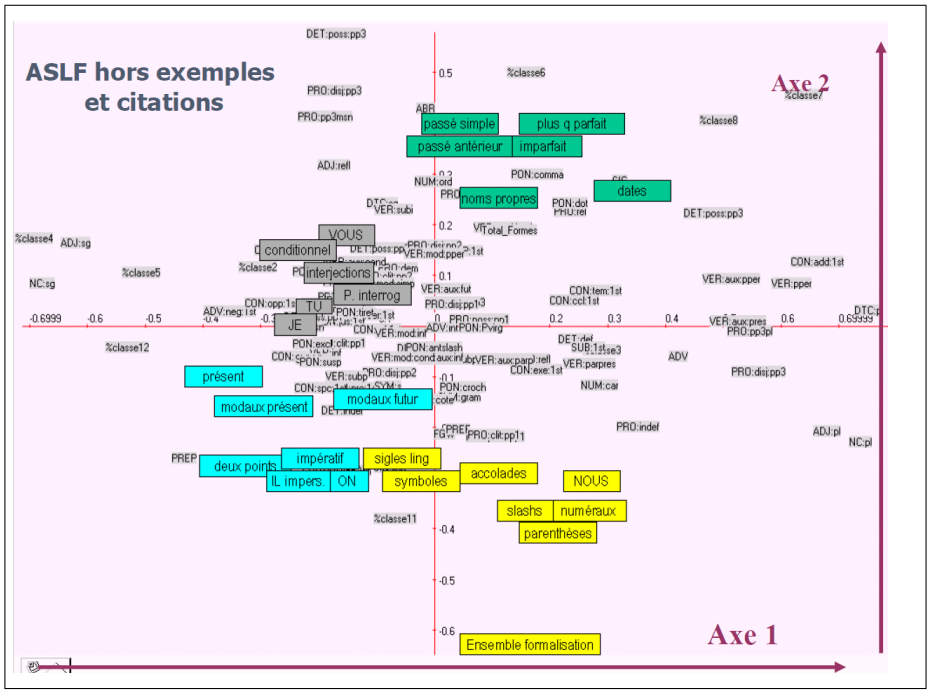

Fig. 4. Analyse en composantes principales [22]

Néanmoins, les méthodes factorielles trouvent leurs limites quand le nombre des unités à comparer (les textes, les énoncés, etc.) devient trop important (au dessus d'une centaine). On peut alors faire appel à une méthode nommée t-SNE (t-distributed stochastic neighborhood embedding) [15]. J'illustrerai son usage avec un exemple.

Dans une étude [12] vise à analyser la circulation des connaissances sur le réseau social Twitter lors d'un événement culture, (la Nuit Européenne des Musées 2016) en étudiant les pratiques des différents tweetos, et de contraster les pratiques des tweetos institutionnels (les musées) et des particuliers. Pendant cet évènement 11 500 tweets ont été envoyés par 7500 comptes Twitter, dont 1750 comptes francophone. La réalisation d'une ACP avec 5 variables sur cette masse de données ne donne pas de résultats interprétables.

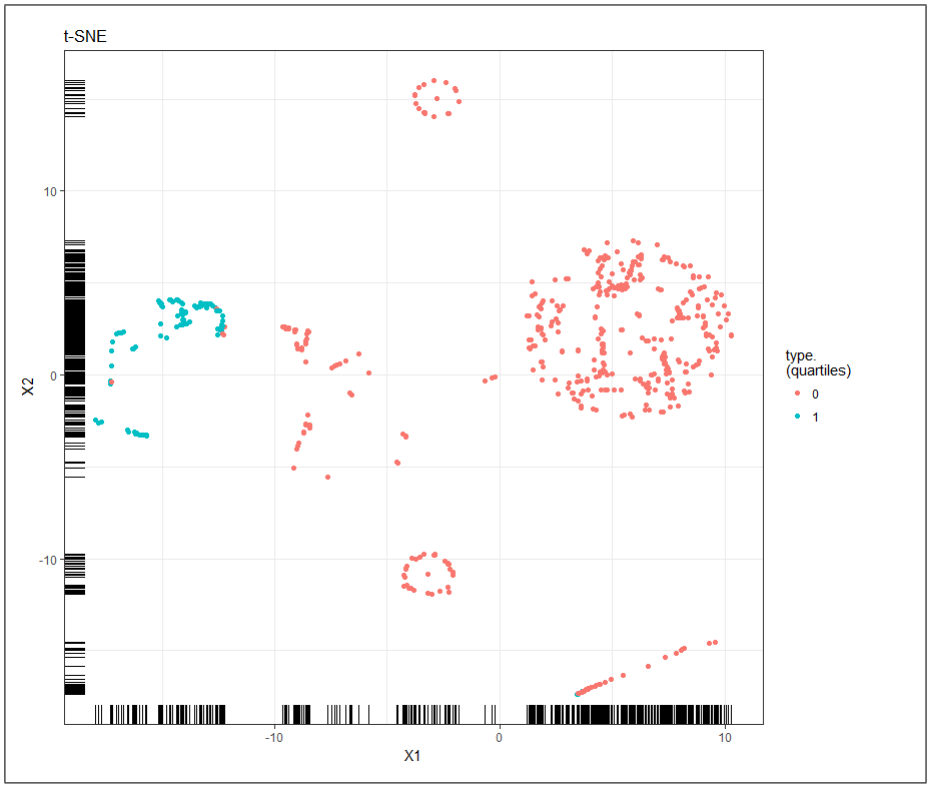

Fig. 5. Analyse t-SNE [12]

Par contre une t-SNE (cf. figure 5) visualise, à partir des 1750 comptes francophones, des regroupements de compte, et la colorisation permet d'identifier une ou plusieurs variables caractéristiques de ces regroupements. On remarque bien un large amas de comptes de particuliers sur la droite et deux anneaux, en bas et en haut de la carte, de comptes de particuliers. Enfin on voit, sur la gauche du graphique, des ilôts (en vert) qui sont des comptes institutionnels, nettement séparés du reste des autres comptes. Une analyse plus attentive, montre que certains comptes de particuliers sont proches des comptes institutionnels, ce qui

laisserait supposer que ces particuliers ont des pratiques discursives proches des comptes institutionnels. Il faut noter que contrairement à une carte factorielle, les axes d'une t-SNE ne sont pas interprétables; il est donc nécessaire d'investiguer de manière plus approfondie les données langagières.

