Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphoniques

Frederik Cailliau[*,**] **— Ariane Cavet**[*,***]

** Sinequa*
12, rue d'Athènes, F-75009 Paris

*** LIPN, Université de Paris 13 - Paris Nord*
99, avenue Jean-Baptiste Clément, F-93430 Villetaneuse

**** UFR de Linguistique, Université Paris 7*
30, rue du château des rentiers, F-75013 Paris

cailliau@sinequa.com, cavet@sinequa.com

RÉSUMÉ. *Dans cet article, nous présentons une méthode pour modéliser le déroulement de conversations provenant d'un centre d'appels. Le système d'analyse des sentiments prend en entrée des transcriptions automatiques, ce qui rend la détection d'entités moins fiable à cause des inévitables erreurs de transcription. Nous évaluons la dégradation subie en termes de précision et de rappel sur un corpus manuellement annoté. Pour y faire face, nous avons défini un grand nombre d'entités évaluatives et de modalités à extraire, auxquelles nous avons attribué des poids d'intensité. Lors du compte de polarité pour chaque tour de parole, les entités neutres sont comptées avec celles à forte polarité. Le plus haut score étant gagnant, cette modélisation permet de visualiser le déroulement émotionnel de la conversation par des courbes positif et négatif.*

ABSTRACT. *This paper presents a way to model conversational speech from call centers. Sentiment analysis on speech transcripts is less reliable because of the unavoidable mistakes made by the automatic transcription. We evaluate the degradation in terms of precision and recall on a manually annotated corpus. To counter it, we defined a great number of evaluative and modality entities to be extracted, and weighted them on intensity. When counting the polarity score for each speech turn, neutral entities are counted with the entities having a strong polarity. For each speech turn, the highest score is taken. This way of processing allows us to represent the emotional course of the conversation by positive and negative curves.*

MOTS-CLÉS : *analyse des sentiments, parole conversationnelle, détection des modalités, fouille de texte.*

KEYWORDS: *sentiment analysis, conversational speech, modality detection, text mining.*

1. Introduction

De nos jours, les centres d'appels sont devenus une interface importante entre le client et l'entreprise. Les grandes quantités d'informations qu'ils véhiculent intéressent en premier lieu les départements commerciaux et marketing. Soucieuses d'améliorer la relation avec le client, les entreprises s'intéressent également au contenu émotionnel de ces conversations. Dans le cadre du projet VoxFactory[1], nous développons une méthode pour sélectionner les conversations présentant un intérêt pour une analyse humaine plus profonde. Les acteurs des centres d'appels, et en premier lieu les téléconseillers mêmes, sont intéressés par toute information qui leur donne un retour sur l'interaction avec le client, dans le but de s'autoévaluer et d'améliorer le service.

Les méthodes du traitement de l'écrit que nous appliquons s'adaptent plutôt bien à la recherche et à la fouille de données dans des corpus de transcriptions automatiques. Le système développé dans le projet Infom@gic ST2.31 (Garnier-Rizet, 2008), en est la preuve : il enregistre, transcrit, analyse et rend accessibles les conversations d'un centre d'appels d'EDF. Son interface donne un accès multimodal aux conversations (Cailliau et Giraudel, 2008) et visualise de multiples informations statistiques sur les conversations (Garnier-Rizet *et al.*, 2010). Grâce à la modélisation du déroulement émotionnel de la conversation, nous mettrons à disposition de l'utilisateur des critères de sélection de conversation supplémentaires qui indiquent si la conversation s'est bien ou mal passée. Cette modélisation est faite par analyse textuelle de transcriptions automatiques. Plus tard dans le projet, nos résultats seront croisés avec les résultats de l'analyse émotionnelle du son (Vaudable *et al.*, 2010).

La détection d'entités avec des patrons morphosyntaxiques reste néanmoins très dépendante de la qualité de la transcription automatique. Sur un corpus de transcriptions d'émissions de radio, Cailliau et Loupy (2007) ont constaté une dégradation significative mais acceptable des groupes nominaux extraits pour la navigation, portant le nombre de groupes nominaux mal formés de 5 % à 10 %. Le taux d'erreur de transcription des mots (WER) est néanmoins bien plus élevé sur les conversations que sur les émissions de radio. Le WER sur ces émissions est de moins de 20 % (Gauvain *et al.*, 2002) et atteignait même les 11,9 % lors de la campagne ESTER (Galliano *et al.*, 2005).

L'adaptation du même système au traitement de la parole conversationnelle a permis de réduire un taux d'erreur initial de 51 % à 21 % avec un temps de 18,9 fois le temps de signal (Gauvain *et al.*, 2004). Or, d'après les résultats d'un test indépendant sur 10 heures de conversations téléphoniques semblables aux nôtres[2], c'est-à-dire provenant d'un centre d'appels EDF, le taux d'erreur varie de 27 %

1. Le projet VoxFactory, labellisé par le pôle de compétitivité Cap Digital, est financé par le FUI6.
2. Fait lors d'Infom@gic ST2.31sur un sous-ensemble des données du projet.

pour les téléconseillers à 33 % pour les clients. Si l'on peut espérer des progrès dans la transcription de ce type de données dans les années à venir, les erreurs de transcription poseront toujours un défi pour l'extraction des connaissances. Dans cet article, nous évaluons la dégradation du système de détection d'entités et proposons une méthode de modélisation de la conversation qui en tient compte.

Après un état de l'art sur la détection des expressions évaluatives dans les conversations orales, nous situons le cadre théorique choisi et le lexique préexistant. Ensuite, nous passons en revue l'ensemble des entités que nous détectons. Nous évaluons alors l'influence des erreurs de transcription sur les entités que nous extrayons, et exposons la méthode que nous mettons en œuvre pour pallier le nombre élevé d'extractions non pertinentes dues aux erreurs de transcription. Avant de conclure nous donnons quelques exemples de conversations et leur visualisation sur un graphe projeté sur deux échelles différentes : le tour de parole et le temps.

2. L'analyse des sentiments

Les travaux en analyse des sentiments connaissent leur essor au début des années 2000 avec un grand nombre d'articles publiés sur ce sujet. Cette floraison d'activité est directement liée à l'avènement du Web 2.0 : désormais les internautes ont la possibilité d'exprimer en quelques lignes leur opinion sur des produits, des films, etc. sur le site de l'éditeur ou du comparateur. Ils écrivent des évaluations qui sont en général courtes et leur attribuent des annotations comme une note (par exemple : 3/5) ou un nombre d'étoiles. Cela fait de ces pages des corpus parfaits pour générer des modèles de classification avec des méthodes d'apprentissage.

Les indices lexicaux sont extraits automatiquement, ce qui donne de bons résultats sur des domaines spécifiques. Néanmoins, Pang *et al.* (2002), après avoir testé trois différentes méthodes d'apprentissage sur des critiques de films, concluent que les résultats sont bons, mais qu'ils n'égalent pas ceux obtenus en classification thématique. Introduisant un peu plus d'analyse linguistique, les indices de Turney (2002) sont des bigrammes correspondant à des patrons morphosyntaxiques prédéfinis (comme par exemple *adjectif nom* et *adverbe verbe*), utilisés pour classifier avec des résultats partagés des avis de consommateurs : de 84 % d'avis bien classifiés pour les voitures à seulement 66 % pour les films. Dave *et al.* (2003) ont fait varier les indices lexicaux entre unigrammes, bigrammes, trigrammes et sous-chaînes (des n-grammes de mots d'une longueur maximale sélectionnés simplement selon un seuil de fréquence). Si les trigrammes sont bien meilleurs que les unigrammes et les bigrammes, une légère amélioration est encore obtenue avec les sous-chaînes. Ce résultat est assez intéressant en ce que les sous-chaînes se rapprochent le plus de ce qu'on pourrait attendre dans un lexique-grammaire constitué à la main.

