# Les offres d'emploi comme texte ?

Romain Loth

# Les offres d'emploi comme texte ?

***Annoter et étudier un corpus pour un projet d'extraction d'information***

Romain Loth
MoDyCo - UMR 7114 CNRS
Univ. Paris Ouest Nanterre La Défense
FR-92001 Nanterre
01 40 97 74 31
rloth@u-paris10.fr

***Résumé***

Ce travail de corpus a été effectué dans le cadre d'un projet d'extraction d'information sur les offres d'emploi. Partant de l'hypothèse qu'il s'agit d'un genre textuel particulier, notre équipe s'intéresse aux phénomènes sémantiques dans ces offres d'emploi. Cela a conduit, entre autres, à l'annotation détaillée d'un petit corpus représentatif des différentes branches professionnelles. Nous avons d'abord élaboré une typologie d'annotation sur un échantillon de 50 documents, en cherchant à suivre au plus près les catégories d'énoncés récurrentes. Une fois la typologie établie, nous avons annoté 150 documents supplémentaires, pour en révéler la structure textuelle et apprendre à classer les descripteurs lexicaux (termes simples et locutions figées) selon leur contexte d'apparition caractéristique.

## 1 Corpus et extraction d'information

### *Projet SIRE*

La publication d'offres d'emploi sur internet a fait naître par son abondance le besoin d'un accès intelligent à ces documents. En pratique, cela équivaut à une indexation selon des descripteurs (clefs lexicales menant au document) plus riches.

Par exemple, les classements actuels s'appuient essentiellement sur les secteurs d'activités et sur des catégories de métiers plutôt rigides, en ignorant ce que les documents précisent à propos des missions et des savoir-faire requis. Or quelqu'un qui cherche un emploi peut tout à fait vouloir chercher parmi des métiers différents, mais se concentrer sur un ensemble de tâches et de compétences qu'il maîtrise. Il faut donc que ces tâches et compétences soient répertoriées dans l'index qui donne accès au document. Le projet SIRE (Sémantique, internet, recrutement et emploi) a donc pour objectif de fabriquer des outils d'extraction automatique des termes les plus pertinents afin de faciliter l'étude statistique du marché du travail et la recherche d'emploi.

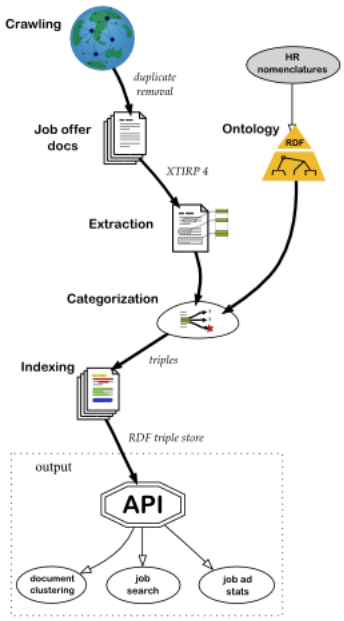


*Fig. 1: Chaîne de traitement*

Notre laboratoire a dans ce projet un double rôle de conseil sur les formalismes adéquats à l'analyse textuelle automatique et de développement de plusieurs modules de la plateforme. Un travail d'annotation manuelle détaillée nous est apparu indispensable en prévision de ces tâches. Il a été mené sur un échantillon de 200 offres d'emploi.

***Approche au corpus en extraction d'information***

A sa manière, le domaine de l'extraction d'information relève tout entier du travail de corpus puisqu'il s'agit d'isoler, étiqueter ou classer des éléments du texte. D'autre part, les annotations manuelles (ou « labeled data ») sont vues comme des ressources rares servant à entraîner les programmes d'extraction. Mais contrairement à la tradition en linguistique, le corpus n'est pas étudié pour lui même : il y a un objectif extérieur concret.

Cette démarche utilitariste implique que le temps passé à annoter manuellement est toujours mis en balance avec le temps gagné si l'on contourne ce besoin. Pour le niveau sémantique où il existe peu de données annotées, l'approche habituelle est de recycler des données extérieures en annotations et d'adapter les algorithmes au peu d'annotations dont on dispose, comme le conseille par exemple Manning *et al* 2008. Le travail d'annotation manuelle qui a été mené ici sur les offres d'emploi est une démarche relativement rare. C'est une tentative de s'inscrire dans les deux traditions, l'une très pragmatique, qui cherche à circonscrire les descripteurs les plus pertinents, et l'autre plus analytique et plus proche de la description linguistique, qui cherche à décrire un genre textuel.