KWIC, KWAC, KWOC Après une première étape permettant d'appréhender des caractéristique du corpus, et par conséquent de pouvoir élaborer quelques hypothèses, une deuxième étape est consacrée à une une fouille plus détaillée des données langagières, afin de confirmer ou de réfuter ces hypothèses, ou à en élaborer d'autres dont la granularité sera plus fine. Les concordances et les calculs de co-occurrence sont des moyens d'analyser finement le comportement d'une ou plusieurs unités linguistiques.

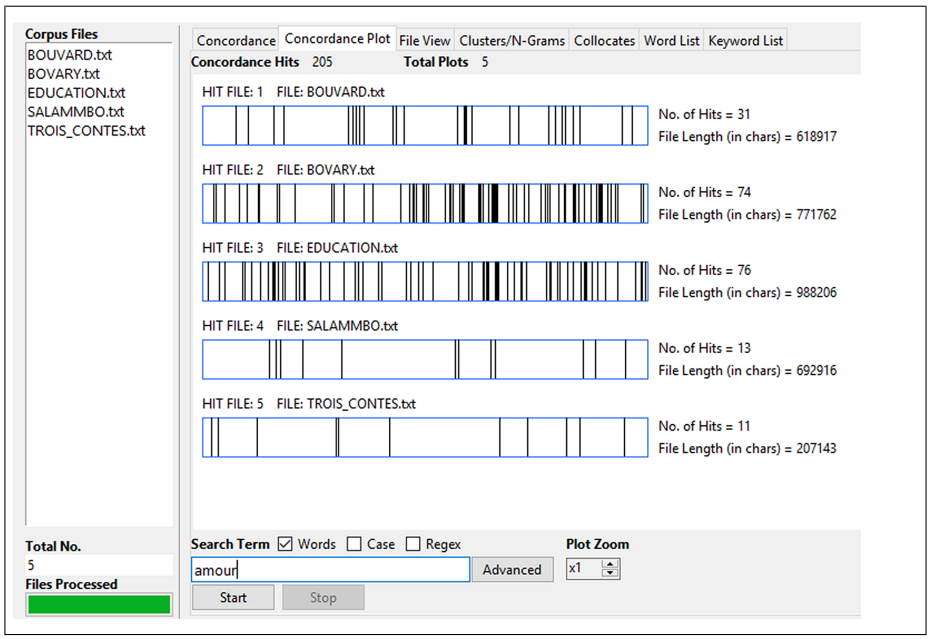

Fig. 6. Concordance topologique [22]

L'usage des concordances s'inscrit dans une longue tradition historique. M. Sekhraoui [23] indique une première concordance de la bible au 13^{e} siècle; trois siècles plus tard Hugo de San Charo mobilisa 500 moines pour construire la concordance d'une bible latine. L'utilisation de logiciels permet un changement d'échelle et fournit un langage de requête qui offre des moyens extrêmement puissant pour explorer une unité linguistique en contexte. Il existe de fait plusieurs type de concordances, je reprendrai la terminologie KWAC, KWIC, KWOC

proposée par [20]. La figure 6 présente un exemple de KWAC et et de KWIC. Je vais plutôt me focaliser sur l'utilisation de deux représentations propres à l'utilisation d'outils d'exploration de corpus. La première (cf. figure 7) est une concordance topologique (KWIP) intégré dans le logiciel AntConc[4] , la seconde (figure 8) une concordance multi-niveaux construite à l'aide du langage de requête CQL [5] intégré dans le logiciel TXM [10]. la concordance topologique permet de visualiser la dispersion ou la "densité" d'une unité linguistique dans un texte et de mettre ainsi en évidence des effets thématiques ou discursifs. La possibilité d'exprimer une requête à l'aide d'un langage comme CQL qui propose de combiner différents niveaux d'annotations (morphologiques, morphosyntaxiques, syntaxiques, voire sémantiques et discursif) fournit une puissance d'exploration ouvrant la voie à des études linguistiques impossible à réaliser hors du numérique. Enfin, il faut insister sur le fait que cette étape d'exploration peut donner lieu