En France la détection de l'opinion a récemment fait l'objet des ateliers DEFT de 2007 et de 2009 (http://deft.limsi.fr/). Les participants de DEFT 2007 ont

travaillé sur quatre types de textes : critiques de spectacles, tests de jeux, relectures d'articles scientifiques et débats sur des textes de loi. Les textes de ces corpus ont été annotés avec des valeurs d'opinion. L'équipe du LIA était classée première à DEFT 2007 avec un système fusionnant les résultats de plusieurs classifieurs (Torres-Moreno *et al.*, 2007). En 2009 il s'agissait de classifier des articles de journaux selon leur subjectivité et de détecter des passages subjectifs dans les documents. La première tâche a été gagnée par l'équipe de l'UCL (Bestgen et Lories, 2009) avec un classifieur SVM, la seconde par le LINA (Vernier *et al.*, 2009a) avec une approche mixte symbolique et statistique.

Pour une vue d'ensemble sur le domaine de l'analyse des sentiments, Pang et Lee (2008) et Tang *et al.* (2009) donnent une synthèse complète de son développement. Ils font également état des autres problèmes à surmonter comme par exemple la distinction entre énoncés subjectifs et objectifs (Pang et Lee, 2004).

Une caractéristique des méthodes d'apprentissage est que les modèles qu'elles engendrent sont très liés à leurs corpus d'entraînement, et plus particulièrement à leurs domaines. Certains travaux comme (Pang *et al.*, 2002) indiquent que la constitution automatique des indices lexicaux donne de meilleurs résultats qu'en utilisant des lexiques constitués manuellement. Ceux-ci souffrent principalement d'une couverture insuffisante. Pour pallier cela, ces lexiques peuvent être constitués avec des méthodes d'apprentissage avant d'être retravaillés comme décrit dans (Wiebe, 2000) et plus récemment dans (Wiebe et Riloff, 2005). Pour rendre le lexique moins dépendant du domaine, il existe des méthodes de *bootstrapping*, qui consistent à créer un lexique adapté plus large à partir des « graines » que sont les entrées du lexique existant (Turney et Littman, 2003 ; Riloff et Wiebe, 2003 ; Whitelaw *et al.*, 2005). Il existe cependant aussi des méthodes automatiques comme celle mise en œuvre par (Esuli et Sebastiani, 2005). Les lexiques ainsi obtenus sont alors généralistes et utilisables dans des contextes moins spécifiques, comme dans l'analyse des blogs.

On trouve principalement deux cadres théoriques en vigueur en analyse des sentiments, même si la plupart des travaux ne font référence à aucun cadre théorique linguistique. Le premier cadre est la théorie de l'évaluation (*Appraisal Theory*), publiée par (Martin et White, 2005) dans la suite de (Halliday, 1994). Elle repose sur quatre types d'attributs qu'ont les adjectifs : attitude (affect, appréciation, jugement), graduation (force, focus), orientation (positif, négatif) et polarité (marqué, non marqué). Ces quatre types ont ensuite de nombreuses options qui permettent de classer finement les adjectifs. Cette théorie a été mise en pratique dans (Whitelaw *et al.*, 2005 ; Bloom *et al.*, 2007a ; Bloom *et al.*, 2007b). La seconde théorie est celle des états mentaux (*Private States Theory*), mise au point par Quirk *et al.* (1995). Les états mentaux recouvrent les opinions, les croyances, les jugements, l'évaluation, les pensées et les sentiments. Cette théorie a été mise en pratique dans (Wiebe *et al.*, 2005 ; Breck *et al.*, 2007 ; Somasundaran *et al.*, 2006), avec de légères adaptations.

Le cadre théorique choisi pour l'annotation dans Blogoscopie[3] s'inspire néanmoins directement de (Charaudeau, 1992), dont la théorie se situe dans la lignée des théories de l'énonciation, référant notamment à (Benveniste, 1970). Ce cadre a été mis au point précédemment sur des critiques de film dans (Charaudeau, 1988). Sur les douze types de modalités que propose Charaudeau (1992) et qui façonnent le discours, cinq expriment une évaluation. Une opinion est exprimée lorsque le locuteur évalue la vérité de son propos et révèle son point de vue. L'appréciation présuppose un fait sur lequel le locuteur donne son sentiment, donc une valeur affective. L'accord/désaccord présuppose un message adressé au locuteur qui demande son adhésion, que le locuteur confirme ou non. L'acceptation/refus présuppose une demande d'accomplissement d'un acte auquel le locuteur répond favorablement ou non. Le jugement est exprimé lorsque le locuteur déclare son approbation ou sa désapprobation à propos d'une action réalisée par son interlocuteur. Ces cinq modalités peuvent avoir plusieurs degrés.

L'approche adoptée dans le projet de recherche Blogoscopie pour la constitution du lexique a été d'annoter d'abord un corpus de billets de blogs (Dubreil *et al.*, 2008). Ces annotations ont ensuite servi de base pour la constitution d'un lexique-grammaire par l'équipe de Sinequa contenant 982 entrées dont essentiellement des adjectifs (493), des verbes (192), des noms (166) et des adverbes (60). Il s'agissait d'un travail de synthèse des annotations du corpus et d'enrichissement pour étendre la couverture. Ce lexique d'évaluations, dont quelques entrées sont illustrées dans le tableau 1, est à la base des extractions des opinions dans les blogs par catégorisation automatique, mises en œuvre par (Vernier *et al.*, 2009b). Il a également été le point de départ pour nos travaux d'analyse émotionnelle des conversations.

Entrée lexicale	**Évaluation + exemple (thématique dans laquelle l'évaluation a été exprimée)**
bien sûr	**Accord/désaccord : accord total**
(adverbe)	Bien sûr (soleil, autobronzants)
	Oui, bien sûr (Russie, démocratie)
	Opinion : conviction
	bien sûr que (planche à découper le saucisson)
	Bien sûr que cela a été commandité (assassinat, Russie)
	bien sûr vous pouvez utiliser (décaféiné, ingrédients)

3. Les partenaires du projet Blogoscopie : le LINA de l'université de Nantes-Atlantique, l'hébergeur de blogs Over-Blog et Sinequa. Blogoscopie est un projet ANR, appel 2006, commencé en décembre 2006, d'une durée de 24 mois.

médiocre	**Appréciation : défavorable**
(adjectif)	médiocre (résultats de foot)
mort de rire	**Appréciation : favorable**
(adjectif)	mdrrrrrrrr (roman, auteur)
	mdr (dictionnaire de mots inventés)
nul doute que	**Opinion : supposition certitude forte**
(conjonction)	nul doute que (névrose, charognards)
pétard mouillé	**Appréciation : défavorable**
(nom)	est un peu un pétard mouillé (film)
protester	**Accord/désaccord : désaccord**
(verbe)	protestaient contre (interdiction)
probablement	**Opinion : supposition certitude moyenne**
(adverbe)	vient très probablement (explication pour les grèves)
	sont probablement (interprétations : associations)
	Ou bien plus probablement (milieux : chape de plomb)
refuser	**Acceptation/refus : acceptation**
(verbe)	Je ne refuse jamais (gâteau au chocolat)
	Acceptation/refus : refus
	On refuse (diplôme)
	je refuse (armement nucléaire)
	Accord/désaccord : désaccord
	refuse (privatisation des universités)

Tableau 1. *Extrait du lexique d'évaluations du projet* Blogoscopie

Au cours des dernières années, l'objet d'étude est passé de textes courts et monothématiques à des textes plurithématiques tels les billets de blogs généralistes. On cherche également à distinguer les passages objectifs des passages subjectifs, à identifier l'objet de l'évaluation et son émetteur, à résumer et à visualiser les résultats. Pour nos travaux, le type de textes étudié nous simplifie, en quelque sorte, la tâche. En effet, nous restons dans un seul et même domaine. Le vocabulaire des conversations n'est pas très diversifié, et les interlocuteurs sont les émetteurs des

messages. Les erreurs faites par la transcription automatique ajoutent néanmoins une difficulté bien particulière.

