# Analyse de forums de discussion pour la relation clients : du Text Mining au Web Content Mining

Camille Dutrey[1], Anne Peradotto[2], Chloé Clavel[3]

[1] EDF R&D / CRIM-INaLCO – camille.dutrey@edf.fr
[2] EDF R&D – anne.peradotto@edf.fr
[3] EDF R&D – chloe.clavel@edf.fr

**Abstract**

In an evolving electricity market, EDF (Électricité De France) wants to provide offers and services fitting to changes in society and customers' expectations. With this aim, it analyzes various types of data representing spontaneous expressions of customers: messages posted on microblogging platforms, open question surveys, etc. Text mining techniques currently used at EDF require clean documents complying with the writing French language conventions (punctuation, spelling, etc.) in particular. This paper presents methods and tools designed to improve the robustness of the analyses carried out on Web data. We develop a prototypic data cleaning program whose main task is to get the text close to the typographical standards. The program is applied to a corpus of forums dealing with the customer relationship with EDF. We show that the method allows us to improve the quality of the grammatical tagging, on which EDF text mining analyses are based.

**Résumé**

Au sein d'un marché de l'électricité en pleine transformation, EDF souhaite proposer des offres et services en phase avec les évolutions de la société et les attentes de ses clients. Pour cela, elle analyse différents types de données représentant leur expression spontanée : messages postés sur des plateformes de microblogage, questions ouvertes d'enquêtes, etc. Les techniques d'exploration de texte actuellement utilisées à EDF attendent des documents non bruités et notamment conformes aux normes du français écrit. Cet article présente les méthodes et outils conçus dans le but d'améliorer la robustesse des analyses effectuées sur des données Web. Nous avons développé un programme prototypique de nettoyage des données dont la tâche principale est de standardiser le texte pour le rapprocher des normes typographiques du français. Le programme a été appliqué à un corpus issu de forums autour de la relation commerciale d'EDF. Nous montrons que cette méthode permet d'améliorer la qualité de l'étiquetage grammatical, étape préalable aux analyses de text mining menées à EDF.

**Mots-clés :** Données bruitées, Nettoyage de texte, Fouille de texte, Fouille du contenu du Web, Corpus Web.

## 1. Introduction

Une meilleure connaissance de ses clients permet à EDF d'améliorer la qualité de ses offres et services sur le marché concurrentiel de l'énergie. Parmi les moyens d'accéder à cette connaissance, l'analyse automatique de l'expression spontanée de sa clientèle paraît essentielle, grâce aux tâches de fouille d'opinion et d'analyse des sentiments. Cette expression spontanée

provient principalement des différents canaux de contact avec EDF (mails, retranscriptions de messages téléphoniques, enquêtes…), mais est également analysée au travers du Web, qui fournit un support privilégié de communication et d'échanges. Le traitement de ces volumes croissants de données textuelles nécessite la mise en place de méthodes d'analyse automatique permettant de détecter les thèmes abordés par les clients, de faire ressortir des motifs de satisfaction ou d'insatisfaction, de mettre en avant leurs attentes. Les enjeux économiques attachés à l'utilisation de ces informations sont ainsi très importants.

Les techniques de text mining utilisées à EDF, décrites dans (Kuznik *et al.*, 2010), reposent sur une étape de traitement linguistique basé sur des modélisations de concepts : concepts métier (comme « la facture » ou « les tarifs ») et concepts d'opinions. Ces concepts viennent enrichir le texte avant la seconde étape, constituée d'algorithmes statistiques permettant de synthétiser l'information issue de gros volumes de texte en effectuant des regroupements automatiques non supervisés (clustering) ou de la catégorisation.

L'outil utilisé à EDF est Luxid de l'éditeur Temis, basé sur une technologie de modélisation appelée « Skill Cartridge » (ou « cartouche de connaissance »). Celle-ci repose sur des lexiques et des règles, avec différents niveaux de priorité, et est enrichie d'une détection contextuelle. L'utilisation de ces méthodes et outils pour l'analyse des textes issus du Web impose une adaptation à ce type d'expression :

1. Les cartouches développées dépendent fortement du corpus : par exemple l'expression de la satisfaction ou du mécontentement se fera différemment si le texte est issu d'un forum Web ou d'une retranscription d'appel téléphonique. Les spécificités rédactionnelles des textes issus du Web doivent être prises en compte.
2. La modélisation de ces cartouches est basée sur du texte conforme aux normes de la langue française : les textes issus d'Internet doivent être « nettoyés ».
3. La constitution du corpus d'analyse est un réel défi : il s'agit d'extraire uniquement le texte pertinent et de concevoir une structure de données adaptée à son contenu et extensible.