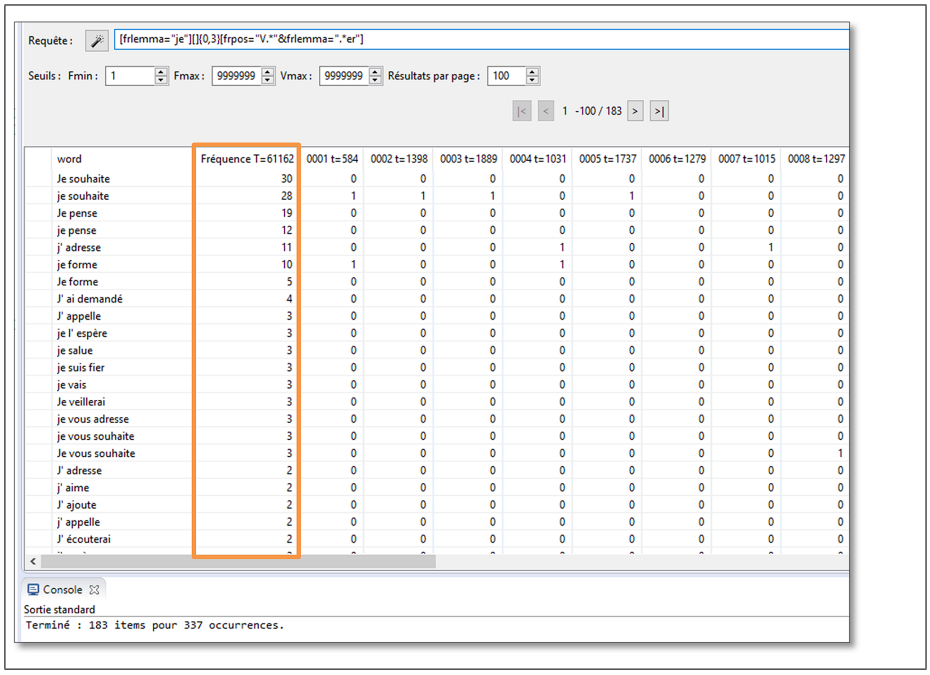

Fig. 7. Concordance multi-niveaux [22]

à l'annotation automatique des structures repérées, initiant ainsi un cycle de recherche incrémental particulièrement puissant.

[4] http://www.laurenceanthony.net/software/antconc/
[5] http://cwb.sourceforge.net/

Calcul de co-occurrences Depuis les travaux de Firth [7], puis de Guiraud [9], l'étude des co-occurents d'une unité linguistique est un moyen d'analyser la thématisation et la signification des unités en discours. Cette approche, comme on le verra, a été généralisée à une plus grande échelle dans les travaux du Deep Learning (voir supra). Je ne vais pas détailler les principes de calcul qui sont bien connu [14],mais aussi controversés [6], mais plutôt mettre en avant les représentations visuelles offertes par l'utilisation d'outils d'exploration. L'exemple des travaux de Née, Sitri et Fleury [18] qui ont étudié les différentes réalisations du pronom "nous" dans un corpus composé de rapports éducatifs annotés manuellement au niveau syntaxique et sémantique. L'étude des co-occurrents révèle ainsi des "préférences" telles que les conjonctions ou des connecteurs suivant la catégorie du prédicat (nous-sujet-dire ou nous-sujet-constat) étudié. Mais un des résultats importants de leur étude est d'insister sur l'apport heuristique des réseaux de co-occurrence (figure 8) représentés visuellement qui leur a permis de détecter un patron lexico-syntaxique du type "connecteur de concession+ nous+prédicat sentiment", dont certaines réalisations sont "en revanche nous restons inquiets (...), mais nous craignons (...), nous restons toutefois attentifs (...).

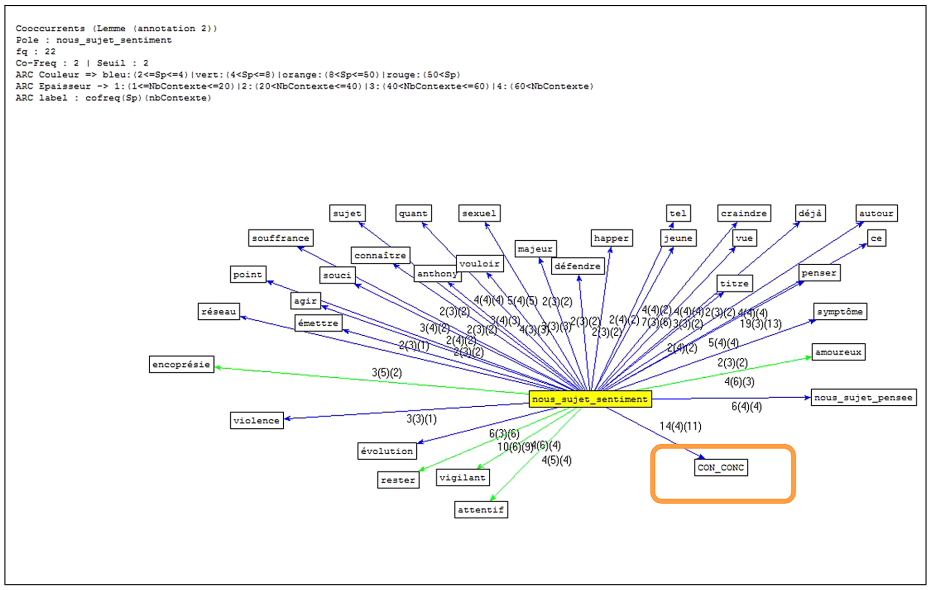

Fig. 8. réseaux de co-occurrences [18]

3 Construire et utiliser des ressources langagières interopérables : Linguistic Linked Open Data

Les données langagières numérisées, plus particulièrement lorsqu'elles sont annotées, représentent une source de connaissances particulièrement précieuse. Néanmoins, la réutilisation de ces ressources par d'autres chercheurs n'est pas immédiate et même souvent impossible. Indépendamment des problèmes de droits, qui relève des aspects juridiques, le principal obstacle à la réutilisation de données et à la capitalisation des connaissances linguistiques tient à l'absence de standards de description. Le Linked Open Data(LOD) , pour la linguistique, le Linguistic Linked Open Data (LLOD), cherche à briser ce verrou. Je vais m'appuyer sur un exemple pour illustrer les gains que l'on peut attendre du LLOD. mais avant de le décrire, je vais brièvement rappeler les points clefs du Web sémantique.

Fig. 9. Le Web sémantique

Le Web Sémantique repose sur trois des éléments qui ont le fait le succès du Web classique. Le protocole http qui permet l'échange de données entre des ordinateurs, un mécanisme d'identification, l'URL (appelée URI dans le web sémantique), un principe de navigation, l'hypertexte. Mais au lieu de décrire les données avec le langage HTML, le web sémantique propose une pile de langages formels fondée sur RDF (RDFS et OWL) pour décrire les données (cf figure 9).

