Lang Resources & Evaluation
DOI 10.1007/s10579-013-9224-5

ORIGINAL PAPER

Spontaneous speech and opinion detection: mining call-centre transcripts

Chloé Clavel · Gilles Adda · Frederik Cailliau · Martine Garnier-Rizet · Ariane Cavet · Géraldine Chapuis · Sandrine Courcinous · Charlotte Danesi · Anne-Laure Daquo · Myrtille Deldossi · Sylvie Guillemin-Lanne · Marjorie Seizou · Philippe Suignard

Abstract Opinion mining on conversational telephone speech tackles two challenges: the robustness of speech transcriptions and the relevance of opinion models. The two challenges are critical in an industrial context such as marketing. The paper addresses jointly these two issues by analyzing the influence of speech transcription errors on the detection of opinions and business concepts. We present both modules: the speech transcription system, which consists in a successful adaptation of a conversational speech transcription system to call-centre data and the information extraction module, which is based on a semantic modeling of business concepts, opinions and sentiments with complex linguistic rules. Three models of opinions are implemented based on the discourse theory, the appraisal theory and the marketers' expertise, respectively. The influence of speech recognition errors on the information extraction module is evaluated by comparing its outputs on manual versus automatic transcripts. The F-scores obtained are 0.79 for business concepts detection, 0.74 for opinion detection and 0.67 for the extraction of relations between

C. Clavel (✉) · C. Danesi · P. Suignard
EDF R&D, 1 Avenue du Général de Gaulle, 92141 Clamart, France
e-mail: chloeclavelpro@gmail.com

G. Adda
LIMSI, Université Paris XI, Bât 508, BP 133, 91403 Orsay Cedex, France

F. Cailliau · A. Cavet
Sinequa, 12 rue d'Athènes, 75009 Paris, France

M. Garnier-Rizet · G. Chapuis · A.-L. Daquo
Vecsys, 3 rue Terre de Feu, 91940 Les Ulis, France

S. Courcinous
Vocapia Research, 3 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France

M. Deldossi · S. Guillemin-Lanne · M. Seizou
TEMIS, 207 rue de Bercy, 75012 Paris, France

Published online: 04 April 2013

opinions and their target. This result and the in-depth analysis of the errors show the feasibility of opinion detection based on complex rules on call-centre transcripts.

Keywords Call-centre data · Automatic speech recognition system · Opinion detection · Business concept detection · Disfluency

1 Introduction

A key challenge of speech processing is to give computers the ability to understand human behavior. The input is low-level information provided by audio samples, which can be very hard to process in the context of human-to-human interactions, such as phone calls for example.

Some approaches focus on the analysis of speech signal. Acoustic features such as prosody, voice quality or spectral features are used in order to develop acoustic emotion recognition systems (Clavel and Richard 2011; Devillers et al. 2010). However, the issue of information extraction on speech is more globally tackled according to the point of view of natural language processing methods focusing on named entities detection and information retrieval. Research has unraveled many aspects concerning this issue with various evaluation campaigns driven in these two fields, for instance the ESTER2 campaign for named entities detection (Galliano et al. 2009), or the TREC 7—Spoken Document Retrieval, SDR—(Garofolo et al. 1999). However, such campaigns are mainly based on broadcast news and have not yet tackled the issue of information extraction on phone conversations, in which spontaneous speech features are more frequent. Moreover, the performance of speech recognition systems falls down on such data and information extraction is thus more difficult. Other approaches, such as the one described in Olsson et al. (2007), search keywords directly in the acoustic signal or in phonetic transcriptions. They can offer solutions to handle speech recognition errors but are difficult to use for the detection of subtler information than keywords such as opinions and sentiments.

Alongside these works on speech transcripts, sentiment analysis and opinion mining on texts are research fields that have been blooming since the year 2000. This is mostly due to the apparition of a new type of corpus: the interactive web. Users comment the products they have bought, review the films they have seen and make their opinions public. The web sites usually equally foresee in a starred notation, which makes the user comments' sites a perfect learning corpus. An overall overview on sentiment analysis and its evolution can be found in Pang and Lee (2008) and Tang et al. (2009). Several methods are in use to distinguish positive from negative. Pang et al. (2002) automatically extract the linguistic clues from movie reviews and have tested three learning methods to classify them. They conclude that if the results are satisfying, they are not as good as the usual text categorization tasks. The clues used by Turney (2002) are bigrams extracted by predefined morpho-syntactic patterns (like *adjective* + *noun* and *adverb* + *verb*). The results are 84 % of good categorizations of product reviews and 66 % of film reviews. When running experiments, Dave et al. (2003) find out that the length of the n-gram should be optimally tuned to optimize the categorization. The longer the

string, the more discriminatory it is, but the less frequent, reducing its relevance. By building a tree of substrings up to a cutoff length, they obtain 88.5 % of accuracy. Linguistically motivated approaches select the linguistic clues manually: a linguist goes through one or different corpora, tags the patterns and makes a lexicon out of it. To counter the critics that these lexicons suffer from an insufficient coverage as expressed in Pang et al. (2002), some approaches build the lexicons automatically. The lexicons are then processed by linguists (Wiebe 2000; Wiebe and Riloff 2005). Others prefer bootstrapping methods using the "seeds" of an existing lexicon to adapt a lexicon to the corpus (Turney and Littman 2003; Riloff and Wiebe 2003; Whitelaw et al. 2005). Since they are less adapted to one type of corpus, these lexicons are reputed to be more suitable for the analysis of general texts, such as blogs, which are not monothematic. This reflects the shift observed during the last years on the type of texts studied in sentiment analysis: from monothematic text types like film and product reviews to multi-thematic ones like blogs and newspaper editorials.

In this paper, the adaptation of opinion mining methods on spontaneous speech transcripts is investigated in the difficult context of call-centre data. The present approach contributes to an important challenge because analyses are driven in a real industrial context with call-centre data provided by the French power supply company EDF.[1] It is motivated by the crucial role that opinion content plays for marketing applications. Indeed, the mining of call-centre data is strategic. It contributes to improve customer insight, and therefore develops loyalty. So far, given the large amount of available data, only few calls are listened to and therefore exploited. Audio analysis and automatic processing of recordings (known as speech analytics methods) provide an answer to this problem. We aim to extract and organize information contained in the phone interactions between EDF and its customers: what are the reasons for the call? How do customer needs and preoccupations evolve? Which opinion do customers have on the company?

The present study comes within the scope of the VoxFactory[2] collaborative project, which aims to analyze client/agent interactions in call-centre data, in continuation of the INFOM@GIC-CallSurf project (Garnier-Rizet et al. 2008). In both projects, the various partners address the needs of two types of users. Call centre supervisors get access to the automatic transcripts and their local statistics through a search and navigation tool (Cailliau and Giraudel 2008), whereas marketing units obtain a global overview of the underlying purposes of customer calls, by cross-topic correlations and topic evolutions in time. The CallSurf project focused on transcription models and the topic analysis of the call. A topic categorization module was thus developed (Bozzi et al. 2009). In the last phase of CallSurf project, a speech analytics pilot based on EDF recordings was developed. The goal of the VoxFactory project is to complement the processing chain behind this pilot with information extraction on opinion and emotion. This paper focuses on the main tool chain (Fig. 1), which consists of two modules: the speech transcription

[1] http://www.edf.com/the-edf-group-42667.html.

[2] VoxFactory is a project of Cap Digital, the French business cluster for digital content in Paris and the Ile de France region (http://www.capdigital.com/vox-factory/).

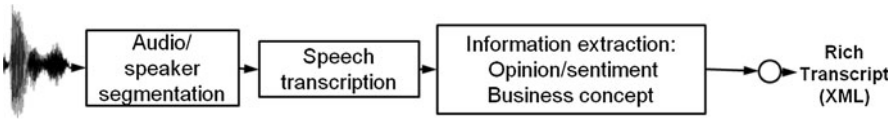

Fig. 1 Processing chain

system combined with the information extraction module. At the same time of this presented work, which focuses on natural language processing-based approaches, an acoustic emotion detection module has been developed (Devillers et al. 2010).

The first step when the conversations between the parties are recorded on the same channel is to separate client and agent speech turns, using automatic speaker segmentation and tracking module. Then an automatic speech-to-text conversion is performed and finally, the information extraction engine mines and extracts, from transcriptions, the required information to index conversations and cluster them into categories. The information extraction relies on a semantic modeling of business concepts, opinions and sentiments.

In this paper, we tackle all the various steps involved in the development of an opinion detection system from the perspective of an industrial application, looking at the joint propagation of errors through the processing chain from the speech transcription to the opinion detection. Our main objective is to grasp the issues of such a strategy. With this aim, the spontaneous speech features occurring in call centre data and their impact both on automatic transcription and on information extraction are highlighted. Two separate studies have been previously carried out on the impact of speech recognition errors on information extraction (Cailliau and Cavet 2010; Danesi and Clavel 2010). The present work extends these studies by providing a description of the speech recognition module, by detailing its behavior and by correlating it with the comparison of information extraction module outputs on manual versus automatic transcripts. We propose also new methods more suited to strong emotional corpora and provide an evaluation on a corpus containing strong emotional events likely to degrade performance and to raise new scientific issues. The paper is organized as follows. Firstly, the methods used for speech recognition and semantic analysis are presented in Sect. 2. The semantic analysis concerns business concept and opinion detection in the speech transcripts. The various corpora, on which the speech recognition system and the information extraction module has been built and tested, are detailed in Sect. 3. Then, we present an in-depth analysis of the evaluation results in Sect. 4. Finally, Sect. 5 unravels the various conclusions and proposes perspectives that can be drawn from this evaluation results.

2 From speech signal to information extraction

2.1 Speech recognition system and speaker segmentation on conversational speech

Transcribing conversational telephone speech is a very challenging task. It has been one of the focal tasks in annual speech recognition benchmarks organized by NIST,

using the SwitchBoard (SWB) family of resources distributed by the LDC (Godfrey et al. 1992). The benchmark tests have demonstrated many of the difficulties encountered in automatic processing of conversational speech (Matsoukas et al. 2002; Stolcke et al. 2000; Ljolje et al. 2000; Hain et al. 2000; Gauvain et al. 2003). The Vocapia research company[3] has worked with the LIMSI academic laboratory[4] in order to develop the transcription system described in this paper.

On the acoustic side, we need to cope with channel variability, and to develop efficient speaker adaptation and accurate pronunciation modeling. Different types of telephone handset affect the speech quality: the background noise (other conversations, music, street noise, etc.), the high proportion of interruptions, overlapping speech (see Sect. 3.2.2) or side conversations.

