Caractérisation lexicale des contributions clients agents dans un corpus de conversations téléphoniques retranscrites

Frederik Cailliau[1,2], Céline Poudat[1]

[1]Sinequa Labs – 12, rue d'Athènes – 75 009 Paris – France

[2] LIPN, Institut Galilée, Université Paris-Nord – 99, avenue Jean-Baptiste Clément
93 430 Villetaneuse – France

Abstract

For the last few years, oral corpus analysis has taken an important position. It is today at stake in descriptive and applicative environments. Amongst the various projects that focus on oral processing, the Infom@gic (ST2.31) project has developed a corpus of transcribed telephonic conversations within a data-mining framework. The present study, composed of 1 268 dialogues between EDF Pro agents and clients, is based on this corpus.
Starting from the hypothesis that the two actor categories are linguistically regulated by stabilised social roles, we have chosen to compare and analyse their lexical characteristics with the DTM software.

Résumé

Bien qu'elle se soit considérablement développée ces dernières années, l'analyse des corpus oraux représente aujourd'hui un enjeu stratégique tant sur le plan descriptif que sur le plan applicatif. Parmi les différents projets intéressés par l'oral et son traitement, le projet Infom@gic (ST2.31) a développé un corpus de conversations téléphoniques retranscrites dans le cadre d'une application de fouille de données. C'est sur ce corpus, constitué de 1 268 dialogues entre agents EDF Pro et clients, que se fonde la présente étude.
En partant de l'hypothèse que ces deux catégories d'acteurs sont linguistiquement régulées par des rôles sociaux stabilisés, nous avons choisi de les contraster et d'observer leurs caractéristiques et leurs fonctionnements lexicaux à l'aide du logiciel DTM.

Mots-clés : fouille de données, corpus oraux, analyse des correspondances.

1. Contexte

Si les progrès de l'informatique et des possibilités de numérisation combinés au développement massif de la linguistique de corpus ont entraîné de nombreux travaux sur corpus numériques écrits, force est de constater que peu d'études portent sur l'observation et l'exploitation de données orales.

Ce phénomène n'est pas étonnant si l'on considère que les corpus sont peu accessibles, en raison de leurs coûts de constitution, de transcription et d'annotation. Malgré des efforts louables (e.g. la TEI), l'annotation des corpus oraux demeure en effet peu normalisée, et les

corpus sont très disparates selon l'objectif qui préside à sa constitution. Ainsi, le format des échanges ou la durée des enregistrements varient substantiellement d'un corpus à l'autre[1].

Les corpus oraux en libre accès sont plutôt rares[2]. Parmi les initiatives de constitution de corpus oraux, on peut mentionner le projet ESLO[3] (l'Enquête Socio-Linguistique à Orléans), qui proposera à terme un corpus de 400h (soit 6 millions de mots) qui sera disponible pour la communauté scientifique (Abouda et Baude, 2006). Comme d'autres corpus (oraux ou écrits), ESLO[4] peut être consulté en ligne par le biais d'un concordancier. La plupart des corpus oraux demeurent néanmoins en accès restreint à l'équipe constituante.

A l'instar de l'écrit, l'oral se décompose en différents genres qui demandent des traitements adaptés. On peut ainsi distinguer les enregistrements d'émissions radio par exemple, dont la bonne qualité de transcription et sa ressemblance aux corpus de type journalistique permet de déployer les technologies aujourd'hui utilisés sur l'écrit sans grande perte de performance sur les transcriptions automatiques (Cailliau et de Loupy, 2007).

C'est sur un genre oral bien spécifique que nous avons choisi de fonder notre étude. Situées à la frontière des genres institués et conversationnels que définit D. Maingueneau (2004), les *conversations téléphoniques agents-clients en centre d'appel* (dorénavant CTCA) s'inscrivent dans un cadre institutionnel défini. Elles impliquent des rôles sociaux déterminés et des scripts relativement stables, bien que la présence de contraintes locales et horizontales (stratégies d'ajustement et de négociations) liées à la nature conversationnelle du genre ne doive pas pour autant être écartée.