Ces langages formels vont garantir l'interopérabilité et la sémantisation des descriptions, l'URI va permettre l'alignement des données entre plusieurs entrepôts de données, qui sont appelés TripleStore. Un dernier élément, le langage de

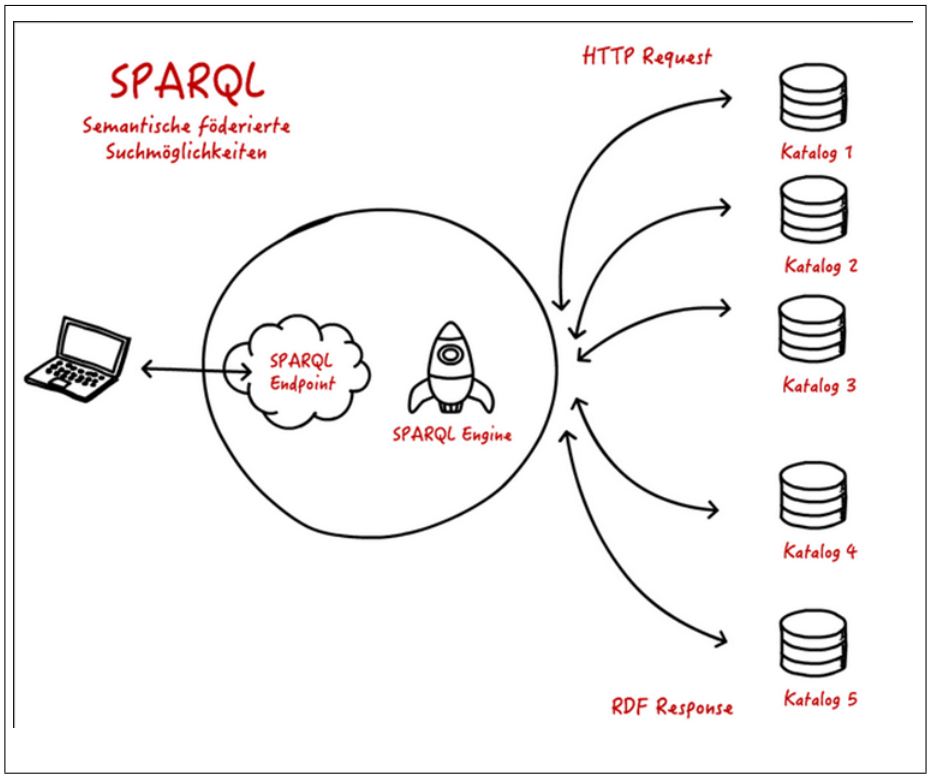

Fig. 10. Le langage SPARQL

requêtes SPARQL permet d'interroger de manière générique différents Triplestore (cf. figure 10)[6] .

3.1 Le projet Cocoon

Le projet Cocoon (http://cocoon.huma-num.fr/exist/crdo) me servira d'exemple pour illustrer l'intérêt du web sémantique. Cocoon pour "COllections de COrpus Oraux Numériques" est une plateforme technique qui accompagne les producteurs de ressources orales, pour créer, structurer et archiver leurs corpus. Les ressources sont constitués de données primaires (audio ou vidéo), d'annotations (transcriptions, traductions, des analyses linguistiques) et de données documentaires (métadonnées). la volumétrie est importante : 10 000 enregistrements, 3 000 transcriptions , 5 000 heures d'écoute. Les objets sont très divers : enquêtes sur 5 continents, 170 langues représentées, descriptions qui relèvent de différents niveaux d'analyse (phonétique/phonologie, syntaxe, socio-linguistiques, etc.).

[6] Ce graphique est emprunté à Gautier Poupeau http://www.lespetitescases.net/

A l'origine (en 2005), les concepteurs du projet choisirent de s'appuyer sur le modèle OLAC (Open Language Archives Community) pour décrire les données et sur le protocole OAI-PMH pour les diffuser. L'intérêt étant de une visibilité accrue grâce aux fournisseurs de service et aux portails thématiques existants. Néanmoins cette solution a trouvé ses limites. D'une part, le modèle OLAC n'a pas évolué depuis 2003, il n'offre pas de moyen pour décrire des personnes (les locuteurs, les chercheurs), les lieux (les terrains d'enquêtes). D'autre part, le protocole OAI-PMH ne permet pas la recherche dynamique par les moteurs de recherche ce qui est un frein à la visibilité des entrepôts.