But the main challenge is on the linguistic side, because conversational speech contains simultaneously three characteristics, which make the recognition process very different from—for instance—transcribing Broadcast: the conversational speech is simultaneously *spontaneous*, *interactive*, and *private*. The conversational speech is *spontaneous*: spontaneity involves various phenomena at the linguistic level. First, the varying speech rate may lead to reduced pronunciations. Second, the presence of disfluency phenomena, typical of spontaneous speech (repetitions, hesitations, see Sect. 3.2.2) (Adda-Decker et al. 2003) even more disrupts the syntactical structure of the message. The lexical and syntactic content is thus difficult to be learnt from text corpora. The conversational speech is *interactive*: during dialog acts, various phenomena may interfere in the speech recognition process (Ten Bosch and Boves 2004). Indeed, interactivity implies the presence of backchannel confirmations to let each interlocutor know that the other person is listening, the breaking off in the syntax, and the production of partial sentences. The conversational speech is *private*: a speaker commonly wants to be understood by one single listener. The latter, interacting with the former, may give feedback about his understanding at any moment. Under these conditions, there is a constant risk of deterioration in the quality of the message elaboration, including pronunciation and syntax, emphasizing the problem induced by the spontaneity of the message.

Another point particularly relevant in our context is variations of linguistic (at lexical and syntactic levels with disfluent speech and insult for example) and acoustic parameters due to *emotion*. Many errors on the speech segments, which contain intense (positive or negative) emotion, are due to the inadequacy of the acoustic and linguistic models.

In order to handle these difficulties, we work both on *acoustic models*, which describe the probabilistic behavior of the encoding of the acoustic–phonetic information in a speech signal, and on *language models*, which allow us to estimate the probability of word sequences. At the language model point of view, the challenge with conversational speech is also to cope with the limited amount of training data: the specificity of conversational speech makes the benefits of using texts or transcriptions coming from other areas (newspapers, texts from the web, etc.) very marginal.

We have adapted the LIMSI-Vocapia CTS (Conversational Telephone Speech) transcription system to the EDF call centre task, updating acoustic, lexical and

[3] http://www.vocapia.com/.

[4] http://www.limsi.fr/index.en.html.

language models using 150 h of task data with fast transcriptions (Garnier-Rizet et al. 2008). This resulting system serves as the baseline in this work. The strategy used for decoding is presented in Fig. 2.

Prior to word recognition, the segmentation step partitions data into different types of audio segments: non-speech segments are identified and removed and speech segments are labeled and clustered separately by gender in order to produce homogeneous clusters according to speaker and background conditions (Gauvain et al. 1998; Barras et al. 2006).

Two versions of the speech recognizer have been developed as described in Sect. 4.1. They differ on the data used for training, development and test of the speech recognizer, which are presented in Sects. 3.3 and 4.1.

Both versions rely on the same algorithm: the 4-g consensus word decoding is carried out in two passes, using tied-state acoustic models and unsupervised adaptation. The word transcription is then enriched with some features in the final version of the output, in order to make the automatic transcripts more readable on the interface and easier for information extraction. In the final XML transcription file, the system indicates for each recognized word its confidence score and its time

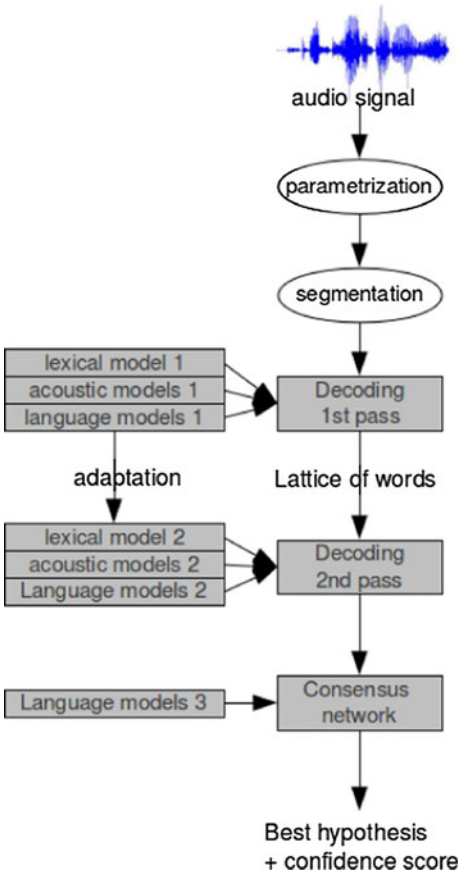

Fig. 2 Multipass decoding strategy with consensus network decoding

codes. The disfluencies are automatically annotated with special tags, denoting when breath and filler words are present in the audio signal. Finally post-processing is done: a specific language model is used to automatically punctuate the output text (inserting commas and final dots—the first letter of each sentence is thus uppercased) and to convert numbers into digits.

The performance of the speech recognition system is here evaluated by first using the usual Word Error Rate (*WER*), which measures the distance between a reference transcript and the hypothesis given by the system:

$$WER = \frac{S + I + D}{N} \times 100$$

where S is the number of substitutions (the reference word is replaced by another word), D is the number of deletions (a word in the reference transcription is missed), I is the number of insertions (a word is hypothesized that was not in the reference), and N is the number of words in the reference. In our application, all the words do not have the same importance.

Second, we have adapted the classical WER, in order to take into account only keywords, and defined a Keyword Error Rate (*KER*) (Park et al. 2008). The definition of KER is identical to the one of WER, using keywords as units. We take here as keywords the words that are involved in the extraction of business concepts and opinions. Keywords may thus be composed of more than one word. We build a list of 834 keywords dealing with contract, technical terms, invoicing, nuclear technology, thanking, prices, etc.

2.2 Semantic analysis of automatic transcripts

The next step consists of extracting information corresponding to business concepts and opinions from the previously obtained transcripts. Our module relies on the definition of semantic rules modeling the information to extract. This task is especially complex in the context of call-centre data. First, the speech recognition errors generated by spontaneous speech features, which are very frequent in conversational telephone speech, tend to bias text-mining analysis. Second, spontaneous speech features engender dysfunctions of linguistic and semantic analysis both at the morpho-syntaxic analysis level and at the information extraction level.

We present in the first paragraph the two different tools used in this paper for semantic modeling. The second paragraph focuses on the models used for business concepts. The third treats the models of opinions and sentiments, and the fourth is about the model of relationship extraction. The fifth paragraph shows an application of the sentiment analysis.

2.2.1 Semantic analysis tools

The semantic modeling used in this paper is based on two technologies: the Skill Cartridge from the TEMIS company[5] and the TMA (Text Mining Agent) from the

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

Sinequa company.[6] It relies on two main steps: the morpho-syntactic analysis and the semantic analysis.

The morpho-syntactic analysis carried out by both technologies is composed of the following steps:

- Tokenization: splits a text into tokens (basic lexical units and punctuation); tokens can be united to multiword expressions (complex lexical units) by the use of dictionaries;
- Segmentation in sentences: the punctuation is interpreted to distinguish sentence delimiters from other punctuation;
- Morphological analysis: a lemma and a Part-Of-Speech (POS) tag are given to each lexical unit by combining morphological rules and dictionary look-up. The POS-tag set is quite different, with 22 tags for Sinequa's and 45 tags for TEMIS' technology. The great difference in the number of tags is explained by the use of many different tags by TEMIS for one and the same Sinequa tag, e.g. Sinequa has one tag for the adjective (ADJ) where TEMIS has many (ADJ_INV, ADJ_SG, ADJ_PL, etc.);
- POS-tag disambiguation: if the lexical unit is grammatically ambiguous (e.g. the word "can" may be a verb or a noun), then it is disambiguated according to two different methods. TEMIS' technology is based on Hidden Markov Models through XeLDA® POS tagging tool,[7] whereas Sinequa has used a Brill model for the work in this article (Brill 1995). The model takes into account only the word part-of-speech categories and their possible sequences as they occur in the corpus.

The semantic analysis relies on lexicons and linguistic extraction patterns (or rules) which describe semantic concepts. These patterns combine different features obtained from the morpho-syntactic analysis (word form, lemma, case, POS tag) or from previous patterns matching. Thereby, the following description of the business concept *Duration* (see Sect. 2.2.2) gives an example of such a rule through the Skill Cartridge formalism:

```
(déjà|depuis|attendre|pendant) / ([0-9]+|(#NUM)+) / ~duree-lex | (un
          | quelque) / instant

| (rester) / #PREP / ligne
```

The pattern matches the following expressions: "j'ai attendu 2 heures" ("I had been waiting for 2 h"), "un instant" ("one moment please"), "restez en ligne" ("hold on"). It calls a POS tag (PREP) and a concept label (duree-lex) that has been defined within an inferior level pattern.

The extraction process consists in several levels. The analyzed text is successively tagged by replacing the matched text by the corresponding concept at each level. As a result, normalized metadata is added to the corresponding text segment, which we call concept labels.

[6] http://www.sinequa.com/.

[7] http://www.xrce.xerox.com/Research-Development/Historical-projects/XeLDA.

The format of expressing text-mining grammars is different for each tool, but their global functioning is the same. Skill Cartridges™ are expressed an XML-based proprietary formalism and TMAs are expressed in generic XML. The grammars are compiled into finite-state automata that respect a predefined execution order, so that previously extracted concept labels can be reused by other grammars. Intex (Silberztein 1994) and Unitex (Paumier 2002) are well-known academic tools that use finite-state automata in a similar way for text annotation by expressing linguistic patterns.

As output of the extraction process, TEMIS technology either produces a XCAS[8] document or stores the extracted information into a database containing a link to the original text, the document's original meta-data and the meta-data added by extraction. Sinequa's technology stores all information in its search engine indexes.

Both tools are used for the semantic modeling in this paper. More specifically, the business models and the marketers' opinion models, presented in Sect. 2.2.2 and in Sect. 2.2.3 respectively, rely on Skill Cartridge™ Technology. The latter is also applied to the implementation of the Appraisal theory of opinion, while TMA is used for the implementation of Discourse theory (Sect. 2.2.3). Skill Cartridge™ technology is also suited to model relationships between opinion and target through the Appraisal theory (Sect. 2.2.4).

2.2.2 Business concept detection through business models

The business concept models are built from an existing Skill Cartridge™ previously defined to analyze customer opinion on open-question inquiries (Danesi and Clavel 2010). We update it in two directions. First, we take into account spontaneous speech features and deal with speech recognition errors. Second, we define new concepts more relevant for call-centre data analysis such as the *Callback* concept, dedicated to identify calls coming from clients who previously were in contact with EDF for the same request. Table 1 stores examples of extracted business entities.