Cet article présente l'ensemble de briques logicielles conçu dans le but de pallier les limites d'une chaîne de text mining confrontée aux données issues du Web : nous avons constitué un corpus (section 2) permettant d'identifier les caractéristiques langagières de ce type de données ; puis, à partir d'une exploration manuelle de ce corpus, nous avons dressé une typologie des phénomènes observables prenant en compte les travaux antérieurs (section 3). Cet article présente ensuite les méthodes développées pour prendre en compte ces spécificités, en amont de la phase d'analyse morpho-syntaxique grâce à un programme de nettoyage des données textuelles (section 4) et en aval de celle-ci grâce à des méthodes d'annotation sémantique d'éléments caractéristiques des écrits numériques (section 5).

## 2. Construire un corpus à partir du Web

L'utilisation du Web comme source de données textuelles soulève des questions en termes de choix – théoriques et techniques – de génération de corpus. D'une part, il est nécessaire de savoir quoi chercher parmi cette densité informationnelle : compte tenu de l'hétérogénéité du contenu et de la structure du Web, les traitements automatiques qui lui sont appliqués amoindrissent la

fiabilité des corpus générés (Duclaye *et al.*, 2006). Ces traitements posent ainsi la question de la validité théorique des corpus, définis dans (Rastier, 2004) comme un regroupement structuré de textes intégraux et documentés, homogène théoriquement (prise en compte des genres et discours) et d'un point de vue pratique (prise en compte d'une perspective applicative). D'autre part, il est primordial de savoir comment chercher l'information (et donc constituer un corpus valide). Compte tenu de la croissance exponentielle et de la perpétuelle mouvance du Web (disparition ou duplication des ressources, précarité et ambiguïté des URLs), une des solutions possibles, et que nous avons retenue, est de cibler a priori les sites Web à aspirer, dans le cadre d'une application précise, comme préconisé par (Beaudoin *et al.*, 2001).

Le corpus Web ainsi conçu, nommé *WebGRC* (pour « corpus issu du Web traitant de la Gestion de la Relation Clients »), est constitué de messages postés sur trois forums de discussion (Yahoo! Questions/Réponses[1], Droit-finances.net[2] et LesArnaques.com[3]). Ce corpus nous permet d'analyser des données écologiques, et donc représentatives de l'expression langagière sur ce type de sites Web, homogènes en termes de généricité des textes et dont le format technique est standardisé afin d'assurer leur pérennité et l'interopérabilité des traitements informatiques futurs. Nous présentons une analyse des choix décisionnels ayant guidé l'élaboration de ce corpus, du point de vue de sa conception (section 2.1) et de sa structuration informatique (section 2.2).

### *2.1. Conception d'un corpus issu de forums Web : le corpus WebGRC*

La première étape de constitution du corpus *WebGRC* a été accomplie par une équipe de sociologues au sein d'EDF. Elle a permis de déterminer *i)* quels étaient les éléments informationnels susceptibles d'intéresser EDF sur la thématique de la satisfaction clients et *ii)* où se trouvaient ces éléments sur le Web : cette analyse manuelle, soutenue par des alertes sur mots-clefs, a mis en exergue le fait que 87 % des discussions à propos d'EDF se trouvaient sur les trois forums que nous avons de ce fait sélectionnés.

Après téléchargement des pages HTML, les sources sont nettoyées pour ne conserver que le contenu textuel pertinent (et les métadonnées associées). Nous utilisons pour cette tâche l'outil de « détourage »[4] Boiler Pipe (Köhlschutter *et al.*, 2010) : cet outil détermine la pertinence d'un bloc textuel et est primitivement développé pour s'appliquer à des sources contenant une grande quantité de texte bien délimité, et peu de balises HTML. Ce n'est pas le cas des textes que nous récupérons : nous avons adapté Boiler Pipe à nos besoins, notamment car les messages courts étaient classés comme non pertinents.

1 http://fr.answers.yahoo.com/

2 http://droit-finances.commentcamarche.net/

3 http://www.lesarnaques.com/

4 Le terme est emprunté à la retouche d'images : il désigne originellement l'opération consistant à n'extraire qu'une partie d'une illustration par séparation de l'objet et du fond ; il est employé ici pour référer à une séparation du texte et du code dans un contexte de document informatique structuré ou semi-structuré, et/ou à une séparation au sein du contenu textuel entre texte pertinent et non pertinent.

### *2.2. Structuration du corpus WebGRC*

Nous avons choisi une structure XML pérenne, extensible et documentée, que nous avons nommée *Atom-Sioc*[5]. *Atom-Sioc* est basé sur des modèles de structuration de contenu et de modélisation des connaissances bien implantés sur le Web, stables et standardisés. Le format structurel retenu est Atom, décrit dans (Nottingham et Sayre, 2005) : conçu originellement pour la syndication de contenu périodique, il est composé d'un vocabulaire générique, adapté à diverses sortes de contenu. Sa généricité peut cependant être source d'imprécision quant à l'encodage de certaines informations spécifiques : nous avons pallié cette limite par l'intégration complémentaire du vocabulaire SIOC, décrit dans (Bojārs et Breslin, 2010), défini en RDF et utilisé pour décrire des objets inhérents aux sites communautaires et la relation entre ces objets.