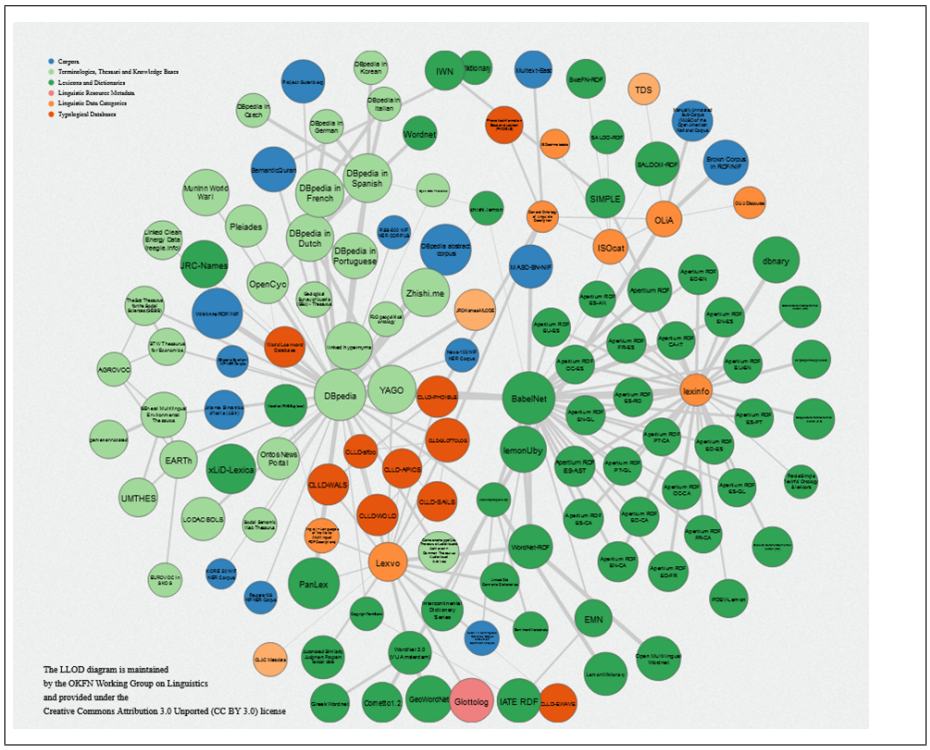

Fig. 11. Le LLOD

Mais de fait, la critique la plus importante, est la notion de silo de données qui est inhérente à ce type d'architecture. En d'autres termes, les données sont limitées aux descriptions réalisées par les gestionnaires du projet, il n'est pas possible de les enrichir avec d'autres descriptions. Or, il existe de nombreux référentiels externes : des thésaurus, des informations biographiques et bibliographiques, des informations géographiques, etc. Ces gisements de données complémentaires

seraient fort utiles pour les chercheurs soucieux de conduire de nouvelles recherches en capitalisant ces données langagières.

En 2015, les gestionnaires du projet décidèrent de convertir leur bases de données en un Triplestore RDF, de décrire leurs données en s'appuyant sur des vocabulaires proposés par le LLOD (figure 11) et de l'aligner avec des Triplestore existants. La migration réalisée et les alignements vérifiés, le gain est très important, car de nombreuses données peuvent être dynamiquement affichées en réponse à une requête d'un utilisateur. Par exemple, il est possible, pour un chercheur mentionné dans le terrain d'enquêtes, d'afficher toute sa bibliographie ou pour une langue rare, d'afficher des informations linguistiques issues d'autres entrepôts. L'effet de levier est très important et ouvre la voix à de nouvelles recherches qu'il n'était pas possible d'envisager auparavant.

Le passage au LLOD ne se réalise pas sans un certain nombre d'exigences, notamment la nécessité d'élaborer un modèle conceptuel (une ontologie) pour décrire les données (cf. figure 12).

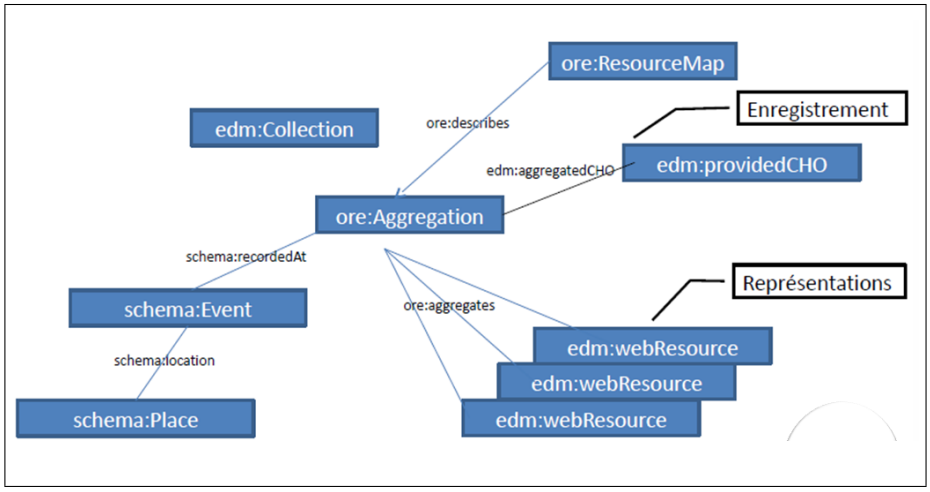

Fig. 12. modele conceptuel du projet Cocoon

C'est un processus qui mobilise des compétences interdisciplinaires et qui exige une certaine rigueur. Le LLOD est aussi confronté à certaines limites. Premièrement, toutes les données langagières, ne sont pas au format RDF. Deuxièmement, celles qui le sont, ne sont pas toujours accessible par un SparqlEndPoint. Enfin, la maintenance d'un triplestore a un cout financier et technique. C'est pour surmonter ce verrou que le CNRS a mis en place la TGIR Huma-Num (http://www.huma-num.fr/).

Il existe d'autres réalisations dans le LLOD. Citons WordNet[7] qui propose un SparqlEndPoint et DBnary[8].

3.2 Quelques éléments de synthèse

Retenons d'abord deux types d'approche du LLOD. Une première approche qui vise à construire des modèles conceptuels (des ontologies) dans des domaines spécialisés. En lexicologie, c'est l'exemple de Lemon (The Lexicon Model for Ontologies)[9]. Une seconde approche, qui vise à construire des ressources langagières en s'appuyant sur un modèle conceptuel et des vocabulaires existants, c'est l'exemple de Cocoon.

On peut noter qu'il n'existe pas de modèle commun, ni de vocabulaires qui seraient considérer comme des standards. On peut y avoir un atout, la flexibilité du LLOD ou une faiblesse, une interopérabilité limitée. Mais malgré cette interopérabilité limitée, les gains en terme de capitalisation et d'agrégation des connaissances sont remarquables.