2.2.3 Opinions and sentiments detection through marketers' models, Discourse theory and Appraisal theory

Sentiment Analysis or Opinion Mining refers to the task of Natural Language Processing whose aim is to point out expressions reflecting the attitude of a speaker or a writer, and to characterize them according to an appraisal typology, from the most basic one (i.e. with only tonality positive/negative distinction) to the most advanced ones (i.e. with modality, attitude, force distinctions, etc.). The three partners, EDF R&D, Sinequa and TEMIS, have built three separate models dedicated to opinion automatic detection: the marketers' models (Table 2), the models based on the discourse theory (Table 3) and the models based on the appraisal theory (Table 5).

[8] XCAS is a format defined by Apache UIMA™ project (Unstructured Information Management Applications) http://uima.apache.org/.

Table 1 Examples of business concepts and associated detected entities (marketers' models)

Concept name	Entity example	English translation	Concept translation
Duration	une minute, s'il vous plaît	please wait a minute	Expressions used when the client is put on hold
	restez en ligne	hold on	
	cela a mis 15 jours	it took 15 days	
Competitors	GDF	GDF	List of competitors
	Poweo	Poweo	
	Suez	Suez	
	fournisseur	Supplier	
Contract	contrat	contract	Expressions related to the contract or to the offers
	souscription	subscription	
	heures creuses	off-peak	
Bill	consommation réelle	Actual consuming	Expressions related to the bill
	duplicate	duplicate	
	facture	billing	
	paiement	payment	
Price	Tarifs	Rates	Expressions related to price or comments concerning the price
	c'est cher	itisexpensive	
Technical	Relevé de compteur	Meter reading	Expressions related to technical field, such as electricity installations and technical help
	branchement	meter installation	
	coupure	power cut	
Callback	EDF m'a appelé	EDF call me	Expressions related to client's callback
	j'ai contacté EDF	I contacted EDF	
EDF	agent	Agent	Expressions related to EDF and EDF agents
	interlocuteur	contact agent	

The marketers' model focuses on concepts currently used by marketers: satisfaction and dissatisfaction concepts as presented in Table 2. These concepts are adapted from the customer satisfaction EDF Skill Cartridge™, as described in Sect. 2.2.2. The satisfaction concept has been broadened for a call-centre context. A list of about 300 words has thus been built for each concept to model dissatisfaction and satisfaction expressed in the corpus through linguistic rules. The marketers' opinion model is implemented through the Skill Cartridge technology, such as the business concepts presented in the previous section.

Table 2 The marketers' opinion model: satisfaction and dissatisfaction

Concept label	Entity example	English translation	Concept description
Satisfaction	Je suis satisfait, c'est parfait	I am satisfied, it's perfect	Expressions related to client satisfaction
Insatisfaction	ça m'énerve, cette situation ne peut pas durer indéfiniment	It's getting on my nerves, this situation should not go on forever	Expressions related to client dissatisfaction

The second model spots discourse features, which are mainly based on the evaluative modalities of the French specialist of discourse analysis (Charaudeau 1992) in the line of of Benveniste (1970). He proposes 12 modalities, of which 5 are evaluative: *opinion*, *appreciation*, *agreement*, *acceptation* and *judgment*. This distinction of modalities is used and adapted by Sinequa to detect five concept classes: *Agreement/Disagreement*, *Favorable/Unfavorable Appreciation*, *Acceptance/Refusal*, *Opinion* and *Surprise*. Sixteen concepts, detailed in Table 3 with examples of matching patterns, are used to specify these classes.

They can be polar (positive or negative), such as *Favorable* and *Unfavorable* in the *Appreciation* class, or scalar (according to the intensity), such as *Approximate Agreement* and *Total Agreement* in the *Agreement/Disagreement* class. The grammar covers slightly more than 1,000 patterns, which have been implemented in TMA grammars without formalizing them into a lexicon. Most of the patterns are simple verbs, adjectives, adverbs and nouns. About a fifth is made of more than a single word. In these complex patterns, part-of-speech tags are used to generalize their application, like in "ça commence ADV* mal" ("it starts ADV* bad"), where ADV* can be any repetition of adverbs. It is worth noting that half of the patterns have been recycled from an opinion lexicon totalling 982 entries built from a manually annotated multi-domain blog corpus (Dubreil et al. 2008). More details on the application of this model can be found in (Cailliau and Cavet 2010).

The third model relies on the Appraisal Theory (Martin and White 2005, Bloom et al. 2007), which analyzes the way opinion is expressed. According to Bloom, an evaluative expression has three main components: a *source*, which expresses an *evaluation* on a *target*. This theory is implemented through the Skill Cartridge™ technology by assigning the following attributes to each extracted evaluative expression:

- Evaluation type: affect or judgment; these two classes result from the adaptation of the Appraisal Theory to call center data.[9]
- Polarity: positive or negative;
- Intensity: strong, normal or low.

The Opinion Mining Skill Cartridge™ uses two evaluative dictionaries, built from Temis expertise: the idioms lexicon and the lexicon of evaluative terms and collocations. The idioms lexicon (about 500 idioms) contains set phrases such as "y mettre du sien" ("to work at it"), "avoir de l'argent à jeter par les fenêtres" ("have money to throw around"). The lexicon of evaluative terms and collocations contains about 3 100 entries. The major parts of speech represented are adjectives (ex: happy), nouns and noun phrases (ex: an aberration), verbs (ex: to please), adverbs and adverbial clauses (ex: happily) and interjections (ex: "well done!"). Attributes are first associated to each lexical item expressing its polarity, its intrinsic intensity and its intrinsic evaluation type(s). For adverbs and some degree adjectives only, a modifier type indicates whether it intensifies, minimizes or flips the initial polarity of the lexical item. The final polarity, intensity and evaluation are fixed at the level

[9] Appreciation and judgment have been gathered because this distinction is not suited to call-centre data.

Table 3 Description of opinions based on discourse theory

Concept class	Concept label	Example	English translation	Description
Acceptance—Refusal	Acceptance	**pourquoi pas** oui	**why not** yes	Speaker's acceptance to a proposal made by the interlocutor
	Refusal	je **refuse** de payer la somme qu'on me demande	I **refuse** to pay the amount demanded	Speaker's refusal to a proposal made by the interlocutor
Agreement—Disagreement	Total agreement	**tout à fait** le relevé compteur date du mois de novembre	**exactly** the meter reading dates from November	Speaker's absolute agreement with the interlocutor
	Approximate agreement	je le **conçois** j'ai compris la situation	I **hear** you I've understood the situation	Speaker's agreement with the interlocutor
	Amending	ce serait **plutôt** pour son appartement qu'il faudrait vérifier	you'd **rather** check for his apartment	Speaker's amending on an information considered as incorrect or insufficient
	Disagreement	je suis **pas d'accord**	I **don't agree**	Speaker's disagreement with the interlocutor
Appreciation	Favorable	Ça c'est **sympa**	that's **nice**	Speaker's satisfaction about a fact or an object
	Unfavorable	je trouve ça **inadmissible**	I think it's **unacceptable**	Speaker's dissatisfaction with a fact or an object
Opinion	Conviction	je vous dis **franchement** c'est trop pour moi	I tell you **straight out** this is too much for me	Speaker's expression of a fact that he considers as true
	Strong certainty	vous avez **sûrement** un fournisseur pour le gaz	you **certainly** have a gas supplier	Speaker's expression of a fact with a strong degree of certainty
	Medium certainty	c'est une estimation **je suppose**	it's an estimate **I assume**	Speaker's expression of a fact that he considers as likely to be true
	Low certainty	je vous **garantis pas** que ce soit ça	I **don't guarantee** that it is right	Speaker's expression of a fact that he considers as unlikely to be true
	Doubt	oui mais je **m'interroge** sur les chiffres	yes but I **wonder** about the figures	The speaker calls something into question

Table 3 continued

Concept class	Concept label	Example	English translation	Description
Surprise	Positive	vous allez avoir une **bonne surprise** depuis hier on l'a reçu	you will have a **good surprise**, we have received it since yesterday	Speaker's positive reaction on something new
	Neutral	ce qui **m'intrigue** c'est que la banque vous facture des frais	the thing which **intrigues** me is that the bank charges you fees	Speaker's reaction on something new, but the word or expression does not tell if the reaction is positive or negative
	Negative	C'est **bizarre** j'ai pas eu le courrier	it's **odd** that I've not received the mail	Speaker's negative reaction on something new

of the entire evaluative expression including the target of the evaluation such as described in Sect. 2.2.4.

The two last models rely on basically different theories—the Appraisal Theory and Charaudeau's evaluative modalities-, which are difficult to compare from a theorical point of view. It would therefore be interesting in another work focusing on opinion models evaluation to compare the variety of patterns used to detect the concepts defined by each theory.

2.2.4 Relationships between opinions and business concepts—the Appraisal theory

According to Bloom et al. (2007), an evaluative expression can be linked with a source, and/or a target (the object of the stance). To go further in the sentiment analysis, the Opinion Mining Skill Cartridge™ relies on this theory to link evaluative expressions with their target by modeling evaluative judgments or speaker emotional state towards products (electricity devices, apparatus, invoices, payments, etc.), persons (service quality level, kindness, efficiency, etc.), or situations. A semantic post-processing handles attributes attached to the lexical items forming the entire evaluative expression, computes the final polarity and intensity and defines the final evaluation type. Thereby, the expression "l'abonnement est euh ben oui plus cher" ("the contract is uh er yes more expensive") is analyzed in Table 4.

Table 5 stores illustrations of judgment and affect entities with their targets, as detected in the test corpus.

2.2.5 Adaptation to spontaneous speech

Dictionaries and linguistic rules have been adapted to take into account spontaneous speech phenomena, such as interjections and hesitation words ("comment dire", "euh", "bah", "ben", "oui", "hein", "ah", "oh", etc.) that can be inserted anywhere within the patterns. Repetition of grammatical words such as determiners and pronouns ("relevé de du de compteur"/"the reading of of of the meter") is also handled within flexible linguistic rules.

3 Call-centre data collection and annotation

3.1 Data collection

Today, most of call centre calls are not recorded on a regular basis but occasionally for training purpose. The CallSurf and VoxFactory projects gave to the partner Vecsys Company[10] the opportunity to collect, transcribe and annotate a significant amount of data recorded in two different EDF call centres. The partners have used the produced corpora for training, development and evaluation purposes. Before

[10] http://www.vecsys.fr/.

