Adaptation des techniques de Text Mining aux données conversationnelles issues de l'oral

Charlotte Danesi

Adaptation des techniques de Text Mining aux données conversationnelles issues de l'oral

Nom : DANESI

Prénom : Charlotte

UFR des Sciences du Langage

Mémoire de master 2 professionnel – 20 ECTS

Mention : Sciences du Langage

Spécialité : Modélisation et traitements en industries de la langue : parole écrit apprentissage

Parcours :Traitement automatique de la langue et de la parole (TALEP)

Sous la direction de Monsieur Thomas Lebarbé

Année universitaire 2009-2010

Adaptation des techniques de Text Mining aux données conversationnelles issues de l'oral

MOTS-CLÉS : Disfluences, parole spontanée, discours oral, erreur de reconnaissance, text mining, règle linguistique, cartouche de connaissance.

RÉSUMÉ

Ce mémoire aborde la problématique du traitement des données issues de l'oral. En effet, les entreprises regorgent de données concernant leurs clients, données issues d'enquêtes de satisfaction, de forums, d'appels téléphoniques… qui ne sont pas exploitables en l'état.
En premier lieu, un rappel des différents travaux existants en matière d'analyse linguistique des données issues de l'oral y est effectué (voir p 75 à 80). Les transcriptions manuelles et automatiques de ces données orales issues de conversations téléphoniques entre agents EDF et clients y sont ensuite analysées (voir p 75 à 80). Enfin, une solution permettant d'adapter une cartouche de connaissance à ces données spécifiques y est proposée (voir p 75).

KEYWORDS : Disfluencies, spontaneous speech, oral speech, recognition errors, text mining, linguistic pattern, skill cartridge.

ABSTRACT

This thesis addresses the problem of processing data derived from the oral. Indeed, businesses are full of data about their customers, data from satisfaction surveys, forums, call-center... Which are not workable.
First, a reminder of existing work on the linguistic analysis of data from the spontaneous speech is proposed (see p 75 à 80). The manual and automatic transcriptions of speech features from telephone conversations between agents and customers EDF are analyzed (see p 75 à 80). Finally a solution is proposed to accommodate a Skill Cartridge to specific data (see p 75).

Remerciements

« Le meilleur ami de "merci" est "beaucoup". »

Michel Bouthot

Je tiens tout d'abord à remercier ma tutrice de stage Mme Chloé Clavel ainsi que notre chef de projet Mme Anne Peradotto, pour leur gentillesse et leur disponibilité. Merci d'avoir pris le temps de répondre à mes questions et de m'avoir accordé votre confiance.

J'aimerais remercier Mme Silvia Fagnoni, chef de groupe, ainsi que l'équipe SOAD de m'avoir intégrée au sein de leur groupe et s'être montrées aussi accueillantes. Egalement, un grand merci à Ludivine Kuznik pour sa gentillesse, sa présence et son soutien tout au long du stage.

Je souhaiterais également remercier Delphine Lagarde et Mathilde Thebault de TEMIS pour leur patience ainsi que pour l'aide apportée dans l'utilisation de l'outil Luxid.

Enfin, je souhaite remercier l'ensemble des professeurs que j'ai pu avoir tout au long de mon cursus universitaire dans les différentes universités (Nice – Aix en Provence – Grenoble).
Et plus particulièrement :

- Mr Lebarbé, mon tuteur de stage, pour sa disponibilité, son amabilité et pour avoir répondu à mes questions et m'avoir conseillé utilement.
- Mr Véronis, de l'université d'Aix en Provence, qui a été un professeur formidable. Il m'a permis de prendre confiance en moi et m'a laissée entrevoir les multiples facettes de notre spécialité. Encore merci pour sa gentillesse, sa disponibilité et son implication indéfectible.

"Mes remerciements les plus sincères à toutes les personnes qui auront contribué de près ou de loin à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire !"

Dédicace

A tous les « Taliens »

Sommaire

Introduction

EDF emploie 8000 téléconseillers[1] qui traitent 25 millions d'appels par an pour le seul marché résidentiel. Jusqu'à présent, seul un petit nombre d'appels était analysé manuellement. Actuellement, compte tenu du volume croissant des appels téléphoniques, l'entreprise est confrontée à la nécessité de développer des technologies capables d'analyser automatiquement ces enregistrements lui permettant ainsi d'améliorer sa connaissance de sa clientèle.

L'objectif de ce stage de fin d'étude du Master Pro TALEP est, dans un premier temps, de réaliser un état de l'art des analyses linguistiques des données issues de l'oral et plus particulièrement de rechercher la meilleure façon de prendre en compte les disfluences présentes dans le discours oral afin de les traiter. Cette phase d'étude va nous permettre d'énumérer et de comprendre les différents phénomènes liés au langage spontané. S'en suivra la mise au point d'une cartouche de connaissance. Celle-ci, devra prendre en compte les caractéristiques liées à l'oral et sera développée dans le but d'extraire l'information importante d'un corpus (voir p 27). Le but étant d'analyser la relation clients/entreprise en repérant, par exemple, les motifs d'appels ou en analysant la satisfaction et de mesurer l'impact de certains événements comme la publicité, les nouvelles offres, la hausse de prix, une situation de crise, etc.

Au niveau méthodologique, il s'agit de proposer une réponse à l'une des problématiques majeures des Centres de Relations Clients :

Comment extraire et structurer efficacement l'information contenue dans le volume croissant des conversations téléphoniques client/téléconseiller ?

Les données conversationnelles téléphoniques constituent une source d'informations particulièrement difficiles à traiter car la parole spontanée[2] contient de nombreuses disfluences (hésitations, bégaiements, répétitions, etc.) et est sujette aux erreurs des systèmes de transcription automatique.

Ce mémoire s'organise en cinq grandes parties :

- La première partie comporte
  - une présentation du groupe EDF et de ses différents départements,
  - une approche du domaine dans lequel nous travaillons, à savoir le Text Mining, avec une présentation de l'outil utilisé : Luxid,
  - une définition de la mission qui m'est confiée et des objectifs.
- La seconde partie recense les travaux existants en matière d'analyses linguistiques de données issues de l'oral. Un état de l'art est proposé afin de délimiter le champs des recherches.
- La troisième partie propose une description de la méthode d'extraction de l'information.

[1] Sa mission est de renseigner, de conseiller et d'aider le client de l'entreprise
[2] Par opposition à la parole préparée des émissions télévisées ou radiophoniques

- La quatrième partie présente et analyse le corpus de conversations téléphoniques CallSurf sur lequel nous avons travaillé tout au long de ce stage. Nous y abordons les différents phénomènes liés à l'oral et la façon de nettoyer et formater le corpus afin de le rendre exploitable pour la suite des analyses.
- La cinquième et dernière partie propose une étude complète des sorties de la cartouche BSM. Nous y présentons également la structure de la nouvelle cartouche adaptée au corpus CallSurf ainsi qu'une étude de ses sorties. Enfin, nous mettons en place un protocole d'évaluation dans le but de valider cette cartouche.

Ont été également réalisés durant ce même stage et figurent en annexe :

- Un guide d'aide à l'utilisation de l'outil Luxid et à la création de cartouches, rédigé en collaboration avec Ludivine Kuznik (Prestataire EDF – Branche Commerce). (voir p 101)
- Un article (en langue anglaise), écrit avec la participation de Chloé Clavel, traitant des résultats des analyses sur la détection de concepts à partir des données issues de l'oral. (ces recherches seront présentées lors de la conférence ACM Multimédia 2010) (voir p 107).

Partie 1

Contexte du stage

Sommaire

Chapitre 1 - Présentation du groupe

Données extraites de documents internes

I. Le groupe EDF et la R&D

Le 8 avril 1946 naît le groupe EDF (Electricité de France), une entreprise complètement intégrée qui produit, transporte, distribue et vend de l'énergie. Au fil des années, EDF, leader européen de l'énergie en Europe (161 560 salariés dans le monde), devient un acteur de référence sur les 4 principaux marchés de l'énergie en Europe (France, Royaume-Uni, Allemagne, Italie). Le groupe, composé de plus de 70 filiales en France et dans le monde, gère un parc de production d'une capacité de 128,2 Gigawatts et fournit énergies et services à plus de 38,2 millions de clients dans le monde dont 28 millions en France. En 2008, il a réalisé un chiffre d'affaires consolidé de 64,2 milliards d'euros.

La division EDF R&D est une ressource stratégique forte pour le groupe puisqu'elle contribue à améliorer et à sécuriser l'avenir industriel de l'Entreprise.

Elle a pour missions principales de :

- Contribuer à la performance des unités opérationnelles,
- Réaliser des études, développer des méthodes et des outils pour les différentes branches et entités du Groupe,
- Identifier et préparer les relais de croissance du groupe EDF.

La R&D compte environ 2000 personnes, réparties sur trois sites français : Chatou (78), Les Renardières (77), Clamart (92) et un site Allemand : Karlsruhe, tous présentés ci-après.
Le site de Clamart regroupe la moitié des forces de la recherche EDF.
En tout, EDF R&D s'organise autour de 17 départements centrés sur différents axes de recherches.

Tableau 1- Activités des différents sites de la R&D

Chatou 	• Chimie : eau, air, environnement • Contrôle commande et systèmes informatisés pour la production d'électricité • Energies renouvelables : photo-voltaïque, etc. • Fonctionnement des centrales : mécanique des fluides, thermodynamique • Maîtrise des risques industriels • Mesures physiques • Météorologie • Radioprotection • Sciences du vivant • Surveillance et maintenance des matériels, robotique
Renardières 	• Contrôle commandes et systèmes informatisés pour la production et les réseaux • Electricité, électrotechnique pour la production et les réseaux • Fonctionnement des process industriels • Fonctionnement des process dans les bâtiments et le secteur tertiaire • Génie civil • Maîtrise de l'énergie, efficacité énergétique • Matériaux, vieillissement des matériaux • Mécanique • Mesures physiques
Clamart 	• Connaissance des clients • Contrôle commande et systèmes informatisés pour les réseaux électriques • Economie des réseaux • Electrotechnique appliquée aux réseaux électriques • Informatique scientifique, algorithmes numériques • Mathématiques appliquées aux prévisions de consommation et à l'équilibre Production/Consommation • Maîtrise des risques, sûreté de fonctionnement • Mécanique, vibrations, acoustique • Neutronique • Sciences humaines et sociales • Statistiques, analyses statistiques
Karlsruhe 	• Economie de l'environnement • Energies renouvelables, dont biomasse, géothermie, etc. • Energie répartie

II. Le département ICAME et le groupe SOAD

Parmi les différents départements, le département ICAME[3], créé début 2002, met en œuvre des compétences diverses (sociologie, marketing, finance, informatique, data mining et statistique) enrichies par une connaissance et une compréhension :

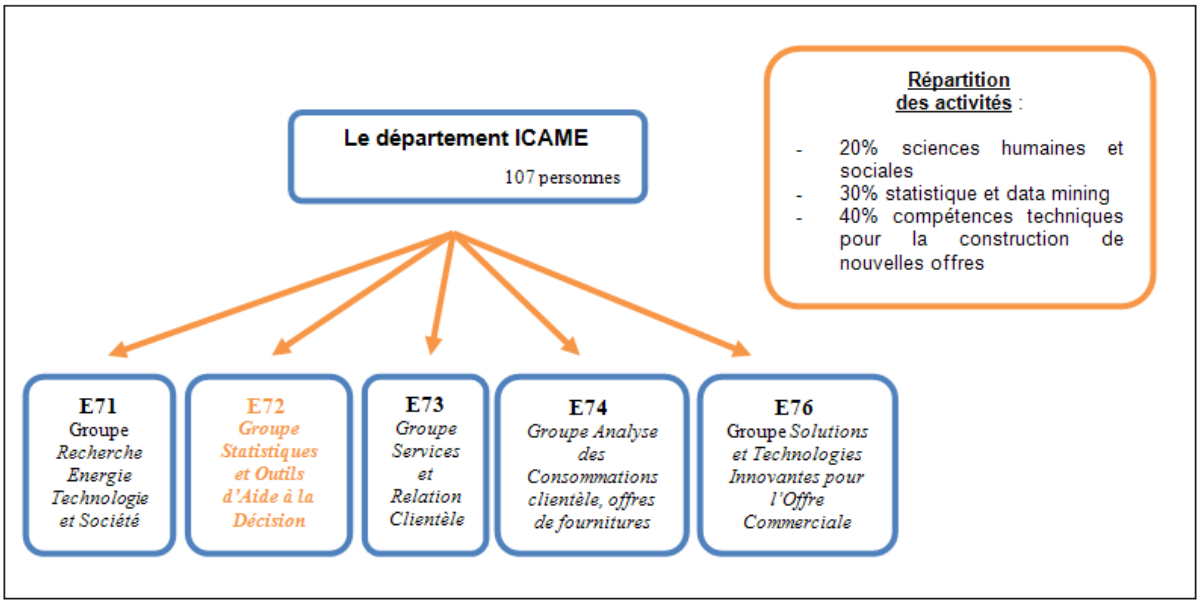
- De l'environnement du Groupe EDF (politique, social, économique, réglementaire et institutionnel),
- Des différents segments de marchés (volume, marges, concurrence, axes stratégiques de développement etc),
- Des clients (perceptions, comportements, consommations, satisfaction, attentes en matière de services),
- Des processus de relation clientèle.

Ces activités sont principalement menées en appui avec la Direction Commerce afin :

- D'apporter à l'entreprise, grâce à son expertise en sciences humaines et sociales et en statistiques, une compréhension de l'évolution du comportement du client,
- D'élaborer de nouvelles offres de fourniture qui adaptent les prix aux inflexions du marché, en s'appuyant sur une modélisation et une simulation de la réponse de la clientèle à de nouvelles structures tarifaires,
- D'innover dans les services en instaurant des dispositifs de gestion locale énergétique et de production locale d'énergie renouvelable, dans un objectif d'éco-efficacité globale,
- De préparer les futurs modes et supports de la relation commerciale (web, …).

Le département est lui-même divisé en cinq groupes :

Figure 1- Organigramme du département ICAME



[3] Innovation Commerciale, Analyse des Marchés et de leur Environnement

L'équipe Text Mining appartient au groupe E72 **"Statistique et Outils d'Aide à la Décision" (SOAD)**, un des trois groupes disciplinaires du Département ICAME. Ce groupe a pour mission de **construire et pérenniser les compétences nécessaires à l'étude et la mise en œuvre des méthodes d'analyse, de fouille et d'enrichissement de données**, principalement issues des systèmes d'information du Groupe EDF.

Les domaines techniques particulièrement explorés sont les suivants :

- L'échantillonnage, la statistique et l'analyse de données dont les courbes de charge,
- Le « data mining » et ses extensions, en particulier aux données complexes (texte, audio, issues du WEB) et aux flux de données,
- L'informatique décisionnelle et l'aide à la décision,
- Les techniques d'extraction et de visualisation de l'information.

Le groupe SOAD mène d'une part, des activités en amont permettant, en réponse à des problématiques du groupe, l'appropriation et l'approfondissement méthodologique et conduit d'autre part, des projets à visée plus opérationnelle (expertise, apport ou transfert méthodologique opérationnel). Il joue aussi un rôle important de dissémination de savoir-faire par l'organisation d'animations statistiques, une veille sur le Data Mining, une contribution à la Direction Commerce et par l'animation du réseau Data Mining du Groupe EDF.

Chapitre 2 - Le Text Mining à SOAD

Données extraites de documents internes

Pendant près de 50 ans, EDF a vécu dans une situation de monopole. La clientèle n'avait pas le choix du fournisseur, et la mission de l'entreprise consistait donc à produire l'électricité au moindre coût en rétrocédant tous les gains de productivité sous la forme de baisse des tarifs. Un grand changement se produit en février 1999. L'ouverture du marché et l'arrivée progressive d'autres fournisseurs d'énergie concurrents pousse l'entreprise à définir de nouvelles stratégies. **La satisfaction, la connaissance des clients et l'optimisation des marges deviennent ainsi des préoccupations majeures.**
C'est pour faire face à l'émergence de ces nouveaux besoins que le Text Mining, notamment, a commencé à être utilisé dans le groupe SOAD. En effet, **les données textuelles représentent environ 80% de l'information dans une entreprise**. Les données issues des contacts avec les clients constituent donc une **source pertinente d'informations** qu'il faut exploiter.

Ces informations permettent de se renseigner sur les sujets suivants concernant la clientèle :

- L'adéquation entre les offres proposées et ses besoins,
- Sa satisfaction et ses attentes,
- Les risques qu'elle parte à la concurrence …

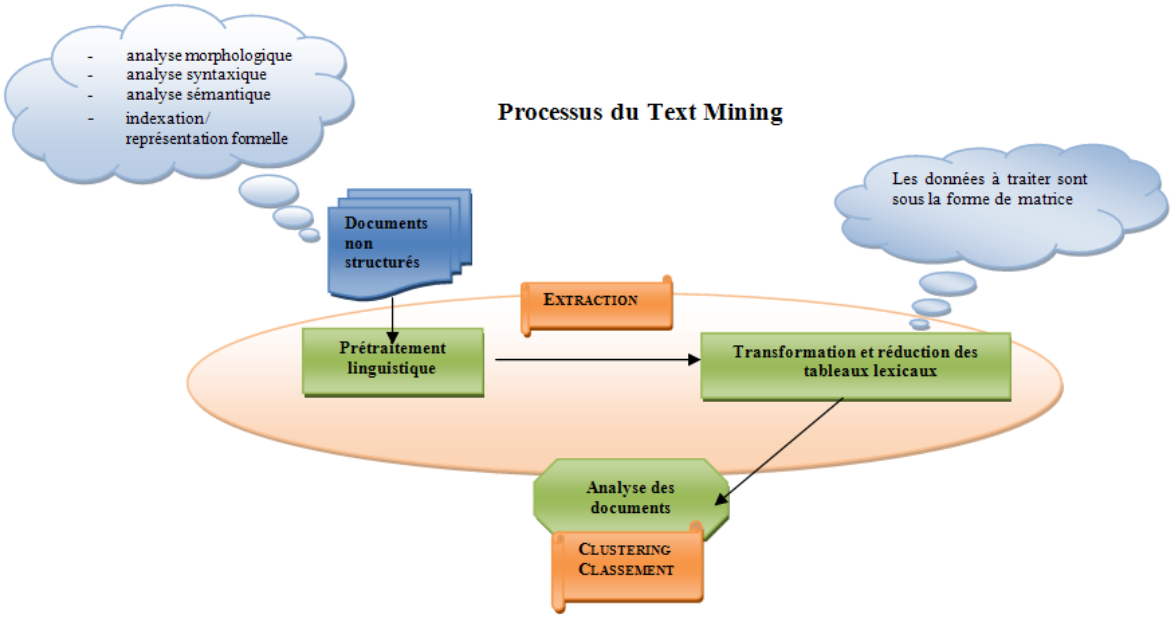
I. Qu'est ce que le Text Mining ?

Le Text Mining, également appelé **fouille de textes ou extraction de connaissances**, est un domaine de recherche dont la première définition est donnée par (R.Feldman, 1995). Le Text Mining, qui est une des disciplines du traitement automatique du langage naturel (TALN), est en pleine expansion car il permet de **traiter un volume important de données textuelles provenant d'internet, de mails, d'enquêtes de satisfaction, de contacts clients**… qui ne peuvent être exploitées manuellement. L'objectif du Text Mining est de faire **ressortir**, dans une masse très importante de données textuelles, **l'information** utile afin qu'elle puisse devenir exploitable informatiquement. Il s'agit donc d'**extraire de la connaissance** de documents sémantiquement proches et de rechercher des relations entre entités textuelles (termes) ou entre documents et de découvrir des tendances, des concepts...

Le processus du Text Mining s'effectue en trois étapes que nous observons également dans le schéma ci-dessous :

- Le prétraitement des données (découpage, nettoyage…),
- L'indexation ou représentation formelle (caractérisation de chaque document par des termes caractéristiques),
- L'analyse des données indexées : classement des documents par thèmes, recherche de relations…

Figure 2 - Processus du Text Mining



II. Les projets Text Mining à SOAD

Au sein du groupe SOAD, les premières études ont pour but de repérer des thématiques d'expression du client dans des enquêtes de satisfaction, des mails entrants, des champs commentaires de la base clientèle à l'aide d'outils que nous verrons par la suite.

La place du Text Mining a donc beaucoup évolué depuis 2002 dans le groupe SOAD. Comme dans beaucoup d'entreprises, la branche commerce d'EDF comprend de plus en plus l'intérêt de l'étude des données textuelles, promettant de beaux jours au Text Mining à EDF, et notamment le développement d'un pôle Text Mining dans leur entité.

De plus, du côté R&D, le groupe SOAD a participé et participe à des projets de grande ampleur comme :

- Infom@gic[4] : « L'objectif du projet Infomagic est d'étudier et de proposer des prototypes logiciels de fonctions avancées d'analyse multimodale de données numériques à des échelles allant d'un seul ordinateur au réseau Internet, aussi bien pour des applications industrielles que pour des applications grand public. »
- DOXA[5] : « Le projet adresse un ensemble de problématiques liées au traitement automatique des opinions et sentiments dans des corpus de données multilingues. »
- Vox factory[6] : « Analyser et de modéliser la qualité et l'efficacité de l'interaction client/télé-conseiller en s'appuyant sur les technologies du traitement automatique de la parole, de la détection des émotions et du text mining. »

Le groupe SOAD participe également au **projet TAO «Travaux amont et Appui à l'opérationnel sur les Outils et méthodes d'analyse de données complexes»**. Ce projet, qui a débuté le 1er janvier 2010 et qui se termine le 31 décembre 2012, traite des données hétérogènes qui concernent les clients : données contenues dans ses systèmes d'information[7], datawarehouse, données d'enquêtes, données externes.