La figure 1 présente la micro-structure du format *Atom-Sioc*, au sein duquel chaque message posté sur un forum correspond à un élément XML `entry` dont les différents éléments fils encodent le contenu textuel du message et les éventuelles métadonnées qui lui sont associées. Afin de conserver une structure exclusivement XML, les ressources sont qualifiées de manière non ambiguë au moyen de propriétés SIOC encapsulées dans des éléments Atom (*cf.* l'élément `link` en figure 1).

```
<entry>
    <id>tag:textmining@der.edf.fr,20110721:dfinances_837_1</id>
    <title>Service client EDF</title>
    <subtitle type="text">Consommation</subtitle>
    <link rel="self" href="#dfinances_837_1" />
    <link rel="related"
        href="http://droit-finances.commentcamarche.net/forum/affich-5227241-service-client-
            edf#topic_question" />
    <link rel="http://www.w3.org/1999/02/22-rdf-syntax-ns#type"
        href="http://rdfs.org/sioc/types#Question" />
    <link rel="http://rdfs.org/sioc/ns#has_container"
        href="http://droit-finances.commentcamarche.net/forum/affich-5227241-service-client-edf" />
    <author><name>labinnt</name></author>
    <updated>2011-03-22T00:00:00+01:00</updated>
    <content type="text" xml:lang="fr">Bonjour, [...]</content>
</entry>
```

*Fig. 1 – Micro-structure Atom-Sioc du corpus WebGRC*

Le format *Atom-Sioc* permet d'encoder toutes les informations liées à un texte issu du Web. Il permet de représenter avec précision tant les messages issus de discussions sur des forums que d'autres types de contenu, comme les messages issus de plateformes de microblogage.

5 En référence aux deux langages de structuration qui la compose.

Le corpus *WebGRC* contient actuellement 18 492 messages uniques, issus de 2 970 discussions de forums, postés entre 2002 et 2011. Les textes composants ces messages s'éloignent des normes du français standard : ils comportent de nombreuses erreurs orthographiques (inversion de deux caractères, suppression des apostrophes, langage abréviatif, etc.) produisant des non-mots, absents du lexique et donc non analysables par un outil basé sur des ressources dictionnairiques. Des erreurs grammaticales sont également présentes, et produisent des mots attestés dans le lexique mais incorrects compte tenu du contexte sémantico-syntaxique de la phrase.

Nous nommons ainsi spécificités rédactionnelles toute manifestation scripturale caractéristique d'un écrit déviant par rapport aux normes en vigueur. Elles correspondent à des phénomènes variés à la fois en termes d'usages et de modifications linguistiques. Afin de mieux les identifier, nous avons réalisé une typologie de ces spécificités, à la fois basée sur une exploration manuelle de notre corpus *WebGRC* et sur les travaux antérieurs.

## 3. Typologie des spécificités rédactionnelles sur le Web

De nombreuses études sur les spécificités rédactionnelles dans le cadre de la communication virtuelle ont été menées sur les forums (Marcoccia et Gauducheau, 2007 ; Mourlhon-Dallies, 2007 ; Tatossian, 2008) mais également sur d'autres types de support, comme les SMS (Panckhurst, 2008) et les salons de dialogue en ligne (Werry, 1996 ; Pierozak, 2003 ; Tatossian et Dagenais 2009).

Les typologies existantes consistent en des catégorisations empiriques et hétérogènes eu égard *i)* à la diversité des phénomènes repérés et *ii)* à leur nature à la fois linguistique et para-linguistique. En dénote une difficulté de hiérarchisation : les typologies analysent des phénomènes équivalents à des niveaux structurels différents selon l'approche choisie et/ou selon la granularité des sous-catégorisations ; elles sont toutefois convergentes dans la représentation des phénomènes, dont la plupart apparaissent systématiquement et indépendamment du support de communication analysé.

Nous avons construit une typologie de ces spécificités rédactionnelles appuyée sur le corpus *WebGRC* (section 3.1) puis nous avons catégorisé les phénomènes présents en fonction des méthodes de traitement envisagées (section 3.2).

### *3.1. Typologie des spécificités rédactionnelles du corpus WebGRC*

Les phénomènes les plus représentés dans la littérature sont les phénomènes abréviatifs, les fautes de frappe, les fautes d'orthographe et les procédés expressifs.