Table 4 The appraisal theory—opinion target detection in the sentence: "l'abonnement est euh ben oui plus cher"

Evaluation	Plus cher (more expensive)
Target	Abonnement (contract)
Polarity	Negative
Intensity	Strong
Opinion type	Jugdment
Target type	Product

starting the process of recordings, EDF R&D sent a statement to the CNIL[11] with the recording protocol.

Two recording campaigns have been conducted: the CallSurf campaign was dedicated to professional customers and the VoxFactory campaign to individuals. During the CallSurf campaign from summer 2006 to early 2007, a recording machine has been installed in one of the EDF call centres in Montpellier. Four seats that correspond to about ten agents have been recorded during 4 months to collect the CallSurf data. For the VoxFactory campaign, the same recording machine has been moved to one of the EDF call centres in Aix-en-Provence to record from December 2009 to February 2010. Recording conditions were the same as during the CallSurf experimentation. The VoxFactory data correspond to the recordings of 16 seats with a total of 36 agents.

For both data sets, the agents are recorded in their working conditions and equipped with their usual headset microphone. We observe a pretty good overall quality but it appears that some parts of the calls can be noisy for different reasons (GSM, free hands use, noisy environment etc.). In order to ensure the best audio quality, the signal is recorded at 64 Kb/s (wav format) and is not compressed. The composition of the calls is heterogeneous: waiting music, recorded messages, telephone rings and speech (dialogue and monologue). The collection is made of two kinds of calls. Besides the traditional scenario involving a client and an agent, sometimes a part of the conversation involves two agents while the customer is put on hold.

One last point to highlight is that clients and agents have been recorded on the same channel. Consequently, overlapping speech appears on the audio signal. This aspect has caused many problems for the manual transcribers and therefore for the transcription system. We will detail further how the overlapping speech has been transcribed and processed. In the future, we would like to investigate a recording of the two channels separately in order to overcome overlapping speech problems. Indeed, call centres will be more and more equipped with VoIP infrastructure with an easier separation of the two channels.

620 hours (5,755 calls) and 1,000 h (8,556 calls) have been collected during the CallSurf and VoxFactory campaigns, respectively. The duration of the calls goes from a few seconds to more than half an hour with a mean of around 7 min. The longest calls usually contain waiting music and silence segments corresponding to

[11] Commission Nationale de l'Informatique et des Libertés. The CNIL's general mission consists in ensuring that the development of information technology remains at the service of citizens and does not breach human identity, human rights, privacy or personal or public liberties.

Table 5 Examples of the relationships between opinions and business concepts extracted by the Opinion Mining Skill Cartridge™ based on the Appraisal theory

Opinion type/Intensity/Polarity	Opinion target			Description
	Situation	Product	Person	
Judgment/Normal/Negative	**Il faut** que ça soit **clair** *That must be clear*	la facture elle va **augmenter alors** *So the bill is going to increase*	Je sentais qu'il y avait **une entourloupe** *I felt there was a dirty trick*	Expression of a social evaluation
Affect/Strong/Negative	je commence **à en avoir sacrément marre** *I'm getting so god damn* fed up	Ils **m'énervent** avec leur **compteur EDF** *they get on my nerves with their electric meter*	il y a un **crétin** qui **m'a raccroché au nez** *a moron hung up on me*	Reference to an emotional state

the customer folder access. The calls, that are shorter than 15 s and longer than 30 min, are removed from the corpus.

3.2 Transcriptions and annotations

3.2.1 Fine and fast transcriptions of speech signal

We proceed to two kinds of transcription: fine and fast transcriptions. The first one is detailed with text and signal alignment at the sentence level or according to the breath in case of long sentences. The fine transcription is the usual transcription for acoustic model training and is mandatory for evaluation purpose (Table 6). It is carried out using Transcriber (Barras et al. 2000).

The fast transcript contains less information and allows us to get a larger amount of data in the same time span. Fast transcription enables us to increase the volume of training data, while maintaining a reasonable cost for the manual transcription process. We use a simple text editor or a patched version of Transcriber for this transcription mode; an example of a fast transcript is shown in the following example:

A/ EDF Pro bonjour *(EDF Pro good morning)*

C/ allô, oui bonjour société YY je vous contacte au sujet d'une facture suspecte *(Good morning, YY company, I'm contacting you about a suspicious bill)*

A/ oui bonjour pourriez-vous s'il vous plaît me fournir votre référence client? *(Yes, could you give me your customer reference please?)*

In this example, YY corresponds to an anonymized proper name. According to EDF requests, all the personal data such as client and agent names, bank account, credit card number, etc., are anonymized. They have been tagged on fine transcripts and directly anonymized on fast transcripts.

The transcription team is composed of two experienced persons who have worked on both campaigns (CallSurf and VoxFactory). The transcription ratio for fine transcription is about 20 (anonymization included), which means that the time needed to transcribe 1 h of signal is about 20 h. The transcription ratio for fast transcription is about 12. The Table 6 summarizes the difference between the two types of transcriptions.

370 hours have been extracted from the CallSurf data set to build a corpus of 20 h with fine transcripts (Call20-188 calls), a corpus of 150 h with fast transcripts (Call150 -1,268 calls) and a second corpus of 200 h with fast transcripts (Call200-1,548 calls).

The VoxFactory dataset has been used to extract a corpus of 50 h with fast transcripts (Vox50) and a specific corpus of 14 h (Vox14) with fine transcripts, whose main statistical features are presented in Table 7 and whose contents will be detailed in the next sections.

Table 6 Comparison between fine and fast transcriptions

Fine transcription	Fast transcription
Text and signal alignment	Text associated to each speaker turn
Tags (noise, breath, laugh …)	No additional tags
Pronunciation information	No pronunciation information
Anonymization (tags)	Anonymization (letters)
Punctuation and case sensitiveness	Punctuation and case sensitiveness

Table 7 Main statistical features of Vox14 in terms of calls, sections, speaker turns and segments

	Number	Mean	Duration (hh:mm:ss)
Calls	77		15:05:31
Sections	155	2.0/file	14:55:26
Turns	17,961	233.3/file	13:31:11
		115.9/section	
Segments	18,369	238.6/file	12:56:30
		1.0/turn	
Speakers	124	2.3/file	
Words	176,997	2298.7/file	
		9.6/segment	
		3.8/s	
Overlap			0:49:18

3.2.2 Manual annotations of conversational speech phenomena

The annotation driven for this paper concerns phenomena representative of telephonic conversational speech, that are overlapping speech and difluencies. This annotation aims to contribute to the analysis of the impact of these phenomena in the processing chain, which is carried out in Sect. 4.

For the overlapping speech annotation, four tags have been used:

- [bc]: this tag indicates backchannel speech: backchannel speech corresponds to interjections (hm, yes, OK,…) used by the interlocutor while the main speaker talks. When it is possible to isolate the words of each speaker, the signal is segmented and each intervention is transcribed. If the speaker's backchannel interventions overlap with the main speaker speech, making the segmentation impossible and the backchannel unintelligible, backchannels are annotated thanks to the tag [bc] inserted in the main speaker's words.[12]
- [Complementary]: this tag is inserted when the interlocutor speaks in the same time of the main speaker, making a short intervention to clarify what is said.

[12] This strategy for backchannel annotation has been chosen in a perspective of overlapping speech segment detection.

- [turn_request]: this tag is inserted when the interlocutor attempts to speak by interrupting the main speaker's speech.
- [overlap]: this tag is used when the last word of the first speaker is superposed on the first word of the second speaker.

The overlapping speech annotation has been carried out on the total transcribed corpus except the backchannel tags which are only used during the fine transcription.

The disfluency tagging consists in annotating speech disfluencies. The definition of Blanche-Benveniste (1990) is the most often cited to describe French disfluencies: a disfluency occurs when the syntagmatic progress is disrupted. Disfluencies include, for example, words and sentences that are cut off mid-utterance, phrases that are restarted or repeated, repeated syllables, grunts or unrecognizable utterances occurring as "fillers", and "repaired" utterances. We separate the two following parts of the disfluency: the "to-repair zone" (tag [dis = del-reg]) and the "corrected zone"; the annotation of this zone depends of the type of disfluency: repetitions (tag [dis = cor-rep]), revisions ([dis = cor-rev]), restart ([dis = cor-rest]), and complex disfluencies ([dis = cor-discomplex]), relying on the Linguistic Data Consortium convention[13] (Shriberg 1994). However, the interval (or break) between the "to-repair" and the "corrected" zones is not so far annotated, such as in the referring document. Indeed, the disfluency annotation task has been simplified for annotation cost reasons. The Table 8 presents examples of each disfluency type.

This annotation is carried out in a separate task on Vox5-neu-fine and Vox5-ang-fine adding a ratio of 6 to the transcription time versus recorded time ratio. Indeed, the disfluency tagging is a complex task. In order to limit annotation errors, the corpus has been first annotated once. Then, another annotator has validated the annotations. In particular, the main difficulties concern the difference between revision and restart, and the "to-repair zone" boundary. Table 9 presents the number of disfluencies for the two subcorpus. Table 10 sums up the various annotations available for each subcorpus.

3.2.3 Post-processing of transcriptions and annotations

The post-transcription and post-annotation processing is essential to guarantee the quality and the homogeneity of the corpus. It consists of a first step of normalization where the transcription structure is checked (encoding, tags, sections contents). Then out-of-vocabulary words are extracted, and validated or rejected. Dictionaries are finally updated. For instance, this allows to process phenomena such as: wrong encoding (ISO instead of UTF-8), unknown proper nouns (spelling to find), "mis-framed" tags, misspellings, etc.

[13] Simple Metadata Annotation Specification, Version 6.2—February 3, 2004 Linguistic Data Consortium www.ldc.upenn.edu/Projects/MDESimpleMDE.