## 2 Méthodologie

***Sources du corpus***

Les 200 offres proviennent de deux « job-boards » : le site de l'APEC et celui de Monster[1]. En utilisant les menus par secteurs d'activité présents sur les sites, nous avons cherché à atteindre une représentativité par branches d'activité[2]. Les pages web ont été téléchargées et ramenées à leur contenu textuel par un petit programme de collecte/filtrage écrit en perl.

***Hypothèses et démarche***

En extraction d'information, on est souvent assez souple pour définir ce qu'on annote et comment on le classe. Une pratique courante est d'annoter directement selon le résultat que la plate-forme devra obtenir[3], mais de nombreuses approches sont possibles. Cette ouverture incite à une exploration de ce qu'il est intéressant de relever, du « repérable » d'un texte. Dans l'idée, notre schéma d'annotations ne devait pas venir a priori, mais résulter d'une confrontation aux textes via un dispositif expérimental.

Pour définir les éléments à relever et leur schéma de classement, nous avons posé deux niveaux imbriqués, qu'on appellera ici « domaines d'information » et « types ». S'éloignant de tradition en annotation sémantique, le premier niveau se voulait centré sur des domaines référentiels idéalisés (« tout ce qui touche à l'entreprise », « tout ce qui touche aux activités », etc.). On verra que ce découpage, utile côté ingénierie, n'est pas sans poser de problèmes : pour certaines distinctions, on ne peut pas se passer de critères énonciatifs indépendants de la référence. Le second niveau a pour rôle de détailler les grandes catégories référentielles et si possible de retomber sur des types homogènes du point de vue de leur forme linguistique. Ce travail a été mené d'emblée sur exemples : on a commencé par annoter 50 offres avec un squelette minimal de classement à deux niveaux à catégories libres (domaine abordé et attributs distinctifs).

1 Le format des offres chez Pôle Emploi est plutôt atypique et demanderai une étude complémentaire.

2 120 offres ont été choisies selon la répartition par secteurs de la population active (enquête emploi 2007 de l'INSEE). 75 autres offres ont été choisies par un tirage aléatoire sur les deux sites. Enfin 5 offres présentant un intérêt linguistique spécifique (particularités de structure ou d'expression) ont été ajoutées à la main.

3 Dans cette optique, c'est une fois les annotations faites qu'on s'intéresse aux traits linguistiques pertinents qui permettront d'arriver à ce résultat automatiquement.

A chaque unité rencontrée, sa description était ajoutée au squelette minimal. Un recodage de ces descriptions libres a été effectué *a posteriori,* en ne gardant que les domaines et attributs les plus courants et/ou significatifs pour l'interprétation du texte, et en les ordonnant entre eux. Ce premier travail a servi à élaborer un schéma d'annotation. Le reste des annotations a ensuite été effectué selon ce schéma arrêté (avec quelques modifications mineures), en ajoutant des informations sur les relations entre unités et sur la structure des paragraphes. Ces 150 documents constituent le corpus final propre, soit 62642 mots et 5189 annotations.

***Glozz et glozz-note***

Les annotations ont été effectuées avec la plateforme Glozz du laboratoire GREYC (Mathet et Widlöcher 2009). L'application permet d'envisager de nombreux travaux de corpus, en particulier autour de l'analyse de discours. En l'occurrence, nous avons traité 3 niveaux :