Afin de mieux connaître les clients et leurs attentes, on peut associer à ces données, des informations concernant leur environnement, leur opinion, leur comportement… Les données disponibles sont de plus en plus nombreuses mais très hétérogènes : données textuelles des systèmes d'information, données d'enquêtes, enregistrements d'appels téléphoniques, mails, logs de navigation sur les sites web EDF, forums et blogs, données météo, cartes, images satellites, réseaux sociaux…

Deux problématiques se posent pour l'utilisation de ces données complexes à des fins de connaissance client.

- Tout d'abord leur **extraction** et leur **gestion**, car ces données sont distribuées, volumineuses et hétérogènes.
- Ensuite, leur **exploitation** :
  - Intégration,
  - Recherche intelligente,
  - Aide à la navigation,
  - Visualisation,
  - Analyse
  - Scoring etc.

[4] http://www.capdigital.com/infomagic/
[5] https://www.projet-doxa.fr/index.php
[6] http://www.capdigital.com/vox-factory
[7] Base de données qui contient des informations sur les clients

Le projet TAO a deux objectifs :

1. Le premier objectif a pour vocation de proposer un appui à la Direction Analyse et Connaissance Client de la DSI Commerce, en ce qui concerne **l'exploitation des données structurées et textuelles** : collaboration étroite sur les méthodes et outils de Text Mining, **adaptation de ces techniques aux corpus audio retranscrits** (call mining), et enrichissement des bases de données Client pour améliorer le scoring.

2. Le second objectif concerne des recherches en amont sur l'exploitation des données complexes précédemment décrites.

III. Outil utilisé par l'équipe

Données extraites de documents Temis

Afin de pouvoir mener à bien l'extraction de connaissances, nous utilisons la technologie Luxid développée par la société TEMIS.

Luxid est un logiciel « **pour le traitement intelligent de l'information qui permet à chaque utilisateur de comprendre, analyser, enrichir et partager l'information afin de la transformer en connaissance pour l'Entreprise** ».[8]

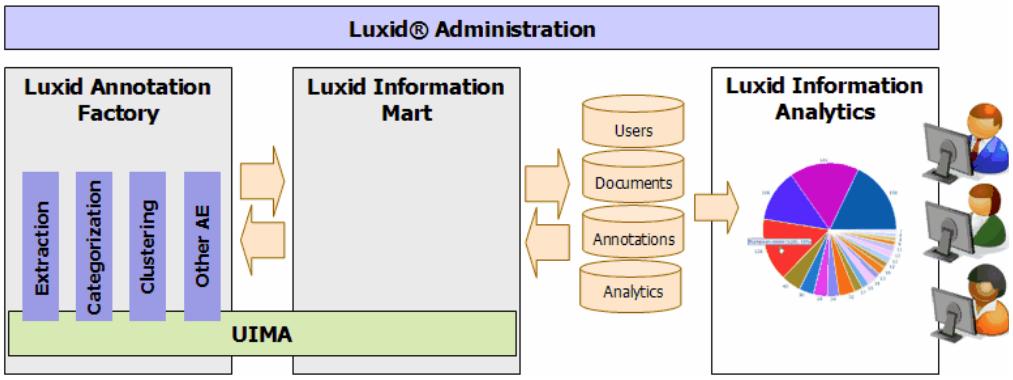
Luxid est structuré en trois couches, que l'on gère via l'interface web Luxid Administration:

- **Annotation Factory,** pour l'installation des différents modules : Skill Cartridges[TM9], modèles de classification, plans d'annotations et scripts.
- **Information Mart,** pour la gestion des workflows (définition des différents processus / modules appliqués pendant la chaîne de traitement).
- **Information Analytics,** portail web conçu pour rechercher et analyser l'information créé par la plateforme précédente. Il sert également pour découvrir et partager la connaissance avec les consommateurs. Dans Luxid Administration, on va pouvoir créer des groupes et leur attribuer des droits.

[8] http://www.generation-nt.com/temis-devoile-version-5-luxid-solution-collaborative-analyse-decouverte-information-strategique-newswire-109701.html

[9] Application qui permet d'améliorer l'extraction d'information à partir de données textuelles

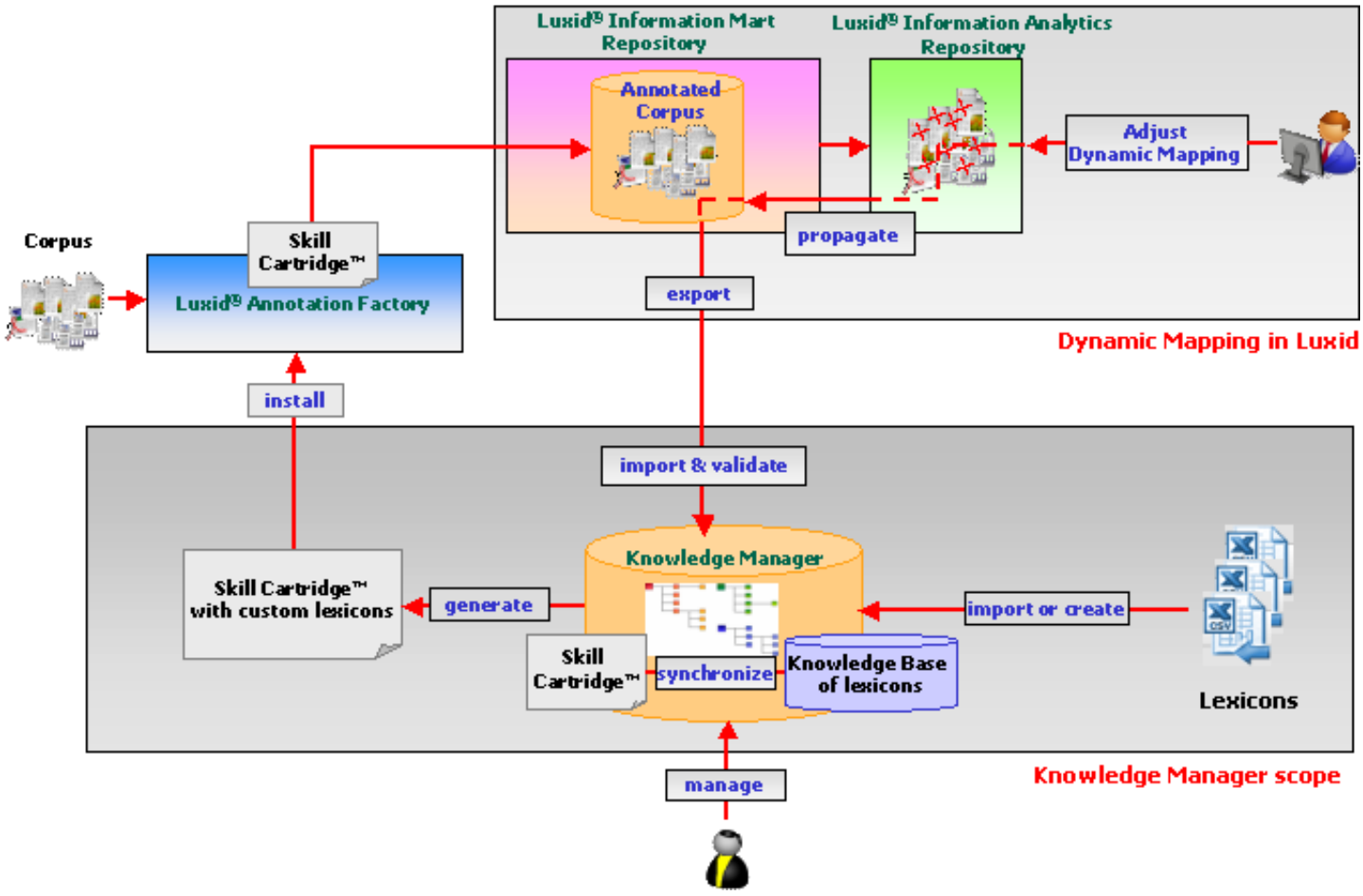
Figure 3 - Luxid Administration



TEMIS propose également un ensemble d'outils de productivité :

- **Luxid® Knowledge Manager,** qui permet de personnaliser les Skill Cartridges™ existantes en ajoutant des listes de vocabulaire et des règles d'extraction.
- **Luxid® Skill Cartridges™ Manager** : environnement de développement qui permet de créer ses propres Skill Cartridges™.
- **Luxid® Training Manager** : qui permet de réaliser du clustering et de développer des outils de catégorisation.

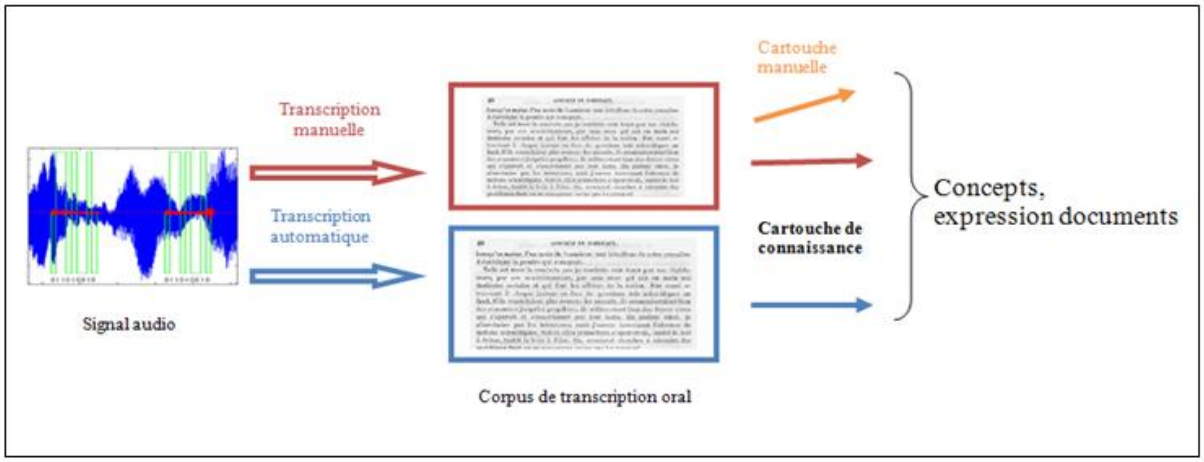
Figure 4 – Architecture de Luxid



Chapitre 3 - Ma mission au sein de TAO

La R&D d'EDF met en œuvre des techniques de Text Mining sur les transcriptions de conversations téléphoniques des centres d'appels. La chaîne de traitement Text Mining (lemmatisation, extraction de concept métier, segmentation, classement) permet de classer ces transcriptions selon différentes thématiques. Que ces conversations soient transcrites automatiquement ou manuellement, les entrées de la chaîne Text Mining diffèrent de celles classiquement traitées par les modules de Text Mining. Il s'agit de **données issues de l'oral** (transcription littérale manuelle et automatique) **contenant de nombreuses disfluences** comme par exemple : les hésitations, les bégaiements sur les amorces de mots, les phrases inachevées, les répétitions de mots ou de groupes de mots.
Ces spécificités liées à l'oral sont difficiles à traiter, notamment lors de l'étape d'extraction de concepts métiers, extraction qui se fait à l'aide de règles linguistiques (Skill Cartridge^{TM}[10]), qui peuvent s'avérer peu adaptées à des données dont le fenêtrage syntaxique diffère de celui des données textuelles classiques.

Figure 5 - De l'oral aux concepts



Cette mission, rattachée au projet TAO et effectuée au sein du groupe SOAD, s'est déroulée en quatre temps :

- Dans un premier temps, nous nous sommes intéressés aux transcriptions manuelles et automatiques, puis nous avons analysé le fonctionnement des cartouches de connaissance[11] sur les deux types de données ainsi que les sorties des cartouches qui nous renvoient les expressions (de satisfaction, de thématique…) détectées (voir p 75 à 80). Le but étant de savoir si les expressions que l'on cherche à détecter sont bien reconnues par le module de reconnaissance malgré les disfluences et les erreurs de reconnaissance.
- Parallèlement à cette phase d'étude, un état de l'art a été réalisé sur les analyses linguistiques de données issues de l'oral (analyse morpho-syntaxique, détection d'entités nommées) (voir p 34 à 39).

[10] Cartouche de connaissance
[11] Règles qui permettent d'extraire des expressions

- Après ces phases d'étude sur la cartouche de connaissance et de recherche sur les analyses linguistiques de données issues de l'oral, une nouvelle cartouche a été développée et adaptée à ces données spécifiques. (voir p 75)
- Enfin nous avons dû réaliser un protocole d'évaluation dans le but de comparer les sorties de la cartouche de la transcription manuelle avec les annotations, réalisées par un annotateur, des expressions détectées comme étant des concepts. (voir p 92)

Partie 2

Etat de l'Art

Sommaire

Chapitre 1 - Spécificités de l'oral

Ces dernières années, de nombreuses études ont été menées sur l'écrit. La plupart des systèmes d'analyse linguistique sont conçus pour l'écrit, mais peu sont prévus pour le langage oral spontané. L'oral reste un domaine peu étudié par les chercheurs tant les disfluences sont présentes et difficiles à traiter.

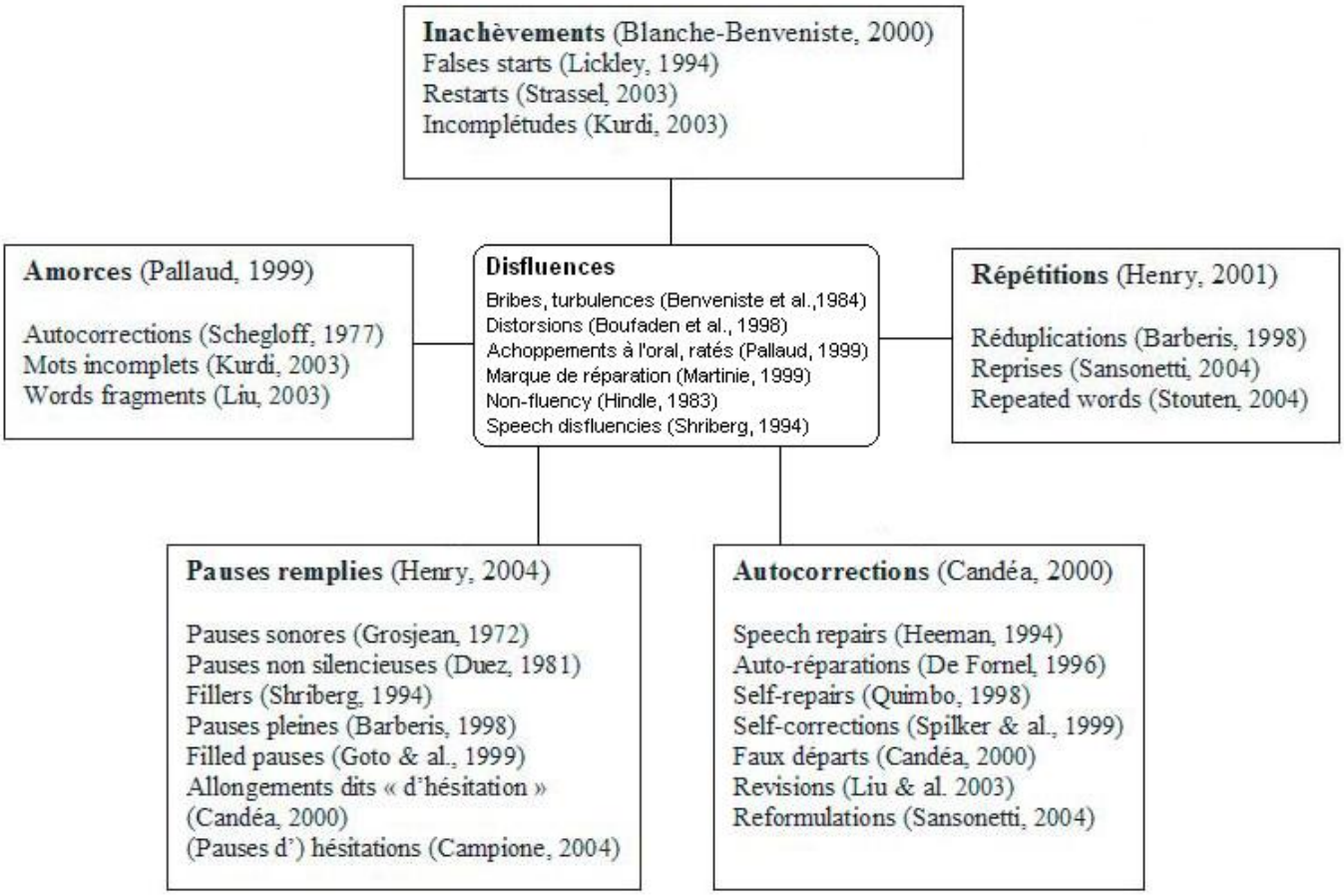
La production du discours oral est soumise au facteur temps. En effet, l'énoncé oral s'élabore en même temps qu'il est construit par le locuteur. Ce processus est similaire à celui du brouillon de l'écrit. A l'oral, les modifications et les corrections présentes dans l'énoncé sont retranscrites lors de la conversion automatique de la parole en texte et par conséquent de nombreuses erreurs sont présentes.

La langue parlée, plus précisément le **langage spontané** d'après (Guénot, 2005), **contient de nombreuses disfluences** qui sont des phénomènes non négligeables de par leur fréquence. Les linguistes ne s'accordent pas sur la définition du terme « **disfluence** » mais une définition revient fréquemment. D'après (Claire Blanche-Benveniste, 1990) c'est un endroit dans un énoncé où le « **déroulement syntagmatique est brisé** ».

De nombreux auteurs ont noté et font remarquer la difficulté de traiter les disfluences issues de l'oral. Ces disfluences sont nombreuses, elles sont toutes répertoriées de façons plus ou moins différentes selon les linguistes et n'ont pas la même terminologie.

Le schéma suivant, inspiré de celui de (Bove, 2008), reprend les équivalences terminologiques des disfluences.

Figure 6 - Equivalences terminologiques des disfluences



(Marie PIU, 2007) identifie cinq phénomènes de disfluences qui peuvent encore être à nouveau sous divisés, comme le fait (Jean-Leon Bouraoui, 2009), en disfluences dites « ponctuelles » et disfluences dites « plus complexes ». Nous rajoutons également une disfluence importante que nous présente (Zellner, 1992) : les hésitations.

Tableau 2 - Les 7 phénomènes de disfluences les plus récurrents

Phénomène de Disfluences	Définitions	Exemples tirés du corpus CallSurf
Les répétitions	Répétition d'un ou plusieurs mots ou reprise à l'identique d'une syllabe, d'un mot ou d'une amorce de mot, de plusieurs syllabes ou de plusieurs mots, sans aucune valeur sémantique (Candéa, 2000).	« Alors **je je** entre tout de suite la mensualisation »
Les autocorrections	Substitution d'un mot ou d'une série de mots à d'autres, afin de modifier ou corriger une partie de l'énoncé (Kurdi, 2003).	« J'ai vérifié elle est pas **de la dans les** documents en attente »
Les amorces	Interruption de morphème en cours d'énonciation (Pallaud, 2002).	« Ils interviennent sur la partie **priva** privative du client »
Les inachèvements	Enoncés auxquels il manque un ou plusieurs éléments pour qu'ils soient grammaticalement bien formés et interprétables sémantiquement (Kurdi, 2003).	« Elle va récupérer, voilà, on résiliera celui-ci qui est toujours actif chez nous »
Les disfluences combinées	Association simultanée d'au moins deux des phénomènes présentés dans ce tableau.	« Vous regardez les 5 derniers chiffres des **chi** des numéros gravés , pas les chiffres qui défilent **hein** »
Les hésitations	« L'hésitation peut se définir comme un comportement d'indécision. » (Zellner, 1992).	« Donc **euh** apparemment, il fallait bien, et le, bien leur dire »
Les marqueurs discursifs	Le marqueur discursif est un décrochement énonciatif métalinguistique (Elsa Pascual, 1995).	« Bon ben alors, une fois qu'on a fait tout ça »

Chapitre 2 – Les difficultés de la reconnaissance automatique

1. L'homophonie

Passer de l'oral à la transcription écrite est un exercice extrêmement difficile. Comme le fait remarquer (Adda-Decker, 2006), la **langue française se caractérise par un grand nombre d'homophones**[12] **qui provoquent des erreurs de reconnaissance**.

Reprenons l'exemple que l'auteur utilise avec un phonème qui peut paraître simple à première vue et qui, sans contexte, devient plus difficile à identifier. Le phonème /*la/* peut s'écrire de nombreuses façons: la, là, l'a, l'as, las et leurs formes acoustiques ne peuvent en aucun cas

[12] De prononciation identique (TLFi)

nous être utile pour identifier la bonne orthographe. Seul le contexte peut aider, mais dans le cas d'un étiquetage morpho-syntaxique, le contexte n'intervient pas.

Nous voyons sur ce second exemple[13] l'illustration du problème de l'homophonie qui se traduit au niveau de la phrase :

Image 1 - Illustration du problème de l'homophonie

Dans cet exemple, ce sont la prosodie de la voix[14] et le contexte qui permettent à l'être humain de désambiguïser les deux transcriptions.

2. L'homophonie liée à l'accent

(Thierry Bazillon, 2008) soulève des critères qui permettent de caractériser la parole dite « spontanée ». Ces critères, qui sont dus à la fois aux différences d'accents entre le nord et le sud de la France et à la confusion, en français, entre les voyelles dites ouvertes et fermées, qui provoquent de nombreuses erreurs de reconnaissance. Les « sudistes », en général, ne font pas la différence entre le [o,e] ouvert et fermé, le contexte devient alors primordial pour déterminer le bon terme ou la bonne conjugaison. Dans le sud, que l'on parle de : la pomme ou la paume, les deux [o] seront des voyelles ouvertes d'où le problème qui se pose pour la transcription de l'oral, dans laquelle le contexte n'intervient pas.