4 Apprentissage supervisé, non supervisé et Deep Learning

4.1 Apprentissage supervisé

Dans la première partie de mon exposé j'ai insisté sur l'importance du processus d'annotation, étape préalable avant tout traitement. Si jusqu'au milieu des années 90, l'annotation automatique était réalisée à l'aide de systèmes symboliques (transducteurs, base de règles), depuis le milieu des années 2000, pratiquement tous les systèmes d'annotation automatique sont fondés sur l'apprentissage supervisé (machine learning). Avant d'en illustrer les gains sur un exemple, j'en rappellerai brièvement les principes. L'apprentissage automatique supervisé nécessite trois étapes.L'étape d'annotation, l'étape d'apprentissage et l'étape d'exploitation. Dans une première étape, une équipe composée de linguistes, éventuellement de chercheurs issus d'autres disciplines annotent un échantillon extrait du corpus qu'il faudra annoter. Cette étape nécessite d'élaborer une grille d'annotation composées des valeurs attribuées aux marquables (un syntagme, une proposition, une phrase, etc., et des critères de décision; Ceci prend généralement la forme d'un guide d'annotation. Parallèlement, la même équipe ou une autre équipe, identifie les traits (features) linguistiques ou sémiotiques qui seront exploités par l'algorithme de classification. Cette étape est cruciale, et plusieurs points conditionnent la qualité du classifieur obtenu : la taille de l'échantillon, les compétences de l'équipe d'annotation, le choix des traits linguistiques. La deuxième étape, l'apprentissage proprement dit, consiste à entrainer le classifieur à partir du corpus annoté dans l'étape précédente. C'est aussi lors de cette étape

[7] (http://wordnet.rkbexplorer.com/sparql/)
[8] http://kaiko.getalp.org/about-dbnary/online-access/
[9] http://lemon-model.net/

que le choix d'un ou de plusieurs algorithmes de classification (arbre de décision, SVM, Naives Bayes, etc.) est effectué. La troisième étape applique le classifieur sur l'ensemble du corpus. Il est fréquent, au vu d'une évaluation conduite à la fin de la troisième étape de recommencer un cycle complet afin d'améliorer la qualité du classifieur (modifications des traits, nouvel échantillon, nouvel algorithme, etc.).

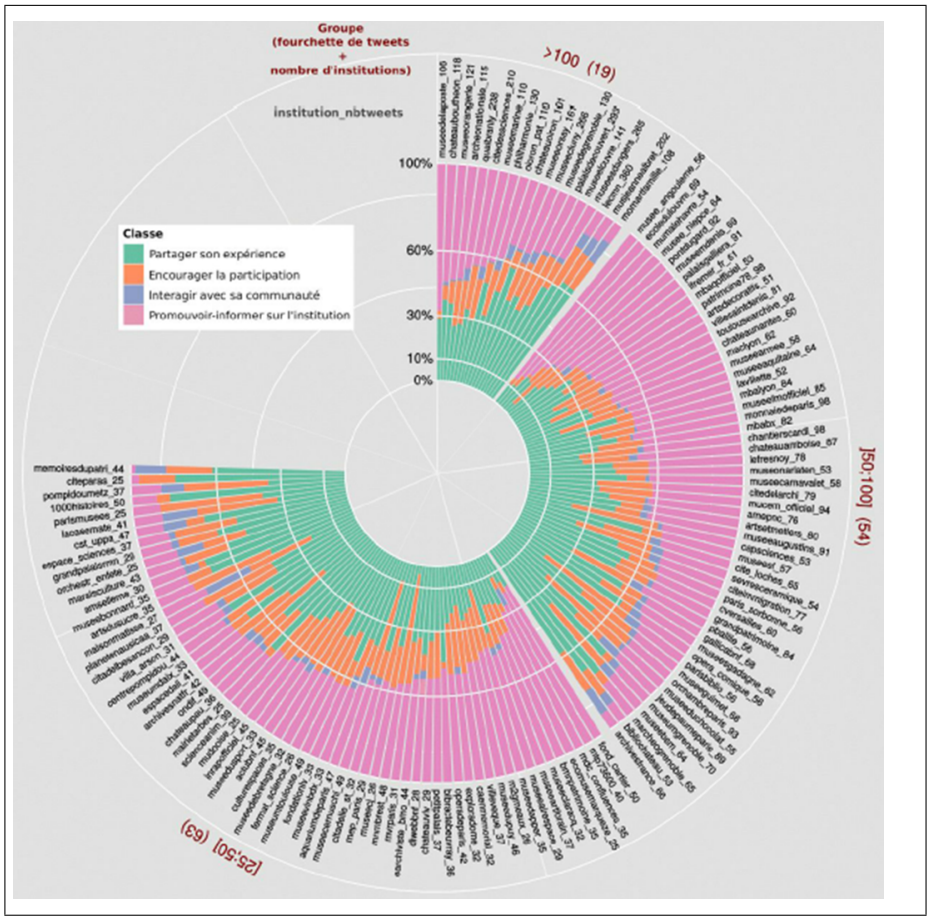

Fig. 13. Classification automatique de tweets [5]

Illustrons ceci sur un exemple. En 2016, une équipe composée de linguistes, de sociologues, de chercheurs en Sciences de l'information et de la Communication, en collaboration avec la Direction des Publiques du Ministère de la Culture, étudie la circulation de l'information et l'engagement des publics sur le réseau social (Rsn) Twitter pendant différents évènements culturels, MuseumWeek

2014, MuseumWeek 2015, la Nuit Européenne des Musées 2016 et les Journées Européennes du Patrimoine 2016. La volumétrie des données collectées sur les Rsn lors d'un événement culturel, même si elle est loin d'atteindre celle des données massives ("big data"), ne permet pas d'utiliser directement des méthodes qualitatives et interprétatives. Pour exemple, le corpus du NDM16 regroupe 7000 comptes et 30000 tweets sur une semaine. Dans le cas de MuseumWeek 2015, c'est 100 000 tweets par jour pendant une semaine. Pour faire face à cette masse de données, l'équipe a décidé d' effectuer une catégorisation des contenus des messages qui repose sur la classification automatique supervisée des tweets à partir d'indices sémiolinguistiques identifiés dans les messages en langue française.