Le corpus que nous mobilisons a été constitué dans le cadre de la sous-tâche ST2.31 du projet Infom@gic[5], qui entre dans sa troisième et dernière année début 2008. Le ST2.31 réunit les compétences de cinq partenaires[6] et se donne comme objectif de valider les approches d'extraction d'information sur des données conversationnelles téléphoniques. La fouille de données sur ce type de données vise à exploiter la masse informationnelle non accessible aujourd'hui. Les raisons d'appels des clients peuvent être très diverses, et l'analyse des sujets les plus fréquemment posés peut être utilisée pour améliorer les services fournis. Grâce aux évolutions technologiques, ce type d'analyse devient possible à grande échelle, et est en passe de devenir un enjeu stratégique pour les grandes entreprises.

Un corpus oral de conversations entre agents et clients appelant à titre professionnel a été enregistré dans les centres d'appel d'EDF par Vecsys, qui a ensuite procédé à la transcription manuelle d'une sélection de conversations. La sélection prise en compte dans cette étude concerne 1 268 conversations, soit 150h d'enregistrement ou encore 1 873 865 mots[7]. Afin de

[1] Voir l'inventaire des corpus oraux disponibles pour le français (P. Cappeau et M. Seijido, DGLFLF, 2005) : http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Inventaire.pdf, http://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Presentation_Inventaire.pdf

[2] Comme le montrent P. Cappeau et M. Seijido (*ibid.*) la plupart des corpus ne sont consultables que « sur place ».

[3] Voir http://www.univ-orleans.fr/eslo/

[4] Grâce au projet Elicop : http://bach.arts.kuleuven.be/elicop/

[5] Voir www.capdigital.com/xwiki/bin/view/Projet/Infomagic

[6] EDF (rd.edf.com), Limsi (www.limsi.fr/Scientifique/tlp), Sinequa (www.sinequa.com), Temis (www.temis.com), Vecsys (www.vecsys.fr)

[7] Calcul intégrant les mots composés du corpus.

protéger la vie privée des clients, Vecsys a anonymisé le corpus en masquant toutes les informations susceptibles d'identifier le client, telles que les noms de personne, les adresses, les noms d'entreprise, les numéros de carte bancaire, etc.

A partir du corpus oral et de transcriptions, le Limsi est en charge de la création des modèles acoustiques et linguistiques pour développer un système de transcription automatique. La technologie Insight Discoverer™ de Temis permet ensuite de procéder à de la fouille de données à partir de la sortie de ce système. Cette tâche peut être optimisée par la prise en compte d'un étiquetage morpho-syntaxique et d'une détection d'entités que permet l'étiqueteur LemmaNG de Sinequa, adapté à l'oral par un entraînement sur une partie du corpus de CTCA étiquetée manuellement.

La présente étude se fonde ainsi sur ce corpus transcrit et annoté, de taille et d'homogénéité discursive exceptionnelle si on le compare aux corpus de conversations téléphoniques existants. A notre connaissance, les corpus existant pour le français sont loin d'être aussi riches. On peut mentionner le corpus de transcriptions *Conversations téléphoniques* édité en 1984 qui regroupait 165 minutes de communications d'ordre privé, professionnel et institutionnel, enregistrées en France de 1982 à 1983 (Schmale 2007a et b pour la réédition), ou encore les *transcriptions de Bielefeld*, qui remontent également à 1984 et proviennent de situations de rédaction conversationnelle (en français) et d'interactions médicales (essentiellement en allemand), l'ensemble ne totalisant qu'une centaine de minutes. Plus récemment, P. Vergely a développé en 2005 un corpus de 750 conversations (soit 106 726 mots) entre chefs de salle et maintenance opérationnelle dans la navigation aérienne, issues de l'environnement réel de travail et de l'environnement simulé (expérimentations effectuées au Centre d'Etudes de la Navigation Aérienne).

Notre corpus d'étude présente l'intérêt d'être fortement régulé sur le plan linguistique. Le scénario est stable : le professionnel client formule un problème ou pose une question, tandis que l'agent y répond et trouver une solution. On peut supposer que le contexte du scénario génère une forte stabilisation thématique (restriction des topics ici aux problématiquesde la fourniture d'énergie, de conseils, de services), de même qu'une grande régulation discursive (registre attendu par une interaction client / agent, probable stabilité des formes de politesse, etc.), qui nous a semblé bien distincte selon la catégorie d'acteur *Agent / Client* considérée. En effet, l'agent tient un discours plus normé en termes de style et de registre que le professionnel client, à l'origine quant à lui des thèmes contenus dans le corpus.