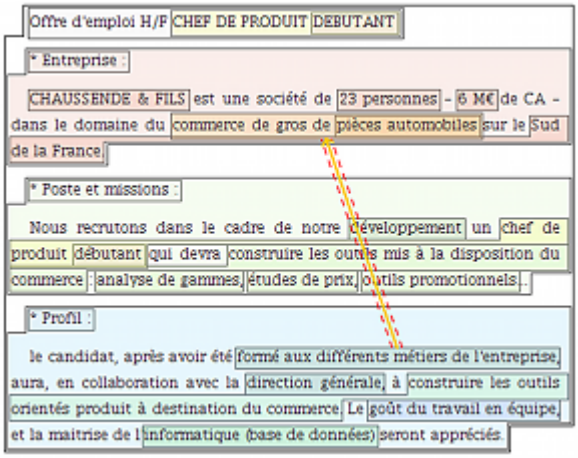


*Fig. 2: Interface d'annotation*

- les blocs/paragraphes de la découpe typique du texte : la « Présentation de l'entreprise », la « Description du poste », le « Profil », etc. formant des séquences au sens de Charolles 1988 ;
- les entités sémantiques : leur typologie a été définie selon la méthode décrite ci-dessus[4] ;
- les relations entre unités (anaphores, ellipses et les relations discursives de présupposition et d'élaboration), mais uniquement quand elles semblaient essentielles pour l'interpétation des éléments extraits. Elles n'ont pas encore été traitées pour le moment.

En pratique on opère *via* une double interface : surlignage coloré (une couleur par domaine référentiel) sur des segments de texte pour créer une annotation et cases à cocher pour la décrire (traits descriptifs dédiés au typage). Les sorties de Glozz se font ensuite à travers un format XML spécial.

```
<unit id="rl_12731688670">
<characterisation>
<type>01-Poste</type>
 <featureSet>
  <feature name="intitule">1</feature>
  <feature name="nom_compo">1</feature>
  <feature name="nom_simple">0</feature>
  <feature name="agent_divers">0</feature>
  <feature name="niveau_hier">0</feature>
  <feature name="service_lieu">0</feature>
  <feature name="theme_objet">1</feature>
  <feature name="theme_vide">0</feature>
  <feature name="autres"/>
 </featureSet>
</characterisation>
<positioning>
  <start><Position index="19"/></start>
  <end><Position index="43"/></end>
</positioning>
</unit>
```

```
<unit id="rl_12731697678">
<characterisation>
<type>09-Expérience</type>
 <featureSet>
  <feature name="anciennete">0</feature>
  <feature name="qualitatif">1</feature>
  <feature name="fonction_xp">0</feature>
  <feature name="secteur_xp">0</feature>
  <feature name="optionnel">0</feature>
  <feature name="autres"/>
 </featureSet>
</characterisation>
<positioning>
  <start><Position index="35"/></start>
  <end><Position index="43"/></end>
</positioning>
</unit>

(...)
```

```
<unit id="rl_12731745142">
<characterisation>
<type>05-Secteur</type>
 <featureSet>
  <feature name="autres"/>
  <feature name="produits">0</feature>
  <feature name="activite">1</feature>
  <feature name="clientele">0</feature>
  <feature name="nom_consacre">0</feature>
  <feature name="explicatif">1</feature>
 </featureSet>
</characterisation>
<positioning>
  <start><Position index="141"/></start>
  <end><Position index="179"/></end>
</positioning>
</unit>

(...)
```

*Table 1: Sortie XML de Glozz*

Nous avons écrit un petit programme appelé glozz-note[5], consacré à la transformation de ce XML en un format tabulé. Au passage, glozz-note opère des corrélations entre attributs pour les recoder en types, puis il extrait le contexte, reconstruit les imbrications des annotations

4 Les frontières dans la construction du syntagme et la recherche d'une convention récurrente pour chaque type sont les critères de cohésions utilisés pour choisir les bornes des entités sémantiques à l'annotation.

5 Ce script est accessible sur demande et peut servir au post-traitement de tout travail de corpus sous Glozz.

(« chemin ») et calcule leur position relativement à la longueur du texte. Cette dernière valeur (« rel. pos. ») est exprimée en pourcentage et s'avère très utile pour comparer la structure textuelle de textes de longueurs différentes.

| ID | DOC | CHEMIN | DOMAINE | TYPE | TEXTE EXTRAIT | POS. | REL. POS. | ATTRIBUTS |
|---|---|---|---|---|---|---|---|---|
| 12731688670 | 127 | A-Titre | 01-Poste | intitule | *CHEF DE PRODUIT* | 19-43 | 4,2% | intitulé /n. composé |
| 12731712118 | 127 | B-Présentation-Entreprise | 03-Etablissement | nom_entreprise | *CHAUSSENDE & FILS* | 57-74 | 8,9% | desc. definie |
| 12731745142 | 127 | B-Présentation-Entreprise | 05-Secteur | activite | *commerce de gros de pièces automobiles* | 141-179 | 21,9% | activité/explicatif |
| 12731753025 | 127 | B-Présentation-Entreprise :: 05-Secteur | 05-Secteur | produit | *pièces automobiles* | 161-179 | 22,2% | produit |
| ... | | | | | ... | | | |
| 12731892030 | 127 | E-Profil-Candidat | 10-Savoirfaire | techniques_spéc | *informatique (base de données)* | 684-714 | 95,5% | infoprog(BDD) |