Les travaux d'analyse des sentiments sur de la parole conversationnelle ne sont pas nombreux. C'est souvent le corpus qui fait défaut comme le signale à juste titre l'appel à contributions du troisième workshop EMOTION[4] à LREC 2010 : la plupart des corpus auraient une durée de moins de 30 minutes et leur annotation ne serait pas optimale. Dans ce domaine, Hollard *et al.* (2005) et Tomokiyo *et al.*, (2005) ont pu travailler sur un corpus d'enregistrements de messages laissés sur le répondeur de l'assistance informatique d'un hôpital public. Le corpus correspond à 5 h 30 de parole. Les auteurs ont étudié les marqueurs lexicaux et phonologiques (ton, débit) et, d'après leurs observations, les deux types ne correspondent pas toujours. Devillers et Vasilescu (2004) combinent les indices lexicaux, dialogiques et prosodiques dans un corpus de 5 000 tours de parole sélectionnés dans des appels d'un centre de transactions boursières. Chaque tour de parole est annoté avec une émotion : colère, peur, neutre, satisfaction, excuse. L'annotation des émotions est faite par un groupe de quarante personnes, dont seulement la moitié a accès aux fichiers audio. Elle bénéficie d'un taux d'accord de 55 % entre les deux groupes, montrant ainsi l'importance des indices lexicaux. Avec les indices lexicaux, les auteurs obtiennent un taux de détection des émotions de 70 % pour les cinq émotions, et 85 % si on réduit les émotions à positif et négatif.

Nos travaux visent à modéliser le déroulement émotionnel d'une conversation en s'appuyant sur les expressions utilisées par les interlocuteurs tout au long de la conversation. Nous nous appuyons sur toutes les expressions indiquant des émotions ou des opinions, car elles nous donnent une indication sur le sentiment des interlocuteurs. Le sentiment est ensuite exprimé sous la forme d'une polarité positive et négative.

Les corpus que nous avons à notre disposition ont des tailles bien supérieures à ceux de l'état de l'art : 350 heures et 1 000 heures ont été enregistrées dans les centres d'appels d'EDF et transcrites dans le cadre des projets Infom@gic ST2.31 et VoxFactory, en utilisant les technologies issues du LIMSI-CNRS et de Vecsys Research.

3. Types d'entités extraites

Nous avons essayé de respecter autant que possible le cadre théorique hérité du projet Blogoscopie. Vu le caractère conversationnel des données, ce cadre nous semble même mieux adapté qu'à l'analyse des blogs. En effet, toute expression évaluative prononcée peut être considérée comme subjective dans le contexte d'un centre d'appels. En revanche, nous avons dû adapter la moitié du lexique et donc les

4. *Third International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*.

grammaires associées aux entrées lexicales. Celles-ci sont au nombre d'environ un millier. L'extrait du lexique original (tableau 1, ci-dessus) illustre les raisons de l'adaptation : certaines entrées sont typiques du langage écrit (*nul doute que*) ou du langage Internet (*mdr*) et d'autres ont peu de chances d'apparaître dans les conversations (*pétard mouillé*). De l'autre côté, nous avons ajouté des entrées typiques du langage conversationnel (voir § 3.1). Cette adaptation a été faite manuellement à partir de tests sur les transcriptions automatiques des appels et une exploration intensive du corpus oral.

Nous avons apporté une légère modification à la liste des modalités évaluatives décrites par Charaudeau et extraites pour le projet Blogoscopie : la classe d'entités *jugement* a disparu, et nous avons ajouté une classe d'entités *surprise*. La suppression de cette classe est due au constat que, dans le cadre de l'analyse de conversations issues de centres d'appels, le *jugement* et l'*appréciation* sont très difficilement différenciables, tant du point de vue du lexique employé (« nul », « catastrophique », etc.) que du point de vue de l'intention du locuteur. En effet, l'approbation et la désapprobation du client sont en général très fortement liées à un sentiment positif ou négatif suite à une action de la part de l'entreprise. Le client a alors tendance à identifier le téléconseiller à l'entreprise, ce qui se traduit par l'emploi du pronom personnel « vous » : « vous faites barrage », « vous m'enverrez la note ». Ces observations nous ont encouragés à verser le contenu de *jugement* dans *appréciation*, et à supprimer la classe d'entités *jugement.*

L'ajout de la classe d'entités *surprise* est dû à un autre constat réalisé lors de l'écoute des enregistrements des conversations. Dans nos conversations, la surprise s'exprime le plus souvent par l'emploi du mot « étonnant ». Le cas prototypique la place plutôt du côté du client, puisque c'est lui qui appelle avec une question. Or, parfois c'est le client qui apporte des informations qui engendrent un sentiment de surprise chez le téléconseiller. Il arrive même que les deux interlocuteurs expriment un sentiment de surprise au sein d'une même conversation. Ces appels non prototypiques traduisent généralement d'un certain déséquilibre entre les interlocuteurs dont les rôles « informateur » et « informé » sont échangés. Ainsi, bien que la *surprise* ne soit pas originellement une modalité évaluative, nous l'avons ajoutée à la liste des types d'entités extraites. Le mot « étonnant » revient assez souvent.

Deux types d'entités n'ont pas d'entrées lexicales correspondantes pour l'instant. Notre liste se veut homogène et symétrique, mais dans les faits, une simple reconnaissance lexicale ne suffit pas toujours à reconnaître toutes les émotions. Par exemple, l'acceptation émotive dans *acceptation/refus* pourrait être exprimée par un « oui » très accentué sur le plan prosodique. Cependant, l'insertion de ce mot dans le lexique sans tenir compte de la prosodie engendrerait trop de bruit : il relève très souvent de la fonction phatique du langage (Jakobson, 1963), et sert donc uniquement à valider le fait qu'un message est correctement parvenu à son destinataire.

Nous avons construit une grammaire d'extraction pour chaque type d'entités. La technologie utilisée est propre à Sinequa et s'appelle TMA (*Text Mining Agent*). Elle permet d'exprimer des patrons textuels avec des critères multiniveaux (expressions régulières, lemmes, catégories grammaticales) par des automates à états finis comme dans Intex (Silberztein, 1993) et Unitex (Paumier, 2002). Il est en outre possible, comme dans les langages de programmation, d'instancier et de manipuler des variables. Une entité n'est donc rien d'autre qu'un mot ou une suite de mots qui correspond aux critères définis dans une grammaire d'extraction.