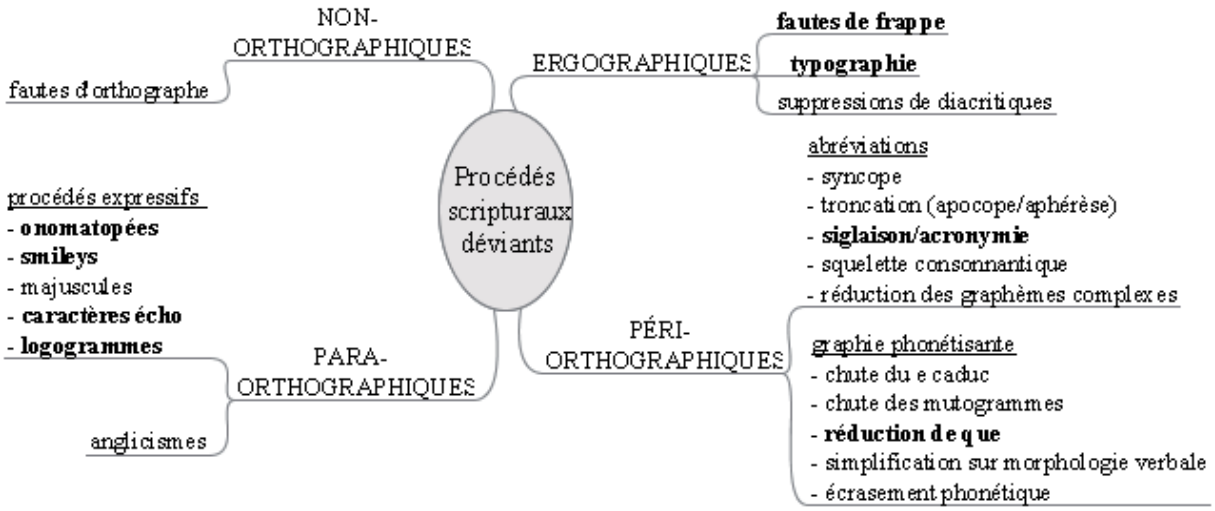


*Fig. 2 – Typologie des spécificités rédactionnelles*

Nous avons développé une typologie, présentée en figure 2, compilant celles développées par (Anis, 2003 ; Krautgartner, 2003 ; Pierozak, 2003 ; Tatossian, 2008) sur des corpus comparables au notre et validée par une analyse humaine sur le corpus *WebGRC*. Nous empruntons ainsi la terminologie mise au point dans (Pierozak, 2003), définissant quatre pôles catégoriels de phénomènes scripturaux : les phénomènes non-orthographiques (qui s'écartent négativement de la norme orthographique), les phénomènes péri-orthographiques (qui s'écartent positivement de la norme en y étant liés), les phénomènes para-orthographiques (qui s'écartent positivement de la norme sans y être liés) et les phénomènes ergographiques (liés à l'ergonomie et ayant une double origine matérielle et cognitive).

### *3.2. Résolution automatique des pratiques scripturales déviantes : positionnement*

Notre typologie met en exergue des phénomènes hétérogènes, que nous avons classés en quatre grands pôles catégoriels. De ce fait, ils n'ont pas la même incidence sur le traitement automatique des textes au sein desquels ils apparaissent et nécessitent des analyses différentes. Nous nous sommes concentrés d'une part, dans le cadre de la chaîne de text mining mise en place à EDF, sur le traitement des spécificités amoindrissant l'efficacité de la tâche d'analyse morpho-syntaxique et d'autre part sur le traitement des spécificités passées sous silence lors de l'extraction d'information dans le cadre de l'analyse d'opinion et de la détection de concepts métiers. Les phénomènes que nous avons traités apparaissent en gras dans notre typologie (figure 2).

Nous considérons que les phénomènes non-orthographiques et que certains phénomènes ergographiques imposent pour leur résolution une tâche de correction orthographique automatique, que nous ne traitons pas : bien que très fréquentes dans ce type de textes, nous constatons que les fautes d'orthographe ne sont pas prédominantes dans notre corpus, notamment car la thématique de nos données (la relation commerciale autour d'EDF) donne lieu à des messages techniques, souvent postés sur des forums spécialisés imposant une expression

correcte. De plus, les méthodes de correction automatique sont paradoxalement susceptibles de bruiter le texte, par exemple en remplaçant une abréviation par une forme étendue qui ne lui correspond pas systématiquement (dans nos corpus, l'abréviation « cb » peut signifier « chèque bancaire » comme « carte bancaire ») ou en faisant disparaître des indices linguistiques propres à ces écrits et utiles pour une analyse ultérieure.

Nous nous concentrons donc sur les éléments des trois autres pôles catégoriels ayant a priori le plus d'impact sur les résultats de l'analyse morpho-syntaxique et sur la détection de concepts : ces éléments peuvent être soit traités par un programme de nettoyage de texte soit identifiés en annotant le texte. Nous nous sommes pour cette deuxième tâche notamment penchés sur la prise en compte des procédés expressifs et plus particulièrement des smileys, considérés comme un des traits les plus caractéristiques du langage internet (Crystal, 2001).