*Table 2: Sortie tabulée de glozz-note*

# 3 Typologie obtenue et discussion

### *Airs de famille et champs d'information*

Les informations de différents domaines s'expriment à travers des formes et des contenus caractéristiques. Ces traits caractéristiques sont relevés à l'annotation sous la forme d'attributs, qu'on regroupe ensuite en types : l'intitulé du poste, la spécialité de l'entreprise, les capacités demandées, *etc.* Toutefois les limites entre les domaines ne peuvent pas être définies de façon absolue. Dans les cas extrêmes, ce n'est pas le « contenu sémantique » qui varie mais uniquement l'angle énonciatif :

(1) Vous effectuez la conception mécanique et le dimensionnement des éléments sous Catia. **(Mission)**
(2) Vous devez impérativement maîtriser les outils de CAO (Catia, Inventor) **(Savoir-faire)**

Les descripteurs sont ici en relation notionnelle les uns avec les autres (ils décrivent une réalité semblable). D'une façon générale, notre méthode basée sur les domaines référentiels n'empêche pas les types sous-jacents de présenter entre eux des airs de famille.

Si l'entreprise et le poste sont bien distincts référentiellement, les autres grands domaines choisis pour décomposer la description de l'emploi le sont moins. Plutôt que des domaines référentiels, ils sont en fait des cadres interprétatifs, facettes d'une même réalité : missions, savoir-faire, études, métier. Pour distinguer à l'annotation, on doit alors prendre en compte la position et la forme énonciative des segments, et parfois s'appuyer sur des représentations subjectives pour trancher entre certains classements ambivalents. Les domaines deviennent un hybride entre classes référentielles et axes d'interprétation. La règle finale est de toute façon toujours très utilitaire : on annote des informations à extraire ou prendre en compte pour indexer le texte, plutôt que des phénomènes linguistiques en eux-mêmes.

### *Typologie circonscrite*

Certaines informations étaient prioritaires pour les usages envisagés : compétences, missions, poste. Mais presque tout texte provenant d'une offre d'emploi s'avère utile au classement qui permettra l'indexation. Les domaines pris en compte sont donc élargis progressivement, pour coller à l'articulation de l'ensemble en texte.

Ces domaines ont l'avantage de s'articuler avec les 5 grandes grilles de classements en ressources humaines : métiers (domaine du Poste), secteurs, compétences (domaines des Savoir-faire et de la Personnalité), missions et contrat.

| Domaine | Principaux types | Fréq. moy. (par offre) | Long. moy. (en mots) |
|---|---|---|---|
| Poste | intitulé du poste, mention du domaine professionnel | 2,36 | 3,98 |
| Mobilité | lieu de travail, déplacements prévus, zone commerciale | 0,85 | 2,67 |
| Etablissement | nom, groupe, site web, données quantitatives (CA, effectifs) | 2,43 | 3,15 |
| Recruteur | nom, spécialisation, site web | 0,23 | 3,27 |
| Secteur | activité, branche, produit ou service vendu | 2,06 | 5,02 |
| Environnement | responsable, interlocuteurs, équipe/service, conditions de travail | 2,09 | 4,63 |
| Missions | fonction principale, tâches, objectifs liés au poste | **7,99** | 8,06 |
| Contrat | type de contrat, durée, horaires, salaire | 1,10 | 3,41 |
| Expérience | [ancienneté + nature de l'expérience] (sous forme phrastique composée) | 1,29 | 9,45 |
| Savoir-faire | connaissance d'un champ, compétences techniques, langues | 2,04 | 3,74 |
| Personnalité | *(pas d'articulation secondaire émergente)* | 3,25 | 2,63 |
| Formation | diplôme, niveau, filière de qualification | 0,86 | 5,26 |
| Contact | e-mail, personne à contacter, adresse, n° de réf., procédure à suivre | 0,79 | 7,20 |

*Table 3: Typologie retenue : domaines, types, annotations/domaine*

Pour résumer, cette typologie prend le parti de nous éloigner un peu de la sémantique des mots pris individuellement mais de nous informer sur le déroulement thématique de l'offre, sur le contexte où une entité sera mentionnée, sur l'importance de cette entité dans le texte et sur son articulation avec les autres en une « mise en scène conceptuelle » (Pottier 1992) de l'emploi décrit.