Nous détectons cinq classes d'entités en plusieurs degrés, totalisant seize types d'entités que nous présentons et définissons dans cette section. Chaque type d'entités a un équivalent émotif, qui indique une forte implication émotionnelle du locuteur et porte le nombre de types d'entités extraites à trente-deux. Cette implication émotionnelle peut s'exprimer sous diverses formes selon le type d'entités : enthousiasme, soulagement, compassion, colère, gêne, méfiance, etc. Elle est marquée par le fait que le message exprime autant l'état émotionnel du locuteur qu'il véhicule du contenu informatif, comme le montrent les exemples qui accompagnent les définitions.

3.1 *Appréciation*

L'*appréciation* se présente sous la forme d'une polarité *favorable*, ou *défavorable*. La polarité *favorable* indique l'expression par le locuteur d'une satisfaction face à un fait ou un objet, comme on peut le voir à travers l'exemple émotivement marqué « **merci beaucoup** », ou dans l'exemple neutre « c'est **intéressant** ». La polarité *défavorable* indique l'expression d'une certaine insatisfaction face à un fait ou un objet, comme c'est le cas dans l'exemple émotivement marqué « J'en ai **ras le bol** ! » ainsi que dans l'exemple neutre « Je trouve que c'est **excessif**. »

3.2 *Acceptation/refus*

L'*acceptation/refus* se présente également, sous forme de polarité : *acceptation* ou *refus*. Il s'agit donc de l'acceptation ou du refus du locuteur face à une proposition faite par son interlocuteur. Comme stipulé précédemment, nous n'avons pas d'exemple émotif pour illustrer l'*acceptation*. L'*acceptation* neutre se reconnaît dans des phrases telles que « C'est **d'accord**. ». Le *refus* émotivement marqué apparaît dans l'exemple « C'est **hors de question** ! » tandis « Je suis vraiment **réticent**. » relève d'un *refus* non émotivement marqué.

3.3 *Accord/désaccord*

L'*accord/désaccord* se présente sous la forme d'une gradation qui s'étend de l'*accord total* exprimé par le locuteur envers son interlocuteur au *désaccord*, en passant par l'*accord approximatif*, qui exprime un accord un peu moins franc que dans le cas de l'*accord total*, et la *rectification*, qui est une correction apportée par le locuteur sur un propos de son interlocuteur qu'il juge erroné ou insuffisant. Un locuteur qui répond « Ah ça **j'imagine** » à son interlocuteur exprime son *accord total*, tout en s'impliquant émotivement, tandis que s'il répond « **Bien entendu** », cet *accord total* ne sera pas marqué émotivement. L'*accord approximatif* est reconnu dans des phrases comme « Je **comprends**… », dans laquelle le locuteur s'implique émotivement, sans pour autant vouloir mettre autant d'enthousiasme dans sa réponse que dans le cas de l'*accord total* émotif. Avec « **Certes**… », le locuteur ne s'implique pas émotivement, et exprime une certaine retenue. La *rectification* s'exprime dans des phrases telles que « Je **ne** vous **dis pas le contraire** ! » dans le cas où le locuteur s'implique émotivement, ou « **Tout de même** ! » dans le cas inverse. Dans ces deux dernières phrases, le locuteur ne peut se satisfaire des propos tenus par son interlocuteur et s'apprête à y apporter une rectification. Enfin, le *désaccord* est identifiable sous sa forme émotivement marquée dans une phrase comme « C'est vraiment **n'importe quoi** » et sous sa forme neutre « C'est **pas cohérent** ce que vous me dites ».

3.4 *Opinion*

L'*opinion* désigne ici le degré de certitude du locuteur. Celle-ci se présente donc sous la forme d'une gradation qui va de la *conviction*, qui exprime une certitude absolue du locuteur, au *doute* qui, au-delà de l'absence de certitude, va jusqu'à mettre en doute la véracité d'un fait. Dans « Je **vous garantis que** je l'ai fait », le locuteur exprime sa *conviction*, dans le but de convaincre son interlocuteur, et s'implique émotivement, contrairement à la réaction « **Évidemment** », qui traduit aussi une *conviction*, dans laquelle le locuteur ne s'implique pas émotivement. Le *doute* se rencontre dans des phrases telles que « Il l'a fait, **soi-disant**… » dans laquelle le locuteur s'implique, ou « Maintenant, **je m'interroge** », dans laquelle le locuteur ne s'implique pas émotivement. La supposition s'exprime en divers degrés intermédiaires : *supposition certitude forte*, dans lequel le locuteur exprime un fort degré de certitude, comme dans l'exemple émotivement marqué « J'**en ai bien peur**… », ou l'exemple neutre « **Visiblement** ça a été fait. », *supposition certitude moyenne*, dans lequel le locuteur exprime un degré de certitude moyen, comme dans l'exemple émotivement marqué « J'**espère que** ça va marcher », ou son équivalent neutre « Je **crois que** ça va marcher », ou *supposition certitude faible*, dans lequel le locuteur exprime un degré de certitude faible, comme dans l'exemple neutre suivant : « Il **y a peu de chances que** ça marche ». Ici, aucun exemple émotif n'a été retenu.

3.5 *Surprise*

Notre dernière classe d'entités est la *surprise*, c'est-à-dire la réaction du locuteur face à un fait qui lui apparaît comme nouveau. Cette réaction peut être de trois types, qui illustrent les trois sous-types pour la classe *surprise*. La *surprise positive* représente une réaction positive du locuteur face à ce fait nouveau, comme dans l'exemple émotivement marqué « **Bingo** ! », ou non marqué « En voilà une **surprise** ! ». Au contraire, la *surprise négative* indique une réaction négative du locuteur face à ce fait nouveau, comme dans l'exemple émotif « **Aïe** ! », ou dans l'exemple non émotivement marqué « Il y a une **anomalie**. ». Enfin, la surprise peut déclencher une réaction chez le locuteur, mais le terme ou l'expression utilisés ne permettent pas de déterminer si cette réaction est positive ou négative. C'est par exemple le cas quand un locuteur prononce la phrase « J'en **reste baba** ! » dans laquelle il s'implique émotivement, ou « Ça c'est **étonnant** ! », dans laquelle il ne s'implique pas. Nous avons nommé ce sous-type la *surprise neutre*.

4. Évaluation de la dégradation de la détection des entités

Pour évaluer l'impact des erreurs de la transcription automatique sur la détection des entités évaluatives, nous avons pris douze conversations que nous avions sélectionnées au cours de la découverte du corpus parce qu'elles concentrent un nombre élevé de ces entités.

Les douze conversations représentent presque 3 heures de parole. Comme le signal des vingt premières secondes de chaque conversation a été anonymisé pour des raisons de confidentialité, nous ne les avons pas prises en compte dans cette évaluation. Cela porte le nombre de mots total de ce corpus d'évaluation à 34 635 mots.

Nous avons calculé pour chaque fichier le WER général et le WER sur les mots qui représentent des entités à extraire. Pour ce faire, nous avons corrigé les transcriptions automatiques pour obtenir des transcriptions de référence et calculé le WER selon la formule [1], où S est le nombre de substitutions, D le nombre de suppressions, I le nombre d'insertions et N le nombre total de mots du corpus.

$$WER = \frac{S + D + I}{N} \quad [1]$$

Afin de calculer la dégradation de la détection des entités nous avons annoté les transcriptions de référence avec les entités extraites par nos grammaires. Cette annotation automatique sert de référence pour la suite des travaux. Nous avons ensuite lancé la détection automatique sur les transcriptions originales et mesuré la précision [2] et le rappel [3].

$$Précision = \frac{\#ExtractionsCorrectes}{\#Extractions} \quad [2]$$

$$Rappel = \frac{\#ExtractionsCorrectes}{\#EntitésDeRéférence} \quad [3]$$

Nous affichons les résultats dans le tableau 2 ci-dessous. La différence entre 100 % et les valeurs de précision et de rappel ainsi obtenue est la baisse causée par la transcription automatique.