## 4. Méthodes génériques de nettoyage de texte

Nous envisageons la tâche de nettoyage comme la succession de procédés de substitution au sein des chaînes de caractères, visant à rapprocher un texte bruité des normes du français écrit standard sans pour autant le décharger de ses spécificités rédactionnelles. Le principal but de ce traitement est d'améliorer les résultats de la tâche d'analyse morpho-syntaxique. Nous utilisons pour cette étape l'étiqueteur XeLDA, initialement conçu par Xerox et intégré à la plate-forme Luxid de Temis, qui possède les fonctionnalités classiques de ce type d'outil : identification de la langue, segmentation en phrases, tokenisation, analyse morphologique et lemmatisation, désambiguïsation syntaxique et attribution d'une catégorie grammaticale, extraction de syntagmes nominaux.

### *4.1. Principe & Fonctionnement*

La chaîne de text mining déjà mise en place à EDF débute par une phase d'étiquetage morpho-syntaxique. Les résultats de cette étape présentent un nombre important d'erreurs directement imputables à une pratique scripturale déviante, caractéristique des écrits spontanés sur le Web.

Les erreurs d'étiquetage sont de deux types :

1. dans le cas d'un non-mot, XeLDA attribue au token un mauvais lemme (potentiellement identique à la forme) et une mauvaise étiquette (exemples « le.Locataire ») ;
2. dans le cas d'un mot attesté dans le lexique, la lemmatisation est correcte mais le mot n'est pas celui attendu compte tenu du contexte (exemples « a » au lieu de « à »).

L'objectif du nettoyage est de combler le manque de robustesse des outils de text mining lorsqu'ils sont appliqués à des données bruitées, en particulier la phase d'analyse morpho-syntaxique. Les tâches plus spécifiquement affectées sont d'une part la segmentation (en phrases et en tokens), d'autre part l'analyse morphologique et la lemmatisation.

Nous apportons une première réponse possible aux problèmes posés par les données textuelles bruitées en développant un programme générique de nettoyage des corpus Web qui, dans le cadre de la chaîne de traitement actuellement efficiente, ne rompt pas la chaîne de constitution de corpus puis d'extraction d'information sur ces derniers.

Le nettoyage s'effectue en amont de l'écriture des textes constituant le corpus et prend la forme d'un programme architecturé selon un système de pipeline (tube) et conçu de manière modulaire : il comporte une série de méthodes ayant chacune une action précise et délimitée, communiquant via le flux standard et permettant ainsi l'intégration de filtres extérieurs.

La table 1 présente un aperçu des dégradations dues à un texte « malformé » : l'on peut voir l'analyse faite par XeLDA sur certains segments mal orthographiés, ces mêmes segments nettoyés par notre programme ainsi que la nouvelle analyse qui en est faite.

| État du corpus | Formes | Lemmes | Part-of-Speech |
|---|---|---|---|
| Avant nettoyage | j ai une | j / avoir / un | Nom / Auxiliaire / Déterminant |
| | locataire.Le | locataire.Le | Nom invariable |
| Après nettoyage | j'ai une | je / avoir / un | Pronom / Auxiliaire / Déterminant |
| | locataire. Le | locataire / . / le | Nom / Sent / Déterminant |

*Tab. 1 – Étiquetage morpho-syntaxique sur des données textuelles bruitées*

Les différents traitements effectués par notre programme StringCleaner sont organisés en cinq phases successives, chacune portant sur un ensemble cohérent de normalisations :

1. **Normalisation des fins de ligne** : insertion des marques de fin de phrase manquantes, remplacement des sauts de lignes par des espaces simples. Elle permet d'améliorer la segmentation du texte : dans la mesure où les formulaires d'édition de texte sur Internet autorisent généralement une rédaction sur plusieurs lignes, le scripteur utilise souvent le retour chariot et le saut de ligne pour signifier un changement de proposition, omettant de placer une ponctuation forte.

   Ex. : *c'est tout* `\n` *Une autre fois* → *c'est tout. Une autre fois*

2. **Identification des URLs** : un balisage de toutes les URLs et adresses mail permet d'isoler ces unités complexes et de les préserver lors des tâches de segmentation notamment. Elles sont pour cela encadrées par des chevrons, caractères absents du corpus puisque réservés en syntaxe HTML et donc préalablement remplacés par leur entité HTML correspondante.

   Ex. : *http://site.domaine.fr* → *<http://site.domaine.fr>*

3. **Opérations sur les nombres**, valeurs et quantités, afin de mieux identifier et extraire ces composants numériques : normalisation des unités, des ordinaux et des décimaux, « collage » ou « décollage » des nombres si nécessaire.

   Ex. : *10E23* → *10,23 euros*

4. **Opérations syntaxiques** : restauration des apostrophes, décollage des virgules et points syntaxiques, normalisation des marques de citation, des points de suspension, des séparateurs de dates et d'acronymes. Cette étape permet de résoudre de nombreux cas d'erreurs d'espacement biaisant la segmentation. La résolution de la segmentation en phrases est primordiale pour la suite de la chaîne de traitement, dans la mesure où la plupart des traitements effectués en aval est effective au niveau phrastique.