C'est en partant de cette hypothèse d'une forte stabilisation linguistique des fonctions sociales des deux acteurs du genre que nous avons procédé à un découpage du corpus en deux sous-ensembles *Agents* et *Clients*, tout en conservant sa division en textes. Cette division nous semble pertinente, dans la mesure où ces rôles sociaux, ou sociolectes, l'emportent naturellement sur les idiolectes, tout comme les genres l'emportent sur les styles. Les seuls traits idiolectaux qu'il serait possible de mettre à jour dans le corpus concernent les agents[8], personnages récurrents des conversations contrairement aux clients, ce qui motive encore une fois ce découpage.

L'inconvénient de cette démarche est qu'elle entraîne l'effacement des tours de parole du corpus, et par conséquent, de la nature fondamentalement contextuelle des énoncés, ce qu'a largement démontré l'analyse conversationnelle dans son ensemble. Plusieurs travaux

[8] Les agents travaillent dans le centre d'appel considéré, et apparaissent donc respectivement à maintes reprises dans le corpus.

d'extraction d'information sur corpus de conversations téléphoniques prenant en compte les conversations dans leur intégralité ont également été menés. Par exemple, les études de Narjès Boufaden et al. (2002, 2004) visent à extraire les informations pertinentes d'un corpus anglais de conversations relatives à des incidents survenus en mer fourni par le Centre de Recherche de la défense Canadienne.

Néanmoins, l'examen des énoncés dissociés Agent / Client qui suivent nous semble relativiser la perte des tours de parole, voire même légitimer la pertinence de notre choix. En effet, l'interrogation du client et les thèmes qui ont motivé son appel demeurent bien apparents, tandis que c'est davantage le discours de l'agent qui est contextuellement dépendant de celui du client :

Agent	Client
edf pro bonjour. oui oui c'est un client particulier professionnel ? d'accord, alors, ça c'est du ressort du service exploitation et non pas du service commercial je me renseigne hein ne quittez pas madame madame ? ouais donc je viens de voir avec responsable là donc il me dit que normalement en dans ce cas-là pour les déclarations d'intention de travaux c'est directement en mairie que vous devez aller chercher les papiers et les remplir et après eux vous aiguillent et vous donnent éventuellement les coordonnées du service exploitation d'edf hein oh je suppose oui après je je sais pas comment eux fonctionnent en en mairie oui vous avez des papiers à à collecter à remplir effectivement et et tout commence à partir de de la mairie et nous après on vous indique le le cheminement nécessaire hein oh alors ça je peux pas vous renseigner madame je sais pas du tout. je vous en prie madame. bonne journée au revoir	bonjour madame établissements X, mais là je vous appelle parce qu'on doit au mois de septembre comme on fait des travaux de terrassement chez un de nos de nos clients et je dois faire une déclaration d'intention de commencement des travaux j'en ai jamais fait comment dois-je faire ? particulier moi je suis une professionnelle lui c'est un particulier. d'accord oui merci. pardon ? oui et je dois je dois faire les le mêmes pour gdf et les eaux du nord ? donc je vais chercher à en mairie ok merci beaucoup de la mairie ok et faut combien de temps vous savez pas ? c'est pas grave je vais aller à la mairie merci beaucoup au revoir. il faut aller à la mairie

Tableau 1 : un exemple de dissociation Agent / Client

Les deux catégories d'acteurs ont donc été soumises à l'analyse lexicale du logiciel DTM[9] après plusieurs filtrages décrits en 2. Les différents axes d'opposition mis à jour en 3. nous permettront de décrire et d'opposer les thèmes d'appel de prédilection des clients, de même que les différentes modalités de réponse des agents.