Prenons l'exemple proposé dans l'article de (Thierry Bazillon, 2008) :

« j'ai été » : *j'étais*
« vous demande » : *vous demandait*
« l'enfant aimait sauter dans l'eau » : *l'enfant aimé sautait dans l'eau*
« je l'ai » : *geler*
« le papa c'est une » : *le pas passé une*
« c'est cool » : *s'écoule*

[13] http://www.rue89.com/2010/03/28/le-massacre-des-sous-titrage-pour-les-sourds-144363
[14] Intonation et débit de parole

3. Les assimilations

Comme nous pouvons le constater, les disfluences affectent considérablement le sens de la phrase. Les auteurs font également remarquer que le problème est dû aux assimilations[15] et aux schwas[16] généralement peu prononcés.

Les exemples suivants, proposés dans l'article de (Thierry Bazillon, 2008), illustrent les «interférences», entre consonnes, engendrées par ce problème :

« envie d(e) passer » : *vite passé*
« pas d(e) sanitaires » : *patte sanitaire*
« coup d(e) fil » : *coûte fils*

Les disfluences, comme les faux départs ou les troncatures, génèrent des homophones.

4. Les hésitations

Les **hésitations peuvent « se définir comme un comportement d'indécision.** » (Zellner, 1992) et provoquent, elles aussi, des erreurs de reconnaissance dans certains cas, comme nous pouvons le voir dans l'exemple suivant tiré du corpus CallSurf (décrit p 63).

Manuelle : « leur euh »
Automatique : « l'heure »

5. Les émotions ou bruits environnants

Les émotions et les bruits environnants perturbent, eux aussi, la conversation ou même la personne (agent ou client). Une émotion, plus ou moins forte, peut changer le timbre de la voix, accélérer la production orale et par conséquent provoquer des erreurs de reconnaissance.

Manuelle : « comme nous n'avons pas reçu de facture, je voulais m'assurer qu'il n'y avait pas eu de problème »
Automatique : « Tout la pompe architecture machines qui n'était pas le problème »

Comme nous pouvons le constater, dans l'exemple ci-dessus tiré du corpus CallSurf (vori p 63), les **disfluences de l'oral affectent la transcription** et par conséquent, lors de l'étiquetage ou du traitement effectué sur ces corpus, les résultats statistiques qui sont retournés ne sont pas forcément bons. (voir p 82)

[15] Modification phonétique subie par un son au contact d'un son voisin – ainsi « je ne sais pas » peut être prononcé comme « shais pas » dans la langue orale familière – qui tend à réduire les différences entre les deux
[16] Ou e muet le terme schwa est employé pour désigner la voyelle neutre, centrale, notée [ə] en alphabet phonétique international

Chapitre 3 - Analyse morpho-syntaxique

De nombreux articles traitent de la reconnaissance automatique, mais peu s'attardent sur les traitements de ces transcriptions issues de l'oral. Comme présenté ci-dessus, les logiciels de transcription rencontrent de nombreux problèmes lors de cette étape qui sont dus aux disfluences. Les disfluences ne sont pas forcément traitées de la meilleure façon qu'il soit, et ne sont pas reconnues comme telles par les logiciels.

Tout au long de cette partie nous allons voir que de nombreuses personnes soulèvent le problème et le révèlent mais n'apportent pas forcément de solutions, alors que d'autres proposent des pistes.

(Abdenour Mokrane, 2008) utilise le principe des chunks qui sont des groupes syntaxiques minimaux et non récursifs. L'exemple suivant, tiré de l'article, nous montre une phrase découpée en chunks : GN, GV, GP.

[cette petite phrase]$_{GN}$ [vous explicite]$_{GV}$ [la notion]$_{GN}$ [de chunk]$_{GP}$

Selon son étude, les disfluences orales n'altèrent pas la structure syntaxique des énoncés. Par exemple, si une disfluence se trouve à cheval sur deux chunks, la structure de l'énoncé n'en sera pas altérée. On pourrait se demander : pourquoi ne pas, tout simplement, supprimer ces disfluences lorsqu'elles sont détectées ? C'est ce qu'ont tenté de faire certaines personnes. Supprimer une disfluence « ponctuelle » en tant que prétraitement à l'analyse morpho-syntaxique peut être alors une solution simple à envisager. Les « heu, ben, hum.. » n'ont pas forcément de sens dans un corpus. Mais attention, dans certains contextes, comme le fait remarquer (Thierry Bazillon, 2008), les hésitations telles que « ben » peuvent avoir une signification, ce sont les formes oralisées de « oui ».

Par conséquent, ces disfluences peuvent orienter et modifier le sens du dialogue.

« heu, tu viens tout à l'heure ? »
« ben c'est gentil mais »

L'exemple suivant permet également d'illustrer le rôle des disfluences dans le sens de la phrase. La suppression des disfluences est un acte dangereux car il peut rendre la phrase incorrecte. Reprenons dès à présent un exemple utilisé dans l'article de (Abdenour Mokrane, 2008) cité plus haut :

Exemple avec disfluences :

Je cherche [un camping près de la gare]$_{REP}$ [euh non]$_{ED}$ un près de la côte pardon

Exemple après suppression des disfluences :

Je cherche un camping près de la gare un près de la côte pardon

Dans cet exemple, nous constatons que le fait d'effacer une disfluence, nous fait perdre une partie importante de l'information qui pourrait conduire à rendre une phrase par exemple : d'humeur générale positive en humeur générale négative.

(André Valli, 1999) propose d'adapter un étiqueteur morpho-syntaxique utilisé normalement pour l'écrit sur des corpus issus de l'oral. L'étiqueteur utilisé est Cordial[17], c'est un étiqueteur morpho-syntaxique qui obtient de bons résultats sur l'écrit. Il utilise à la fois des informations statistiques et des règles explicites. De plus, il contient un lexique interne très développé. Un prétraitement modifie la transcription manuelle (enlèvement des pauses silencieuses ou remplies, des amorces de mots interrompus, des indications d'événements non linguistiques, etc.). Cordial étiquette alors cette version aménagée (Habert, 2006). Toutes ces caractéristiques font de Cordial un étiqueteur robuste face aux distorsions qui pourrait être pertinent pour nos transcriptions de centre d'appel.

Chapitre 4 - Extraction d'information

La littérature et les différentes campagnes d'évaluation ont abordé le problème de l'extraction d'information sur des données issues de l'oral sous trois angles distincts présentés ci-dessous.

1. La détection d'entités nommées

« Les entités nommées sont des entités du monde « réel », dont la forme linguistique est une représentation directe dénuée d'ambiguïté. » (Favre, 2007). Il s'agit de rechercher des mots ou groupes de mots correspondant à des noms de personnes, des noms d'organisations ou d'entreprises, des noms de lieux, des quantités, des distances, des valeurs, ou des dates.

La détection d'entités nommées est une tâche assez proche de celle que nous effectuons au sein de SOAD pour la détection de concepts, bien que les patrons linguistiques et les règles soient plus simples dans le cas de la détection d'entités nommées. Les problèmes rencontrés dans ce cadre pourront donc être rencontrés dans le cas de la détection de concepts métiers ou d'opinions à EDF.

2. La recherche documentaire

Dans le cas de la recherche documentaire, « les informations sont matérialisées sous forme de documents dans une ou plusieurs modalités. Un ensemble de documents est appelé corpus et la tâche consiste à extraire d'un corpus l'ensemble des documents correspondant au besoin de l'utilisateur, exprimé sous forme d'une requête. » (Favre, 2007)

3. Le suivi de thème

Dans le cas du suivi de thème, il s'agit de « détecter des coupures thématiques en utilisant aussi bien le contenu linguistique que le contenu audio » (Favre, 2007) et de détecter la nouveauté dans le flux d'information.

[17] http://www.synapse-fr.com/

Partie 3

Méthode d’analyse linguistique

Sommaire

Chapitre 1 – La technologie Skill Cartridge™

I. Qu'est ce qu'une cartouche de connaissance ?

La technologie Skill Cartridge™ a été mise en place par la société TEMIS[18]. Il s'agit d'une application permettant d'améliorer l'extraction de l'information à partir de données textuelles. **Une Skill Cartridge™, appelée également cartouche de connaissance, peut se définir comme une hiérarchie de composants de linguistique tels que des lexiques ou des règles d'extraction.** Son but est de décrire l'information à extraire pour un métier, un domaine spécifique ou une thématique en modélisant l'information sous forme de patrons linguistiques. L'utilisation de cette technologie s'effectue sur un type de corpus spécifique et ne peuvent se généraliser à tous les domaines. La construction d'une Skill Cartridge™ nécessite néanmoins d'avoir une bonne connaissance des données à traiter et d'avoir une idée claire de ce que l'on souhaite extraire.

II. Pourquoi construire une cartouche de connaissance ?

Une Skill Cartridge™ devra permettre, non seulement de filtrer l'information en éliminant l'information inutile, mais également d'organiser l'information à l'intérieur de concepts/sous-concepts.

L'objectif de la construction d'une cartouche est de **mieux connaître les besoins et les demandes des clients afin de répondre à leurs attentes dans les plus brefs délais.**

Exemples d'applications à EDF :

1. Repérer et extraire des informations pertinentes d'un gros corpus ou d'un système d'information
   - Extraire tous les clients qui abordent des problématiques environnementales dans un corpus.

2. Faciliter l'apprentissage d'un modèle de classement
   - Le groupement de lexiques ou d'expressions sous un même concept peut améliorer l'affectation d'un document à une catégorie prédéfinie.

3. Améliorer les résultats d'une classification
   - Modélisation de la satisfaction/insatisfaction permettra de gérer les problèmes liés à la négation et ainsi distinguer des clients satisfaits des clients insatisfaits.

4. Normaliser des données en vue d'une meilleure classification
   - Création des règles permettant de mieux qualifier des données.

[18] http://www.temis.com/

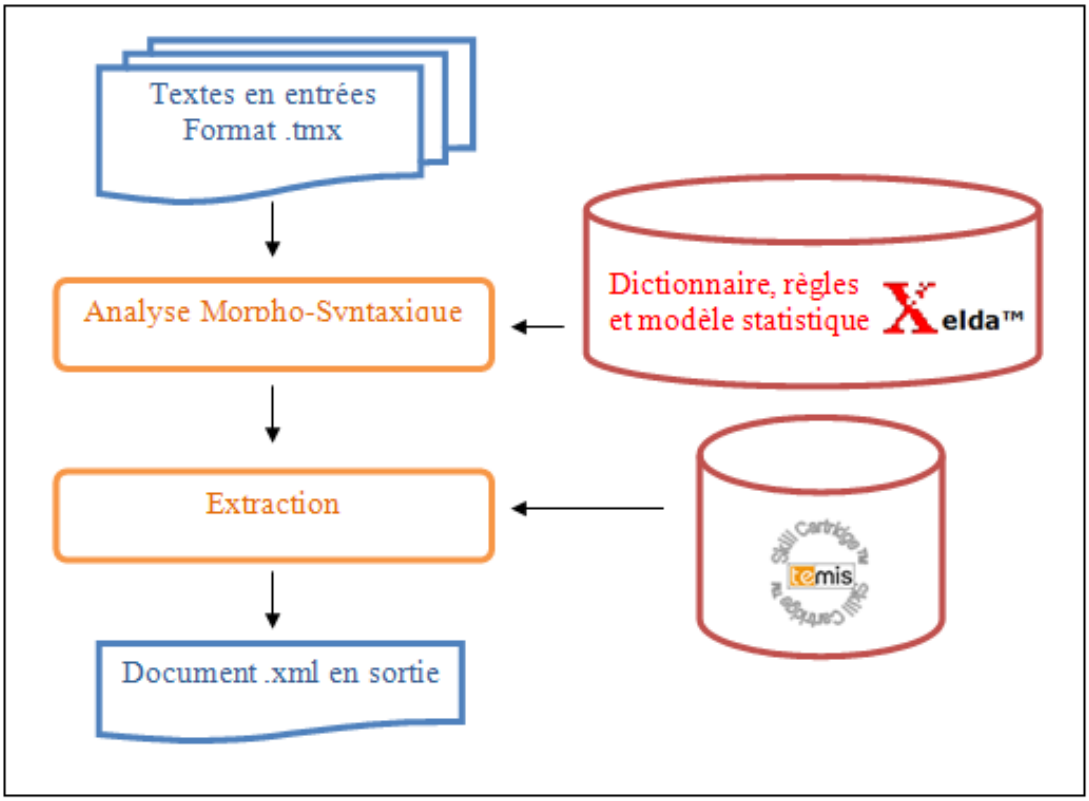
Chapitre 2 - Principe et organisation d'une cartouche de connaissance

I. Organisation globale

L'objectif premier de la cartouche est **d'extraire des concepts d'un corpus volumineux de données.**

Le processus d'extraction se fait en plusieurs étapes qui sont représentées dans le schéma suivant :

Figure 7 - Structure d'une Skill Cartridge™



En entrée, les données textuelles sont préalablement transformées en format TMX [19] et sont analysées par XELDA.

XELDA[20] est un **moteur linguistique multilingue qui modélise et normalise des documents non structurés, en vue d'une exploitation automatique de leur contenu**. Il permet la réalisation des tâches suivantes :

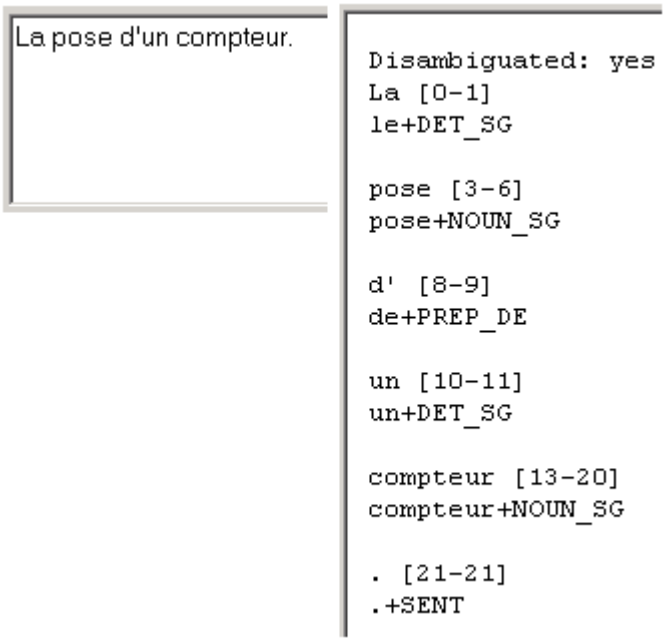
- Identification de la langue (à partir des premiers caractères)
- Tokenisation[21]
- Segmentation en phrase
- Confrontation au dictionnaire
- Identification des lemmes possibles pour chaque forme
- Affectation des différentes étiquettes grammaticales possibles
- Désambiguïsation sur la base d'un calcul statistique

Nous utilisons l'exemple suivant : « La pose d'un compteur. » afin d'illustrer le fonctionnement de ce moteur linguistique. La première balise nous informe sur la langue identifiée, puis la phrase est découpée en tokens <surface-form> et en lemmes [22]<base-form>. Une étiquette <part-of-speech> est alors attribuée aux mots. Enfin la fin de la phrase est identifiée grâce à la ponctuation.

```
<language iso639_1="fr" iso639_2_B="fre" iso639_2_T="fra" name="French" />
<surface-form>La</surface-form> <part-of-speech>DET_SG</part-of-speech> <base-form>le</base-form>
<surface-form>pose</surface-form> <part-of-speech>NOUN_SG</part-of-speech> <base-form>pose</base-form>
<surface-form>d'</surface-form> <part-of-speech>PREP_DE</part-of-speech> <base-form>de</base-form>
<surface-form>un</surface-form> <part-of-speech>DET_SG</part-of-speech> <base-form>un</base-form>
<surface-form>compteur</surface-form><part-of-speech>NOUN_SG</part-of-speech><baseform>compteur</base-form>
<surface-form>.</surface-form> <part-of-speech>SENT</part-of-speech> <base-form>.</base-form>
```

[19] Format XML suivant une norme propre à TEMIS
[20] http://www.xrce.xerox.com/Research-Development/Historical-projects/XeLDA
[21] Découpage du texte en unités lexicales
[22] Le lemme est l'unité autonome constituante du lexique d'une langue.

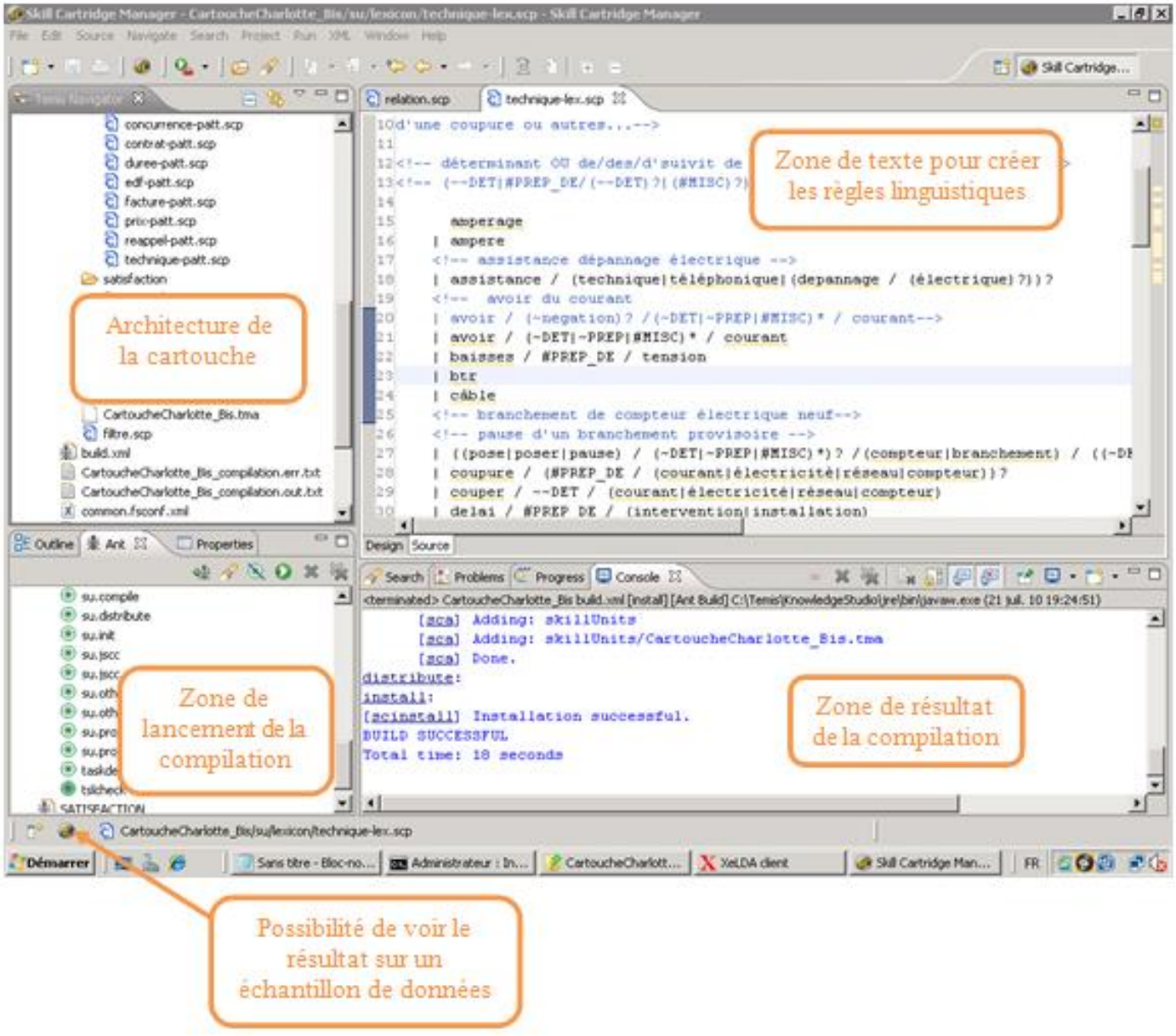
L’interface nous permet de mieux visualiser le résultat obtenu après analyse.



Une fois l’étape XELDA terminée, la Skill Cartridge™ intervient pour procéder à l’extraction de l’information, à l’aide des règles linguistiques.

La copie d’écran ci-dessous illustre l’interface du Skill Cartridge™ Manager.

Figure 8 - Interface du Skill Cartridge™



La Skill Cartridge™ comporte quatre fichiers qui sont primordiaux pour la création d'une cartouche.

- Les *.scp sont les fichiers qui comportent à la fois les règles et les lexiques. Le nombre de fichiers *.scp est fixé par le linguiste.

Image 2 - *.scp

*edf.scp

```
<?xml version="1.0" encoding="UTF-8"?>
<component name="Lexicon" xsi:noNamespaceSchemaLocation="http://www.temis-grou
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<concept name="edf" display="always">
    <e>
<!-- EXPLICATION DU CONCEPT -->
<!-- ce concept récupère tous n-grammes qui sont en rapport avec le thème EDF.

          edf
        | commercial / edf
        | agent / (edf)?
        | interlocuteur / edf
        | edf / :pro
        | edf / branche / commerce

    </e>
</concept>
</component>
```

- Le *.scu est un fichier qui permet de déclarer, à l'aide des balises <include>, les *.scp appelés par la cartouche.

Image 3 - *.scu

BSM .scu

```
<?xml version="1.0" encoding="UTF-8"?>
<skillunit xsi:noNamespaceSchemaLocation="http://www.temis-
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<!-- Ordre et application des composants -->
<apply order="Lexicon,Rules"/>

<!--tools-->
<include ref="outils/metatag.scp"/>
<include ref="outils/outils.scp"/>

<!--lexique-->
<include ref="satisfaction/negatif/negatif-lex.scp"/>
<include ref="satisfaction/positif/positif-lex.scp"/>
```

Remarque :

On attribue aux fichiers *.scp un nom de composant qui détermine l'ordre de lecture des différents fichiers. Cet ordre de lecture est déclaré dans la balise <apply> du *.scu. Généralement, deux niveaux sont utilisés : le niveau « Lexicon » et le niveau « Rules ».