	Nbre mots	**WER**			**Nbre entités**	**Détection**	
		Général	Hors entités	Entités		Préci-sion	Rappel
1	1 629	34	33	15	51	96	86
2	1 803	44	41	72	83	72	52
3	5 867	39	38	36	166	77	73
4	1 201	32	31	28	44	88	87
5	5 873	24	23	35	223	88	72
6	3 955	44	43	38	91	81	67
7	2 714	33	32	31	78	76	78
8	790	24	24	23	23	91	91
9	666	21	19	33	28	96	86
10	708	35	32	55	40	96	65
11	5 640	48	47	39	187	80	69
12	3 789	38	37	31	132	79	73
Moy.		37	32	47		83	72

Tableau 2. *WER, précision et rappel de la détection d'entités*

Sur notre sélection, la moyenne du WER général est assez élevée par rapport à l'état de l'art. Elle n'est néanmoins pas très représentative car on constate de grandes disparités entre les fichiers individuels. La moyenne du WER sur les entités

est encore plus élevée, mais elle n'est pas représentative non plus : sept conversations sur douze ont un meilleur WER sur les entités que sur les autres mots. Il n'existe donc pas de lien systématique entre le WER sur les entités et celui sur les autres mots ou les mots en général.

Il convient de rappeler ici que nous traitons des données en provenance de centres d'appels, et que le son n'est pas toujours de bonne qualité. La ligne peut être assez bruitée, notamment quand les clients appellent depuis un portable, et le son se dégrade également quand les locuteurs élèvent la voix. De plus, les conversations ont été enregistrées sur un seul canal et, quand le client et le téléconseiller parlent en même temps, il est impossible d'obtenir de bonnes transcriptions, alors que nous avons transcrit la voix qui prenait le dessus.

Les erreurs de transcription influent énormément sur les performances de la détection des entités, agissant à la fois sur le rappel et sur la précision.

La baisse du rappel indique que la transcription automatique contient moins d'entités que la référence. L'origine de cette baisse vient du fait que la transcription a remplacé des mots qui correspondaient à un patron à extraire par des mots qui ne correspondaient pas à ce patron. Nous donnons quelques exemples de ce phénomène dans le tableau 3.

Phrase de référence	**Transcription automatique**
tout à fait	tout ça c'est
C'est **débile débile débile**	c'est des billes délit débité
votre recouvrement de **merde**	votre recouvrement de mais
ça m'a coûté deux cents euros ces **conneries**	ça m'a coûté deux cents euros économique
vous les avez envoyé **chier**	vous qui est d'abord oui j'ai

Tableau 3. *Exemples d'entités non repérées à cause d'une erreur de transcription*

La baisse de la précision indique qu'une partie des entités reconnues dans les transcriptions automatiques le sont à tort. La transcription a donc remplacé du texte qui ne contenait pas de patrons à extraire par du texte qui en contenait. Nous donnons quelques exemples de ce phénomène dans le tableau 4.

Phrase de référence	Transcription automatique
l'installation électrique est aux normes	l'installation électrique **étonnant**
là là ça y est maintenant moi je vous mets	la place et **malheureusement** je mets
vous avez qu'à marquer un post-it	vous avez qu'à marqué **impossible**
il y a y a y a	il y **aïe aïe aïe**
c'est ma passion	c'est **pas possible**

Tableau 4. *Exemples d'entités repérées à cause d'une erreur de transcription*

Nous avons également identifié un phénomène marginal : dans un nombre très limité de cas, la transcription a remplacé du texte correspondant au patron d'une entité par du texte correspondant au patron d'une autre entité. Nous donnons les quatre exemples trouvés dans notre corpus dans le tableau 5.

Phrase de référence	Transcription automatique
pas malin	**pas mal**
mais **n'importe quoi**	mais **d'accord** ce sera
avec quelqu'un qu'on **aime**	avec quelqu'un **quand même**
Je vous **engueule**	je vous **embête**

Tableau 5. *Exemples d'entités mal repérées à cause d'une erreur de transcription*

Les coefficients de corrélation linéaire entre le WER et les mesures de précision et de rappel confirment nos constats. Le coefficient est de – 0,92 pour la corrélation entre le rappel de la détection des entités et le WER sur les mots correspondant à un patron d'extraction. Le coefficient est de – 0,65 pour la corrélation entre la précision de la détection des entités et le WER sur les mots ne correspondant pas à un patron d'extraction. Il est normal que ce second coefficient soit plus bas : le coefficient prend également en compte la mauvaise transcription de mots qui ne sont

pas transcrits en patrons correspondant à des entités. Ces corrélations sont visualisées dans la figure 1.

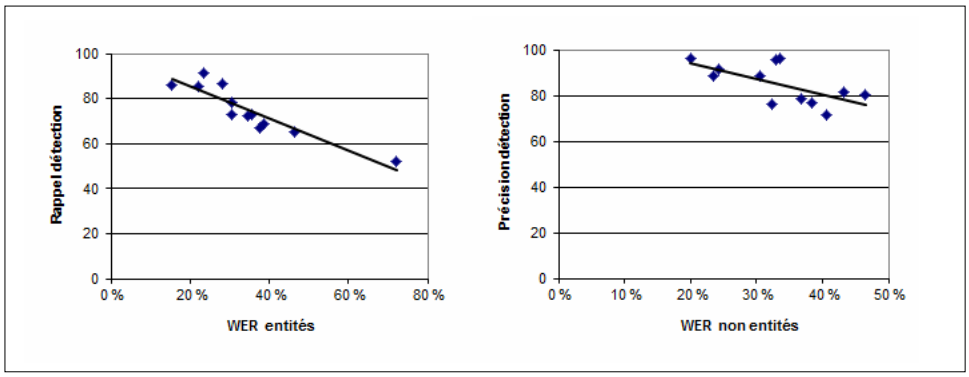

Figure 1. *Corrélations entre WER et rappel (gauche), et WER et précision (droite)*

5. Modélisation de la conversation

5.1 *Méthode*

L'objectif est de modéliser le déroulement émotionnel d'une conversation en polarités positive et négative. Comme vu ci-dessus, l'identification des expressions évaluatives est très vulnérable aux erreurs de transcription. Il est donc impossible de se fier aveuglement aux entités évaluatives détectées pour simplement compter le nombre d'entités qui manifestent explicitement un événement positif ou négatif. Pour cette raison, nous avons développé la méthode de modélisation suivante.

Nous commençons par détecter les entités décrites dans la section 3. Nous partons du principe qu'une densité élevée de ces entités indique un passage intéressant et que certaines entités sont plus importantes que d'autres. Nous avons classé les types d'entités en positif, négatif et neutre, et attribué à chaque type un score d'intensité que nous avons déterminé et affiné de façon empirique. Les entités les moins prononcées et les neutres obtiennent un score de 1, alors que celles qui se trouvent aux extrémités de la polarité positif/négatif obtiennent un score de 2. Ces scores sont multipliés par deux si l'expression est considérée comme émotive. Cette attribution de poids est illustrée dans le tableau 6.