2. Corpus et méthodes

2.1. Corpus

Le corpus est constitué de 1 268 fichiers, totalisant 7,75 Mo, 150h d'enregistrement ou encore 1 873 865 mots, soit une moyenne de 1 478 mots, ou environ 7 minutes par conversation. Chaque fichier contient une conversation agent/client et/ou agent/agent. Est également considérée comme conversation le monologue de l'agent en cas d'appel sortant du centre d'appel qui tombe sur le répondeur du client. L'enregistrement démarre dès que l'agent décroche le téléphone. Nous avons relevé un total de 34 messages laissés sur répondeur par l'agent.

[9] http://ses.enst.fr/lebart/

Les enregistrements ont été réalisés au cours de l'année 2006, et l'essentiel des transcriptions en 2007. Un grand nombre de thèmes y sont abordés, comme par exemple des changements d'adresse, des demandes de raccordement ou des problèmes de paiement. Chaque conversation dispose d'une micro-structure assez figée, qu'on pourrait décrire comme une séquence salutations, présentation et/ou identification du client, question, résolution et remerciements. La partie question-résolution peut compter plusieurs thèmes. C'est cette partie qu'il faudrait idéalement réussir à isoler pour obtenir de meilleurs résultats en classification automatique.

2.2. Filtrages

L'étiquetage morphosyntaxique et l'extraction des entités sont des traitements précieux pour la fouille de données, qui permettent d'éviter de créer du bruit lors de la construction des classes. L'étiqueteur morphosyntaxique de Sinequa, spécialement entraîné sur un corpus de transcriptions manuellement étiquetées dans le cadre du projet, a été utilisé pour procéder à deux filtrages thématique et stylistique. En ce qui concerne les thèmes, les catégories grammaticales des mots qui ont été conservées incluent les noms, les verbes (avec leurs participes passés), les adjectifs et les adverbes. Les catégories mobilisées pour caractériser les styles incluent quant à elles les interjections, les conjonctions de coordination et de subordination, les négations, les pronoms et les superlatifs. La désambiguïsation des catégories s'effectue sur la base de règles construites par apprentissage. La lemmatisation prend en compte les mots composés, tel *EDF Pro*, qui sont renseignés dans les lexiques de l'étiqueteur. Cela crée des différences qui ne sont pas toujours visibles entre les analyses avec et sans lemmatisation.

2.3. Analyses statistiques

Nous avons mené plusieurs analyses factorielles des correspondances sur le corpus, à l'aide des différents sous-ensembles de descripteurs pris en compte, ce qui nous a permis d'explorer les champs d'opposition thématiques du corpus.

L'ensemble des tests statistiques a été réalisé avec le logiciel DTM développé par Lebart.

3. Analyses lexicales

Plusieurs analyses ont été menées avec et sans lemmatisation. Ces dernières utilisent les filtrages décrits en 2.2, appliqués aux corpus clients et agents séparément ainsi que sur l'ensemble des données. Seuls les résultats montrant des oppositions significatives sont discutés ci-dessous.

3.1. Axes des clients

Observons d'abord les points qui ont la plus forte contribution aux deux premiers axes de la carte factorielle associée aux clients avec un filtrage thématique, et le cosinus carré le plus élevé (*fig. 1*) : *mandat, pénalité, rappel, retard, moitié, échéancier, paiement, encaisser, chèque,* etc. s'opposent à *coffret, raccordement, jaune, branchement, provisoire, technique, bleu, bâtiment, ampère,* etc. Le financier s'oppose ainsi de manière nette au technique.

Le second axe, moins interprétable, opposerait à ses extrémités les lieux de l'intervention, du plus local (*machine, système, brancher, disjoncteur*, etc.) au plus global (*cabinet, société, rue, boulevard*, etc.).

La carte factorielle issue du lexique fait ainsi apparaître les pôles principaux de demandes des professionnels clients.