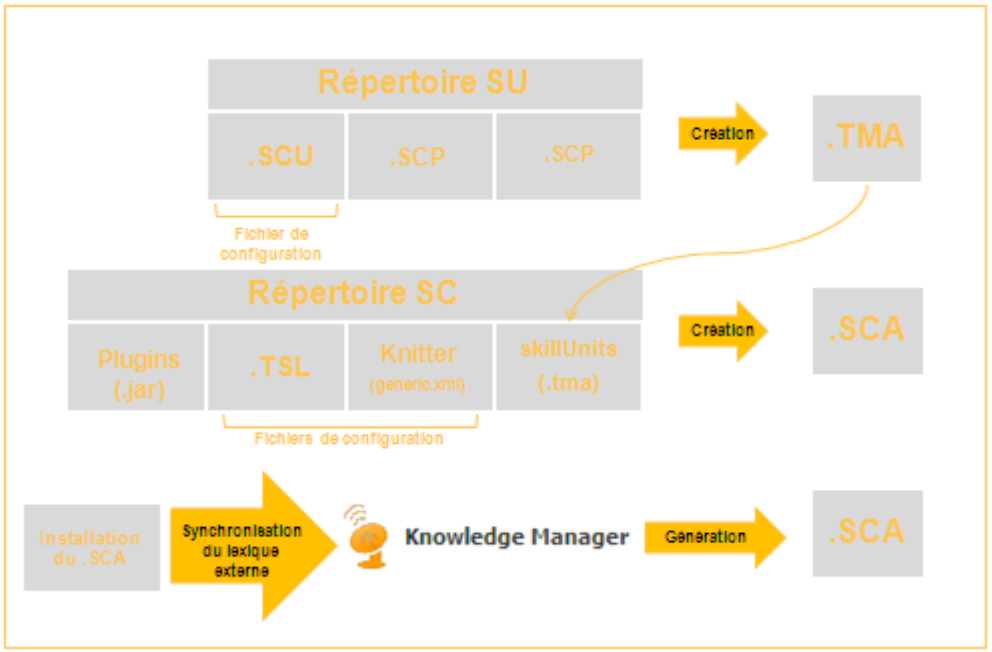
```
5<!-- Ordre et application des composants -->
6<apply order="Lexicon,Rules"/>
```

- Le *.tma est le résultat de la compilation des fichiers *.scu et .*scp. Il comporte l'ensemble des règles linguistiques traduites en langage machine.

- Le *.sca est un fichier compressé de la cartouche. Il comprend :

    - le *.tma,
    - les knitters[23],
    - les pluggins.

Image 4 - Chaine de création d'une cartouche



Remarque :

La dernière ligne de la figure précédente, nous indique que l'on peut :

- Personnaliser la Skill Cartridge™ par le biais du Knowledge Manager,
- Ajouter du lexique grâce à l'utilisation des dictionnaires,
- Affiner les résultats de l'extraction grâce au Static mapping[24].

[23] Moteur conçu pour cartographier les concepts générés par la Skill Cartridge™
[24] Processus permettant de corriger les résultats de l'extraction de la Skill Cartridge™

II. La notion de concepts

Comme nous venons de le voir précédemment, la détection de concepts s'effectue en deux étapes :

- Analyse morpho-syntaxique (Xelda tagger),
- Analyse sémantique.

Durant cette dernière étape, le corpus est étiqueté par des concepts et des sous-concepts prédéfinis. **Ces concepts sont des ensembles de règles linguistiques qui répertorient toutes les formulations possibles d'une même information à l'échelle de la phrase.** Ils sont organisés en niveaux, ce qui permet de définir un ordre dans lequel les règles sont appliquées.

La syntaxe minimale d'un concept est la suivante :

```
<concept name= "Nom_concept">
<e>
</e>
</concept>
```

Lorsqu'un concept est détecté dans un corpus, celui-ci est automatiquement remplacé par le concept. Il n'est alors plus possible de repérer l'expression initiale. Il est donc impératif de ne remplacer que les éléments qu'à coup sûr. Lorsqu'un doute subsiste, il est possible de mettre le lexique dans une macro et non un concept. **La macro permet de déclarer les éléments sans pour autant les remplacer immédiatement par une étiquette.**

La syntaxe minimale d'une macro est la suivante :

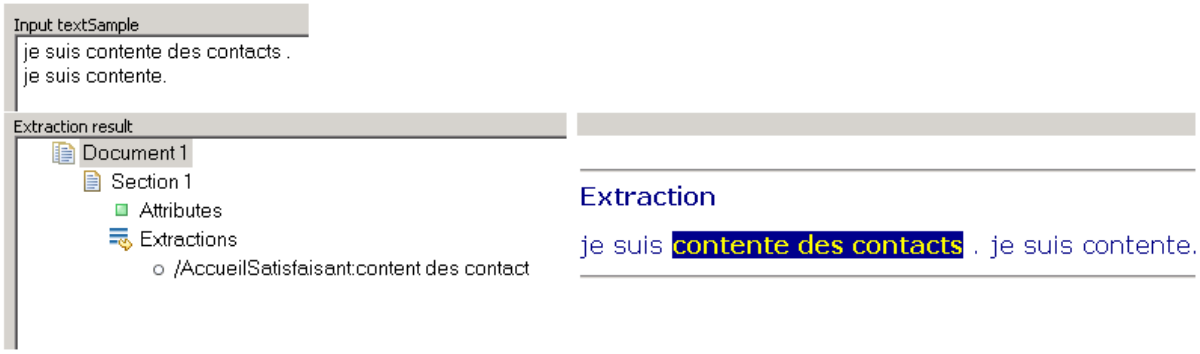
```
<macro name = "Nom_macro">
<e>
</e>
</macro>
```

Considérons l'expression « Je suis contente des contacts. » qui serait formulée par un client. Nous souhaitons extraire en priorité les raisons de la satisfaction des clients « contente des contacts. ».

Figure 9 - Détection de concepts – règle A

La première règle récupère du lexique dans des concepts au niveau 0 : « contente » et « contacts » sont remplacés par les concepts « Positif » et « Contact ». Le niveau 1 permet d'assimiler le « Positif suivi de Contact » comme étant le concept « AccueilSatisfaisant ».

On récupère bien ainsi l'information qui nous intéresse.



Mais si une cliente venait par ailleurs à ne dire que « Je suis contente. », nous aimerions que cette expression soit également extraite. Or avec la règle A, on ne récupère que le lexique positif « contente ». Il nous faut donc créer une nouvelle règle « être + lexique Positif ».

Le problème étant que cette expression est également présente dans la première phrase « Je suis contente des contacts ». Si la règle est instaurée au niveau 1, elle entre en conflit avec la règle décrite dans le schéma précédent. On récupèrera alors « suis contente » et non plus « contente des contacts. », comme illustré ci-dessous.

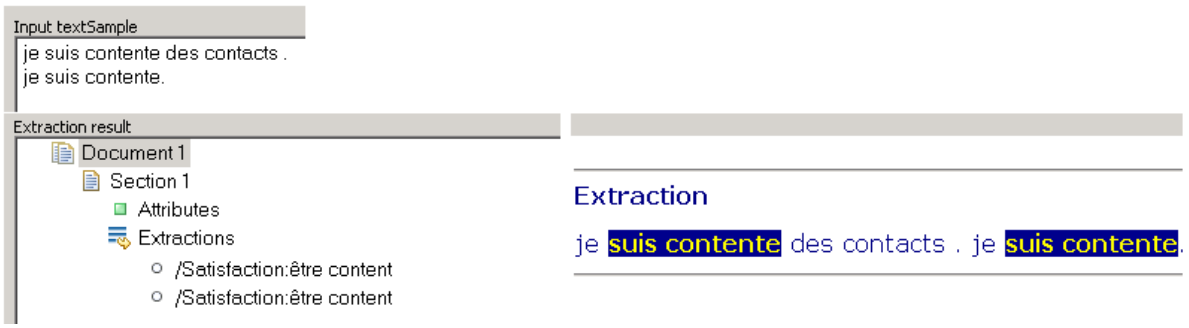
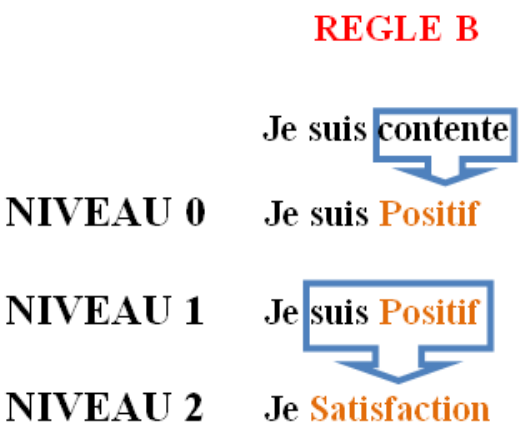
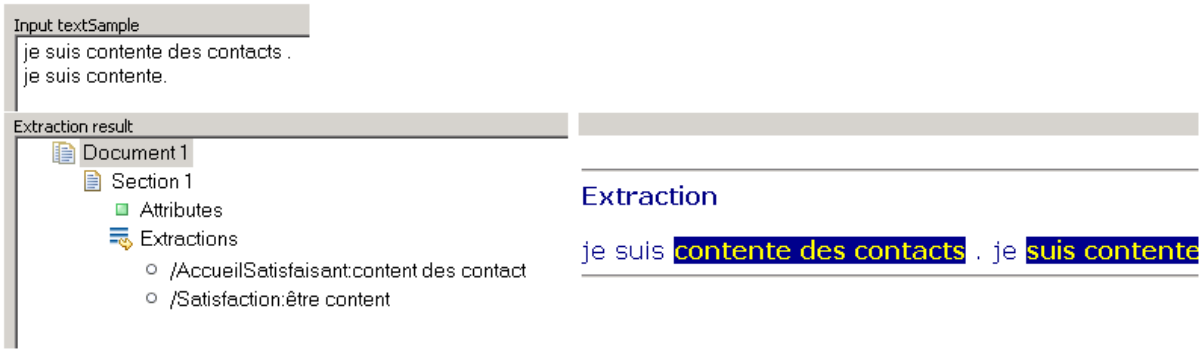


Figure 10 - Détection de concepts – règle B



Il faut donc écrire cette dernière règle au niveau 2, puisque le concept « AccueilSatisfaisant » aura bloqué le concept « Positif » et ce dernier ne sera plus récupérable.
Pour éviter ce conflit nous récupérons « être + lexique Positif » comme étant un concept « Satisfaction » à un niveau supérieur.

En créant la règle de cette façon, nous récupérons bien les deux expressions qui nous intéressent sans provoquer de conflits.



III. La syntaxe des règles

Les règles de la cartouche sont élaborées avec des opérateurs proches de ceux utilisés dans les expressions régulières.
Nous les présentons dans le tableau ci-dessous :

Tableau 3 - Les opérateurs

Signification	Symboles	Exemples
Commentaire	< !-- -->	<!-- relation clientèle est satisfaisante -->
Concept père	~	~contact-lex Pour appeler le concept ou la macro contact-lex
Concept fils	~~	~~Positif-general Pour appeler tous les concepts fils du concept Positif-general (concept fils = concept contenu dans un concept).
Tag	#	#DET_SG Pour appeler le tag XELDA « déterminant singulier » (voir p 103)
Ou	|	(bon | bien) Pour gérer la présence du « bon » **ou** « bien » dans l'expression.
Négation	^ !( )	^ !(impossible) Rmq : il est impossible d'appeler autre chose que des lemmes ou des formes.
0 ou 1 occurrence	?	(~EDF)? Le concept EDF est facultatif.
0 ou n occurrences	*	(~contrat) * Il est possible de rencontrer 0 ou n fois le concept contrat.
1 ou n occurrences	+	(#ADV) + Il faut minimum un tag adverbe.
Forme	:	(:est | :sont) : permet de faire une recherche sur la forme de surface.
Lemme		(contacter | appeler | démarcher) L'absence des : permet de faire une recherche sur les lemmes.
Séparateur	/	il / avoir Si l'on recherche l'expression « il aurait », on sépare le pronom du verbe par un slash (séparateur d'unité).

Ces expressions régulières permettent d'élaborer les règles linguistiques présentes dans la cartouche. Voyons dès à présent l'exemple d'une règle qui détecte le concept « Positif-general ».

Image 5 - Règles linguistiques

```
<concept name="Positif-general"  display="always" level="2">
        <e>
                < !--cela  se passe très convenablement -->
(1)     (ça|cela|tout|:choses) / se / passer / (~ADV|~intensifie-lex)* /
        (bien|correctement|convenablement)
                <!-- C'est très bien -->
(2)     | ~PRON / (:est|:sont) / (~~quantifieur)+ / (bon|bien) / \.

        </e>
</concept>
```

Dans cet exemple, nous pouvons remarquer que nous sommes en présence d'un concept nommé « Positif-general » de niveau 2 et qui a pour display « always ». Le champ **display permet de gérer l'affichage du concept dans le résultat de l'extraction en fonction de ses relations avec les autres concepts.** L'attribut « always » indique, dans le cas présent, que le concept sera toujours apparent à la sortie.

Dans ce concept, nous avons deux règles avec deux commentaires qui sont représentés par la balise < !--*commentaire*-->.

- Dans la première règle : l'expression doit commencer par l'un des trois lemmes proposés ou par la forme « choses » au pluriel. Le premier mot doit être suivi des lemmes « se » et « passer ». Les concepts (ADV ou intensifie-lex) contenus dans la troisième parenthèse sont facultatifs (0 ou plusieurs occurrence(s)). L'expression détectée se termine obligatoirement par l'un des trois derniers lemmes.

- Il en est de même pour la seconde règle dans laquelle s'enchaînent l'une des deux formes spécifiées du verbe « être », le concept fils « quantifieur » et un point en fin de phrase.

-

Chapitre 3 - Présentation de la cartouche BSM

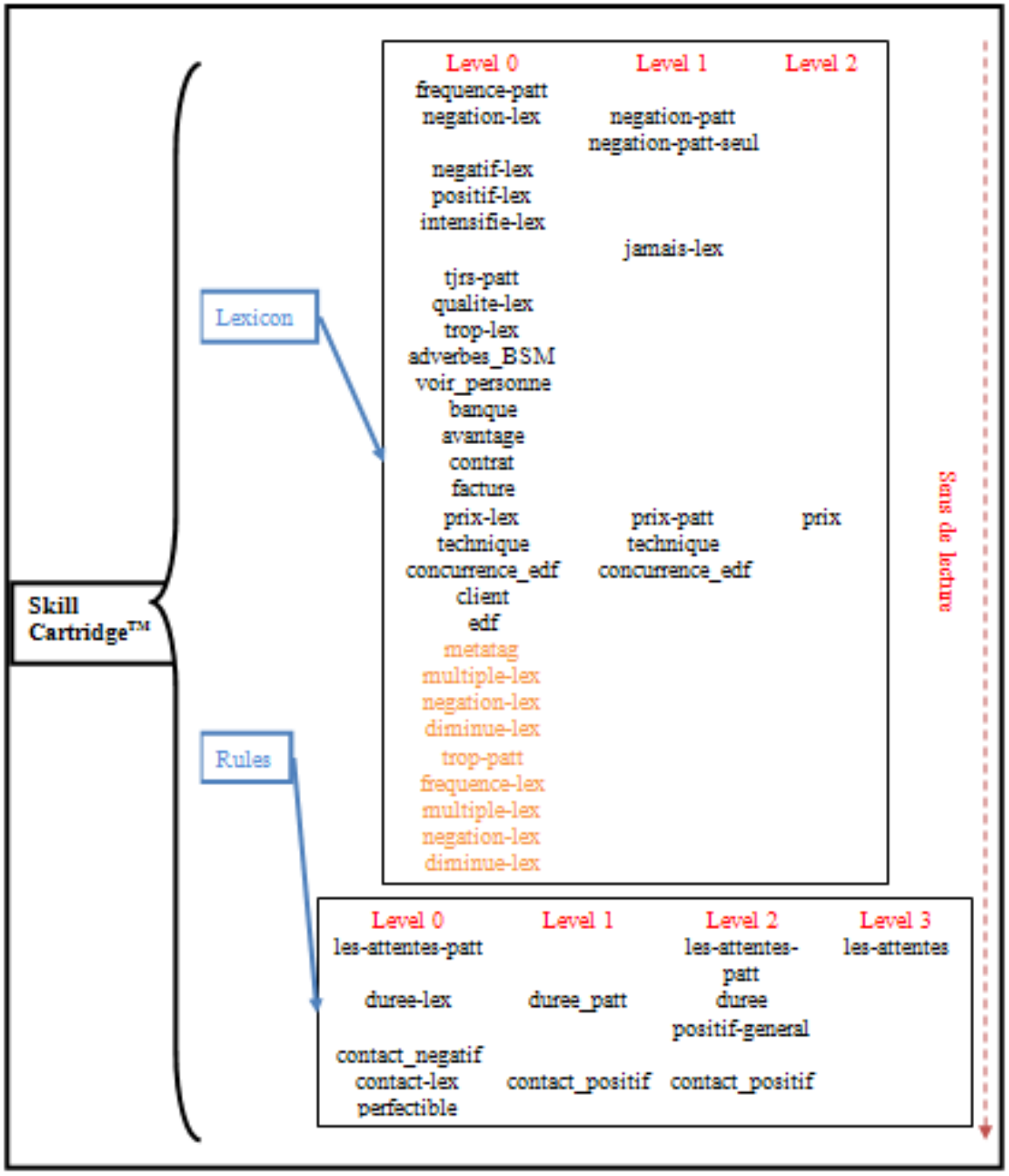
La cartouche EDF, que nous présentons dans la partie suivante, est la cartouche Baromètre Satisfaction Marché (BSM).
Elle a été créée dans le but **d'améliorer le traitement des réponses aux questions ouvertes de satisfaction** (classement des réponses en catégories prédéfinies).

Nous allons vous présenter, dans un premier temps, les différents concepts présents dans la cartouche BSM et dans un second temps, l'architecture générale de la cartouche (orange = macro).

Tableau 4 - Description des différents concepts présents dans la cartouche BSM

Concepts	Descriptions	Exemples
Durée	Permet de modéliser la durée. Ce concept contient 3 sous-concepts (courte, moyenne, longue)	Plusieurs mois, trop long, attendre deux heures, minute, seconde
Positif général	Permet de modéliser la satisfaction des clients	Je suis satisfait, c'est satisfaisant, pas de problème
Les attentes	Permet de modéliser les expressions qui expriment les attente des clients	EDF devrait, je souhaite, je désire, EDF pourrait
Adverbes	Récupérer des adverbes spécifiques	Commercialement, techniquement, pour l'instant
Concurrence EDF	Permet de récupérer les concurrents	GDF, fournisseur, powéo, Suez
Contrat	Permet de modéliser les expressions liées aux contrats	Abonnement, contrat, heure pleine
Contrat_négatif	Permet de modéliser les expressions traduisant l'insatisfaction sur les contrats	Mauvais contact, personne ne m'a contacté
Contrat_positif	Permet de modéliser les expressions traduisant la satisfaction sur les contrats	Contact, relation, très bon contact, relationnel
Facture	Permet de modéliser les expressions liées aux factures	Duplicata, estimation, frais, sur-estimation
Prix	Permet de modéliser les expressions liées aux prix	Abusif, cher, ça coûte, trop cher
Technique	Permet de modéliser les expressions relevant du domaine technique	Ampérage, branchement, baisse, coupure
EDF	Permet de modéliser les expressions liées à EDF	Agent, interlocuteur, edf, edf pro

Figure 11 - Architecture de la cartouche BSM



Cette cartouche est organisée en deux composants dont l'ordre de lecture et de traitement des fichiers dans Luxid est le suivant :

- Lecture des fichiers « Lexicon » et application des règles de niveau 0 à 2,
- Lecture des fichiers « Rules » et application des règles de niveau 0 à 3.

1. Lexicon

Les fichiers « Lexicon » contiennent des lexiques qui sont placés soit dans des macros (en orange) soit à l'intérieur de concepts. Les éléments, dont nous sommes sûrs de l'étiquette, sont placés dans des concepts et ceux dont nous sommes moins sûrs, sont placés dans des macros. Les macros permettent de déclarer les éléments sans pour autant les remplacer immédiatement par une étiquette.

2. Rules

Les fichiers « Rules » comportent de nombreuses règles qui font appel aux lexiques contenus dans les composants « Lexicon ». De nombreux niveaux sont présents dans cette cartouche, ce qui permet d'éviter les conflits dans la détection de concepts comme nous avons pu le voir précédemment (voir p 52).

Partie 4

Présentation et Analyse du corpus

CallSurf

Sommaire

Chapitre 1 - Qu'est ce que CallSurf

Le **corpus Callsurf** a été développé durant le projet Infom@gic (voir p 23). Ce corpus comporte **89 fichiers d'enregistrements de conversations téléphoniques entre agents EDF et des clients professionnels** qui proviennent des centres d'appel d'Aix en Provence. Environ 10 heures de conversations ont été **retranscrites manuellement et automatiquement** dans le but d'évaluer le système de reconnaissance de la parole. La transcription automatique consiste en une conversion automatique de la parole en texte effectuée par Vecsys Company fondée sur LIMSI-CNRS et Vecsys Recherche (Martine Garnier-Rizet, 2008). **Le taux d'erreur de mots**[25] **est de 29% pour ce corpus** (Vecsys Research travaille à l'amélioration de ces performances). La transcription manuelle est réalisée par Vecsys et contient 10770 tours de parole, sachant qu'un tour de parole correspond à une portion de parole associée à un seul locuteur (agent ou client).