Ex. : *bonjour.voilà j ai une facture E.D. F. datée du 25 / 08 / 2011 → bonjour. voilà j'ai une facture E.D.F. datée du 25/08/2011*

5. **Normalisation des espaces blancs** : suppression des espaces blancs surnuméraires. Cette tâche n'influe pas directement sur les tâches ultérieures mais améliore la lisibilité humaine du corpus et son traitement automatique (notamment en termes de conversions de format).

   Ex. : *bonjour à tous → bonjour à tous*

La gestion des points est une tâche complexe : notre programme de nettoyage distingue les points ayant un rôle syntaxique des autres. Cette méthode n'a pas été étendue aux autres marqueurs de fin de phrase du français car dans notre cas d'utilisation de l'analyseur XeLDA ce dernier les gère correctement indépendamment de leur contexte typographique.

Nous avons présenté les différentes méthodes du programme StringCleaner, regroupées en fonction de leur champ d'action (opérations syntaxiques, sur les nombres, etc.). Bien qu'il s'agisse d'une version prototypique, l'application de ce programme à un corpus en amont des traitements linguistiques montre une amélioration des tâches de segmentation et d'étiquetage.

### *4.2. Résultats*

Nous avons mesuré l'impact de notre programme de nettoyage sur la segmentation et l'étiquetage de l'analyseur morphosyntaxique XeLDA. Nous avons pour cela généré le corpus *WebGRC* en deux versions, comportant chacune les mêmes 18 492 documents : une version « nettoyée » résultat d'une génération filtrée par StringCleaner, comportant la série de normalisations décrite en section 4.1, et une version « brute » exempte de ces pré-traitements. Ces deux versions ont subi une analyse morpho-syntaxique complète effectuée par XeLDA.

L'évaluation des résultats de cette analyse s'appuie essentiellement sur le traitement des mots inconnus : en effet, une des stratégies utilisées par XeLDA face aux mots inconnus est qu'il effectue l'analyse d'une forme en ajoutant au lemme et à l'étiquette grammaticale qu'il lui attribue un label indiquant un score de confiance plus faible que pour un étiquetage exact :

1. le label « Guessed » indique la reconnaissance d'indice(s) morphologique(s) permettant à XeLDA de prédire l'appartenance d'un mot à une catégorie : la lemmatisation est opérée ;
2. le label « Open » indique qu'aucun indice n'a été perçu par XeLDA : ce dernier ne peut pas lemmatiser le token (lemme et forme sont alors identiques) et la catégorie la plus probable, compte tenu du contexte, lui est assignée. Il s'agit généralement de l'étiquette « nom commun » : en effet, lorsque XeLDA échoue dans la reconnaissance d'une forme, il en propose une version lemmatisée en lui associant un score de confiance faible et lui attribue l'étiquette « nom commun ».

Les résultats présentés en table 2 comparent les versions « brute » (référence) et « nettoyée » du corpus *WebGRC* ; ils montrent une diminution du nombre de mots inconnus. La diminution significative du nombre de tokens annotés « Open » confirme l'efficacité du programme de nettoyage StringCleaner.

| Label XeLDA | Nb de labels dans WebGRC « brut » | Diff. nb de labels dans WebGRC « brut » / « nettoyé » |
|---|---|---|
| Guessed | 45 420 | - 0,58 % |
| Open | 16 002 | - 40,80 % |

*Tab. 2 – Diminution du nombre de mots inconnus de XeLDA*

Les résultats présentés en table 3 évaluent le gain en nombres d'étiquettes supplémentaires assignées lors de l'analyse. Cette comparaison nous a servi de base pour une analyse manuelle des différences.

| Tag XeLDA | Nb de tags dans WebGRC « brut » | Diff. nb de tags dans WebGRC « brut » / « nettoyé » |
|---|---|---|
| Noms | 445 939 | - 0,91 % |
| Verbes | 313 845 | 0,27 % |
| Prépositions | 257 956 | 0,44 % |
| Déterminant | 217 059 | 0,52 % |
| Conjonctions | 126 941 | 1,01 % |
| Adverbes | 122 414 | 0,96 % |
| Adjectifs | 100 390 | 0,03 % |
| EOS | 76 701 | 35,49 % |
| Virgule | 69 486 | 4,76 % |
| Numérique | 28 916 | 12,54 % |
| Symbole | 3 958 | 20,19 % |

*Tab. 3 – Amélioration des tâches de segmentation et d'étiquetage*

L'amélioration des résultats de l'analyse morpho-syntaxique est plus ou moins marquée selon les groupes d'étiquettes. Nous pouvons noter une diminution du nombre de tokens étiquetés « nom commun », ce qui correspond naturellement à la diminution du nombre de mots non reconnus par XeLDA au vu de la stratégie de l'analyseur présentée supra.