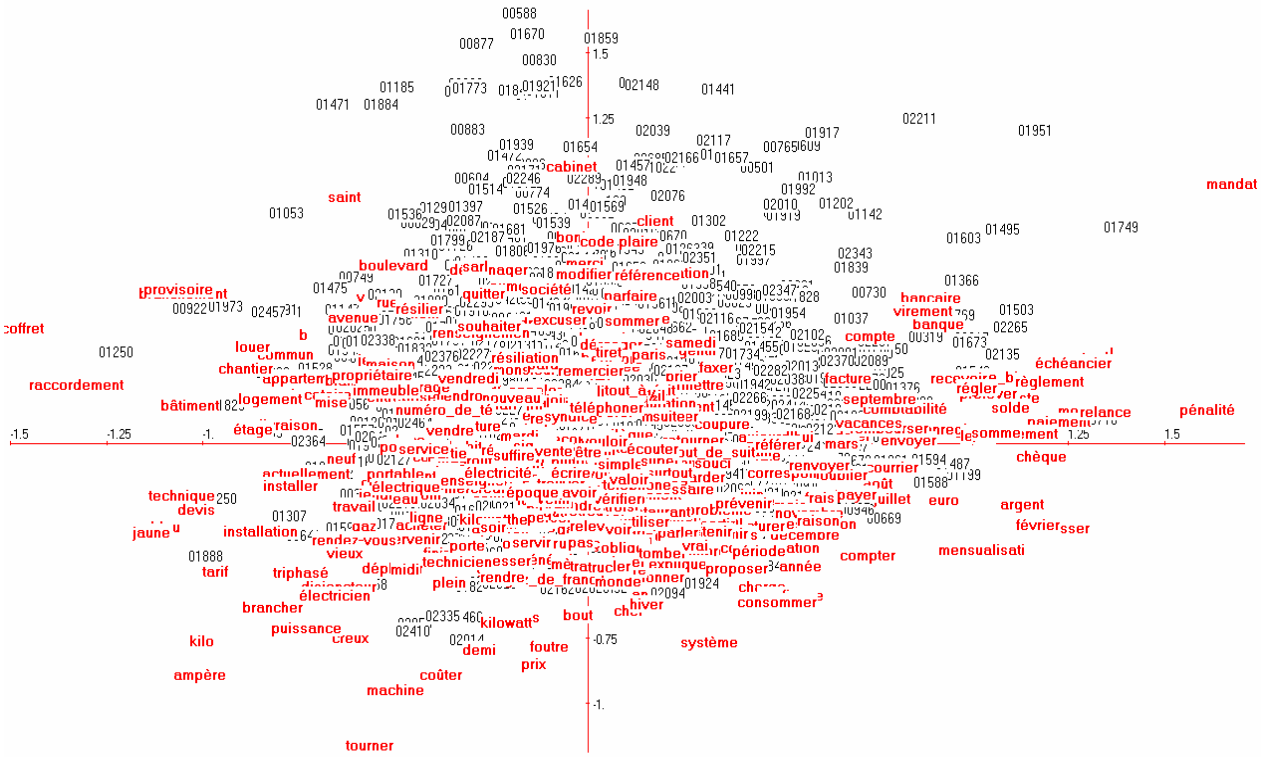

Figure 1 : Carte factorielle du lexique associé aux clients – Lemmes pleins

Notons que si l'on observe les données sans filtrage par lemmes (Figure 2), en ne considérant que les 1 000 premières formes lexicales les plus répandues, on retrouve cette opposition du financier (en haut à gauche) au technique (en bas à droite) sur les deux premiers facteurs.

En revanche, on voit apparaître deux pôles thématiques supplémentaires, qui étaient peu visibles précédemment. Sur le versant positif des deux axes, on peut observer un pôle qui paraît renvoyer à un ensemble de questions très courtes, relatives à des demandes de codes, références ou numéros divers (siret, téléphone, etc.), assortis de marques de politesse. La majorité des appels est d'ailleurs concentrée sur ce pôle, ce qui nous paraît tout à fait conforme à la structure des appels observée dans le corpus. En effet, la phase qui concerne l'identification du client ou ses références occupe une partie importante dans le corpus.

Le versant négatif des deux axes est caractérisé par la thématique de l'aspect pratique de l'intervention : *installation*, *immeuble*, *syndic*, *accès*, *rendez*(-vous), etc., associés à quelques termes plus techniques : *branchement*, *raccordement*.