		Polarité	Poids non émotif	Poids émotif
Appréciation	Favorable	positive	2	4
	Défavorable	négative	2	4
Acceptation/refus	Acceptation	positive	2	4
	Refus	négative	2	4
Accord/désaccord	Accord total	positive	2	4
	Accord approx.	positive	1	2
	Rectificatif	négative	1	2
	Désaccord	négative	2	4
Opinion	(tous)	neutre	1	2
Surprise	Positif	positive	2	4
	Négatif	négative	2	4

Tableau 6. *Poids d'intensité et polarités assignés aux types d'entités*

Chaque tour de parole obtient un score de polarités positif et négatif, grâce à la somme des poids d'intensité des entités rencontrées. Les poids d'intensité des entités neutres sont alors ajoutés au plus haut score entre les scores de polarités positive et négative. De cette façon, la densité locale des modalités exprimées renforce la polarité du tour de parole.

Nous illustrons ce calcul en l'appliquant sur l'exemple de la figure 2, dans lequel nous avons mis en gras les entités détectées.

> […] bon je vous **embête** parce que c'est bien pour rien mais **quand même** c'est **scandaleux suppression** de de **considérer** gens moi je suis **désolé** il y a des gens qui […]

Figure 2. *Extrait d'un tour de parole (client)*

Le détail de la détection d'entités est donné dans le tableau 7. Le nom de chaque entité est spécifié, avec sa polarité et son poids d'intensité. Les poids d'intensité

sont additionnés et donnent un score de polarité négative de 19, auquel on a ajouté le score de polarité neutre de 1, puisque le score de polarité positive est de 0. Le patron *désolé* fait en effet partie de deux grammaires et compte deux fois dans le calcul.

Patron	Entité	Polarité	Intensité
embête	Appréciation : défavorable, émotif	négative	4
quand même	Accord/désaccord : rectificatif	négative	2
scandaleux	Appréciation : défavorable, émotif	négative	4
suppression	Appréciation : défavorable	négative	2
considérer	Opinion : supposition certitude forte	neutre	1
désolé	Appréciation : défavorable, émotif	négative	4
désolé	Acceptation/refus : refus	négative	2

Tableau 7. *Entités reconnues dans l'exemple de la figure 2*

Pour adoucir l'impact des détections isolées et pour passer du tour de parole au passage, nous prenons la moyenne sur une fenêtre glissante, selon la formule donnée en [4].

$$f_{(p)} = \frac{1}{1+2L} \sum_{i=p-L}^{p+L} v_{(i)} \qquad [4]$$

Ce calcul a l'effet d'un filtre passe-bas : il atténue les valeurs élevées tout en réduisant fortement le bruit causé par des valeurs basses. Nous l'appliquons sur une fenêtre de cinq tours de parole, donc la demi-longueur L de la fenêtre est égale à 2, ce qui est la meilleure valeur d'après nos observations. Ce calcul permet de visualiser le déroulement émotionnel de la conversation sous forme de courbes positive et négative.

Afin d'obtenir une meilleure vision du déroulement de la conversation, nous avons projeté cette représentation sur une échelle temporelle, grâce au minutage des tours de parole dont nous disposions dans les transcriptions automatiques.

Les deux types de courbes, sur échelle du tour de parole et sur échelle temporelle, sont illustrés ci-dessous.

5.2 *Exemples*

Nous avons sélectionné les exemples 2 et 3 de notre corpus annoté pour illustrer notre modélisation. L'exemple 2 (figure 3) est une conversation qui est très émotionnelle au début et qui se normalise en différentes étapes. Elle dure environ 10 minutes. Elle commence par une cliente très énervée, qui mène un échange très déséquilibré avec la téléconseillère. Celle-ci reste très calme et posée, en dépit de l'agressivité subie. Au douzième tour de parole, la cliente passe le combiné à une personne de son entourage. Elle est un peu moins agressive mais la situation reste tendue. Elle emploie des expressions comme « pas normal », « pas logique ». Cette personne se calme peu à peu, et au trente-huitième tour de parole la téléconseillère appelle un tiers, professionnel également, pour vérifier les propos de la cliente. Les sept tours de parole suivants, au cours desquels aucune entité négative n'est détectée, représentent la prise de contact avec le tiers. Cette conversation dure jusqu'à ce que l'enregistrement soit coupé. Pendant toute la conversation, la téléconseillère est restée calme.

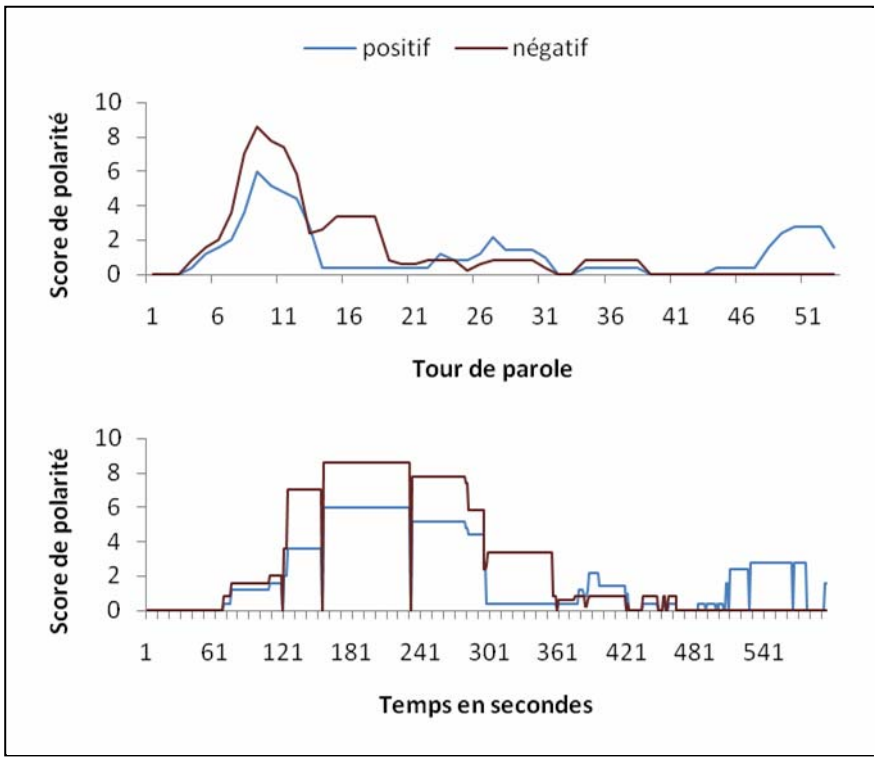

Figure 3 : *Déroulement émotionnel d'une conversation, sur échelles de tour de parole et de temps (exemple 2)*

L'appel de l'exemple 3 (figure 4) provient d'une cliente dont le fils a raté le rendez-vous avec le technicien. La conversation dure un peu moins de 30 minutes. La majeure partie de la communication se passe entre des téléconseillères afin de fixer un rendez-vous, avec quelques courts allers-retours avec la cliente pour lui proposer des créneaux horaires. Celle-ci parle d'un ton excédé tout au long de la conversation, mais son temps de parole limité n'a pas permis de faire émerger des passages même si son vocabulaire était très négatif : « arnaque », « insupportable », « trucs de fou », « société de fou », « inadmissible », « impossible », « scandaleux ». À la fin, après la prise finale du rendez-vous, la cliente prend la parole, la monopolise jusqu'à la fin de l'appel et se plaint de la situation en employant le même vocabulaire qu'auparavant.