L'annotation du corpus d'entrainement a été réalisée par une équipe composée de deux linguistes et de 10 animateurs de communauté en ligne ("community managers", CM). La même équipe a conçu le modèle, ainsi que les traits qui ont permis de calculer la classe d'un tweet. Les traits retenus, une vingtaine, sont des traits linguistiques majoritairement lexicaux, mais ils incluent aussi les marques de ponctuation et des traits spécifiques aux tweets (par exemple, la présence/absence de hashtags dans les tweets) ainsi que certaines métadonnées comme l'identité des auteurs des tweets. Le modèle propose 4 catégories (classes) de tweets : interagir entre comptes, encourager à contribuer, promouvoir un musée et exprimer une expérience. Un échantillon de 1000 tweets a été annoté par des experts (CM) du domaine culturel en fonction des catégories prédéfinies dans l'étape précédente. Cette étape a été nécessité environ 200 heures de travail.

Dans un deuxième temps, un classifieur a été construit par apprentissage sur un échantillon puisé dans les tweets de deux événements culturels (MuseumWeek2014 et MuseumWeek 2015) [5]. Le classifieur construit est fondé sur les modèles Naives Bayes et SVM, avec vote à l'unanimité. Cela signifie que les deux modèles de classification doivent prédire la même catégorie pour un tweet ; dans le cas contraire, le tweet n'est pas catégorisé. Dans un troisième temps, le classifieur a été appliqué au corpus de tweets pour catégoriser l'ensemble des tweets. En résumé, le modèle construit permet de classer les tweets en catégories de type communicationnel sur la base du contenu textuel des tweets. La figure 13 illustre une partie des résultats, il s'agit de la classification des tweets originaux envoyés par les institutions pendant la Museum Week 2015. Cet exemple est une bonne illustration des gains et des limites de ce type d'approche. En termes de gain, c'est la possibilité d'annoter des corpus de données langagières de grande taille. C'est aussi sa reproductibilité et la stabilité des annotations. Le classifieur a ainsi été appliqué sur 4 corpus différents avec des résultats tout à fait convaincant [8]. La principale limite tient essentiellement à l'effort, en termes de disponibilité et de compétences des ressources humaines, qu'il faut fournir pendant la première étape d'annotation de l'échantillon. C'est justement l'objectif du Deep Learning, de faire l'économie de cette étape considérée comme étant encore trop chronophage et dépendante de savoirs linguistiques qui introduiraient des biais.

4.2 Deep Learning

Initiée à la fin des années 1980 avec la naissance des premiers réseaux de neurones artificiels multicouches, eux mêmes reprenant un concept datant de la fin des années 1950 (perceptron, etc.), cette approche avait donné des résultats plutôt décevant dans le domaine du Traitement Automatique des langues (TAL), alors que des applications dans le domaine de la reconnaissance faciale avait débouché sur des applications réelles. On était donc resté sur l'idée que ce type d'approche était plutôt dédié aux traitements de données de bas niveau (comme le sont les pixels d'une image numérisée).

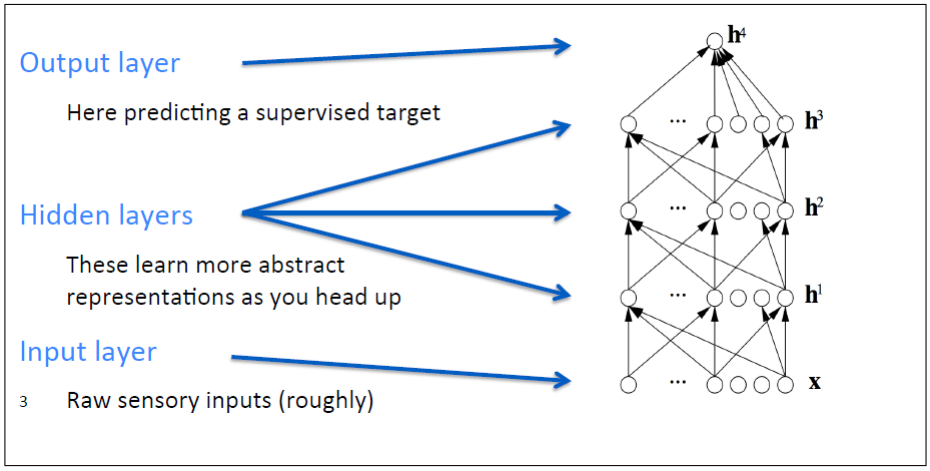

Fig. 14. Réseaux récursifs [24]

Un tournant important s'est produit après les années 2010, conséquence de plusieurs innovations importantes. En terme statistique, c'est d'abord la conception de réseaux récursifs multicouches [24] (figure 14). C'est la proposition par [16] d'un nouveau mode de représentations des mots dans un texte associé à la mise à disposition de corpus de données comme Word2vec et Glove. Pour faire bref, cette représentation permet de prendre en compte des dépendances longues entre des mots dans une phrase. C'est la proposition par [4] d'une architecture et d'un algorithme d'apprentissage nécessitant beaucoup moins de ressources de calcul. C'est enfin l'exploitation de la technologie des cartes graphiques utilisées pour les jeux vidéos pour implémenter les réseaux récursifs de neurones. Ces différentes avancées ont débouché sur plusieurs résultats remarquables. Dans le domaine de l'analyse syntaxique, le groupe de recherche de Stanford dirigé par Chris Manning a fait état de nettes améliorations dans le rattachement prépositionnel; dans le domaine sémantique, c'est la classification des relations qui est notablement amélioré; dans le domaine discursif, c'est la détection des

paraphrases et l'analyse des sentiments qui sont améliorés. Tous ces résultats ont été accompagnés de déclarations quelquefois fracassantes sur le fait que cette technologie pourra dans les années à venir se passer des linguistes et des informaticiens spécialisés dans le TAL. Pour ma part, je reprendrai la déclaration faite par Chris Manning lors de sa conférence introductive de l'ACL (Association for Computational Linguistics) en 2015 : "Don't panic !".