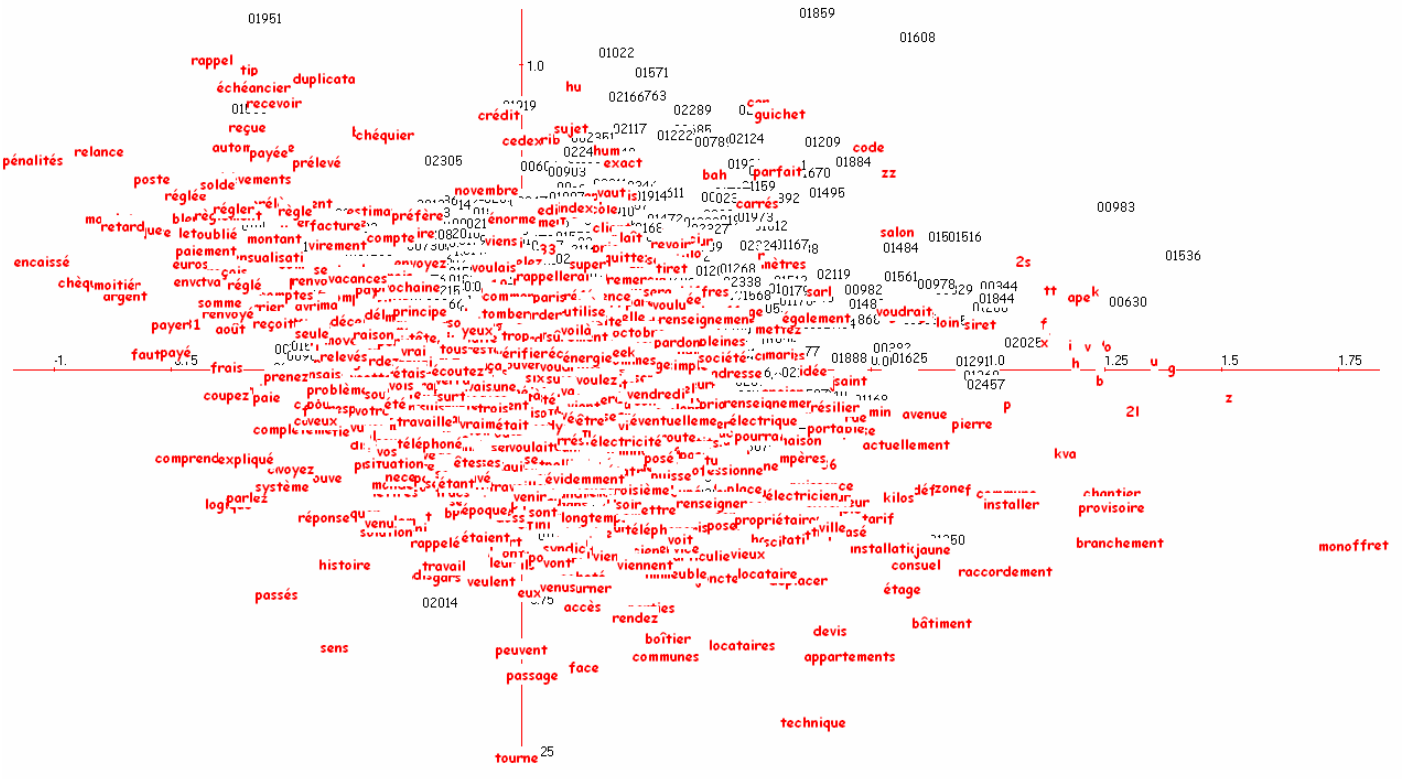

Figure 2 : Carte factorielle du lexique associé aux clients (1 000 premiers mots) – sans lemmatisation

3.2. Axes des agents

La carte factorielle issue du lexique associé aux agents sans lemmatisation (Figure 3) montre un premier pôle discursif qui émerge dans les dialogues des agents, caractérisé par la présence de marqueurs de registre familier (*tu, te, as, vas, salut, etc.*) qui ont la plus forte contribution à l'inertie du nuage projeté sur le premier axe (versant négatif du premier axe) : ces marqueurs sont liés à la présence de conversations agent/agent imbriquées dans les conversations elles-mêmes, comme le confirme la présence de *transfère* et *transférer* (l'appel client reçu). En effet, quand les agents s'appellent entre eux, le style devient informel.