Table 8 Examples of disfluencies

Disfluency	Example
Repetition	alors [-dis = del-reg] je [dis = del-reg-] [-dis = cor-rep] je [dis = cor-rep-] entre tout de suite la mensualisation
Revision	j'ai vérifié elle est pas [-dis = del-reg] de la [dis-del-reg-] [-dis = cor-rev] dans les documents en attente [dis = cor-rev-]
Restart	[-dis = del-reg] elle va récupérer, [-dis-del-reg-][dis = cor-rest] on résiliera celui-ci qui est toujours actif chez nous
Complex	[-dis = del-reg] je vais peut-être/je vais/euh/je dois [dis = del-reg-] [-dis = cor-discomplex] je dois [dis-discomplex-] je dois changer de numéro

Table 9 Number of disfluencies for each sub-corpus

	Vox5-neu-fine		Vox5-ang-fine	
	Total	Mean per call	Total	Mean per call
Repetition	454	13.75	712	26.37
Restart	238	7.21	331	12.25
Revision	385	11.66	671	24.85
Complex	204	6.18	538	19.92
Total	1,281	38.81	2,252	83.40
Number of files	33		27	

Table 10 Overview of the various corpora of the CallSurf and the VoxFactory campaigns

Name	Duration	Nb of calls	Fine trans.	Fast trans.	Disfluencies
EDF (Professionals)					
Call20-fine	20	98	x		
Call150-fast	150	1,268		x	
Call200-fast	200	1,548		x	
Call10-fine	10	90	x		
EDF (Residentials)					
Vox14-fine					
Vox5-neu-fine	5	33	x		x
Vox5-ang-fine	5	27	x		x
Vox4-joy-fine	4	17	x		
Vox50-fast	50	319		x	
Automatic transcription (Residentials)					
Vox1000-auto	1,000	8556			

3.3 CallSurf and VoxFactory corpora

3.3.1 Statistical description of the "experimental" corpus: Vox14-fine

The 14-h calls composing the Vox14-fine corpus have been selected among the data collected during the VoxFactory campaign in order to train acoustic models for emotion detection [Devillers, et al., 2010]. For the present study, a part of this corpus corpus is used to adapt the transcription system and the other part to evaluate the transcription performance on emotional data as described in Sect. 3.3.2. The corpus is divided in three sub-corpora:

- Vox5-neu-fine, which is composed of 5-h calls selected at random. It is considered as the neutral sub-corpus.
- Vox5-ang-fine, which is composed of 5 h of calls containing manifestations of anger, expressed at the linguistic and acoustic levels.
- Vox4-joy-fine, which contains 4 h of calls containing manifestations of positive emotions, expressed at the linguistic and acoustic levels.

The selection of the calls composing Vox5-ang-fine and Vox4-joy-fine has been made manually by randomly listening calls among the 1,000 h of VoxFactory corpus. Fine transcripts are provided for all the calls of the Vox14-fine corpus. Table 8 presents the main statistical features of Vox14-fine considering call, section, speaker turn, segment and word levels. Each call is indeed segmented into sections in order to separate portions to transcribe from irrelevant portions (waiting music, inaudible voice). Duration of calls is between 2 and 30 min (77 calls in total) but most of the calls last between 5 and 14 min (57 calls). A speaker turn is delimited by a speaker change and contains one or several segments. A segment is a part of a turn and is usually ended by a breath. The segment duration is the total duration of speech segments. Speaker turns are short and contain barely more than one segment in the mean. The overlap duration is the total duration of overlapping speech segments (excluding backchannels) and corresponds to about 5.4 % of the call duration. We can also notice that there are more than two speakers by call. This is due to the fact that there may be more than one agent trying to answer client's request or there may be more than one client by call trying to explain the problem. Regarding the percentages of female and male speakers in the corpus, we can note that women are overrepresented for both the client and the agent speakers (see Fig. 3).

Vox5-neu-fine and Vox5-anger-fine have been annotated with disfluency tags. Table 9 quantifies the four types of disfluencies described previously (repetitions, restarts, revisions, complex disfluencies) for each sub-corpus. For both corpora, the most frequent disfluencies are repetitions followed by revisions. As we could have expected, there are more disfluencies in the anger corpus than in the neutral one, 2,252 versus 1,281. Besides, the client produces as many disfluencies as the agent in the neutral corpus but he produces twice as many disfluencies as the agent in the anger corpus.

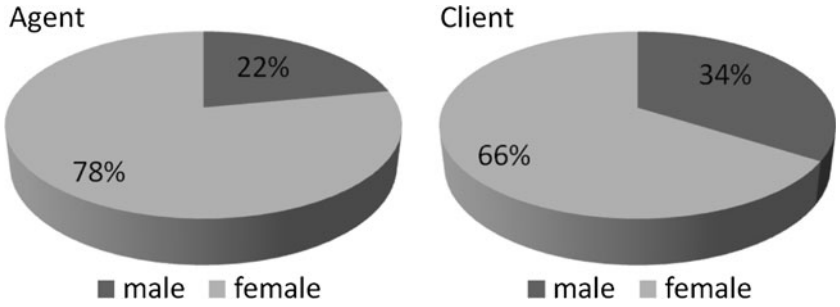

Fig. 3 Percentages of female and male speakers in the corpus according to the speaker role (client or agent)

3.3.2 Overview of CallSurf and VoxFactory corpora

The table below summarizes the various corpora that have been produced during the CallSurf and VoxFactory campaigns for the training, development and test of the automatic speech recognition system and that will be described in the next section. An automatic transcription of the 1,000 h collected during the VoxFactory campaign is also available.

A sub-corpus of Vox14-fine, the Vox9 corpus (9 h), will be used for the evaluation of both the automatic speech transcription system and the semantic analysis. The remaining corpus is used in the development set of the adapted transcription system (version 2 described in Sect. 4.1.2). This evaluation is presented in the following section.

4 Experiments and results

4.1 From V1 to V2 automatic speech transcription system

Two different versions of the decoding system are implemented and evaluated. Version1 (V1) is the CallSurf system (see Sect. 2.1) built on the professional customer data (see Table 7), without any VoxFactory data. V1 is dedicated to evaluate the performance of a system trained and developed on near but not identical domain data. Version2 (V2) is the VoxFactory system. It is an adaptation of V1 using the VoxFactory data from the residential market. V2 is dedicated to evaluate the benefits of adding in-domain data (VoxFactory data). The following two sections describe the lexical, acoustic and language models of the V1 and V2 systems.

4.1.1 Version 1 models

A development set (dev) representative of the CallSurf data set was selected to tune the system. The Version1 dev set (dev1) was composed of 5 h extracted from Call20-fine (see Table 7).

The acoustic models (AMs) of the V1 system are the ones used in the CallSurf system. The models used in the first pass are gender specific models, which have been trained on about 100 h of CTS data from previous projects unrelated to the VoxFactory or CallSurf tasks, and 9 h of speech from the CallSurf data (6 h for female training and 3 h for male training). The AMs used in the second pass are gender independent. They were trained with a flexible training procedure on pooled data consisting of 150 h from general CTS data and of the Call20-fine and Call150-fast corpus and then were adapted with the CallSurf data. These two sets of AMs, which have 18 k and 20 k contexts, both use MLLT (Maximum Likelihood Linear Transform) and are trained using SAT (Speaker Adaptive Training).

The system vocabulary, a 40 k word list, was created selecting the most probable words from the training corpus, in order to minimize the out of vocabulary (OOV) rate on the development data. The OOV rate for Version1 on dev1 is 0.8 %. A semi-automatic grapheme to phoneme conversion was done to phonetize the vocabulary with a 36 phone set and to create the pronunciation dictionary. The pronunciation counts, necessary to calculate the pronunciation probabilities, were determined during acoustic training.

The Language Model (LM) training data is composed of texts mainly of two kinds: manual transcripts of audio data and texts. Different sources were used, including the CallSurf training data (Call20-fine + Call150-fast), as well as some internal conversational telephone speech transcripts (CTS) and some broadcast news (BN) transcripts and web pages (see Table 8).

Component LMs were built from each training source using the Kneser–Ney discounting method (Kneser and Ney 1995), without any cut-offs. Then the LMs were interpolated, with the interpolation coefficients being automatically calculated so as to minimize the perplexity on the development data. The perplexity of the resulting 4-g LM is 35 on dev1.

4.1.2 Version 2 models

As VoxFactory development set (dev2) we used 10 h of CallSurf together with some VoxFactory data (2 h neutral and 2 h anger with detailed transcripts, and 10 h of quick transcripts, extracted from Vox50-fast). Due to the small amount of available data, we did not include data from the joy sub-corpus.

The acoustic models in the V2 system first pass are the same as the ones used in the V1 system first pass. For the second pass, the AMs are gender dependent models trained on pooled data described above, plus the Call200-fast corpus (totalizing 350 h of CallSurf data). They have been created by adapting the speaker independent models with gender-specific task related data. The AMs now have 20 and 22 k contexts.

In the V2 system the language model training data was extended with some additional CallSurf data (Call200-fast) as well as the VoxFactory data (Vox1000-auto automatically transcribed by the CallSurf system and 40 h of fast transcripts extracted from Vox50-fast). As for V1, CTS and BN data were also used, to which some other call-centre conversations (CCS) from the AMITIES project (Hardya et al. 2006) were added. A 45 k word list was selected, minimizing the OOV rate on the dev2 set; we have used all the available text data, and have included the business

Table 11 Data sets used for the construction of the two versions of the VoxFactory speech recognizer (V1 and V2)

	V1 data sets	V2 data sets
AM	**CallSurf** (Call20-fine, Call150-fast = 170 h) + **CTS** (100 h)	**CallSurf** (Call20-fine, Call150-fast, Call200-fast = 370 h) + **CTS** (100 h)
LM	CallSurf (15 h from Call20-fine, Call150-fast = 1.6 M words) + CTS (3 M words) + BN (600 M words)	**CallSurf** (Call20-fine, Call150-fast, Call200-fast—3 M words) + **VoxFactory** (Vox1000-auto, 40 h from Vox50-fast = 10 M words) + **CCS** (990 k words) + **CTS** (4 M words) + **BN** (800 M words)
Dev	CallSurf (5 h from Call20-fine) = 63 k words	**CallSurf** (Call10-fine) + **VoxFactory** (Vox2-neu-fine, Vox2-ang-fine, 10 h from Vox50-fast) = 279 k words

concepts or topics that should be detected. The OOV rate on dev2 is 0.5 % and the perplexity of the 4-g LM is 45.

Table 11 summarizes the characteristics of the different models and development sets used in both systems.

4.2 Automatic Speech Transcription evaluation on Vox9-fine

Three hours are extracted from each subcorpus of Vox14-fine corpus (neutral, anger and joy) to build the evaluation corpus Vox9-fine. The Vox9-fine corpus is decoded by both versions of the speech recognizer. Table 12 shows the WER with both systems for the different test data subsets. The V1 system has a WER of 36.4 % whereas the updated V2 system obtains a WER of 33.8 %.

Besides, we have previously carried out an evaluation of V1 system on CallSurf data and have obtained a WER at 30.3 %. Consequently, we observe a diminution (20 % relative) of the V1 system performance when processing Vox9-fine data (WER at 36.4 %). This degradation is due to the presence of emotional speech, which infers errors due to the presence of a greater proportion of disfluencies and other spontaneous phenomena, but also because of the transition from professional to residential customers. However this degradation is very limited and the WER observed is similar to the one observed in Garnier-Rizet et al. (2008). Adding in-domain transcription reduced the WER to a level, which is identical for the neutral subset (29.6 %) to the one observed on pure CallSurf data (30.3 %).