# 4 Forme syntaxique des éléments relevés

***Analyse en n-grammes***

Pour l'écriture des règles d'extraction (qui s'appuient avant tout sur des motifs lexico-syntaxiques) il était nécessaire d'étudier les constructions syntaxiques sous-jacentes aux expressions annotées (et à leur contexte d'apparition immédiat). Pour ce faire, le texte du corpus a été étiqueté en utilisant un analyseur basé sur un algorithme de Brill.

Le résultat tabulé des annotations permet de les classer en listes d'exemples pour chaque type. On trouvera, par exemple, concernant l'intitulé de poste :

(3) Secrétaire[N] facturation[N]
(4) Acheteur[N] distribution[N]
(5) Consultant[N] cabinet[N] de[P] recrutement[N]

L'étude de ces listes à travers les suites d'étiquettes (ou n-grammes) les plus communes permet de décrire les constructions attendues en terme de formats comme [N-N] et sa variante en [N-N+P+N]. Le signe « - » indique ici une parataxe : on remarquera que l'usage de la préposition est limité dans les intitulés de poste.

Ce genre de diagnostic est complété par des relevés lexicaux, notamment sur le contexte

immédiat des extraits, à la recherche de locutions introductrices caractéristiques. Une règle d'extraction est alors envisageable. Elle prend la forme d'une séquence avec une locution de déclenchement et une construction syntaxique attendue (par exemple « *nous recherchons sur* + [$N_{lieu}$] » pour extraire le lieu de travail).

***Diversité des formes observées***

Pour certains types, on attend une forme très rigide : le contrat, l'adresse, la zone de mobilité réservent peu de variété. D'autres types sont des désignations semi-figées (on retrouve toujours les mêmes termes avec des variations minimales) : c'est le cas de l'intitulé du poste, des interlocuteurs mentionnés, du nom du secteur d'activité, des filières de formation, des qualificatifs de personnalité.

La description de l'entreprise, les critères de mobilité, de savoir-faire, d'expérience et de formation requise, opèrent autour de formes phrastiques archétypales déclinées :

(6) 3 années d'expérience dans le génie climatique
(7) expérience de 2 à 5 ans en administration des ventes et logistique

Enfin, le type le plus riche est celui des missions, qui correspondent à la description des activités du poste. L'expression de l'activité se fait autour d'un noyau verbal ou déverbal, avec différentes possibilités d'expensions. Le noyau central le plus courant est le syntagme verbal transitif classique, schématisable par la séquence [V(activité)+(D+N(objet)+qualificatifs)]:

(8) Assister[V] le[D] directeur[N]
(9) Accompagner[V] des[D] centres[N] régionaux[A]
(10) Dispenser[V] les[D] enseignements[N] en[P] comptabilité[N]
(11) Vous[Pro] fidélisez[V] les[D] clients[N]

Le schéma incite les rédacteurs à s'appuyer sur les collocations entre verbe et nom (*suivre* + *dossiers, fidéliser* + *client, garantir* + *conformité...*). Ce noyau ordinaire offre toutefois un point de départ pour élaborer la phrase et compléter par des détails plus spécifiques à l'entreprise ou à l'exercice du métier[6].