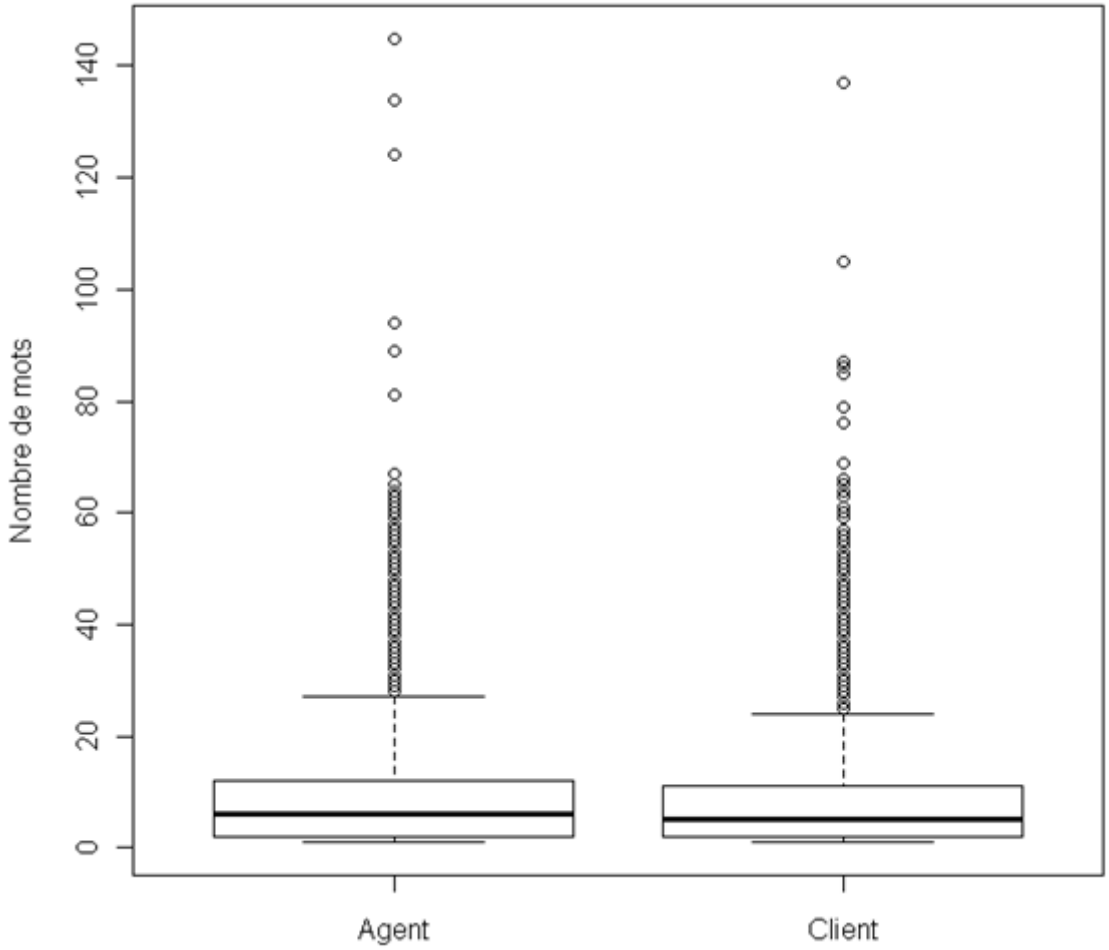
Tableau 5 - Transcription Manuelle/Automatique par tours de parole

	Speaker	timeStart	timeEnd	Manuelle	Auto
1er tour de parole	spk1	1.419	3.048	XXXX XXXX EDF Pro bonsoi	XXXX EDF pro bonsoir
2e tour de parole	spk2	3.229	4.457	oui , bonjour Madame ,	oui bonjour madame
3e tour de parole	spk2	4.812	10.443	je reçois ce matin une lettre d	que je reçois ce matin une lettre
4e tour de parole	spk1	10.443	10.986	hm hm .	
5e tour de parole	spk2	11.439	16.932	euh toutes les factures sont im	le toutes les factures sont imméc
6e tour de parole	spk1	16.932	19.727	euh oui . alors je vais prendre	oui alors je vais prendre votre ré
	spk2	19.727	22.666	oui euh 0 4 521 ,	oui zéro quatre cinq cent vingt et

Le nombre de mots par tours de parole est pratiquement équivalent entre l'agent (8,9) et le client (8,4). La boite à moustache[26], ci-dessous, montre la répartition du nombre de mots par tour de parole pour l'agent et pour le client.

[25] Unité de mesure servant à mesurer la performance d'un système de reconnaissance de la parole
[26] moyen rapide de figurer le profil essentiel d'une série statistique quantitative (Wikipedia)

Figure 12 - Nombre de mots par type de locuteur



Cette figure nous confirme que le nombre de mots par tours de parole pour les agents comme pour les clients est similaire.

Tableau 6 - Répartition du nombre de mots par type de locuteur

	0%	25%	50%	75%	95%	100%
Nombre de mots par agent	1	2	6	12	27	145
Nombre de mots par client	1	2	5	11	24.5	137

Nous constatons que 50% des conversations contiennent 5 à 12 mots. Il nous est possible d'affirmer que nous sommes en présence de statistiques atypiques lorsque le nombre de mots est supérieur à 27 pour les agents et à 24,5 pour les clients.

Tout au long de notre projet et au cours différentes analyses du corpus qui seront effectuées, nous utiliserons comme transcription de référence la transcription manuelle.
Après étude plus approfondie du corpus, nous avons remarqué que certains tours de parole n'avaient pas d'équivalences en automatique ou en manuelle. Effectivement 697 tours de parole provenant de la transcription manuelle n'ont pas d'équivalence en automatique et 551 tours de parole provenant de la transcription automatique n'ont pas d'équivalence en manuelle.

Ces différences sont dues à deux raisons :

- Aux **times codes,** plus précis dans les transcriptions manuelles que dans les transcriptions automatiques (voir p 67).

  Manuelle: <Turn speaker="spk1" startTime="1.419" endTime="3.048">
  Automatique : <Word stime="1.36" dur="0.30" conf="0.641">

- Aux **hésitations** (voir p 37) qui ne sont pas retranscrites en automatique et n'ont, par conséquence, pas d'équivalences.

  Manuelle : « hm hm »
  Automatique : « »

Chapitre 2 – Disfluences et erreurs de reconnaissance du corpus CallSurf

Le **corpus CallSurf contient un très grand nombre de disfluences et d'hésitations**. En effet lorsque les gens s'expriment à l'oral, le discours est construit au fur et à mesure de la conversation et de nombreux réajustements sont effectués. Ces phénomènes sont spécifiques à l'oral comme nous avons pu le voir précédemment.

Les hésitations et les disfluences, présentes en grand nombre dans le corpus, **peuvent engendrer des erreurs de reconnaissance. Celles-ci provoquent du bruit et des silences** au niveau de la détection de concepts.

- **Le bruit signifie une détection abusive de concepts due aux erreurs de reconnaissance. (Indicateur de mesure : Précision)**
- **Le silence est l'ensemble des concepts qui auraient dû être détectés et qui ne le sont pas. (Indicateur de mesure : Rappel)**

Regardons les exemples suivants qui illustrent ces erreurs de détection de concepts.

Tableau 7 - Erreurs de détection de concepts

	Manuelle	Automatique	Concept détecté	Bruit / Silence
Hésitation	leur euh	l'heure	Durée « heure »	Bruit
Hésitation	han donc, avec le relevé du compteur	ans donc avec le relevé du compteur	Durée « ans »	Bruit
Hésitation / Répétition	le compteur euh euh à poser	le compteur le le à poser	Ø	Silence

Prenons l'exemple du tableau : « le compteur euh euh à poser ». L'hésitation, retranscrite en automatique comme une répétition, empêche la cartouche de détecter l'expression « compteur à poser ». Par conséquent un concept sera manquant et nous aurons du silence dans la détection de concepts.

Afin de comprendre le comportement des disfluences et plus précisément des hésitations, nous avons étudié leurs modes d'apparition pour pouvoir, par la suite, améliorer nos scores sur la détection de concepts. **Le corpus contient 4285 hésitations.**
Le tableau, ci-dessous, répertorie le nombre d'hésitations dans la transcription manuelle. La transcription automatique ne contient malheureusement pas d'hésitations car elles ont été partiellement supprimées par Vecsys research qui nous fournit la transcription automatique.

Tableau 8 - Hésitation dans la transcription manuelle selon les tours de parole

	Agent	Client	Entre Agents	Total
Euh	973	1188	205	2366
Hein	327	188	41	556
Ah	163	159	49	371
Hm	167	133	39	339
Ouais	138	119	73	330
Ben	124	167	32	323
TOTAL	1892	1954	439	4285

Nous pouvons dès lors constater que les hésitations sont présentes en grand nombre, que ce soit sur les tours de paroles des agents comme sur ceux des clients.

L'hésitation la plus fréquente du corpus est le « euh » qui intervient principalement entre les phrases. Ces **hésitations** sont utilisées, la plupart du temps, comme **des mots de liaisons et n'empêchent pas la détection de concepts.**

Les disfluences ne sont pas les seules caractéristiques du discours spontané, le **corpus contient** également

- Des « **back channels** » qui sont généralement des interjections (hum hum, OK, oui, d'accord…) ou des répétitions du second locuteur pour acquiescer aux propos du locuteur principal.
- Des **chevauchements** qui sont utilisés pour marquer la prise de parole anticipée. Le dernier mot du premier locuteur est superposé au premier mot du second locuteur.

Les disfluences et les hésitations ne sont pas les seules responsables des erreurs de détection de concepts dans le corpus. Les erreurs de reconnaissance, dues ou non aux disfluences, provoquent également des erreurs de détection de concepts, ce qui entraîne du bruit dans la sortie de la cartouche.

Lors de notre étude du corpus, nous avons répertorié quatre sortes d'erreurs de reconnaissance récurrentes dans le corpus CallSurf.(voir p 35 à 37) Les deux premières sont de simples erreurs de reconnaissance et les deux suivantes sont des erreurs de reconnaissances dues aux disfluences.

Tableau 9 - Les erreurs de reconnaissance dans le corpus CallSurf

Erreurs de reconnaissance	Exemples
Nom propre *(noms de famille, de ville…)*	Manuelle : « Madame Maubert » Automatique : « Madame ampère »
Homophonie	« mois / moi » « en / an »
Hésitation	Manuelle : « c'est long hein » Automatique : « c'est loin »
Emotion et/ou bruit environnant	Manuelle : « ils vont résilier virtuellement le contrat » Automatique : « ils ont résilié sexuellement le contrat »

Remarque :

Dans le premier exemple, le concept « technique » est détecté de façon incorrecte à l'aide du mot « ampère ». Comme on peut le constater dans cet exemple, l'erreur est localisée mais dans la plupart des cas, les erreurs comme celles-ci, font un effet boule de neige et l'erreur de reconnaissance se répand à la phrase. C'est ce qu'illustre l'exemple suivant tiré du corpus CallSurf :

Manuelle : « et de toute façon c'est le CRAM, c'est la le le Crédit Agricole d'Ile de France »
Automatique : « toute façon c'est quand c'est prévu donc de francs dix »

Chapitre 3 – Nettoyage et Formatage du corpus

1. Transformation des données en tableau

Le corpus CallSurf est un corpus qui n'est pas directement exploitable. En effet, il contient deux sortes de fichiers différents (XML et TRS) qui ne sont pas constitués de la même façon. Les transcriptions manuelles et automatiques ne contiennent pas les mêmes balises et c'est pour cette raison que nous ne pouvons pas les opposer et les analyser sans les mettre sous un même format.

Notre but premier est donc de convertir nos fichiers en un tableau à l'aide de scripts Perl. Ce tableau permet d'améliorer la lecture et l'analyse des fichiers.
Nos transcriptions manuelles, qui sont des fichiers XML, comportent de nombreuses données mais seules quelques unes nous intéressent.

Image 6 - Transcription Manuelle

```
<Turn speaker="spk1" startTime="1.419" endTime="3.048">
<Sync time="1.419"/>
<Event desc="nom.personne.clientagent" type="entities" extent="begin"/>
 Sylvie Altérini
<Event desc="nom.personne.clientagent" type="entities" extent="end"/>
 EDF Pro bonsoir .
</Turn>
```

Comme vous pouvez le remarquer, la transcription manuelle est divisée en tours de parole qui sont représentés par les balises <Turn></Turn>. Cette balise comporte plusieurs attributs dont un qui nous indique la personne qui parle. Le speaker 1 (spk1) correspond à l'agent EDF, le deuxième speaker (spk2) correspond au client et le reste (spk[3 à 7]) correspond aux conversations entre agents EDF. Le deuxième attribut important est le temps. Le startTime indique, de façon précise, le commencement de la phrase et le endTime, la fin de la phrase.

L'architecture du fichier qui comporte la transcription automatique est très proche de celle de la transcription manuelle. Mais il existe quelques différences.

Image 7 - Transcription Automatique

```
<SpeechSegment ch="1" sconf="1.00" stime="1.18" etime="2.98" spkid="3" lang="fre" lconf="1.00" trs="1">
<Word stime="1.36" dur="0.30" conf="0.641"> Sylvia </Word>
<Word stime="1.66" dur="0.38" conf="0.951"> Altérini </Word>
<Word stime="2.04" dur="0.25" conf="0.951"> EDF </Word>
<Word stime="2.29" dur="0.21" conf="0.950"> pro </Word>
<Word stime="2.50" dur="0.44" conf="0.951"> bonsoir </Word>
</SpeechSegment>
```

La balise <SpeechSegment></SpeechSegment>, qui est équivalente à la balise <Turn> en manuelle, nous informe, entre autres, sur les tours de parole des locuteurs et le temps. Contrairement à la transcription manuelle, le découpage de la conversation ne se fait pas en phrases mais en mots. Chaque mot contient un attribut « stime » qui indique à quel moment le mot est dit. Comme nous pouvons le constater, la précision de ce chiffre est inférieure à celle de la transcription manuelle car le temps est en centième alors qu'il est en millième pour la transcription manuelle.
Lors de la création du programme permettant d'aligner les transcriptions, nous nous appuyons sur la transcription manuelle, car elle est notre référence mais aussi parce qu'elle est plus précise au niveau des times codes. Il a donc fallu récupérer les times codes de la transcription manuelle pour les comparer avec ceux de la transcription automatique. Si les times codes se situent bien entre ceux de la transcription automatique, les mots sont récupérés et imprimés dans la sortie.

En sortie de ce programme nous avons le tableau suivant :

Tableau 10 - Structuration des données EDF

Speaker	timeStart	timeEnd	Manuelle	Auto
spk1	1.419	3.048	XXX EDF Pro bonsoir .	XXX EDF pro bonsoir
spk2	3.229	4.457	oui , bonjour Madame ,	oui bonjour madame
spk2	4.812	10.443	je reçois ce matin une lettre du 22 août	que je reçois ce matin une lettre du vingt-deux août
spk1	10.443	10.986	&hm &hm .	
spk2	11.439	16.932	euh toutes les factures sont immédiate	le toutes les factures sont immédiatement retournée
spk1	16.932	19.727	euh oui . alors je vais prendre votre ré	oui alors je vais prendre votre référence monsieur s'
spk2	19.727	22.666	oui euh 0 4 521 ,	oui zéro quatre cinq cent vingt et un
spk1	22.666	25.111	alors 0 4 521 .	alors zéro quatre cinq cent vingt et un
spk2	25.413	27.103	673 ,	cent soixante treize
spk1	27.103	28.522	673 .	six cent soixante-treize
spk2	28.522	29.669	404 ,	quatre cent quatre
spk1	29.669	30.786	404 .	quatre cent quatre
spk2	30.786	32.295	355 ,	trois cent cinquante-cinq
spk1	32.295	32.959	oui .	et
spk2	32.959	34.317	263 .	deux cent soixante-trois
spk1	35.192	37.408	d'accord . je vous demande un instant j	d'accord je vous demande un instant je vais accéder
spk1	49.869	51.589	alors , il s'agit du cabinet XXX ?	alors il s' agit du Cabinet XXX
spk2	51.589	52.012	oui .	oui
spk1	52.012	55.785	euh 2 rue des fourmis à *Chennegy	deux rue des Fourmis à Chennegy juste Chennegy c'

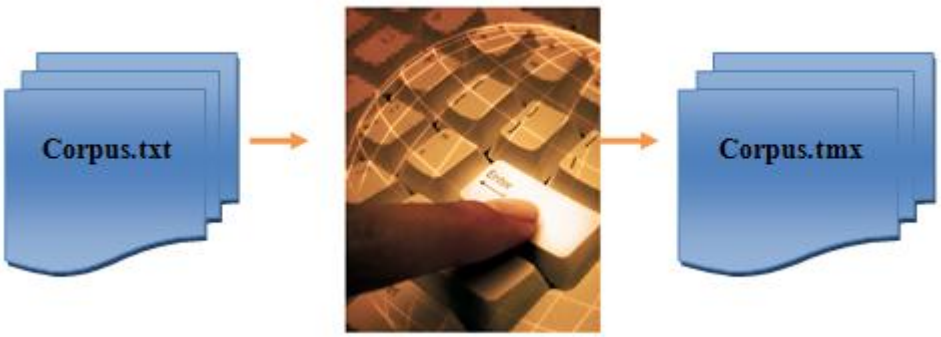
Ce tableau comporte cinq colonnes :

- Colonne 1 correspond au locuteur (spk1, spk2…),
- Colonne 2 correspond au timeStart,
- Colonne 3 correspond au timeEnd,
- Colonne 4 correspond à la transcription manuelle
- Colonne 5 correspond à la transcription automatique.

Il est à présent possible d'analyser les conversations grâce à ce parallèle.

2. Transformation des fichiers nettoyés en TMX

Une fois l'étape précédente effectuée, nous devons transformer en format TMX notre corpus afin qu'il soit compatible avec la Skill Cartridge™. A cet effet, un script nous est fourni par TEMIS, pour de nous permettre de transformer nos fichiers.



http://www.la-souris-apprivoisee.fr/images/clav

Mais pour que cette étape soit réalisable, il faut, en fichier d'entrée, un corpus en format TXT qui contienne un séparateur de champs telle que des tabulations (voir p 67) et qui dispose d'au minimum 2 colonnes (une pour l'identifiant et une pour les données textuelles).
Nous avons donc utilisé les fichiers TXT qui comportent les cinq colonnes (voir Tableau 10) en entrée du programme de TEMIS et nous obtenons en sortie de celui-ci un fichier structuré, comme suit, en TMX compatible Luxid :

Image 8 - Fichier TMX

```
<doc id="2">
 <features>
  <ft>/Metadata/Source/00008372AM</ft>
 </features>
 <features zone="Speaker">
  <ft>/Metadata/Speaker/spk2</ft>
 </features>
 <features zone="timeStart">
  <ft>/Metadata/timeStart/5.01</ft>
 </features>
 <features zone="timeEnd">
  <ft>/Metadata/timeEnd/10.752</ft>
 </features>
 <text zone="Auto">
  <data> vous bonjour madame je je vous appelle parce que j'ai besoin d'avoir un rendez-vous pour une mise en service </data>
 </text>
 <text zone="Manuelle">
  <data>oui , bonjour Madame . euh , je vous rappelle parce que j'ai besoin d'avoir un rendez-vous pour une mise en service .</data>
 </text>
</doc>
```

Partie 5

Adaptation de la cartouche BSM aux données CallSurf : la cartouche CallSurf

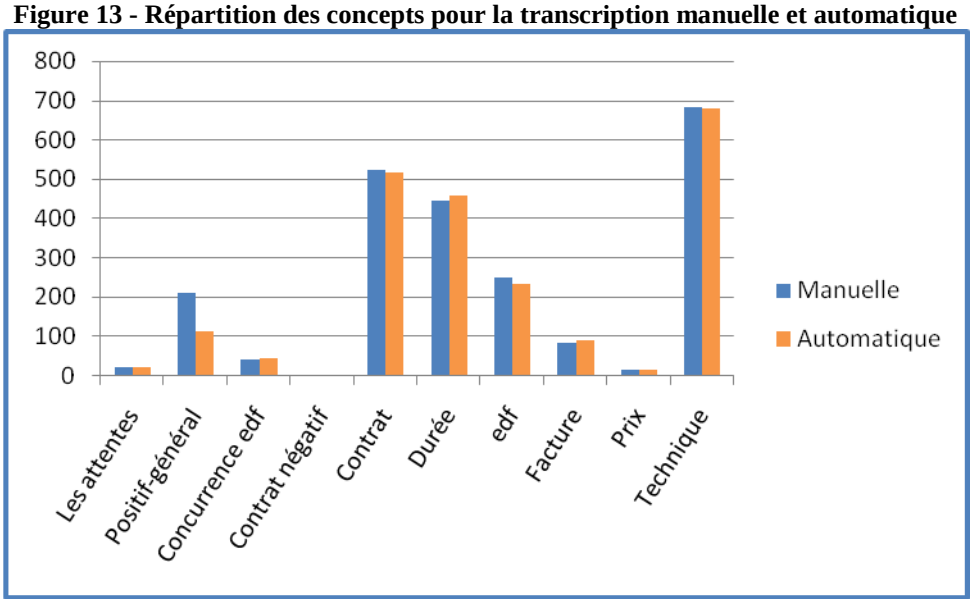
Sommaire

Chapitre 1 - Etude des sorties de la cartouche initiale : BSM

Nous analysons les résultats de la cartouche BSM (voir p 56) sur le corpus CallSurf. Rappelons qu'une **cartouche doit être spécifique à un corpus** or cette cartouche n'a pas été conçue dans le but d'extraire des concepts sur ce corpus. Nous allons étudier les résultats de la cartouche sur les transcriptions manuelles et automatiques qui contiennent **10770 tours de parole** (voir p 63). **16,3% des tours de parole contiennent au moins un concept détecté pour la transcription manuelle contre 15,7% sur la transcription automatique avec 2288 concepts détectés sur la transcription manuelle et 2163 concepts pour la transcription automatique.** Le nombre de concepts détectés entre les deux transcriptions est pratiquement équivalent. Il est intéressant de remarquer que **1660 concepts sont identiques entre les transcriptions manuelles et automatiques.**

Afin de confirmer cette observation, comparons la répartition des différents concepts qui sont détectés pour la transcription manuelle et automatique.