## 5. Extraction de spécificités rédactionnelles des corpus Web

Certaines spécificités rédactionnelles ne peuvent pas être traitées en amont de l'analyse morpho-syntaxique, comme les procédés para-orthographiques : n'appartenant pas au lexique de la langue française écrite, ces éléments langagiers ne sont pas reconnus par les systèmes faisant

appel à des ressources lexicales de type dictionnaires. Nous avons donc choisi d'agir en aval du nettoyage, en annotant sémantiquement ces items. Ce choix permet également de prendre en compte ces spécificités sans dénaturer les textes, sans masquer les phénomènes discursifs propres à ce type d'écrits.

### *5.1. Principe & Fonctionnement*

EDF utilise la chaîne de traitement Luxid, en amont de laquelle nous faisons intervenir notre programme de génération de corpus Web enrichi du pré-traitement décrit en section 4.

Nous avons développé au sein d'un des modules de cette chaîne une cartouche de connaissances permettant d'identifier des éléments lexicaux spécifiques aux textes issus du Web, porteurs d'opinions ou de concepts métier et passés sous silence par les extracteurs classiques déjà mis en œuvre. Il s'agit d'un ensemble de composants de modélisation sémantique décrivant l'information à extraire d'un texte sous forme de « concepts ». Ces derniers sont représentés par un ensemble de règles linguistiques hiérarchisées (permettant ainsi une extraction possible à plusieurs niveaux d'abstraction) agissant au niveau phrastique et sous-phrastique et permettant d'annoter des mots ou des segments pour les enrichir d'une information sémantique.

En nous basant sur la typologie présentée en section 3.1, nous nous sommes concentrés dans un premier temps sur l'identification de spécificités linguistiques et principalement les procédés expressifs.

### *5.2. Détection d'expressions Web*

Cette cartouche analyse par exemple les smileys, marqueurs saillants de l'opinion exprimée. Leur identification constitue la première étape de leur utilisateur en tant que marqueurs d'opinion. Elle est essentiellement basée sur l'adéquation entre des ressources lexicales, permettant d'identifier des formes précises, et des patrons permettant de repérer leurs éventuelles variantes. La table 4 présente les annotations sémantiques concernant les smileys les plus fréquents.

| Concept | Sous-concept | Exemple |
|---|---|---|
| Smileys | Content | :-) |
| | Triste | :-(( |
| | Sceptique | :-/ |
| | Surpris | 8-O |

*Tab. 4 – Exemple d'identification de smileys avec une cartouche de connaissances*

Ces annotations peuvent ensuite être utilisées comme marqueurs d'opinion et/ou dans la construction de relations sémantiques entre concepts, grâce à l'identification hiérarchisée de concepts permise par Luxid. Les patrons créés dans une cartouche de connaissance combinent, selon leur niveau d'application au texte, une modélisation proche des expressions régulières ainsi que des concepts identifiés à des niveaux inférieurs. Il est également possible de poser des conditions de détection sur des étiquettes morpho-syntaxiques : cette fonctionnalité rend

d'autant plus nécessaire la phase de pré-traitement décrite en section 4 et permettant d'améliorer l'analyse morpho-syntaxique.

Parmi les procédés expressifs les plus fréquents, nous avons également annoté sémantiquement des items d'argot Web (« lol », « mouhahaha », etc.) et la ponctuation forte soumise au phénomène des caractères écho (« !!!!!! » et « ??????? »), encodant une exclamation ou une interrogation emphatiques et donc marqueur d'opinion. Notre méthode a permis d'annoter 5 306 de ces phénomènes.

## 6. Conclusions & Perspectives

Nous avons vu que l'intégration du Web comme source d'information confrontait les outils de text mining à de nouveaux cas d'usage, notamment en termes de généricité des textes et de particularités techniques. Ces données possèdent de nombreuses spécificités les différenciant des corpus usuels sur lesquels sont opérées les tâches d'extraction d'information : les outils de text mining actuellement utilisés à EDF ont besoin d'être adaptés à ce nouveau type de structure et de contenu. Devant ce fort besoin d'adaptation, nous avons élaboré un corpus Web autour de la relation commerciale d'EDF pour lequel les choix de conception ont été guidés par une prise en compte de ses caractéristiques structurelles et informationnelles. Ce corpus, dont le format *Atom-Sioc* est bâti sur des standards, nous permet d'analyser des données écologiques afin de faire émerger les spécificités inhérentes à ce type de texte.

Nous avons mis en place un ensemble de briques logicielles prenant chacune en compte un aspect précis des caractéristiques des corpus Web et résolvant une partie des problèmes engendrés en analyse linguistique automatique : nous nous sommes attachés au traitement des spécificités structurelles et rédactionnelles éprouvant la robustesse des analyses linguistiques, en particulier la tâche d'analyse morpho-syntaxique. Nous avons pour cela appliqué au corpus un programme de nettoyage de textes, intervenant en amont de l'action de l'analyseur morpho-syntaxique XeLDA, puis une méthode d'annotation sémantique en concepts intégrée à la chaîne de text mining Luxid et permettant de repérer formes canoniques et variantes de ces phénomènes.