On observe aussi une propension des rédacteurs à appliquer le schéma syntagmatique même dans les cas où il paraîtrait moins efficace sémantiquement : le verbe est toujours présent mais ce n'est plus lui qui porte l'activité. La reprise systématique des mêmes constructions participe de l'aspect stylistiquement morne du texte des offres (« langue de bois »). Ainsi, dans les exemples qui suivent, on relève la même séquence syntaxique [V+D+N+P+N], mais avec des rôles sémantiques décalés : [$V_{support}$+$N_{déverbal}$(activité)+ P + N(objet)]

(12) Assurer[V] la[D] gestion[N] des[P] sinistres[N]
(13) Assurer[V] le[D] montage[N] des[P] dossiers[N]
(14) Assurer[V] le[D] reporting[N] des[P] ventes[N]

Au-delà de la perspective d'extraction, l'étude des séquences n-grammes est ainsi utile pour mettre en lumière des phénomènes phraséologiques, phénomènes dont la distribution pourrait être considérée dans une optique de description du genre textuel et de ses variantes (Biber *etal* 1998). D'autre part, en poussant plus en détail les analyses, on pourrait tenter de classer les

6 Lexicalement, les descriptifs de missions puisent beaucoup dans les jargons de métiers : noms d'instruments, de procédés, d'interlocuteurs.

lexèmes du genre en « primitives » transversales qui entrent dans la composition de nos types : agent, activité, niveau/qualité, produit, instrument, lieu, *etc*.

## 5 Organisation du texte

***Aspects pragmatiques de la construction du texte***

L'offre d'emploi est un message public, à la forme contrainte par un environnement légal et par les enjeux sociaux du recrutement. Comme pour la plupart des textes écrits, il y a un interlocuteur attendu : on sait par avance beaucoup sur la personne qui est « ciblée » par l'offre. La visée principalement informative est ainsi complétée par une visée perlocutoire : emporter la conviction et présenter l'entreprise sous un jour favorable. L'organisation du message puise dans les registres de l'annonce/proclamation, de la publicité, du signalement/descriptif.

De plus, ce texte de quelques paragraphes est inséré sur un tableau d'affichage, dans un journal ou un site web. Une offre est entourée d'autres offres, et dans sa présentation habituelle sur internet le texte est complété par des liens et un encadré/résumé. La lecture d'une offre est donc en même temps une activité de navigation. Cet environnement paratextuel favorise une construction rigide, qui permet une comparaison plus rapide entre les offres.

L'ordre d'exposition dans une offre d'emploi est ainsi déterminé selon un schéma traditionnel de composition du texte, autour de 3 séquences décrites par Kessler et Bèze 2009 : la *Présentation de l'entreprise*, la *Description du poste*, et le *Profil du candidat*. Deux autres séquences optionnelles complètent le schéma : l'*Annonce de poste* (sous la forme d'un court paragraphe avant la *Description du poste* plus détaillée) et les *Contacts* (situés en fin de texte, mais de plus en plus rares sur les offres « en ligne »).

***Séquentialité des types***

Les domaines et leurs types sous-jacents s'égrennent en suivant ces séquences traditionnelles. La figure 3 place chaque annotation selon sa position relative (le début du texte est à gauche du schéma et la fin à droite).