For both systems the subset, on which the recognition error rate is the lowest, is the neutral one, closely followed by the anger one. The joy subset seems more difficult to be transcribed. Given the small amount of data, the larger WER observed on the joy data could be due to the presence in this subset of a speaker on which the speech recognition was significantly degraded. ASR error rates are indeed known to differ greatly between speakers (Goldwater et al. 2010). Speech recognition on the anger subset has the same accuracy as on the neutral one. This is a satisfying result

Table 12 Word error rates and confidence intervals obtained on the different test data subsets of Vox9-fine, using the two versions of the VoxFactory speech recognizer (V1 and V2)

	#words in the ref.	V1 WER	V2 WER
All	108 k	36.4 % ± 2.4	33.8 % ± 2.6
Neutral	36 k	33.1 % ± 3.0	29.6 % ± 3.2
Anger	38 k	32.4 % ± 3.3	30.8 % ± 3.4
Joy	34 k	44.5 % ± 5.0	41.6 % ± 5.3
Agent	58 k	35.5 % ± 3.1	32.5 % ± 3.2
Client	50 k	37.2 % ± 2.8	35.0 % ± 2.9
Machine	384	76.3 % ± 22.9	72.0 % ± 19.9
Male	36 k	41.1 % ± 5.5	38.1 % ± 5.9
Female	72 k	34.0 % ± 2.3	31.6 % ± 2.5

from an applicative point of view, because it is especially interesting, to detect the speech segments where anger is present.

The results also show that the female speakers are better recognized than the male speakers. This result is in accordance with the study made in Adda-Decker and Lamel (2005). Besides, the agents are better recognized than the clients and the systems seem to have much more difficulties to deal with the few sequences corresponding to the answering machines; this is due to the fact that the answering machines are not transcribed in the CallSurf training data, and thus are not included in the acoustic or linguistic training data.

4.2.1 Keyword Error Rate

Table 13 firstly presents the results of the KER, defined in Sect. 2.1, obtained on the Vox9-fine for a global list, which groups together all keywords referring to EDF vocabulary, and then focuses on two topic lists: invoicing terms (e.g. billing, payment) and technical terms (e.g. meter, power). The KER is calculated for the V1 and V2 systems.

We observe that KER on the global list improves similarly as the WER when using V2 instead of V1 system. On the specific invoicing keyword list, the V1 and V2 systems obtained the same results (28.7 vs. 29.2 %), but larger differences could be observed on the technical lists with a reduction of 15 % relative. This improvement is due to the addition of specific training data from the VoxFactory corpus, containing terms, which are not present in CallSurf data. Furthermore these terms have been explicitly included in the V2 system vocabulary.

4.2.2 Error analysis

4.2.2.1 Typical recognition errors We provide an analysis of word confusion pairs in order to go further in the analysis of the recognition task. A total 23,264

Table 13 KER for three lists of EDF keywords; D stands for substitution, I for insertion and S for substitution errors

List name	#keywords in the ref	V1 KER	V2 KER
Global list	5,013	38.4 %	33.7 %
		D: 193	D: 138
		I: 55	I: 56
		S: 1677	S: 1497
Invoicing	188	29.2 %	28.7 %
		D: 4	D: 3
		I: 1	I: 1
		S: 50	S: 50
Technical	171	41.5 %	38.0 %
		D: 2	D: 1
		I: 1	I: 1
		S: 68	S: 63

word confusion pairs are thus produced. But few of them are frequent: 52 % confusion pairs appear only once, 11 % twice, and 6 % three times. The confusion pairs that appear more than 30 times correspond to less than 6 % of confusion pairs. As a majority, they correspond to words having one or two phonemes. But longer words are also subjected to errors. For example, the word "prélevée" ("deducted") is substituted 14 times by one of the following words: "prélevé" ("deducted"), "relever" ("to read the meter"), "élevée" ("high bill") and "prélever" ("to deduct").

A first type of errors corresponds to homophones or to words with phonetic proximity, which are words sharing the same or a similar pronunciation, such as noticed in the in-depth study of ASR errors types provided by Goldwater et al. (2010). The last example perfectly illustrates this type of error. It illustrates also the difficulty of language models to deal with gender agreement in the case of homophones in such data (confusion between "prélevée" and "prélevé"). Indeed, in call-centre data, the context required to gender agreement is frequently very far (in previous speaker turns), not precisely defined (because of errors present in the context) or dependant on the interlocutor gender, such as in the following example:

Reference: vous avez moins consommé que ce que vous avez éte PRÉLEVÉE[14]
Hypothesis: vous avez moins consommé que ce que vous avez éte PRÉLEVÉ[14]

Another examples of homophone errors are the confusions between "deux" (two) and "de" (of) and between "qu'elle" (that she) and "quel" (whose). We find also frequent errors due to the schwa and to the confusions between [e] and [ɛ], such as "avez" (have) and "avait" (had).

A second type of error is the proper names, which are either out of vocabulary words or under-represented words. Proper names contained in the data are persons, names, cities, organization, etc. The following example illustrates an error on a city name confused with the noun "neighbours":

Reference: ça fait bientôt trente ans que j'habite a **BEAUVOISIN**[15]
Hypothesis: ça fait bientôt trente ans que j'habite à **VOS VOISINS**[16]

A third type of errors concerns telephonic conversational speech features, such as acoustic conditions (recording conditions produced by GSM, free-hands phones, conversations in background, saturation), speech overlaps (which favor the long error sequences—see the following paragraph), backchannels, and disfluencies (see Sect. 3.3). For instance, the hesitations ("hum" or "euh" in French) often appear in confusion pairs (8 % of the confusion pairs). This observation is in accordance with

[14] "You have less consumed than you has been debited".

[15] "I have lived in Beauvoisin for thirty years".

[16] "I have lived at your neighbours for thirty years".

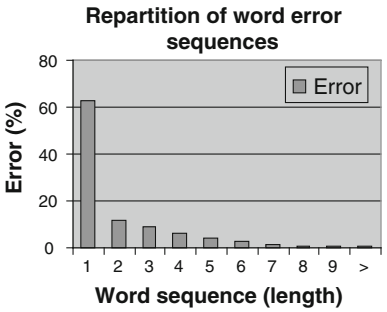

Fig. 4 Repartition by length of word error sequences for the V1 system

the study of Goldwater et al. (2010) where the disfluency factor in ASR errors is analyzed.

4.2.2.2 Sequence of errors A second in-depth analysis of the speech recognition errors is carried out by analyzing sequence of errors produced by the V1 and V2 systems on the Vox9-fine corpus. Indeed, the information extraction system described in Sect. 2.2 extract concepts and opinions from complex word sequences. It is thus important to wonder whether errors are uniformly distributed or whether they frequently propagate themselves in groups or sequences of errors,[17] likely to degrade further analysis, such as in the following example: the speaker stammers and gets confused, which provokes four error sequences with a length from 4 to 13 words ("Uh, because I am not, I am not on off-peak. So, each time, it's the thing, it must always draw". The transcript is too confused to be translated but we can notice the apparition of the nouns "cheque", "car", and of the verb "to consume", which can be detected as business concept).

```
Reference:  (%hesitation)PARCE QU'EN FAIT je SUIS pas je SUIS PAS SUR
            LES   HEURES CREUSES   donc À CHAQUE FOIS C' EST LE MACHIN
            IL DOIT      TIRER  EN      PERMANENCE QUOI

Hypothesis:              AU CENTRE TU VOIS je SAIS pas je **** ***  ***
            ***   VAIS   REPRENDRE donc * ****** **** ** *** ** ******
            JE CONSOMME  CHÈQUE VOITURE JE         DONC
```

Figure 4 shows the repartition of the number of words contained in error sequences for the V1 system (repartition for the V2 system is similar). We observe that errors appear as a majority (60 %) on isolated words and that only 12 % of the errors are located on a succession of two words. This allows us to check that the propagation of errors is not frequent and will not hamper the semantic analysis.

4.2.3 Towards the use of confidence scores

In this paragraph, we want to investigate the use of confidence score as a parameter of the semantic analysis evaluated in the next paragraph: expressions containing too

[17] Long sequences of errors could be explained by bad acoustic conditions (phone noise, saturation), a bad articulation, presence of disfluencies, etc.

many words with low confidence score could be differently processed by the semantic analysis. Such research trails have been previously deeply explored in some works such as in Hazen et al. (2000). We want here to check if the computed confidence score is relevant for further uses.

The confidence score is computed a priori on each word: the ASR engine provides it without considering the reference and could thus be used in the processing chain. A score between 0 and 1 is assigned to each word returned by the speech transcription system depending on how likely it is correct. Inspired by Gillick et al. (1997), the confidence score is here estimated by a logistic regression based on features extracted from confusion sets (Allauzen 2007).

By averaging the confidence score of all words of the conversation, a global confidence score is obtained at the conversation level. The following figures present the global confidence score according to the word error rate, for the two versions of the ASR and for the test corpus (48 conversations of Vox9-fine). The correlation between the two measures confirms that the higher the confidence score is, the higher the quality of the transcription is (Fig. 5).

In Table 14, on the same corpus, the global confidence score are computed for words correctly recognized (C), as well as for the inserted (I) and for the substituted (S) words.

In both versions, the confidence score for C words is higher than the confidence score for I and S words. It means that the confidence score, which is the "engine point of view", is lower when the engine commits an error. The confidence score increases by 6 % from version 1 to version 2 for correctly recognized words. Although it increases also by 12 % for I words and by 18 % for S words, the confidence scores of v2 seem more reliable.

This shows that using confidence score, at a global or local level, makes sense. The score could be used for subsequent treatments during the detection of opinions and concepts.