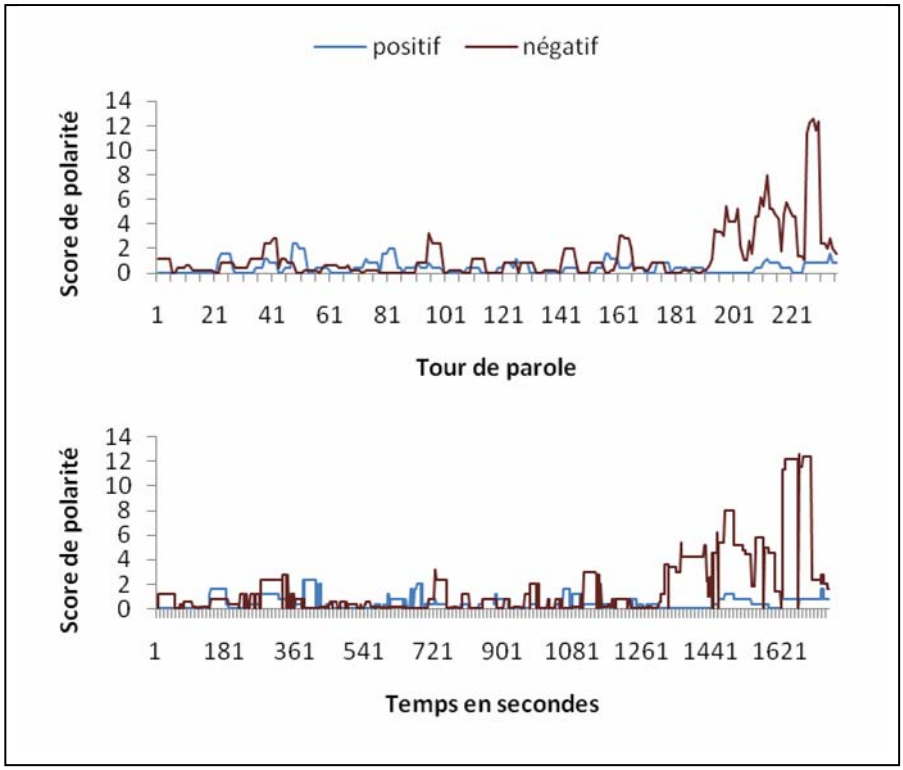

Figure 4 : *Déroulement émotionnel d'une conversation, sur échelles de tour de parole et de temps (exemple 3)*

L'exploitation d'une telle modélisation est évidente : elle peut être transformée en critères de sélection avec quelques simples heuristiques. Il sera par exemple possible de chercher toutes les conversations qui se sont globalement mal passées, ou bien celles qui ont mal commencé mais bien fini, ou l'inverse.

D'après nos premières observations, chaque conversation dont le score de polarité négative dépasse le seuil de 4 présente un intérêt pour les services d'amélioration de la qualité de la relation avec le client. Avec ce seuil, le système sélectionne 178 conversations sur 7 564, soit 2,35 % des conversations du corpus. Cette heuristique choisit des conversations qui représentent donc des pics d'activité émotionnelle, mais ignore celles qui ne se passent pas très bien dans l'ensemble. Ces conversations peuvent être repérées par deux autres heuristiques que nous expérimentons également. La première repère une courbe négative qui est relativement haute sur toute la conversation : sur un certain pourcentage (par exemple un tiers) de la conversation, le score de polarité dépasse un seuil à définir. La seconde heuristique exploite la fréquence relative des expressions émotives dans les conversations : au-delà d'un certain seuil, la fréquence d'expressions émotives dénote la présence d'une émotivité supérieure à la normale.

Ces trois méthodes restent à évaluer plus précisément, mais sont complémentaires pour sélectionner différents types de conversations problématiques : celles contenant un pic d'activité émotionnelle et celles dont l'activité émotionnelle est importante mais répartie sur l'ensemble de la conversation.

6. Conclusion et perspectives

Si les méthodes de traitement de l'écrit s'adaptent bien au traitement de l'oral, certains types de données posent un vrai défi. Dans cet article, nous avons évalué la dégradation que subit un système de détection d'entités quand il est confronté à des transcriptions automatiques de conversations. Nous avons ensuite proposé une méthode d'exploitation des entités qui visualise le déroulement émotionnel d'une conversation tout en intégrant activement cette dégradation. Les entités évaluatives que nous détectons sont des modalités qui cadrent dans une théorie plus large d'énonciation. Nous utilisons cette détection très générale dans un but très spécifique.

Nos travaux ont montré qu'il est possible de modéliser le déroulement d'une conversation téléphonique par une détection du sentiment exprimé par les interlocuteurs, même sur des transcriptions automatiques. Au cours de nos expérimentations, l'affinage des poids a été assez délicat. Il est apparu qu'il fallait différencier les expressions en fonction de leur charge émotive pour mieux capter l'état d'esprit des interlocuteurs : cela fait la différence entre mécontentement et colère, entre satisfaction et soulagement. En outre, il nous était impossible d'attribuer une polarité positif/négatif aux entités d'opinion pourtant importantes

pour modéliser le sentiment. En effet, il aurait fallu une analyse fine de l'objet sur lequel porte l'opinion pour déterminer sa polarité de façon automatique. Cette analyse étant dépendante du domaine et du métier, nous ne nous sommes pas engagés dans cette voie.

Les prochaines étapes seront de transformer la modélisation du déroulement de la conversation en critères de sélection pour un moteur de recherche et d'évaluer la pertinence des extraits et des conversations ainsi sélectionnés. Celles-ci pourront ainsi être cherchées et analysées par des professionnels comme les téléconseillers eux-mêmes ou leurs superviseurs. Le but de notre modélisation est en effet de sélectionner, dans une grande masse de données, les conversations qui peuvent être utilisées dans une démarche d'amélioration de la relation entre le client et le téléconseiller.

La sélection des conversations problématiques en combinaison avec une extraction des groupes nominaux à la volée permettra d'identifier grossièrement les thèmes abordés dans les passages qualifiés de problématiques. Ce sera un début d'identification de la raison de l'énervement du client et pourra, après analyse, servir à l'amélioration du service en général.

À l'avenir, l'évolution technique permettra l'enregistrement de conversations téléphoniques sur deux canaux, ce qui donnera une meilleure qualité de transcription, la possibilité d'exploiter le taux de recouvrement et de paramétrer notre analyse selon que le locuteur est le client ou le téléconseiller. Cela devrait nettement augmenter les performances générales du système.

Remerciements

Les auteurs remercient Éliane Cheung et Mélodie Soufflard pour leur participation à l'évaluation, ainsi que EDF R&D et Vecsys pour la mise à disposition des corpus de transcriptions.

7. Bibliographie

Benveniste, E., « L'appareil formel de l'énonciation », in *Langages*, vol. 5/17, p. 12-18, 1970. Repris dans *Problèmes de linguistique générale, II*, Gallimard, 1974, p. 79-88.

Bestgen Y., Lories G., « Un niveau de base pour la tâche 1 (corpus français et anglais) de DEFT'09 », *Actes de l'atelier de clôture du 3e DEfi Fouille de Textes*, DEFT 2009, Limsi, 2009.