Figure 13 - Répartition des concepts pour la transcription manuelle et automatique



Nous constatons que le comportement global de la détection de concepts est similaire entre la transcription manuelle et automatique.
La différence se situe principalement au niveau du concept « Positif général » (213 concepts pour la transcription manuelle et 111 concepts pour la transcription automatique).
En effet, ce concept est un concept d'opinion modélisé par des règles plus complexes que celles qui servent à détecter les thématiques. De plus, la modélisation de ce concept prend en compte les signes de ponctuation qui ne sont, pour l'instant, pas présents dans la transcription automatique. D'autres concepts, tels que «attentes», «contact positif» et «contact négatif » ne sont pas bien représentés dans ce corpus.

Après cette première analyse générale, nous analysons maintenant les tours de parole dans lesquels la détection de concepts est différente entre la transcription manuelle et la transcription automatique.

Nous considérons ici que la détection de concepts obtenue sur la **transcription manuelle est la référence**. Il est dès lors possible de calculer le rappel et la précision. En effet, nous avons pu constater (voir p 64), que le nombre moyen de mots sur un tour de parole variait de 2 à 12 mots. Ces résultats permettent d'affirmer que lorsque deux concepts sont détectés à l'identique sur la transcription manuelle et automatique, il y a de grandes chances pour que ce soit les mêmes expressions qui soient à l'origine de la détection de concepts

Calculons, dès à présent, le rappel et la précision.

- **Le rappel est défini comme le nombre de tours de parole contenant les mêmes concepts détectés pour les transcriptions manuelles et automatiques, divisé par le nombre de tours de parole dans lequel, au moins, un concept de la transcription manuelle a été détecté.**

$$\text{Rappel} = \frac{1317}{(1484+276)} = 74{,}8\%$$

- **La précision est définie comme le nombre de tours de parole contenant les mêmes concepts détectés pour les transcriptions manuelles et automatiques, divisé par le nombre de tours de parole dans lequel, au moins, un concept a été détecté sur la transcription automatique.**

$$\text{Précision} = \frac{1317}{(1484+210)} = 77{,}7\%$$

Nous obtenons ainsi un rappel de 74,8% et une précision à 77,7%.
Ce qui signifie **qu'il y a peu de bruits et de silences dans nos détections de concepts. Ces résultats confirment que le comportement global de la cartouche sur la transcription automatique est proche de celui obtenue sur la transcription manuelle.**

Toutefois, lorsque l'on étudie les concepts, des différences entre les deux transcriptions peuvent être mises en évidence. Le Tableau 11 montre le nombre de concepts qui sont identiques/différents entre la transcription manuelle et automatique. Une étude approfondie de ces différences est effectuée par ailleurs (voir p 75).

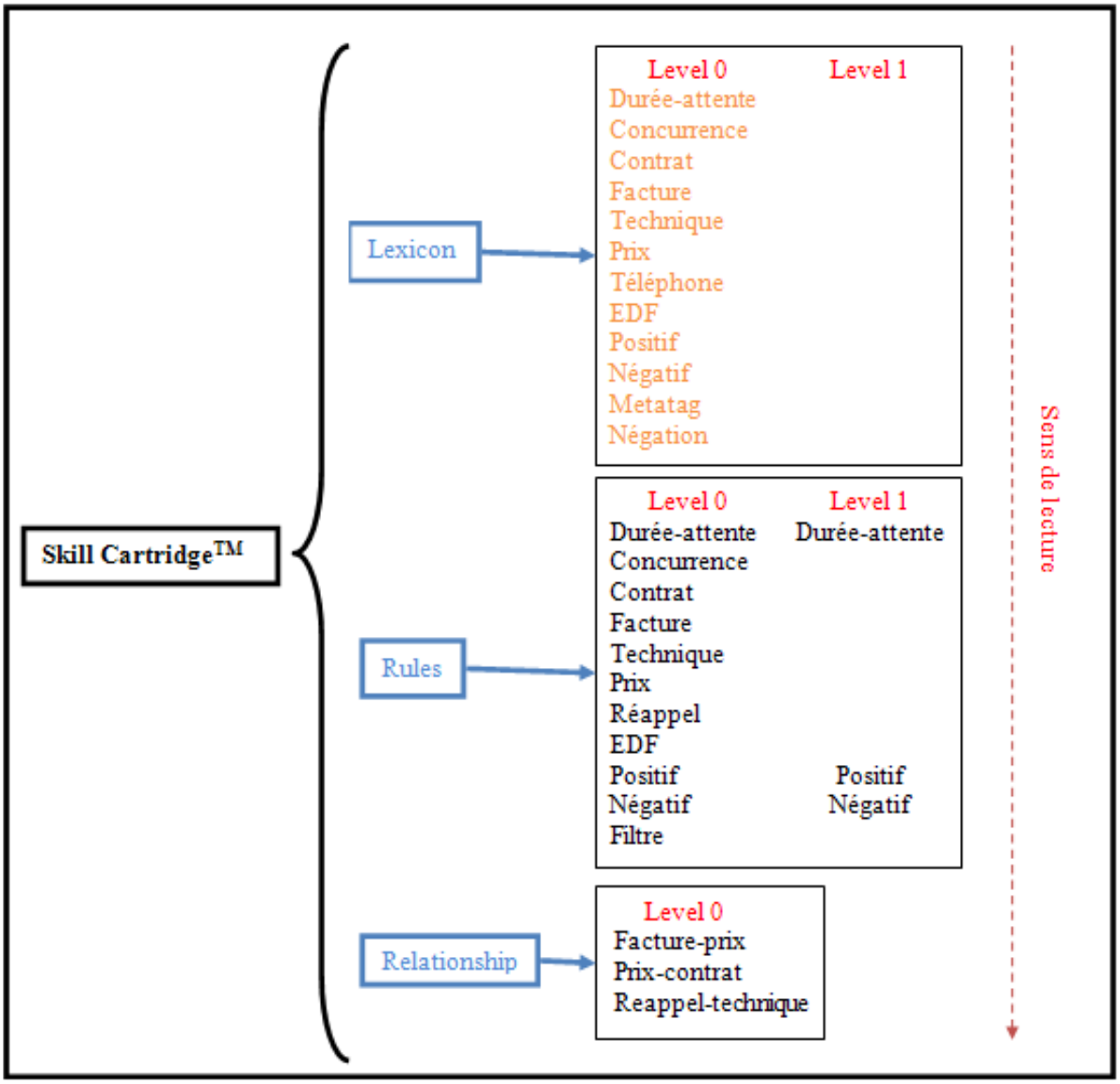
Tableau 11 – Nombre de concepts détectés dans la transcription manuelle VS automatique

Nombre de concepts qui sont détectés en manuelle et pas en automatique	Nombre de concepts qui sont détectés en automatique et pas en manuelle
277	207

Chapitre 2 – Organisation générale de la nouvelle cartouche : CallSurf

Lors de la conception de cette nouvelle cartouche, nous avons décidé d'agir différemment de la cartouche BSM. Il faut savoir que le fait de mettre un lexique dans un concept bloque ce lexique (voir p 52).

Figure 14 - Architecture de la cartouche CallSurf



Cette nouvelle cartouche est organisée de façon simple. Comme nous pouvons le constater sur la Figure 14, la cartouche est organisée en trois composants dont l'ordre de lecture et de traitement des fichiers dans Luxid est le suivant :

- Lecture des fichiers « Lexicon » et application des règles de niveau 0,
- Lecture des fichiers « Rules » et application des règles de niveau 0 suivi du niveau 1,
- Lecture des derniers fichiers « Relationship » et application des règles de niveau 0.

Afin d'expliquer rapidement l'architecture de la cartouche CallSurf, voyons l'exemple suivant sur le concept « durée-attente ».

1. Lexicon

Les fichiers « Lexicon » contiennent des lexiques qui sont placés dans des macros (en orange). Les macros permettent de déclarer les éléments sans pour autant les remplacer immédiatement par une étiquette.
Dans le cas de notre concept « durée-attente », la macro contient un lexique de durée comme suit :

Image 9 - Macro Durée-attente

```
<macro name="duree-lex" display="no">
    <e>
        minute
      | :mn
      | :mns
      | :secondes?
      | :s
      | heure
      | :h
      | semaine
      | mois
      | journée
    </e>
</macro>
```

2. Rules

Les fichiers « Rules » comportent de nombreuses règles qui font appel aux lexiques contenus dans les composants « Lexicon ». Dans cette nouvelle cartouche seuls trois fichiers comportent des règles avec deux niveaux. Ce sont les fichiers positif, négatif et durée-attente. Nous y avons instauré des niveaux afin d'éviter les conflits dans la détection de concepts comme nous avons pu le voir précédemment (voir p 52).

- *Niveau 0*

Dans ce premier niveau, nous détectons différentes expressions en contexte telles que :

- « j'ai attendu 2 heures »
- « un instant »
- « restez en ligne »

Dans ces mêmes expressions, nous faisons appel, lorsque cela est nécessaire, au lexique contenu dans la macro du fichier durée-attente. Cette macro, portant le nom « duree-lex », est appelée de la façon suivante :

```
<concept name="duree-attente-rules" level="0" display="never">
    <e>
    (déjà|depuis|attendre|pendant) / ([0-9]+|(#NUM)+) / ~duree-lex / (et / demi)?
    | (un | quelque) / instant
    | (rester) / #PREP / ligne
```

➢ *Niveau 1*

Ce second niveau comporte seulement deux règles. La première fait appel au niveau précédent avec un contexte plus large. La seconde permet de détecter, par exemple, l'expression « demander d'attendre un peu ». Cette règle ne peut être placée au niveau précédent car nous ne voulons pas détecter l'expression isolée « un petit peu » ou « un peu ».

```
<concept name="duree-attente" level="1" display="always">
        <e>
        (demander / (~DET|~PREP|#MISC)*)? /(attendre|patienter)? / ~~duree-attente-rules
        |(demander / (~DET|~PREP|#MISC)*)? /(attendre|patienter) / un / (petit)? / peu
```

Nous procédons de la même façon pour les fichiers positif et négatif.

3. Relationship

En plus des deux composants que nous connaissons : « Lexicon » et « Rules », nous avons également mis en place des « relationships ». Le « relationship » est un composant déclaré dans le *.scp (voir p 51). Le *.scp permet de déterminer l'ordre de lecture des différents fichiers.

```
<apply order="Lexicon,Rules,Relationship"/>
```

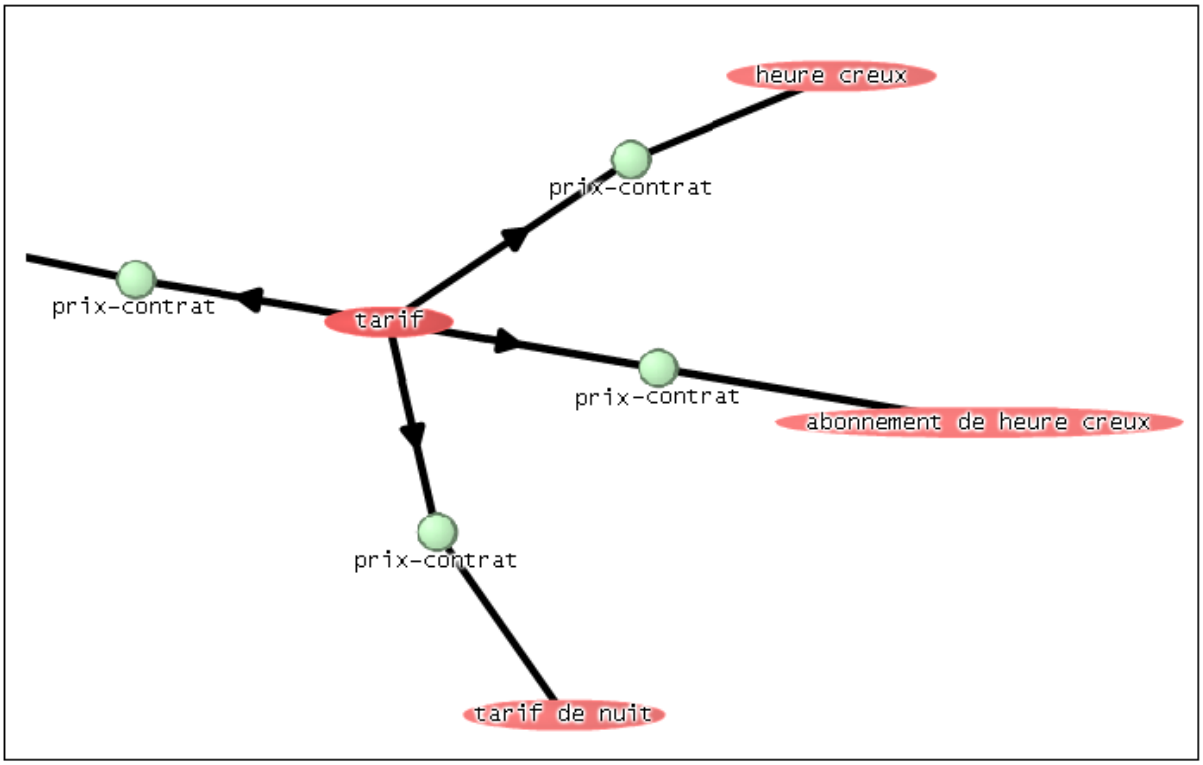
Ce composant comporte des règles qui permettent de mettre en relation différents concepts. Ces règles doivent comporter au minimum un « subject » suivi d'un « object ».

Image 10 - Le composant RelationShip

```
<concept name="Relationship" display="always">
    <concept name="prix-contrat">
        <e>
        <!-- montant de l' abonnement  -->
          (/Role/Subject:~~prix) / (~DET|~PREP|#MISC)* / (/Role/Object:~~contrat)
        </e>
    </concept>
</concept>
```

Dans l'exemple que vous pouvez voir ci-dessus, nous avons créé une relation qui comporte comme sujet le concept « prix » suivi soit d'un déterminant, d'une préposition ou du tag Xelda « misc », tous facultatifs. Le tout suivi du concept « contrat » qui a un rôle d'objet.
Cette relation est créée dans le but d'être visible dans la sortie du Luxid® Knowledge Browser sous la forme d'un arbre. Grâce à cet arbre, il est désormais possible de voir si le concept « prix » est souvent suivi du concept « contrat ».

Image 11 - Sortie dans le LIA des RelationShips



Comme vous pouvez le constater dans l'exemple ci-dessus, la relation « prix-contrat » est matérialisée par un rond vert. Le sujet, qui est le concept « prix », pointe sur le concept objet « contrat ».

Chapitre 3 - Adaptation de la cartouche au corpus CallSurf

Comme vous avez pu le voir précédemment (voir p 35), le corpus CallSurf contient de nombreuses disfluences que l'on doit prendre en compte pour améliorer la cartouche. Pour cela, l'étude du corpus a été indispensable afin de mieux comprendre le fonctionnement des disfluences et leurs modes d'apparition.

Pour adapter la cartouche à un corpus oral, nous avons apporté cinq types d'amélioration :

- Assouplissement des règles linguistiques
- Prise en compte des erreurs de reconnaissance
- Prise en compte des tags XELDA
- Prise en compte des différentes orthographes et combinaisons d'un mot
- Enrichissement des concepts métiers
- Création d'un filtre

I. Assouplissement des règles linguistiques

Nous nous sommes tout d'abord attardés sur les répétitions et les hésitations qui sont courantes dans un corpus oral.
Nous avons crée la règle suivante :

(~DET|~PREP|#MISC)*

Cette règle nous laisse la possibilité d'avoir plusieurs répétitions du déterminant et de la préposition :

« relevé de du de compteur »

Le tag « MISC » est un tag « miscellaneous » qui signifie divers. Ce tag repère et étiquette les onomatopées, les hésitations et les mots de liaison. Grâce à ce tag, on laisse la possibilité à la personne d'hésiter sans que cela perturbe la détection de concepts dans la cartouche. Cette règle offre la possibilité de récupérer et de prendre en compte un grand nombre de disfluences, en voici un exemple :

« branchement de euh du de compteur électrique neuf »

II. Prise en compte des erreurs de reconnaissance

Nous avons également dû traiter le problème des erreurs de reconnaissance. Prenons trois exemples afin d'illustrer ce phénomène et de permettre de mieux comprendre la façon dont on le traite.

Le premier exemple, voir ci-dessous, est dû à une erreur de reconnaissance de l'expression « l'abonnement ». Dans la transcription automatique, celle-ci apparaît sous la forme « la bonne non ». La prononciation phonétique des deux expressions reste très proche, ce qui explique l'erreur de reconnaissance. Il est donc facile, dans la cartouche, de créer une règle linguistique qui permette de récupérer ce trigramme.

la / :bonne / non

Dans cette règle, nous avons appliqué un contexte strict. Nous imposons à ce que le déterminant soit suivi de la forme « bonne » et non du lemme et de la négation.
Le second exemple, nous présente le même cas de figure que le premier. Il est possible de trouver, dans la transcription automatique, l'adverbe « alors » à la place du terme « heure ». Dans ce cas, nous devons prendre en compte l'erreur de reconnaissance pour diminuer le silence dans la détection de concepts.

(heure|alors) / (creux|plein) / (et / (creux|plein))?

Dans les exemples suivants, nous avons cherché à récupérer le morphème[27] [poze] qui n'est pas forcément bien retranscrit que ce soit du côté de la transcription automatique comme du côté manuelle.
Mais nous avons remarqué deux phénomènes différents dans les quatre exemples que nous vous proposons.

- Le premier étant dû à une erreur de reconnaissance,
- Le second est lié à une erreur morpho-syntaxique liée à XELDA.

1. Automatique : « et non pas avoir j'ai le consuel on va demander une demande la pension compteur »
2. Manuelle : « ben non , apparemment euh j'ai le consuel , on me demande uniquement de la pose un compteur »
3. Manuelle : « puis euh à partir de là ce ce sera très rapide . à savoir qu'il faut compter à peu près 10 jours ouvrés entre la demande et la pause de de du compteur . »
4. Manuelle : « vous pose le compteur »

Afin de pouvoir récupérer le n-gramme[28] souhaité, nous devons proposer plusieurs orthographes à ce mot afin qu'il soit détecté. Cette démarche est à faire avec prudence car si la règle devient trop souple, du bruit[29] sera récupéré. Il faut donc préciser le contexte pour ne pas avoir à faire face à ce problème.
Nous avons donc proposé la règle suivante :

(pose|pause|pension) / (et / raccordement) ? /(~DET|~PREP|#MISC)* / compteur

Nous demandons à ce que l'un des trois lemmes (pose ou pause ou pension) soit suivi, facultativement, de « et raccordement » puis d'un ou plusieurs déterminants facultatifs et du lemme compteur qui est obligatoire. Dans cette règle, nous prenons en compte l'erreur de reconnaissance en proposant dans la règle de repérer le lemme « pension ».
Le contexte étant strict, nous n'aurons pas de bruit lors de la détection de concepts. Il ne sera, par conséquent, pas possible de détecter la phrase « pension d'invalidité ».

Nous expliquons le second phénomène qui se passe au niveau du tag dans la partie suivante.

III. Les tags XELDA : Information non négligeable

Continuons avec l'exemple précédent :

« vous pose le compteur »

Lors de la création de la règle linguistique, il est important de mettre dans la règle, le lemme « pose » qui est reconnu par Xelda en tant que NOM et de mettre le lemme « poser » qui lui est reconnu en tant que VERBE.

[27] Unité minimale ayant un sens
[28] Chaine de caractère composée de n symboles (Claude Shannon)
[29] Détection abusive de concepts

Pour illustrer ce problème, prenons le quatrième exemple. La forme « pose » porte une étiquette de VERBE et par conséquent si nous n'avions pas mis le lemme « poser » dans la règle, la forme « pose » conjuguée n'aurait pas été détectée.

vous+PRON_P1P2
poser+VERB_P1P2
le+DET_SG
compteur+NOUN_SG

Suite à ce problème, nous complétons la règle précédente avec le lemme du verbe « poser ».

(pose|poser|pause|pension) / (et / raccordement) ? /(~DET|~PREP|#MISC)* / compteur

Pour mieux comprendre ce problème délicat et difficile à repérer sur les erreurs faites sur les étiquettes grammaticales du corpus, voyons d'autres exemples significatifs.

Pour le concept d'insatisfaction, nous avons découvert que l'analyse de la phrase suivante, faite par XELDA, est erronée.

« j'ai un un problème »

```
Disambiguated: yes
j' [0-1]
je+PRON_P1P2

ai [2-3]
avoir+VAUX_P1P2

un [5-6]
un+DET_SG

un [8-9]
un+NUM

problème [11-18]
problème+NOUN_SG
```

Nous pouvons remarquer que le déterminant « un » n'est pas bien reconnu lors de la répétition. Par conséquent, il faut prendre en compte cette erreur d'étiquetage et modifier la règle de façon à ce qu'elle récupère malgré tout l'expression « avoir un problème ».
Nous avons donc proposé la règle suivante qui permet de récupérer un déterminant suivi, ou non, du mot « un ». Suite à cette modification, nous avons réduit le silence et le concept sera détecté.