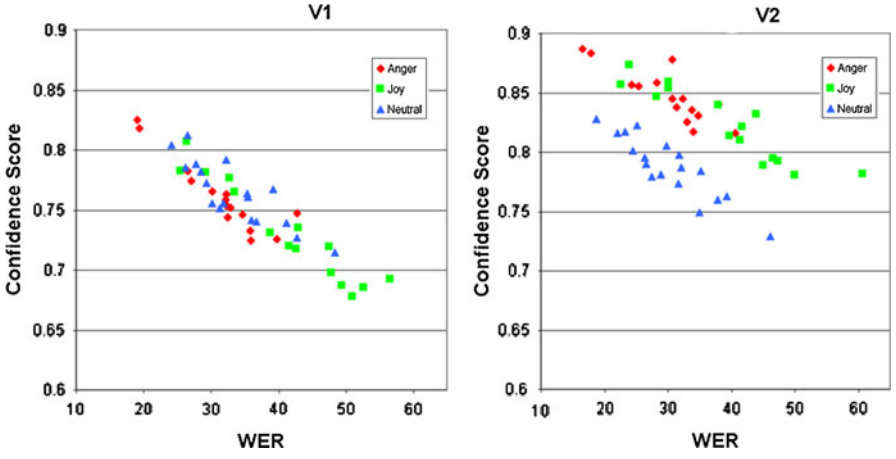

Fig. 5 Confidence Score according Word Error Rate (WER) with the two versions of the ASR

Table 14 Confidence score for correctly recognized (C), inserted (I) and substituted (S) words, with the two versions of the ASR

	v1	v2
C words	0.816	0.864
I words	0.636	0.710
S words	0.595	0.706

4.3 Impact of speech recognition errors on semantic analysis

The impact of speech recognition errors is here evaluated on both the business concepts detection and the opinion and sentiment analysis. The outputs of the various semantic analyses on automatic transcripts are compared to those obtained on manual transcripts. This analysis is carried out on each transcription version v1 and v2 of the Vox9 corpus described in Sect. 3.3 and distinguishes the three types of information modeled in this paper: business concepts, opinions and relationships.

The automatic transcript is aligned with the manual one by considering the manual speaker segmentation. Automatically transcribed speech portions, which don't correspond to any manually transcribed speaker turn, are thus not considered. It includes speaker turns annotated as overlaps between speakers, as described in Sect. 3.3.

The entities detected on the manual transcripts are considered as the reference and the recall and the precision are computed at the level of the speaker turn. Two versions of the calculus of the recall and precision are possible. In the first version used in Danesi and Clavel (2010), only speaker turns containing exactly the same detected concepts on manual and automatic transcripts are considered as correct. In the second version used in Cailliau and Cavet (2010), the recall is defined as the number of the extracted entities that are identical between the manual and automatic transcripts of each speaker turn, divided by the number of entities extracted from the manual transcripts. The precision is defined as the number of the extracted entities that are identical between the manual and automatic transcripts of each speaker turn, divided by the number of entities extracted from the automatic transcripts. Since we evaluate the impact of speech recognition errors according to various entity types, the second version has been chosen.

The following tables present the recall and precision that are obtained for each transcription for each type of information described in Sect. 2.2: business concepts modeled by marketers (Table 15), opinion/sentiment related concepts modeled by

Table 15 Impact of speech recognition errors on the extraction of business concepts (marketers' models): precision and recall (manual/v1 and manual/v2)

Corpus	Nmanual	Recall man./v1	Recall man./v2	Precision man./v1	Precision man./v2	F-score man./v1	F-score man./v2
Anger	2164	0.75	0.79	0.81	0.84	0.78	0.81
Neutral	1959	0.78	0.81	0.79	0.83	0.78	0.82
Joy	1649	0.66	0.70	0.74	0.75	0.70	0.72
Global	5772	0.74	0.77	0.78	0.81	0.76	0.79

Table 16 Impact of speech recognition errors on the extraction of opinion and sentiment entities (merged results of marketers, discourse theory-based and Appraisal theory-based models): precision and recall (manual/v1 and manual/v2)

Corpus	Nmanual	Recall man./v1	Recall man./v2	Precision man./v1	Precision man./v2	F-score man./v1	F-score man./v2
Anger	2,776	0.74	0.76	0.80	0.80	0.77	0.78
Neutral	2,687	0.69	0.72	0.77	0.78	0.73	0.75
Joy	2,715	0.59	0.63	0.73	0.74	0.65	0.68
Global	8,178	0.67	0.70	0.77	0.77	0.72	0.74
Marketers	1,104	0.66	0.69	0.90	0.91	0.76	0.79
Discourse	3,114	0.69	0.71	0.77	0.77	0.73	0.74
Appraisal	3,960	0.66	0.69	0.73	0.75	0.69	0.72

Table 17 Impact of speech recognition errors on the extraction of relationships (Appraisal theory): precision and recall (manual/v1 and manual/v2)

Corpus	Nmanual	Recall man./v1	Recall man./v2	Precision man./v1	Precision man./v2	F-score man./v1	F-score man./v2
Anger	775	0.67	0.71	0.68	0.68	0.67	0.69
Neutral	744	0.65	0.70	0.67	0.72	0.66	0.71
Joy	802	0.51	0.57	0.59	0.63	0.55	0.60
Global	2321	0.61	0.66	0.65	0.68	0.63	0.67

marketers, through the Discourse theory and through the Appraisal theory (Table 16) and the relationships between opinions and business concepts modeled through the Appraisal theory (Table 17). *Nmanual* represents the number of entities extracted on manual transcripts. In table 16, the results of the three modeling methods (marketers' models, the Appraisal theory, the discourse theory) used for opinion and sentiment extraction (see Sect. 2.2) are first merged and then detailed.

Concerning the dependence of the results on the emotional factor, the results are better on the neutral and anger sub-corpora than those obtained on the joy sub-corpus. Indeed, the performance of the speech recognition system is by far the worst on this sub-corpus, with a WER at 41.6 % for the v2 (see Sect. 4.1).

Comparing v1 and v2 speech recognition systems, the results are slightly better for v2. In particular, recall on the relationships extraction (Table 17) globally progresses with 0.04 from 0.63 to 0.67.

At last, the precision is higher than the recall for each subcorpus and each type of information. This means that the missed entity detections (silence) generated by speech recognition errors are more frequent than the false entity detections (noise). As explained in Cailliau and Cavet (2010), a low recall happens when the automatic transcription replaces words corresponding to an entity pattern with words that don't correspond to an entity pattern, as shown in Tables 18 and 19.

Table 18 Examples of opinion entities present in the manual transcripts but not in the automatic transcripts

Manual transcript	English translation	Automatic transcript
putain mais c'est incroyable	**Dammit** it's incredible	mais c'est incroyable (v1, v2)
qui me **plait** absolument **pas**	which **does not please** me at all	y connaît absolument pas (v1,v2)
ça devient vraiment le **bordel**	it is becoming such a **mess**	ça puis enfin moi quand (v1,v2)
c'est **malheureux** j'aurais dû prendre le prénom	it's a **shame** I should have taken the first name	laissez aujourd'hui du prendre le prénom (v1)
		laissez mal rouge du prendre le prénom (v2)

Table 19 Example of business entities present in the manual transcripts but not in the automatic transcripts

Manual transcript	English translation	Automatic transcript
ah parce que Tempo il était un peu moins **cher** l'**abonnement**	Ah, because, Tempo, it was a bit cheaper the subscription	parce que Tempo était un peu moins cher ah bon ben (v1)
		parce que le Tempo et Tempo. Oui. Abonnement (V2)

Table 20 Examples of entities present in the automatic transcripts but not in the manual transcripts

Manual transcript	English translation	Automatic transcript
ça me va très bien Gaz de France	Gaz de France suits me fine	**sympa** très bien Gaz de France (v1, v2)
pour faire mon virement	to make my transfer	pour faire **mentir** (v1, v2)
quartier Jauffret à Gassin	district Jauffret in Gassin	quartier les **offrez** à **agacent** (v1, V2)

A low precision indicates that some words included in entities, which are not present in the manual transcripts are wrongly added during the automatic transcription process. Table 20 shows examples of this phenomenon.

In the light of these explanations, the fact that the precision is higher than the recall seems consistent. As our grammars and lexicons do not cover a very wide number of words and expressions compared to the variety of the data, automatic transcription errors are more likely to turn entities into non-entities than the opposite. This confirms the observation made in Sect. 4.2: key-word deletions are more numerous than insertions.

Besides, the impact of speech recognition errors is weaker on business concept detection, as shown by the F-score (0.79) achieved on the global corpus, when opinions and relationships detection achieve an F-Score respectively at 0.74 and 0.67. This is due to the fact that the rules or the grammar used to model opinions and sentiments are more complex than those used for business concepts modeling. Indeed, the business concepts are based on keywords that are well transcribed (see Sect. 4.2.1). Table 21 stores the average number of words contained in the detected

Table 21 Average number of words in the detected entities according to the type of information for manual transcript

Business	1.44
Opinion/sentiment	1.41
Relationships	3.34

entities on manual transcripts for each type of information. It shows that the linguistic patterns used to model relationships between the opinions and their targets are more complex than those used to model business concepts or opinion and sentiments. It explains the deterioration of the results for the detection of relationships. An error, which occurred on one of the words of the relationship can reverse the meaning of the entities and provoke bad detections. The longer the entities are, the more probable are errors on it. Besides, the target of the relationship can be a person or a situation frequently modeled by pronouns, which are more difficult to get recognized by the speech recognition system because they are short (one or two phonemes).

Table 16 shows that the three opinion models (marketers, discourse and appraisal) follow the same behavior, what the impact of speech recognition errors concerns, except for the precision of the marketers' models. Indeed, the marketers' model focuses on satisfaction concept, which is more specialized and generates thus less noise. In addition, occurrences of this model are less represented in the corpus with 1,104 entities extracted against between 3,000 and 4,000 entities for the two other models.

We have presented here the impact of speech recognition errors. No quantitative evaluation of semantic models has been carried out because no manual annotation of the data in opinions is so far available. However, some issues created by our semantic models on call-centre data are interesting to notice. A qualitative analysis of semantic extractions shows two main difficulties. The first difficulty is about building generic semantic rules when modeling opinion in business data. For example, in the sentence: "le jour nuit se met pas en route" ("the day night mode doesn't work"), the expression "jour nuit" corresponds to a system for controlling electricity consuming. However, "nuit" is here tagged as the third person of the verb "nuire" ("harm") and therefore polarized as negative. The second main difficulty is about the processing of disfluent speech such as in the following example: "c'était le c'était le con le euh l'état je sais pas quoi climatique" ("it was the it was the con the uh the statement climatic something"), "con" is considered as a French insult, while it corresponds here to the beginning of the word "contract". Such difficulties are inherent of an opinion detection system and we will keep on tackling these issues in further works by developing hybrid methods combining natural language processing with machine learning methods.

5 Conclusions

This paper presents the processing chain developed to perform opinion detection on call-centre data. This chain allows us to have access to high-level information, such as the relationship between opinions and their target, which is crucial information

not only for marketing applications but also an important challenge for the industrial and scientific community. The results presented in this paper demonstrate that opinion detection on call-centre data is a promising task.