A droite de ce premier pôle, toujours au négatif à gauche, on distingue une zone dans laquelle la plupart des documents sont concentrés. Cette zone est marqué par les introductions à la conversation comme les salutations (*bonjour*, (bonne) *soirée*, (au) *revoir*). De l'autre côté du premier axe, côté négatif on retrouve le pôle financier observé dans le sous-corpus clients (*virement*, *retard*, *bancaires*, *échéancier*, *prélevée*, *régularisation*, etc.)

En positif sur le deuxième axe, les termes manquent de cohérence, même si on peut distinguer quelques termes plus techniques (*triphasé*, *raccordement*).

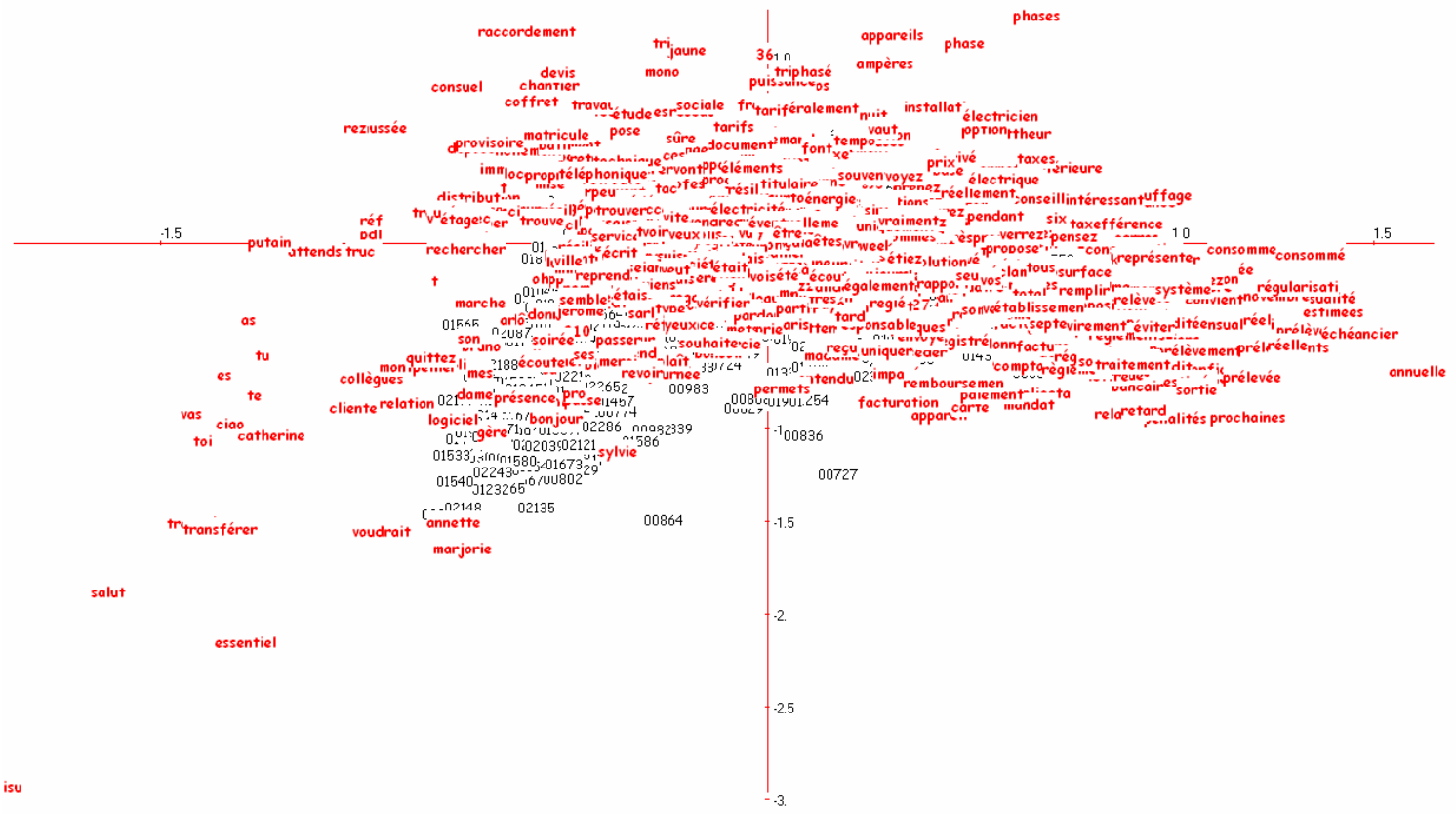

Figure 3 : Carte factorielle du lexique associé aux agents (1 000 premiers mots) – sans lemmatisation