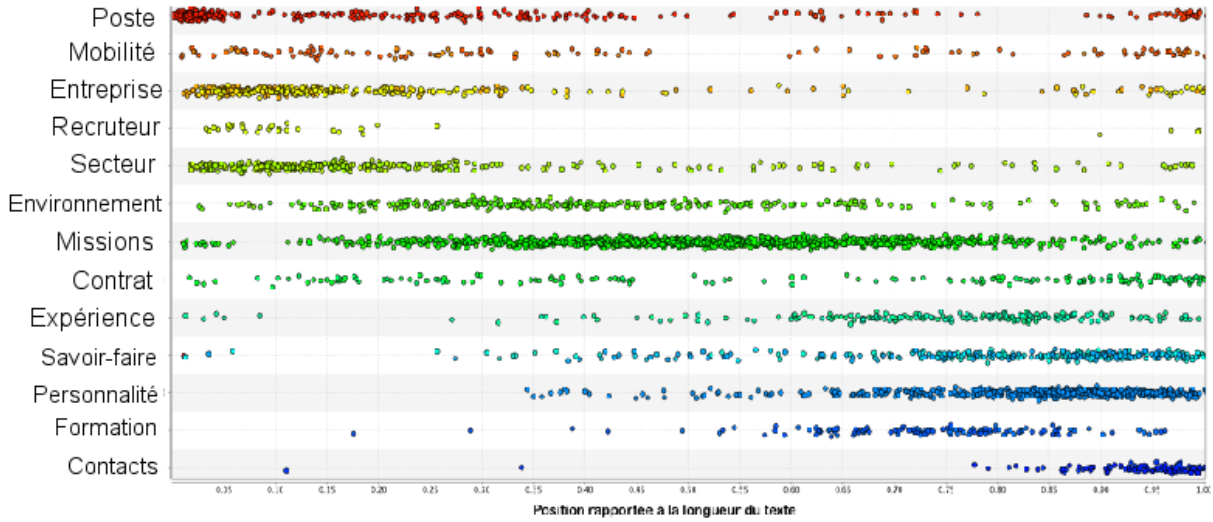


*Figure 3: Distribution des annotations/domaine au fil du texte (synthèse sur tout le corpus)*

L'étude de la position relative de ces champs d'information à travers le corpus permet de

confirmer que la plupart d'entre eux ont une place préférentielle dans le déroulement du texte. En même temps que la position, c'est l'univers thématique qui avance, le propos abordant successivement chaque domaine de la description de l'emploi.

La charpente générale a un effet sur la taille et la dispersion des infos à extraire : le milieu du texte contient une ou quelques phrases très cohésives qui décrivent les activités (annotations plus longues), tandis que les périphéries du texte présentent des compléments d'informations plus autonomes (annotations plus courtes).

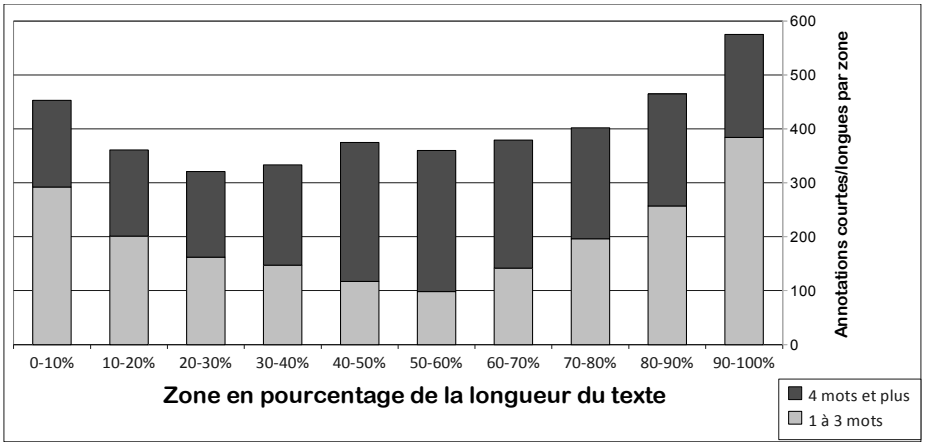


*Figure 4: Longueur des annotations au fil du texte*

Cela nous permet de proposer une nouvelle caractérisation des séquences traditionnelles de l'offre d'emploi : la *Présentation de l'entreprise* prend la forme d'une sorte d'exorde (*cf.* Barthes 1970), qui capte l'attention et met en contexte l'emploi décrit par des informations brèves, présentées toujours de la même façon. Le *Descriptif du poste* constitue une partie un peu plus libre dans son contenu et plus rédigée. A partir d'un modèle très ouvert de période à plusieurs noyaux phrastiques, l'auteur y détaille les tâches ou les objectifs selon « son » point de vue sur le métier. Le *Profil du candidat* condense ensuite d'autres informations exprimées brièvement et de façon plus codifiée : diplôme et expérience requis, énumération des compétences (savoir-faire et traits de personnalité). Cette dernière partie rappelle la péroraison antique, en réévoquant rapidement à travers des qualificatifs conventionnels la teneur des exigences exprimées dans l'offre.

# 6 Conclusion

***Enseignements sur la méthode***

La construction d'une typologie des énoncés observés paraissait nécessaire pour guider l'analyse des offres d'emploi et l'extraction d'informations pertinentes. La méthode suivie, inspirée du repérage d'entités nommées, pose certains problèmes au niveau des critères de référentialité utilisés pour délimiter un domaine. Cela nous a conduit à poser des domaines *ad hoc,* selon un mélange de critères liés à la référence et aux représentations. Toutefois, ces domaines s'avèrent efficaces pour l'étude des offres et de leur genre textuel. Ils participent de la description de l'emploi de façons complémentaires, sont disposés à travers le texte selon un schéma rigide et recouvrent des régularités de forme.