~PRON / avoir / (~DET|#MISC)* / (:un)* / ~negatif-lex

Nous avons également remarqué que les temps composés n'étaient pas forcément bien étiquetés. En effet, nous avons fait analyser la phrase « EDF avait téléphoné hier. » par XELDA. A première vue, nous pourrions penser que nous avons un pronom suivi du verbe « téléphoner » et d'un adverbe.

```
Disambiguated: yes
EDF [0-2]
EDF+NOUN_INV

avait [4-8]
avoir+VAUX_P3SG

téléphoné [10-18]
téléphoner+PAP_SG

hier [20-23]
hier+ADV

. [24-24]
.+SENT
```

Nous constatons alors que l'étiquetage du syntagme n'est pas fait de cette façon. En effet, pour XELDA, deux verbes, dont un auxiliaire, sont présents dans cet exemple. Il faut donc prendre en compte l'auxiliaire dans la règle linguistique sinon il ne sera pas possible de détecter le concept « reappel ». Nous avons donc créé la règle suivante qui oblige la présence d'un auxiliaire avant le concept « verbeTelephonerPasse-lex ».

(~PRON|~edf-lex) / ~AUX / ~verbeTelephonerPasse-lex

Lors de la création des règles, nous avons découvert un autre comportement « déroutant » mais qui cette fois-ci n'est pas dû à XELDA mais à Luxid. En effet, Luxid ne prend pas en compte les accents. Or cela est primordial pour certains mots.

Prenons les deux exemples suivants :

« Il est sur la ligne »
« Il est sûr de lui »

Dans les deux cas, XELDA différencie, de façon correcte, l'adjectif de la préposition mais pas Luxid. Nous sommes donc obligés de forcer le tag de XELDA pour que l'on puisse récupérer la forme que l'on souhaite. Pour cela nous indiquons le tag que nous souhaitons à la suite du lemme, de la façon suivante :

sûr#ADJ_SG
sûr#ADJ_PL

Il est dès lors possible de récupérer « sûr » comme un terme positif et de diminuer de façon considérable le bruit lié à cette extraction.

IV. Prise en compte des différentes orthographes et combinaisons d'une phrase

Outre les disfluences, les erreurs de reconnaissance… il faut prendre en compte les différentes orthographes ou combinaisons possibles d'un mot ou d'une phrase. Pour cela, il faut les lister, ce qui est indispensable pour réduire considérablement le silence dans la détection de concepts.
Prenons les exemples suivants tirés des concepts « technique » et « concurrence » qui illustrent les différentes possibilités d'orthographes sur une expression.

(microcoupure| micro-coupure|(micro / coupure))
(sur / tension)|sur-tension|surtension
:provalis
:provalys

Dans cet exemple, nous pouvons constater que le terme « microcoupure » sera récupéré quelle que soit son orthographe, il en va de même avec les autres exemples.
Il en est de même pour les combinaisons de phrases. Les clients, ne connaissant pas forcément les termes techniques, utilisent différentes expressions pour exprimer la même chose. Celles-ci doivent donc être répertoriées. Les règles suivantes, tirées du concept « technique », nous révèlent un échantillon des différentes possibilités d'expressions que l'on veut détecter.

(numéro|référence)? /(#COORD)? /(~DET|~PREP|#MISC)*/ (numéro|référence)?
/(~DET|~PREP|#MISC)* / (devis)

Grâce à cette règle, il est possible de détecter les expressions suivantes :

- numéro de devis
- référence ou un numéro de devis
- numéro ou une référence de devis
- devis

Il en est de même pour la règle suivante :

(relever|relève) / (~DET|~PREP|#MISC)* / (numéro)?/(~DET|~PREP|#MISC)* /(compteur|(matricule /compteur)|consommation)

Grâce à cette règle, il est possible de détecter un nombre important d'expressions comme celles-ci :

- relevé de compteur
- relevé du numéro de compteur
- relevé du numéro de matricule compteur
- relèvera le consommation

V. Enrichissement des concepts métiers

Comme nous avons pu le dire précédemment, nous sommes partis d'une cartouche existante qui n'était pas adaptée au corpus. Nous avons donc dû, non seulement modifier et compléter les règles des patterns linguistiques, mais aussi compléter les lexiques afin de détecter le plus de concepts possibles. Pour cela, nous avons d'abord récupéré, trié et complété une liste de mots positifs et une de mots négatifs d'environ 300 mots chacune. La liste de lexique positif a été triée en 4 listes différentes (adjectif, nom, verbe, avec-contexte). Les mots ne s'utilisant pas forcément toujours de la même façon, chaque liste a ses règles propres.
Prenons l'exemple suivant tiré du corpus CallSurf :

« OK donc j'ai rentré la mensualisation Monsieur donc à hauteur de deux cent vingt-cinq euros par mois »

Il n'est pas intéressant de récupérer les termes « OK, bon, bien » car ils signifient « oui » dans la plupart des cas. C'est une affirmation que l'on peut classer dans la fonction phatique. Elle « peut donner lieu à un échange profus de formules ritualisées, voire à des dialogues entiers dont l'unique objet est de prolonger la conversation » (Jakobson, 1963). Par conséquent, cela ne signifie pas obligatoirement que le client est satisfait. Alors que cet exemple « voilà , c'est OK. » signifie que le client est d'accord avec l'agent. Il est donc en quelque sorte satisfait.
Dans le tableau suivant, nous vous présentons les concepts détectés par la cartouche CallSurf avec leurs exemples.

Tableau 12 - Description des différents concepts de la cartouche

Concepts	Descriptions	Exemples
Durée-attente	Permet de modéliser la durée, l'attente du client au téléphone.	Attendez une seconde, restez en ligne, ça fait 15 jours, patientez quelques temps
Concurrence	Permet de récupérer les concurrents d'EDF.	GDF, fournisseur, powéo, Suez, choisir éventuellement un autre fournisseur
Contrat	Permet de modéliser les expressions liées aux différents contrats et offres que proposent EDF.	Abonnement, contrat, demandé la la résiliation, abonnement en heures creuses
Facture	Permet de modéliser les expressions liées aux factures, aux différents moyens de paiement.	Duplicata, paiement des factures, sur-estimation, facture de la lettre de relance
Prix	Permet de modéliser les expressions liées au prix. (Pas au sens premier du prix = 36€…)	Tarif, cher, ça coûte, trop cher, était moins cher, prix
Technique	Permet de modéliser les expressions relevant du domaine technique, comme les différentes installations ou l'assistance technique.	Ampérage, branchement, coupure, relevé de compteur, assistance téléphonique, numéro de matricule du compteur
Réappel	Permet de modéliser le réappel entre clients et agents.	EDF m' a appelé, j' ai contacté, personne m' a appelé
EDF	Permet de modéliser les expressions liées l'agent EDF.	Agent, interlocuteur, EDF, releveur
Positif	Permet de modéliser la satisfaction des clients et les expressions positives	Je suis satisfait, c' est OK, pas de problème, parfait, c' est pas grave
Négatif	Permet de modéliser l'insatisfaction des clients et les expressions négatives	Il y a un souci, ça m'énerve, pas conformes, merde, on hallucine, il y a une incompréhension

VI. Création d'un filtre

Lors de la création de la cartouche, nous nous sommes rendus compte que certaines expressions étaient détectées malgré la mise en place de contextes stricts dans les règles linguistiques. Afin de diminuer le bruit de la détection de concepts, nous avons créé un fichier « filtre ». Ce fichier est un concept « stop ». C'est à dire que lorsque l'on veut empêcher la détection d'une expression précise dans un contexte précis, il faut, pour que cela soit possible, placer celle-ci dans ce filtre.
Ce fichier est un composant « Lexicon », il sera, par conséquent, appliqué avant tous les composants « Rules » (voir p 78).

```
<concept name="filtre" level="0" display="never">
```

Ce concept est de niveau 0 et a pour attribut « never » (voir p 55). Le « never » indique que le concept ne sera jamais affiché en sortie de la cartouche. Nous utilisons cette technique de concept stop car la négation, symbolisée « ^!( ) », (voir p 54) ne peut être utilisée sur les tags, les concepts et les macros.
Voici quelques exemples pour illustrer ce concept « stop ».

```
(~VERB|~AUX) / ~positif-avec-contexte / ~PAP
```

Dans cet exemple, nous ne voulons pas récupérer la phrase « la lettre est bien partie » mais nous voulons continuer à détecter le verbe ou l'auxiliaire suivi du concept « positif-avec-contexte ». D'où la création de la règle, dans le fichier « filtre », qui contraint le concept « positif-avec-contexte » à être suivi d'un participe passé.

Le terme « expert » fait partie de la liste de mots dits positifs mais nous ne voulons surtout pas détecter le bi-gramme « expert international » qui provoquerait du bruit dans notre détection de concepts. Pour cela, nous rajoutons, à notre fichier « filtre » une seconde règle qui empêche la détection de cette expression dans ce contexte.

```
expert / international
```

Prenons un dernier exemple. Le terme « excellent » fait également parti de la liste de mots positifs. Mais dans le contexte suivant « excellente journée » ou « excellente fin de journée », nous ne voulons pas que cette expression soit détectée. En effet, elle est considérée comme une expression de politesse de fin de conversation qui ne dépend pas forcément de l'opinion générale de celle-ci. Nous réduisons donc le bruit en rajoutant la règle suivante :

```
excellent / fin / de
excellent / journée
```

Chapitre 4 – Evolution des résultats

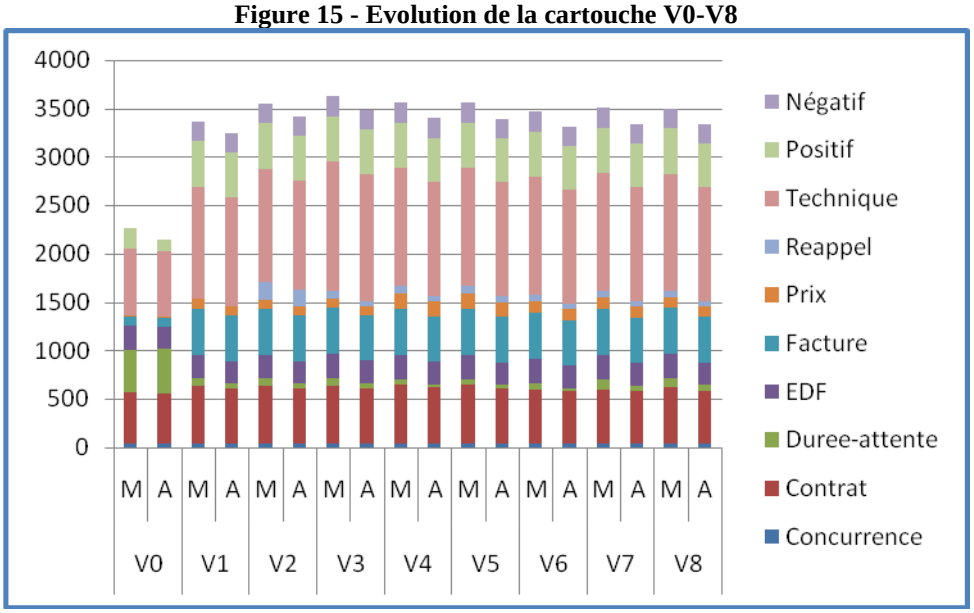
I. Comparaison des transcriptions manuelles et automatiques

Lors des différentes étapes de modification de la cartouche, nous avons enregistré l'évolution des résultats. La Figure 15, nous permet d'observer l'évolution de la cartouche entre les différentes versions. Nous sommes partis de la version V0, qui est la sortie de la cartouche BSM, pour arriver à la version V8 qui est la sortie de la cartouche CallSurf.
Nous pouvons faire deux observations sur cette figure :

- **L'augmentation importante du nombre de concepts détectés (environ 2200 à 3500 concepts).**
- **La similarité du comportement entre la transcription manuelle et automatique persiste dans le temps.**

Ce dernier point est très satisfaisant car, malgré les disfluences et les erreurs de reconnaissance présentes dans la transcription automatique, la cartouche détecte, le plus souvent, les concepts dans les deux transcriptions.

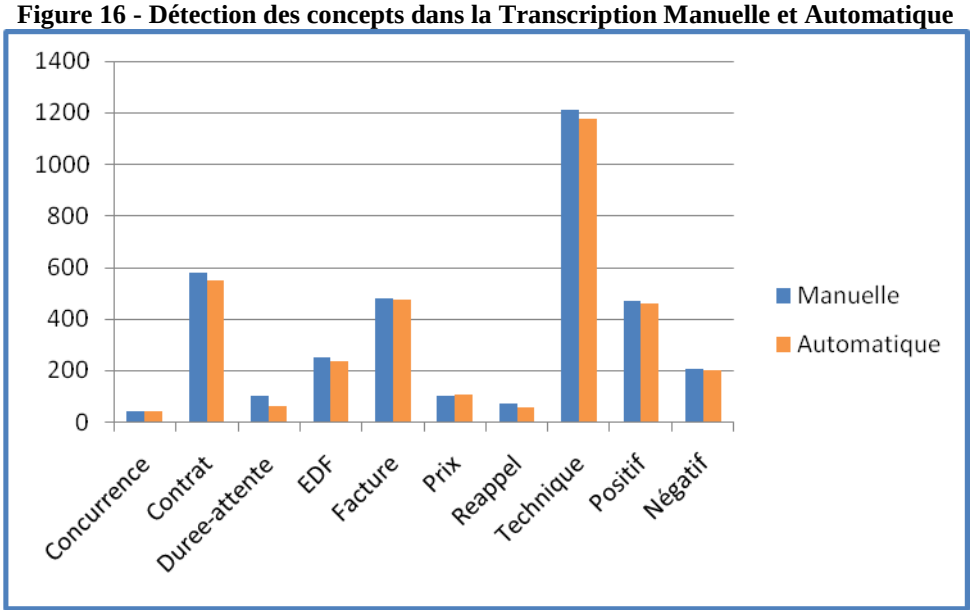
Figure 15 - Evolution de la cartouche V0-V8



Sur ce corpus de 10770 tours de parole, **3501 concepts ont été détectés pour la transcription manuelle contre 3343 pour la transcription automatique**. Ces chiffres confirment les résultats présents sur la figure. En effet, le **nombre de concepts détectés dans les deux transcriptions est pratiquement équivalent** soit une différence de 158 concepts.

Afin de confirmer cette observation, nous comparons la répartition des différents concepts détectés dans la transcription manuelle et automatique.

Figure 16 - Détection des concepts dans la Transcription Manuelle et Automatique



La figure ci-dessus confirme nos premiers résultats. L'écart que nous avions (voir p 75) au niveau de la détection du sentiment positif passe de 213 concepts pour la transcription manuelle à 469 concepts détectés et de 111 concepts pour la transcription automatique à 456 concepts.

Les différences les plus significatives que nous pouvons observer se situent au niveau des concepts « contrat » et « technique ».

Tableau 13- Différence de détection entre les Transcriptions

Concepts	Transcription Manuelle	Transcription Automatique
Contrat	556	533
Technique	1213	1174

La différence de détection de concepts entre les deux transcriptions est relativement faible par rapport au nombre de concepts détectés. Nous obtenons une difference de 7% pour le concept « technique » et de 1,9% pour le concept « contrat ». Nous pourrions ne pas étudier ces résultats de plus près car la différence n'est pas significative. Mais le fait que ce soient des concepts métiers nous interpelle car nous utilisons des lexiques pour détecter ces concepts. Après étude, nous avons pu constater que la plupart des concepts présents dans la transcription manuelle mais pas dans l'automatique proviennent d'erreurs de reconnaissance. Prenons les exemples suivants afin de mieux comprendre la présence de ces silences au niveau de la détection de concepts.

Manuelle : « rendez-vous »
Automatique : « rangées juste »

Le concept « technique » est détecté en manuelle mais pas en automatique, et la différence de transcription est telle que nous ne pouvons pas traiter ce cas particulier.

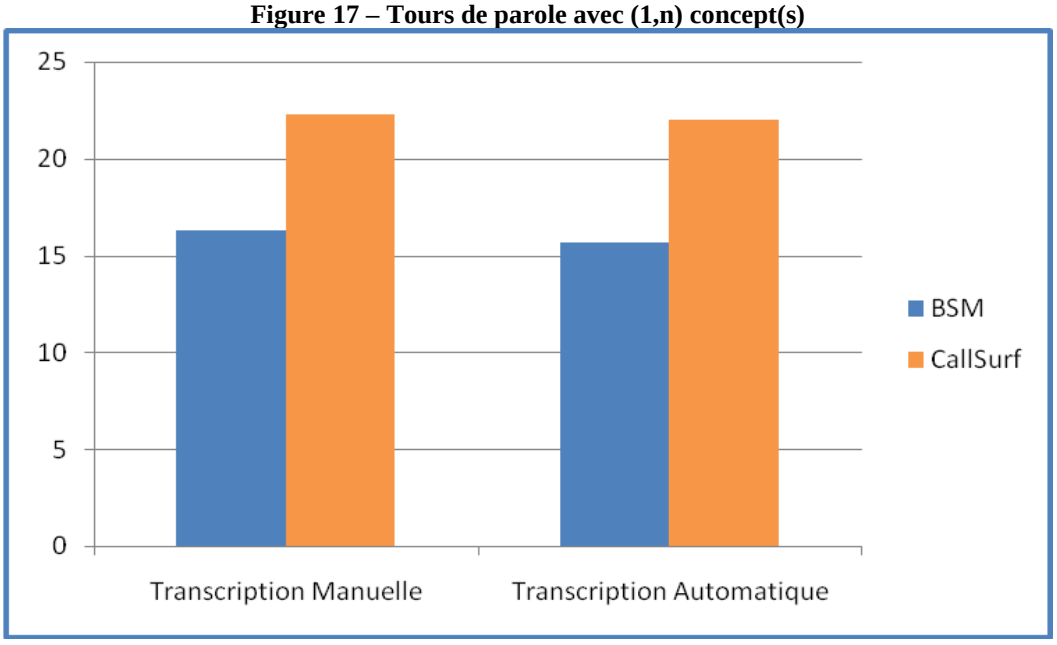
Manuelle : « souscrivons des contrats »
Automatique : « souscrit bon des contrats »

Il en est de même pour le concept « contrat ». Dans la transcription manuelle, une expression est détectée : « souscrivons des contrats » alors que dans la transcription automatique nous avons deux expressions détectées : « souscrit » et « des contrats ».

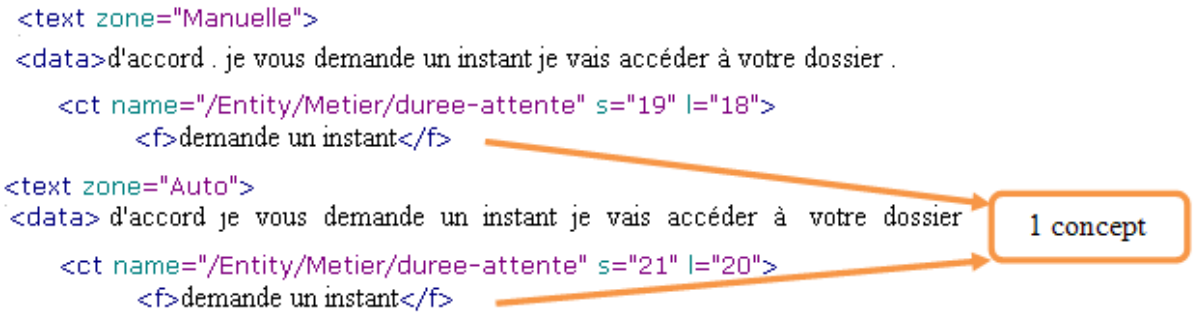
Nous pourrions prendre en compte ces erreurs dans les règles de la cartouche. Il nous suffirait, pour cela, de créer des règles avec un contexte strict et bien défini afin de ne pas obtenir de bruit ou de silence, mais ce ne serait pas sans danger. En effet, la cartouche deviendrait trop spécifique au corpus. De plus, dans le cas présent, l'erreur de détection de concepts n'est pas due aux disfluences mais aux erreurs de reconnaissance. Il n'est donc pas possible de prendre en compte tous les cas particuliers du corpus.

Sur l'ensemble des 10770 tours de parole, **22,3%** (contre 16,3% avec la cartouche BSM) **contiennent au moins un concept détecté dans la transcription manuelle contre 22% dans la transcription automatique.** (Voir la figure ci-dessous)

Figure 17 – Tours de parole avec (1,n) concept(s)



2506 concepts sont identiques entre les transcriptions manuelles et automatiques. Ce chiffre est obtenu de la façon suivante : un concept détecté en manuelle et le même détecté en automatique compte pour un concept.



Nous considérons que la détection de concept obtenue dans la transcription manuelle est la référence. Nous calculons alors le rappel et la précision.

- Rappel (voir p 75) :

$$\text{Rappel} = \frac{1849}{(2122+323)} = 75{,}6\%$$

(Contre 74,8% pour la cartouche BSM)

- Précision (voir p 75) :

$$\text{Précision} = \frac{1849}{(2122+247)} = 78\%$$

(Contre 77,7% pour la cartouche BSM)

Nous obtenons ainsi un rappel de 75,6% et une précision à 78%. Grâce aux modifications faites dans la cartouche et à l'adaptation des règles linguistiques face aux données orales, nous avons pu diminuer le bruit et le silence liés aux erreurs de reconnaissance et aux disfluences dans la détection de concepts.

Le nombre de tours de parole qui n'ont pas d'équivalence et le nombre de tours de parole qui ont des concepts mal détectés est faible (rappel et précision élevés). Ces résultats confirment que le comportement global de la cartouche CallSurf, dans la transcription automatique, est proche de celui obtenu dans la transcription manuelle.