Le Deep Learning n'est pas générique, en d'autres termes, chaque problème exige la conception d'un modèle spécifique qui, jusqu'à présent, nécessite les savoirs des linguistes. Par contre, il est certain que les tâches des linguistes et des spécialistes en TAL vont profondément se transformer. Il leur appartiendra de spécifier avec les statisticiens et les spécialistes de l'intelligence artificielle le type de données langagières qu'il faut injecter dans les grands corpus de données et d'en évaluer les gains.

5 Conclusion

Au terme de cet exposé je voudrais insister sur les point suivants:

- La nécessité d'une approche interdisciplinaire;
- Le choix des données qui composent un corpus;
- Le choix des observables et du modèle d'annotation;
- Le choix des modèles de représentation et leur interopérabilité;
- Le choix des outils d'exploration et de traitements;
- Le choix des outils de visualisation.

L'utilisation d'outils de traitements n'est pas sans risques. J'ai déjà souligné le fait qu'un outil impose un certain point de vue sur les données à traiter et même dans certains cas des hypothèses fortes, comme le fait par exemple la Classification Ascendante Hiérarchique (CAH). Un deuxième risque tient à la surinterprétation que les outils visuels imposent. En effet, une cartographie, un réseau, de par les dispositions spatiales qu'ils proposent peuvent être investis d'une sémantique (le haut versus le bas, la droite versus la gauche, etc.) qui biaisent les interprétations qu'en font les chercheurs. Un troisième risque, notamment dans le cas de l'apprentissage supervisé est la tendance à ignorer les cas peu nombreux ou inversement à renforcer les régularités très fréquentes.

De mon point de vue, ces risques ne sont pas rédhibitoires, ils exigent de la part des chercheurs qui utilisent les outils numériques une rigueur accrue, qui n'est pas un des moindres enjeux scientifiques soulevés par leur utilisation.

Enfin concernant les enjeux, deux me semblent être prioritaires. Le premier relève du LLOD et de la nécessité de construire des ressources interopérables qui s'enrichissent mutuellement. Le second concerne le Deep Learning. Les linguistes doivent s'impliquer dans les équipes qui conçoivent et évaluent les modèles, ce qui impose qu'ils en comprennent les principes.

Références

1. Akrich, M.: Les formes de la médiation technique. Réseaux 60 (1993)

2. Benzécri, J.P.: L'Analyse des Données. Dunod (1973)
3. Biber, D.: Variation across Speech and Writing. Cambridge University Press (1998)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research 12(Aug), 2493–2537 (2011)
5. Courtin, A., Juanals, B., Minel, J., de Saint Léger, M.: A tool-based methodology to analyze social network interactions in cultural fields: The use case "museumweek". In: Proceedings of the 6th International Conference on Social Informatics (SocInfo'14). pp. 144–156 (2014)
6. Dunning, T.: Accurate methods for the statistics of suprise and coincidence. Computational Linguistics 19(1), 61–74 (1993)
7. Firth, J.: Papers in lingusitics 1934-1951. Oxford University Press (1957)
8. Foucault, N., Courtin, A.: Automatic classification of tweets for analyzing communication behavior museums. In: LREC 2016. pp. 3006–3013 (2017)
9. Guiraud, P.: Problèmes et méthodes de la statistique linguistique. Presses Universitaires de France (1960)
10. Heiden, S., Magué, J.P., Pincemin, B.: Txm : une plateforme logicielle open-source pour la textométrie- conception et développement. In: Actes des 10 Journées Internationales d'Analyse statistique des Données Textuelles, JADT. pp. 1021–1032 (2010)
11. Jouet, J.: Pratiques de communication et figures de la médiation. Réseaux 60, 99–120 (1993)
12. Juanals, B., Minel, J.L.: Information flow on digital social networks during a cultural event: Methodology and analysis of the european night of museums 2016 on twitter. SMS+Society Special Issue (2017)
13. Latour, B.: Petites leçons en sociologie des sciences, la clé de Berlin. Le Seuil, Paris, France (1993/1996)
14. Lebart, L., Salem, A.: Analyse statistique des données textuelles. Dunod (1988)
15. Van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research 9, 2579–2605 (2008)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
17. Müller, C., Strube, M.: Multi-level annoatation of linguistic data. In: Braun, S., Kohn, K., Mukherjee, J. (eds.) Corpus technology and lanaguage pedagogy : New resources, new tools, new methods. Peter lang (2006)
18. Née, E., Sitri, F., Fleury, S.: L'annotation du pronom "nous" dans un corpus de rapports éducatifs. objectifs, méthodes, résultats. In: JADT 2014. pp. 495–506 (2014)
19. O'Donnel, M.: The uam corpustool: Software for corpus annotation and exploration. In: Bretones Callejas, C., al. (eds.) Applied linguistics now : Understanding Language and Mind. pp. 1433–1447 (2008)
20. Pincemin, B.: Concordances et concordanciers : de l'art du bon kwac. In: XVII colloque d'Albi Langages et signification. pp. 33–42. CAL-CPST (2006)
21. Poudat, C.: Etude constrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres. Ph.D. thesis, Université d'Orléans (2006)
22. Poudat, C., Landragin, F.: Explorer un corpus textuel. Deboeck (2017)
23. Sekhraoui, M.: Concordance: historique, méthode et pratique. Ph.D. thesis, Université Paris 3 (1995)

24. Socher, R., Perelygin, A.and Wu, J., Chuang, J., Manning, C., Ng, A., C., P.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)
25. Tognini-Bonelli, E.: Corpus linguistics at work. John Benjamin (2001)
26. Widlöcher, A., Mathet, Y.: La plateforme glozz : environnement d'annotation et d'exploration de corpus. In: Actes de la 16 conférence sur le Traitement Automatique des Langues Naturelles. ATALA (2009)