Ces méthodes permettent d'une part d'améliorer la qualité des textes et de ce fait la qualité des analyses linguistiques et d'autre part d'identifier avec succès des procédés scripturaux caractéristiques de l'expression Web. Elles permettent de constituer une base plus précise et plus riche pour les étapes de fouille d'opinion et d'analyse des sentiments qui interviennent par la suite, grâces auxquelles EDF souhaite acquérir une meilleure connaissance de ses clients, notamment sur la thématique de la satisfaction et du mécontentement.

Les perspectives de cette étude sont nombreuses, et parmi elles :

1. Concernant les méthodes de nettoyage de textes, nous souhaitons les développer sur les phénomènes non traités actuellement et les compléter avec une correction automatique contrôlée, particulièrement pour traiter les problèmes d'inversion de caractères (« masi » pour « mais ») et de déplacement de l'espace entre deux tokens (« lep aiement » pour « le paiement »).
2. Concernant les méthodes d'extraction, nous envisageons de poursuivre le développement de la cartouche de connaissances dédiée à l'identification des expressions Web.

Ces perspectives ont pour objectif commun de couvrir l'ensemble des phénomènes rédactionnels présents dans notre typologie.

Nous souhaitons également élargir le corpus *WebGRC* à des données issues d'autres moyens de communication : nous l'avons notamment enrichi d'un sous-corpus composé de messages postés sur la plateforme de microblogage Twitter, sur lequel nous testons actuellement nos méthodes pour déterminer leur caractère générique.

## Références


Anis J. (2003). Communication électronique scripturale et formes langagières : Chat et SMS. In *Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux Technologiques*. Université de Poitiers.

Beaudoin V., Fleury S., Habert B., Illouz G., Licoppe C. and Pasquier M. (2001). TyPWeb : décrire la Toile pour mieux comprendre les parcours. In *Actes du Colloque International sur les Usages et les Services de Télécommunications (CIUST'01)*, pp. 492–503, Paris, France.

Bojārs U. and Breslin J. G. (2010). SIOC Core Ontology Specification.

Crystal D. (2001). *Language and the Internet*. Cambridge University Press (September 24, 2001).

Duclaye F., Collin O. and Pétrier E. (2006). Fouille du Web pour la collecte de données linguistiques: avantages er inconvénients d'un corpus hors-normes. In *6èmes Journées Francophones Extraction et Gestion des Connaissances*, pp. 53–64.

Kohlschütter C., Fankhauser P., and Nejdl W. (2010). Boilerplate Detection using Shallow Text Features. In *Proceedings of WSDM'10*, New York, USA. ACM.

Krautgartner K. (2003). Techniques d'abréviation dans les webchats francophones. *Linguistik online journal*, 15(3) : 47–67.

Kuznik L., Guénet A.-L., Peradotto A. and Clavel C. (2010). L'apport des concepts métiers pour la classification des questions ouvertes d'enquête. In *Proceedings of TALN'10*.

Marcoccia M; and Gauducheau N. (2007). L'analyse du rôle des smileys en production et en réception : un retour sur la question de l'oralité des écrits numériques. *Regards sur l'internet, dans ses dimensions langagières. Penser les continuités et discontinuités. GLOTTOPOL, Revue de sociolinguistique en ligne*, 10 : 39–55.

Mourlhon-Dallies F. (2007). Communication électronique et genres du discours. *Regards sur l'internet, dans ses dimensions langagières. Penser les continuités et discontinuités. GLOTTOPOL, Revue de sociolinguistique en ligne*, 10 : 11–23.

Nottingham M. and Sayre R. (2005). The Atom Syndication Format.

Panckhurst R. (2008). *Polyphonies, pour Michelle Lanvin*, chapter Short Message Service (SMS) : typologie et problématiques futures. LU.

Pierozak I. (2003). Le français tchaté. *Une étude en trois dimensions – sociolinguistique, syntaxique et graphiuqe – d'usages d'IRC*. PhD thesis, Université d'Aix-Marseille I.

Rastier F. (2004). Enjeux épistémologiques de la linguistique de corpus. Texto ! Dits et inédits.

Tatossian A. (2008). Typologie des procédés scripturaux des salons de clavardage en français chez les adolescents et les adultes. In Durand J., Habert B. and Laks B., eds, *Actes du 1*er *Congrès mondial de linguistique française*, pp. 2337–2352, Paris, France. Congrès Mondial de Linguistique Française, Institut de Linguistique Française.

Tatossian A. and Dagenais L. (2009). Procédés abréviatifs dans les salons de clavardage en français : une comparaison entre adolescents et adultes. *Crisolenguas*, 2(1) : 29–44.

Werry C.C. (1996). *Linguistic and interactional features of Internet Relay Chat*, chapter I. Linguistic perspectives, pp. 47–64. John Benjamins Publishng Compagny.