On the one hand, the first adaptation of the speech recognition system to call-centre data clearly improves the performance and leads us to a WER at 30.3 % on the CallSurf corpus (Garnier-Rizet et al. 2008). The performance on the VoxFactory corpus, which is evaluated in the present paper, is subjected to degradation but after adaptation, the obtained WER of 33.8 % on emotional speech is still satisfying.

On the other hand, the impact of speech recognition errors on business concept and on opinion detection is quite weak with a F-score at 0.79 for business concepts and at 0.74 for opinion and sentiment detection. When dealing with more complex linguistic patterns—relationships between opinions and their target, the impact is a little bit stronger: the F-score falls at 0.67. However, the adaptation on the VoxFactory corpus allows us to weaken the impact of speech recognition errors (the F-score with the CallSurf system was at 0.63).

The in-depth analysis of the speech recognition errors allows us to identify some tracks to explore in order to improve the information extraction results.

Concerning perspectives for the speech recognition system, we could investigate a better handling of spontaneous speech phenomena, which are correlated with emotion in speech; this will lead to improvement of the speech transcription process on segments where opinions or emotions are expressed. Improvement in the speaker segmentation and identification processes could also allow us to define separate linguistic models for the client and the agent during the semantic analysis.

What the semantic analysis concerns, it will be interesting to combine the natural language processing methods with prosodic analysis at the information extraction level in order to handle the issue of degrade transcripts on strong emotional segments or to disambiguate ironic speech. Another track to be explored is to investigate hybrid methods combining machine learning methods with linguistic ones to assess the polarity and the intensity of an extracted opinion.

Besides, the correlation of confidence score and WER that is highlighted in this paper will lead us to investigate the use of this score to parameterize the rules of the linguistic models. Another lead, that we would like to explore, concerns the use of the N best outputs of ASR by information extraction modules such as done by Baumann et al. (2009) for semantic processing in dialog systems.

Acknowledgments This work was partly financed by CAP DIGITAL, the Business Cluster for digital content through the VoxFactory project.

References

Adda-Decker, M., & Lamel, L. (2005). Do speech recognizers prefer female speakers? In *Proceedings of the 9th international conference on speech communication and technology* (Interspeech'05) (pp. 2205–2208).

Adda-Decker, M., Habert, B., Barras, C., Adda, G., & Boula De Mareuil, P. (2003). A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. In *Proceedings of disfluency in spontaneous speech workshop* (pp. 67–70), Göteborg, Sweden.

Allauzen, A. (2007). Error detection in confusion network. In *Proceedings of the 8th annual conference of the international speech communication association* (Interspeech' 07) (pp. 1749–1752), Antwerp, Belgium.

Appelt D., Hobbs J., Bear J., Israel D., Kameyama M., & Tyson, M. (1993). FASTUS: A finite state processor for information extraction from real-world text. In *Proceedings of the international joint conference on artificial intelligence* (pp. 1172–1178), Chambéry, France.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2000). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication (special issue on speech annotation and corpus tools), 33*(1–2).

Barras, C., Zhu, X., Meignier, S., & Gauvain, J.-L. (2006). Multi-stage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing, 14*(5), 1505–1512.

Baumann, T., Buß, O., Atterer, M., & Schlangen, D. (2009). Evaluating the potential utility of ASR N-best lists for incremental spoken dialogue systems. In *Proceedings of interspeech* (pp. 1031–1034), Brighton.

Benveniste, E. (1970). L'appareil formel de l'énonciation. In *Problèmes de linguistique générale II* (pp. 79–88). Gallimard, 1974.

Blanche-Benveniste, C. (1990). *Le français parlé : Etudes grammaticales*. Paris: Didier-Erudition.

Bloom, K., Garg, N., & Argamon, S. (2007). Extracting appraisal expressions. In *Proceedings of HLT-NAACL* (pp. 308–315).

Bozzi, L., Suignard, P., & Waast-Richard, C. (2009). Segmentation et classification non supervisée de conversations téléphoniques automatiquement retranscrites. In *Proceedings of TALN—Traitement Automatique des Langues Naturelles*, Juillet 2009.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics, 21*(4), 1–37.

Cailliau, F., & Cavet, A. (2010). Analyse des sentiments et transcription automatique: modélisation du déroulement de conversations téléphoniques. *Revue Traitement Automatique des Langues, 51*(3), 131–154.

Cailliau, F., & Giraudel, A. (2008). Enhanced search and navigation on conversational speech. In *Proceedings of SIGIR workshop of searching spontaneous conversational speech*, Singapour.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Paris: Hachette Education.

Clavel, C., & Richard, G. (2011). Recognition of acoustic emotion. In C. Pelachaud (Ed.), *Emotional interaction system*, John Wisley.

Dubreil E., Vernier M., Monceaux L., & Daille, B. (2008). Annotating opinion—evaluation of blogs. In *Proceedings of the LREC workshop on sentiment analysis: metaphor, ontology and terminology (EMOT-08)*, Marrakech, 2008.

Danesi, C., & Clavel, C. (2010). Impact of spontaneous speech features on business concept detection: A study of call-centre data. In *Proceedings of the ACM multimedia workshop on searching spontaneous conversational speech*, Firenze, Italy.

Dave, K., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the twelfth international world wide web conference* (pp. 519–528).

Devillers, L., Vaudable, C., & Chastagnol, C. (2010). Real-life emotion-related states detection in call centers: A cross-corpora study. In *Proceedings of interspeech* (pp. 2350–2353), Makhuari, Japan.

Galliano, S., Gravier, G., & Chaubard, L. (2009). The ESTER2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of interspeech* (pp. 2583–2586), Brighton.

Garnier-Rizet, M., Adda, G., Cailliau, F., Gauvain, J.-L., Guillemin-Lanne, S., & Lamel, L. (2008). CallSurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Proceedings of the sixth international language resources and evaluation (LREC) European Language Resources Association (ELRA)* (pp. 2623–2628), Marrakech, Morocco.

Garofolo, J., Auzanne, C., & Voorhees, E. (1999). The TREC spoken document retrieval track: A success story. In *Proceedings of text retrieval conference* (TREC) (Vol. 8, pp. 16–19).

Gauvain, J.-L., Lamel, L., & Adda, G. (1998). Partitioning and transcription of broadcast news data. In *International conference on speech and language processing* (Vol. 4, pp. 1335–1338), Sydney, Australia.

Gauvain, J. L., Lamel, L., Schwenk, H., Adda, G., Chen, L., & Lefevre, F. (2003). Conversational telephone speech recognition. In *Proceedings of ICASSP* (pp. 212–215), Hong Kong, 2003.

Gillick, L., Ito, Y., & Young, J. (1997). A probabilistic approach to confidence estimation and evaluation. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (Vol. 2, pp. 879–882), Munich, Germany.

Godfrey, J. J., Holliman, E. C. & Mcdaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP* (Vol. 1, pp. 517–520), San Francisco.

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication, 52*(3), 181–200.

Hain, T., Woodland, P. C., Evermann, G., & Povey, D. (2000). *The CU-HTK Hub5e Transcription System*. College Park, MD: In Proceedings of NIST Speech Transcription Workshop.

Halliday, M. (1994). *Introduction to functional grammar* (2nd ed.). London: Edward Arnold.

Hardya, H., Bierman, A., Bryce Inouye, R., Mckenzie, A., Strzalkowski, T., Ursu, C., et al. (2006). The Amities system: Data-driven techniques for automated dialogue. *Speech Communication (issue Spoken Language Understanding in Conversational Systems), 48*(3–4), 354–373.

Hazen, T., Burianek, T., Polifroni, J., & Seneff, S. (2000). Recognition confidence scoring for use in speech understanding systems. In *Proceedings of the ISCA ASR2000 tutorial and research workshop* (pp. 213–220). Paris.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (Vol. 1, p. 181).

Ljolje, A., Hindle, D., Riley, M., & Sproat, R. (2000). The AT&T LVCSR-2000 system. In *Proceedings of NIST speech transcription workshop*, College Park, MD.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation, appraisal in English*. London, New York: Palgrave Macmillan.

Matsoukas, S., Colthurst, T., Kimball, O., Solomonoff, A., Richardson, F., Quillen, C., et al. (2002). The 2001 Byblos English large vocabulary conversational speech recognition system. In *Proceedings of ICASSP* (Vol. 1, pp. 721–724).

Olsson, J. S., Wintrode, J., & Lee, M. (2007). Fast unconstrained audio search in numerous human languages. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (Vol. 4, pp. 77–80), Honolulu, Hawaii.

Ostendorf, M., (2009). Transcribing human-directed speech for spoken language processing, In *Proceedings of interspeech* (pp. 21–26), Brighton, United Kingdom.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. In *Foundation and trends in information retrieval* (Vol. 2(1–2), pp. 1–135). Hanover, USA: Now Publishers Inc.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Acl-02 conference on empirical methods in natural language processing* (Vol. 10, pp. 79–86), Morristown, NJ.

Park, Y., Patwardhan, S., Visweswariah, K., & Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. In *Proceedings of Interspeech* (pp. 2070–2073).

Paumier S. (2002). *Manuel d'utilisation d'Unitex, Universitéde Marne-la-Vallée*. http://www-lipn.univ-paris13.fr/~rozenknop/Cours/MICR_REI/Seance2/ManuelUnitex1.2.pdf. Accessed 31 Jan 2012.

Poibeau, T. (2002). Extraction *d'information à base de connaissances hybrides,* PhD Thesis, Université Paris-Nord, March 2002.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Longman.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the conference on empirical methods in natural language processing—theoretical issues in natural language processing. association for computational linguistics* (Vol. 10, pp. 105–112), Morristown, NJ.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD Thesis, University of Berkeley, California.

Silberztein, M. (1994). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson.

Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Rao Gadde, V. R., Plauche, M., et al. (2000). The SRI March 2000 Hub-5 Conversational speech transcription system. In *Proceedings of NIST Speech Transcription Workshop*, College Park, MD.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*(7), 10760–10773.

Ten Bosch, L., & Boves, L. (2004). Survey of spontaneous speech phenomena in a multimodal dialogue system and some implications for ASR. In *Proceedings of Interspeech* (pp. 1505–1508), Korea.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424), Morristown, NJ.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315–346.

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on information and knowledge management*-CIKM (pp. 625–631), Bremen, Germany,.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence* (pp. 735–740).

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th international conference on computational linguistics and intelligent text processing* (*CICLing-05*)*, invited paper, springer LNC* (p. 3406). Berlin: Springer.