On peut faire l'hypothèse que l'austérité stylistique des offres d'emploi facilitait ce travail de typologie. La thématique des textes est fixée et extrememement homogène. Un découpage du

texte en « lieux » de la progression thématique rejoint alors une analyse en domaines dénotés : la composition formelle du document est symétrique à une décomposition du propos en « contenus » ou dimensions de classement des informations à extraire dans une optique d'indexation. Le sens du texte, imaginable comme la seynète conceptuelle décrivant un emploi, peut-être analysé comme l'assemblage linéaire de ces éléments fixes : ils sont les briques que le genre s'autorise.

***Objectifs pratiques***

A la fois pour le laboratoire et pour l'entreprise, travailler en partenariat contraint à changer des habitudes pourtant essentielles (quantité de différences entre les partenaires : échelle de grandeur des traitements à effectuer, organisation des équipes, nature des objectifs, rythmes de conception et de fabrication, *etc*.). Cependant, l'étude empirique de corpus reste un point central pour les deux démarches.

Dans cette étude, nous sommes partis du principe que la réponse à une demande industrielle de traitement de documents écrits peut être une bonne occasion d'être plus proches du corpus et d'envisager des chaînes d'analyse propres à étayer ou non des hypothèses de linguistique. En l'occurrence, ce travail sur un échantillon de 200 offres a permis d'élaborer une typologie pragmatique des énoncés récurrents, d'obtenir des listes d'exemples classés et d'interroger le rôle du lexique et de la syntaxe dans les résultats recherchés. Ces éléments permettent d'élaborer plus facilement des règles d'extractions pour la suite du projet et nous éclairent quelque peu sur les principes sous-jacents à cette forme textuelle très caractéristique qu'est l'offre d'emploi.

## 7 Remerciements

Le projet SIRE est financé par l'Union Européenne dans le cadre du fonds FEDER (Fonds européen de développement régional). C'est un projet de recherche sur 3 ans lancé par la société Lingway en 2009, en consortium avec la société Proxem et le laboratoire MoDyCo, UMR 7114 du CNRS et de Paris Ouest Nanterre La Défense. Cette recherche donne lieu à une thèse en cours sur la sémantique des offres d'emploi, à l'Ecole Doctorale 139 « Connaissance, Langage, Modélisation ».

## 8 Bibliographie

Barthes, R. (1970) “L'ancienne rhétorique, aide-mémoire”, *Communications* n° 16 : 172-223.

Battistelli, D., et J.-L. Minel (2006) “Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes.” Dans *Compréhension des langues et interaction*, 299-336. Paris: Lavoisier.

D. Biber, S. Conrad, et R. Reppen (1998) *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Charolles, M. (1988) “Les plans d’organisation textuelle: périodes, chaînes, portées et séquences”, *Pratiques* n° 57 : 3-13.

Charolles, M. (1995) “Cohésion, cohérence et pertinence du discours”, *Travaux de linguistique* n° 112 : 125-151.

Condamines, A. (2006) “Modes de construction du sens en corpus spécialisé”, *Cahiers de grammaire* : 75-88.

Fort, K., M. Ehrmann, et A. Nazarenko (2009) “Vers une méthodologie d’annotation des

entités nommées en corpus ?", dans *Actes de TALN 2009,* Senlis, France.

Loth, R., D. Battistelli, F. Chaumartin, H. de Mazancourt, J.-L. Minel, et A. Vinckx (2010) "Linguistic information extraction for job ads (SIRE project)", dans *Actes de RIAO 2010,* Paris, France.

Manning, C. D, P. Raghavan, et H. Schütze. (2008) *An introduction to information retrieval*. Cambridge: Cambridge University Press.

Marchal, E., et D. Torny. "Des petites aux grandes annonces: le marché des offres d'emploi depuis 1960." *Travail et Emploi* 95 (2003): 59-71.

Mathet, Y, et A. Widlöcher (2009) "La plate-forme d'annotation Glozz : environnement d'annotation et d'exploration de corpus", dans *Actes de TALN 2009,* Senlis, France.

Péry-Woodley, M. P. (2005) "Discours, corpus, traitements automatiques." Dans *Sémantique et corpus*, A. Condamines (ed.), Londres: Hermes.