Bloom K., Garg N., Argamon, S., « Extracting appraisal expressions ». *Proceedings of the human language technology conference of the North American chapter of the association of computational linguistics (HLT-NAACL 2007)*. Rochester, New York, USA, 2007a, p. 308-315.

Bloom K., Stein S., Argamon S., « Appraisal extraction for news opinion analysis at NTCIR-6 », *Proceedings of the sixth NTCIR workshop meeting on evaluation of information access technologies : Information retrieval, question answering, and cross-lingual information access*, National Institute of Informatics, Tokyo, Japan, 2007b, p. 279-285.

Breck E., Choi Y., Cardie C., « Identifying expressions of opinion in context », R. Sangal, H. Mehta, and R. K. Bagga (Eds.) *International Joint Conference On Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, CA, 2007, p. 2683-2688.

Cailliau F., de Loupy C., « Aides à la navigation dans un corpus de transcriptions d'oral », *actes de TALN 2007*, Toulouse, 2007, p. 143-152.

Cailliau F., Giraudel A., « Enhanced Search and Navigation on Conversational Speech », *Proceedings of Searching Spontaneous Conversational Speech (SSCS 2008)*, SIGIR 2008 workshop, Singapour, 2008.

Charaudeau P., « La critique cinématographique : faire voir faire parler », *La Presse, Produit, Production, Reception*, Paris, Didier Érudition, 1998, p. 47-70.

Charaudeau P., *Grammaire du sens et de l'expression*, Paris, Hachette Éducation, 1992.

Dave K., Lawrence S., Pennock D., « Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Review », *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

Devillers L., Vasilescu I., « Détection des émotions à partir d'indices lexicaux, dialogiques et prosodiques dans le dialogue oral », *actes de JEP*, Fez, 2004.

Dubreil E., Vernier M., Monceaux L., Daille B., « Annotating opinion – evaluation of blogs », *Proceedings of the LREC workshop on Sentiment Analysis : Metaphor, Ontology and Terminology (EMOT-08)*, Marrakech, 2008.

Esuli, A., Sebastiani, F., « Determining the semantic orientation of terms through gloss classification », *Proceedings of the 14th ACM international Conference on information and Knowledge Management (CIKM'05)*, ACM, New York, NY, 2005, p. 617-624.

Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J.-F., Gravier G., « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News », *Proceedings of the European Conf. on Speech Communication and Technology (Interspeech)*, Lisbonne, 2005.

Gauvain J.-L., Lamel L., Adda G., « The LIMSI Broadcast News Transcription System », *Speech Communication*, 37(1-2), 2002, p. 89-108.

Gauvain, J.-L., Adda G., Lamel L., Lefèvre F., Schwenk H., « Transcription de la parole conversationnelle », *Traitement automatique des langues*, vol. 45/3, Lavoisier, Paris, 2004, p. 35-47.

Garnier-Rizet M., Adda G., Cailliau F., Guillemin-Lanne S., Waast-Richard C., « CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content », *Proceedings of LREC 2008*, Marrakech, 2008.

Garnier-Rizet M., Cailliau F., Guillemin-Lanne S., « Search by Content, Navigation and Knowledge Extraction on Call Center Conversational Speech, for Marketing and Strategic Intelligence », *Proceedings of RIAO*, Paris, 2010.

Halliday M., *Introduction to Functional Grammar*, Edward Arnold, second edition, 1994.

Hollard S., Tomokiyo M., Tufelli D., « Une Approche de l'expression orale des émotions : étude d'un corpus réel », *actes des quatrièmes Journées de la Linguistique de Corpus*, Lorient, 2005.

Jakobson R., « Linguistique et poétique », *Essais de linguistique générale*, Paris, Éditions de Minuit, 1963, p. 215-217.

Martin J.R., White P.R.R., *The Language of Evaluation, Appraisal in English*, London & New York, Palgrave Macmillan, 2005.

Pang B., Lee L., Vaithyanathan S., « Thumbs up ? : sentiment classification using machine learning techniques. » *Proceedings of the Acl-02 Conference on Empirical Methods in Natural Language Processing, vol. 10*, Morristown, NJ, 2002, p. 79-86.

Pang B., Lee L., « A Sentiment Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cut », *Proceedings of ACL*, 2004, p. 271-278.

Pang B., Lee L., « Opinion Mining and Sentiment Analysis. », *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 2008, p. 1-135.

Paumier S., *Manuel d'utilisation d'Unitex.* Université de Marne-la-Vallée, 2002.

Quirk R., Greenbaum S., Leech G., Svartvik J., A comprehensive grammar of the English language, Harlow : Longman, 1985, p. 1779.

Riloff E., Wiebe J., « Learning extraction patterns for subjective expressions. » *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, vol. 10*, Theoretical Issues In Natural Language Processing, ACL, Morristown, NJ, 2003, p. 105-112.

Silberztein, M., *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, Paris, 1993.

Somasundaran S., Wiebe J., Hoffmann P., Litman D., « Manual annotation of opinion categories in meetings. » *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, ACL Workshops, ACL, Morristown, NJ, 2006, p. 54-61.

Tang H., Tan S., Cheng X., « A survey on sentiment detection of reviews. » *Expert Systems with Applications*, vol. 36/7, 2009, p. 10760-10773.

Tomokiyo M., Chollet G., Hollard S., « Studies of emotional expressions in oral dialogues : towards an extension of Universal Networking Language », Jesús Cardeñosa, Alexander Gelbukh, Edmundo Tovar (eds.) : *Universal Networking Language : advances in theory and applications*, Mexico City, 2005.

Torres-Moreno J-M., El-Bèze M., Béchet F., Camelin N., « Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? Application au défi DEFT 2007 », *actes de l'atelier de clôture du 3e DEfi Fouille de Textes*, DEFT 2007, AFIA, 2007.

Turney P. D., « Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews. », *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, 2002, p. 417-424.

Turney P. D., Littman, M. L., « Measuring praise and criticism : Inference of semantic orientation from association », ACM Trans. Inf. Syst. 21, 4 (Oct), 2003, p. 315-346.

Vaudable C., Rollet N., Devillers L., « Annotation of affective interaction in real-life dialogs collected in a call-center », *Third International Workshop on EMOTION*, LREC Workshop, Malta, 2010.

Vernier M., Monceau L., Daille B., « DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique », *Actes de l'atelier de clôture du 3e DEfi Fouille de Textes*, DEFT 2009, Limsi, 2009a.

Vernier M., Monceaux L., Daille B., Dubreil E., « Catégorisation des évaluations dans un corpus de blogs multi-domaine », *Revue des Nouvelles Technologies de l'Information (RNTI)*, RNTI-E-17, p. 45-70., 2009b.

Whitelaw C., Garg, N., Argamon, S., « Using appraisal groups for sentiment analysis », *Proceedings of the 14th ACM international Conference on information and Knowledge Management (CIKM '05)*, ACM, New York, NY, 2005, p. 625-631.

Wiebe J., « Learning Subjective Adjectives from Corpora », *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on innovative Applications of Artificial intelligence (AAAI/IAAI)*, AAAI Press, 2000, pp. 735–740.

Wiebe J., Wilson T., Cardie C., « Annotating expressions of opinions and emotions in language », *Language Resources and Evaluation*, vol. 39/2-3, 2005, p. 165-210.

Wiebe J., Riloff, E., « Creating Subjective and Objective Sentence Classifiers from Unannotated Texts », *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05)*, Invited Paper, Springer LNCS vol. 3406, 2005, Springer-Verlag.