Remarquons que l'absence de lemmatisation fait apparaître des mots ici non porteurs d'information ici (*pro* pour *EDF Pro*) et que d'autres mots apparaissent au singulier et au pluriel à proximité l'un de l'autre (tarif, tarifs). Dans la figure avec lemmatisation, les revoir et salut réapparaissent, car ces mots sont étiquetés comme nom dans les lexiques de l'étiqueteur, et non pas comme interjection. Cela veut dire qu'il ne suffit pas seulement d'adapter les modèles de langage de l'étiqueteur à de l'oral, mais que le lexique doit être également adapté. Il ne suffit pas d'ajouter des interjections spécifiques à l'oral, mais il faut intervenir sur le lexique déjà existant pour ajouter des ambiguïtés comme par exemple interjection pour *salut*.

4. Conclusion

Dans le monde idéal de la fouille de donnée sur des transcriptions conversationnelles téléphoniques, on arriverait à séparer le corps de l'appel de ses autres parties pour filtrer le bruit provoqué par les formalités du discours. En attendant les avancées technologiques, notre approche permet déjà de le limiter partiellement.

Les analyses exploratoires que nous avons menées démontrent ainsi l'intérêt de distinguer clients et agents, et valident notre première hypothèse d'une plus grande dépendance contextuelle du discours de l'agent, plus marqué par les formalités du discours que par les thèmes que gère le service client. Au contraire, le discours client est lexicalement plus stabilisé, ce qui le rend plus ad hoc pour détecter les thèmes.

Références

Abouda L. et Baude O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO. In Rastier F., Ballabriga M. (dir.), *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du colloque international d'Albi*, juillet 2006. Publiés par Carine Duteil et Baptiste Foulquié. Paris : Texto, 2006. Supplément de Texto ! septembre-décembre 2006 [en ligne], Vol. XI, n°3-4. Disponible sur : http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Actes_ALBI-06.pdf. (Consultée le 24/10/07).

Boufaden N., Lapalme G. et Bengio Y. (2002). Segmentation en thèmes de conversations téléphoniques : traitement en amont pour l'extraction d'information. In *Actes de Traitement Automatique de la Langue (TALN 2002)*. Juin 2002. Nancy, France. Tome I, pp. 377-382.

Boufaden N., Bengio Y. et Lapalme G. (2004). Approche statistique pour le repérage de mots informatifs à partir de textes oraux. In *Actes de Traitement Automatique de la Langue (TALN 2004)*. Avril 2004. Fez, Maroc, pp. 71-80.

Cailliau F. et de Loupy C. (2007). Aides à la navigation dans un corpus de transcriptions d'oral. In *Actes de TALN 2007*. Toulouse, pp. 143-152.

Maingueneau D. (2004). Retour sur une catégorie : le genre. In J.-M. Adam, J.-B. Grize et M. Ali Bouacha, *Texte et discours : catégories pour l'analyse*. Editions Universitaires de Dijon, pp.107-118.

Schmale G. (2007a). Communications téléphoniques I : Conversations privées. Un corpus de transcriptions. *Beiträge zur Fremdsprachenvermittlung*, 12, Sonderheft.

Schmale G. (2007b). Communications téléphoniques II : Conversations en contexte professionnel et institutionnel. Un corpus de transcriptions. *Beiträge zur Fremdsprachenvermittlung*, 12, Sonderheft.

Vergely P. (2004). *Analyse linguistique de l'expression du dysfonctionnement technique : le cas des échanges entre chefs de salle et maintenance opérationnelle dans la Navigation aérienne*. Thèse de doctorat, Université de Toulouse.