II. Analyse par type de locuteur

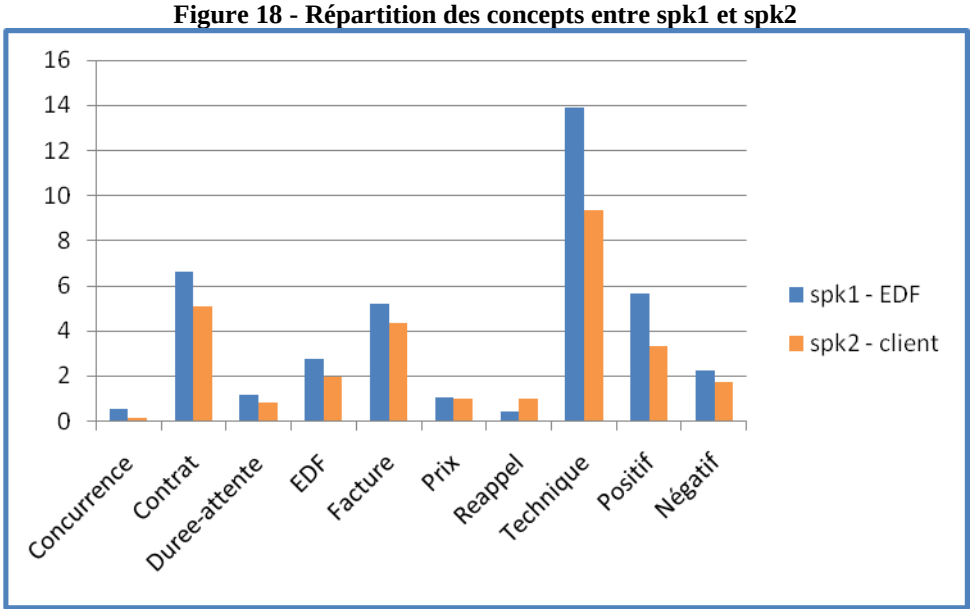
Lors des études précédentes, nous avons toujours analysé et comparé les transcriptions manuelles avec les transcriptions automatiques.

Mais certaines questions demeurent :

> *Est-ce que les agents utilisent plus de vocabulaire spécifique que les clients ?*
> *Est-ce que la différence est significative ?*
> *Faut-il à l'avenir construire deux cartouches différentes : une pour l'agent et une seconde pour le client ?*

Nous avons donc crée deux nouveaux corpus. **Un corpus de spk1 (EDF) qui contient 5280 tours de parole et un corpus de spk2 (client) qui contient 4273 tours de parole.** Les tours de parole manquants sont ceux des conversations entre agents qui, dans le cas présent, ne nous intéressent pas. Une fois ces corpus analysés par la cartouche CallSurf, nous avons obtenu la répartition des concepts comme suit :

Figure 18 - Répartition des concepts entre spk1 et spk2



Lors de l'analyse, nous aurions pu penser que les clients auraient tendance à plus utiliser des expressions d'opinions que l'agent, mais comme nous pouvons le constater, sur la figure précédente, ce n'est pas le cas.

En effet, la différence se situe principalement au niveau du concept « positif ». Cette différence est due à l'attitude positive de l'agent, il doit être poli et doit encourager le client. L'exemple suivant, tiré du corpus CallSurf, illustre cette attitude chez l'agent EDF.

Manuelle : Donc lui s'il appelle le même jour que vous ou le lendemain , c'est parfait , c'est juste un transfert administratif .

Pour les concepts métiers, les résultats ne sont pas surprenants. Les agents EDF emploient les termes qui conviennent pour décrire une situation, à la différence des clients qui vont hésiter sur l'emploi de certains termes. Ces hésitations provoquent dans certains cas des erreurs de reconnaissance et dans d'autres des disfluences.
Le tableau suivant nous démontre que les hésitations du client ont un impact direct sur la détection de concepts :

Tableau 14 - Rappel et Précision par Speaker

	spk1 – EDF	spk2 - client
Rappel	77,6%	75,1%
Précision	81,7%	75,2%

Comme nous pouvons le constater, la précision et le rappel chez le client, sont inférieurs à ceux de l'agent. Par conséquent, le corpus du spk2 contient plus de disfluences qui provoquent du bruit ou des silences selon les cas. Pour accompagner et confirmer ces résultats voici le taux d'erreur de mot (voir p 63) chez le client et l'agent qui est respectivement de 35% et de 29,9%.

Chapitre 5 - Protocole d'évaluation

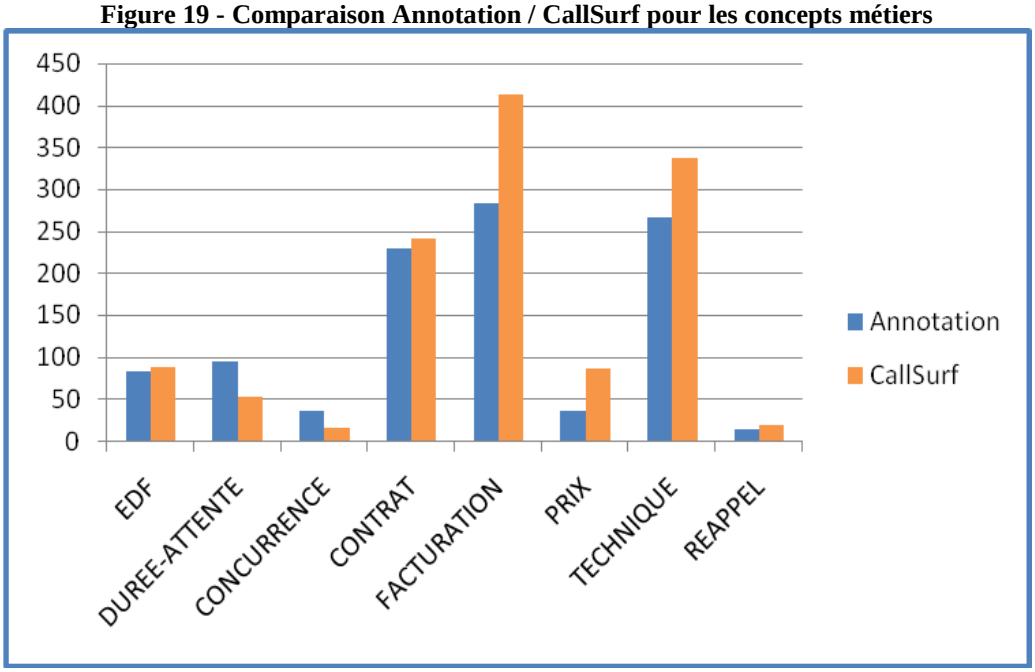
Afin de pouvoir réaliser un protocole d'évaluation, nous avons fait annoter par un annotateur un **nouveau corpus**. Ce corpus, de **33 fichiers, contient 5894 tours de parole entre agents EDF et particuliers**.
Lorsque nous avons comparé nos résultats avec ceux de l'annotateur et des différences dans la détection de concepts sont apparues.
Afin de pouvoir les analyser, nous avons séparé les résultats en deux groupes :

- La détection de concepts métiers
- La détection d'opinions

I. Détection de concepts métiers

Comme nous pouvons le constater sur la figure ci-dessous, quelques différences sont présentes entre l'annotation et la détection automatique de concepts.

Figure 19 - Comparaison Annotation / CallSurf pour les concepts métiers



Les différences sont dues, dans la majorité des cas, à la difficulté de différencier certains concepts entre eux, tant la frontière est mince.
Prenons l'exemple des concepts « contrat »/« prix ». L'annotateur détecte le lemme tarif comme étant un concept « contrat » alors que la cartouche CallSurf détecte ce lemme comme étant un concept « prix ». Cela provoque, par conséquent, un déséquilibre dans la détection de concepts entre la cartouche CallSurf et l'annotation. De plus, certains lemmes, comme prix, ne sont pas annotés par l'annotateur or dans la cartouche CallSurf, ce sont des expressions détectées comme étant des concepts « prix ». Par conséquent, il manque, en moyenne, 40 concepts « prix » pour l'annotation faite manuellement.

Pour le concept « duree-attente », la différence est due au contexte. En effet, l'annotateur a pris en compte le contexte au sens large. C'est-à-dire qu'il prend en compte les tours de parole précédents pour aider à détecter les concepts.

Agent : je vais me renseigner pour voir si le fait de changer de RIB
Agent : coupez pas

Grâce au contexte, nous pouvons constater que l'expression « coupez pas » signifie tout simplement « restez en ligne » ou « patientez un instant ». Il n'est donc pas possible de créer une règle permettant de détecter cette expression comme étant un concept de « duree-attente » car nous aurions trop de bruit avec le concept « technique » dans le sens « couper le courant ». En voici un exemple :

l'électricité vous vous nous coupez pas ?

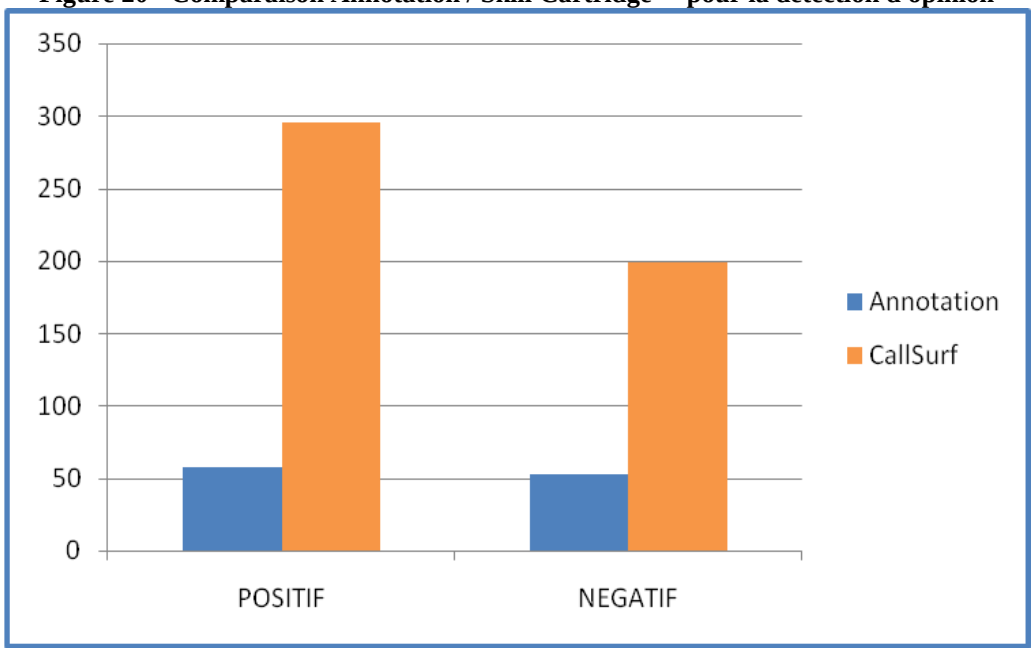
En essayant la cartouche CallSurf sur un nouveau corpus, nous nous sommes rendus compte d'une part que le lexique différait selon que l'on s'adresse à un client professionnel ou à un particulier, et d'autre part qu'il nous faudrait actualiser de façon automatique les listes de lexiques. En effet, avec le temps, le lexique des corpus évolue, de nouveaux noms de contrats apparaissent, de nouveaux concurrents arrivent sur le marché …

Sur la figure 19, nous pouvons constater qu'il y a une différence au niveau du concept « concurrence ». Cela est dû au terme ERDF présent, dans ce nouveau corpus, sur 16 tours de parole. Ceci qui équivaut à la différence Annotation/ CallSurf.

II. Détection d'opinions

Pour la détection d'opinions, la différence de détection de concepts est plus importante entre la cartouche CallSurf et l'annotation mais nous pouvons justifier cette différence.

Figure 20 - Comparaison Annotation / Skill Cartridge™ pour la détection d'opinion



D'après (Olena Zubaryeva, 2009) **« la distinction entre une opinion positive ou mixte peut parfois être sujette à interprétation et donc varier d'un individu à l'autre ».** Pour pallier le problème de la subjectivité, il faudrait se baser sur des annotations faites par plusieurs annotateurs. Il est aussi difficile de circonscrire ce qui relève de l'opinion, de l'émotion, de l'attitude et de l'humeur.
Prenons un exemple dans la cartouche CallSurf. L'expression « au pire » y est détectée comme étant une expression négative. En effet, elle a comme synonyme « à la limite, à la rigueur » ce qui justifie bien le concept qui lui est attribué. Or cette expression n'a pas été annotée comme étant un concept « négatif » par l'annotateur et il en est de même pour de nombreuses autres expressions.

Prenons un autre exemple suivant :

c'est super, j'adore

Est-ce de l'ironie ? Pour la détecter sans se tromper, il faudrait traiter l'ironie par l'analyse acoustique couplée de l'analyse linguistique.

La détection d'opinion est un sujet complexe qui nécessite d'être approfondi au travers d'une étude spécifique/dédiée.

Conclusion

Lors de la création de la nouvelle cartouche CallSurf, nous avons adapté les règles linguistiques à un corpus spécifique. Ces nouvelles règles prennent en compte les disfluences et gèrent les erreurs de reconnaissance.

Le résultat de la détection de concepts, dans la transcription automatique d'appels provenant des centres d'appel, est similaire à celui de la détection de concepts dans la transcription manuelle. La précision et le rappel pour les transcriptions sont respectivement de 75,6% et 78%. **Les résultats montrent une augmentation significative du nombre de concepts détectés et une diminution du bruit et du silence**.

Mais la création de le cartouche CallSurf a également ouvert la voie à une nouvelle réflexion sur l'interprétation des opinions. En effet, le protocole d'évaluation a montré qu'il reste encore beaucoup de chemin à parcourir pour détecter l'opinion des gens dans des données issues de l'oral. La détection d'opinion dans un corpus demeure extrêmement subjective.
Ce sujet, de part sa complexité, mériterait d'être traité à part et constituerait, à lui seul, le sujet d'une thèse :

Comment détecter une opinion ?
Qu'elle est la frontière entre la notion de satisfaction et une attitude positive ?
Faut-il prendre en compte le contexte ?
Faut-il utiliser l'analyse acoustique pour détecter l'ironie ?

Ces 6 mois de stage furent très riches en apprentissage et en contenu. J'ai eu la chance de pouvoir travailler en collaboration avec des chercheurs et des ingénieurs, notamment chez LIMSI et Vecsys.

Au cours de ce stage, EDF m'a donné l'opportunité d'assister à une conférence internationale, les JADT « Statistical analysis of textual data » à Rome. Grâce à cette expérience, j'ai pu découvrir de nombreuses applications dans le domaine du TAL.
Cette conférence ainsi que les premiers résultats de mon travail et les découvertes effectuées durant ce stage m'ont amenée à écrire et à soumettre un article au jury de sélection de la conférence internationale ACM Multimedia qui se déroulera le 29 octobre à Florence. Cet article ayant été accepté, mon prochain défi sera donc de le présenter en anglais, je ne serais plus spectatrice mais bien actrice, lors de cette prochaine conférence.
Bien entendu, tout au long du stage, j'ai été confrontée à certaines difficultés que j'ai pu surmonter. La difficulté principale fût la mission qui m'a été donnée sur le développement de la cartouche. N'ayant pas de commanditaire ni de demande précise, il a fallu que je m'adapte, que je comprenne ce qu'on attendait précisément de moi et que je choisisse les expressions et les thèmes que je souhaitais extraire.

Je ressors, de ces 6 mois de stage, grandie sur le plan professionnel (une première vraie expérience dans mon domaine, et une reconnaissance pour le travail effectué) et personnel (j'ai gagné en confiance et en assurance) :

- J'ai développé mes compétences sur le Text Mining,
- J'ai découvert un sujet qui me passionne : les disfluences à l'oral et les erreurs de reconnaissance,
- Pour conclure, la problématique du traitement de ces données issues de l'oral, m'a fortement intéressée.

Bibliographie

Abdenour Mokrane, N. F.-Y. (2008). Cascades de transducteurs pour le chunking de la parole conversationnelle : l'utilisation de la plateforme CasSys dans le projet EPAC. *TALN*. Avignon.

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de. Université de Paris-Sud.

André Valli, J. V. (1999). Etiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée* , 113-133.

Bove, R. (2008). *Analyse syntaxique automatique de l'oral : étude des disfluences.* Université d'Aix-Marseille 1.

Candéa, M. (2000). Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Université Paris III.

Claire Blanche-Benveniste, M. B. (1990). *Le français parlé : Etudes grammaticales.* Paris.

Elsa Pascual, M.-P. P.-W. (1995). Ma définition dans le texte. *Atelier Texte et Communication : Journées "Le texte de type consignes".*

Favre, B. (2007). *Résumé automatique de parole pour un accès.* Avignon.

Guénot, M.-L. (2005). Parsing de l'oral : traiter les disfluences. *TALN*. Dourdan.

Habert, B. (2006). Portrait de linguiste(s) à l'instrument. Dans H. S. Guillot C, *A la quête du sens : études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia* (pp. pp. 124-132). Lyon: ENS.

Jakobson, R. (1963). *Essais de Linguistique générale.* Trad.fr.Paris : Seuil.

Jean-Leon Bouraoui, N. V. (2009). Traitement automatique de disfluences dans un corpus. *TALN*. Senlis.

Kurdi, M. Z. (2003). Contribution à l'analyse du langage oral spontané. Université de Grenoble I.

Marie Piu, R. B. (2007). Annotation des disfluences dans les corpus oraux. *RÉCITAL*. Toulouse.

Martine Garnier-Rizet, G. A.-L.-L.-R. (2008). CallSurf - Automatic transcription, indexing and structuration of call center. *LREC.*

Olena Zubaryeva, J. S. (2009). *Evaluation de modèles de classification.* Neuchâtel (Suisse).

Pallaud, B. (2002). Les amorces de mots comme faits autonymiques en langage oral.

R.Feldman, I. D. (1995). Knowledge Discovery in texts', Proceeding of the Conf. On Knowledge Discovery (KDT). *AAAI* .

Thierry Bazillon, V. J. (2008). La parole spontanée : transcription et traitement. *TAL*, (p. 47 à 76).

Véronis, A. V. *Etiquetage grammaticale des corpus de parole : problèmes et perscpectives.*

Zellner, B. (1992). *Le be - begayage et euh...., l'hésitation en français spontané.* PARIS 7.

Table des annexes

Annexe 1 - Guide d'utilisation de Luxid

Annexe 2 - Les tags de Xelda

Tag	Description	Exemple
+ADJ2_INV	special number invariant adjective	gros
+ADJ2_PL	special plur. adjective	petites, grands
+ADJ2_SG	special sing. adjective	petit, grande
+ADJ_INV	number invariant adjective	heureux
+ADJ_PL	plural adjective	gentils, gentilles
+ADJ_SG	singular adjective	gentil, gentille
+ADV	adverb	finalement, aujourd'hui
+CM	comma	,
+COMME	reserved for the word "comme"	comme
+CONJQUE	reserved for the word "que"	que
+CONN	connector subordinate conjunction	si, quand
+COORD	coordinate conjunction	et, ou
+DET_PL	plural determiner	les
+DET_SG	singular determiner	le, la
+MISC	miscellaneous	miaou, afin
+NEG	negation particule	reserved for ne
+NOUN_INV	number invariant noun	taux
+NOUN_PL	plural noun	chiens, fourmis
+NOUN_SG	singular noun	chien, fourmi
+NUM	numeral	treize, 13, XIX
+PAP_INV	number invariant past participle	soumis
+PAP_PL	plural past participle	finis, finies
+PAP_SG	singular past participle	fini, finie
+PC	clitic pronoun	[donne-]le, [appelle-]moi, [donne-]lui
+PREP	preposition (other than à, au, de, du ...)	dans, après
+PREP_A	preposition "à"	à, au, aux
+PREP_DE	preposition "de"	de, d', du, des
+PRON	pronoun	il, elles, personne, rien
+PRON_P1P2	1st or 2nd person pronoun	je, tu, nous

+PUNCT	punctuation (other that comma)	: -
+RELPRO	relative/interrog. pronoun (exept "que")	qui, quoi, lequel
+SENT	sentence final punctuation	. ! ? ;
+SYM	symbols	@ %
+VAUX_INF	infinitive auxiliary	être, avoir
+VAUX_P1P2	1st or 2nd pers. aux., any tense	suis, as
+VAUX_P3PL	3rd pers. plur. aux., any tense	seraient
+VAUX_P3SG	3rd pers. sing. aux., any tense	aura
+VAUX_PAP	past participle auxiliary	eu, été
+VAUX_PRP	present participle auxiliary	ayant
+VERB_INF	infinitive verb	danser, finir
+VERB_P1P2	1st or 2nd pers. verb, any tense	danse, dansiez, dansais
+VERB_P3PL	3rd pers. plur. verb, any tense	danseront
+VERB_P3SG	3rd pers. sing. verb, any tense	danse, dansait
+VERB_PRP	present participle verb	dansant
+VOICILA	reserved for "voici", "voilà"	voici, voilà

Annexe 3 - Impact of spontaneous speech features on business concept detection: a study of call-center data.

Tableaux

Figures

Images

Table des matières

MOTS-CLÉS : Disfluences, parole spontanée, discours oral, erreur de reconnaissance, text mining, règle linguistique, cartouche de connaissance.

RÉSUMÉ

Ce mémoire aborde la problématique du traitement des données issues de l'oral. En effet, les entreprises regorgent de données concernant leurs clients, données issues d'enquêtes de satisfaction, de forums, d'appels téléphoniques… qui ne sont pas exploitables en l'état.
En premier lieu, un rappel des différents travaux existants en matière d'analyse linguistique des données issues de l'oral y est effectué (voir p 75 à 80). Les transcriptions manuelles et automatiques de ces données orales issues de conversations téléphoniques entre agents EDF et clients y sont ensuite analysées (voir p 75 à 80). Enfin, une solution permettant d'adapter une cartouche de connaissance à ces données spécifiques y est proposée (voir p 75).

KEYWORDS : Disfluencies, spontaneous speech, oral speech, recognition errors, text mining, linguistic pattern, skill cartridge.

ABSTRACT

This thesis addresses the problem of processing data derived from the oral. Indeed, businesses are full of data about their customers, data from satisfaction surveys, forums, call-center... Which are not workable.
First, a reminder of existing work on the linguistic analysis of data from the spontaneous speech is proposed (see p 75 à 80). The manual and automatic transcriptions of speech features from telephone conversations between agents and customers EDF are analyzed (see p 75 à 80). Finally a solution is proposed to accommodate a Skill Cartridge to specific data (see p 75).