Acronym	DATCHA
Name of the project	KNOWLEDGE EXTRACTION FROM LARGE CORPORA OF HUMAN-HUMAN CONVERSATION DATA FROM WEB CHAT SERVICES *Extraction de connaissances à partir de vastes corpus de conversations WEB de type "chat" client-opérateurs*
Appel	Appel à projets générique 2015
Challenge	Axe 7 - Société de l'Information et de la Communication
Duration	42 months
Requested funding	409.8K €

Abstract

The goal of the DATCHA project is to perform knowledge extraction from very large databases of WEB chat conversations between operators and clients in customer contact centers. Extracting knowledge from chat corpus is a challenging research issue. Simply applying traditional text mining tools is clearly sub-optimal as it takes into account neither the interaction dimension nor the particular nature of this language which shares properties of both spoken and written language. The DATCHA project will address scientific issues including *intra-conversation analysis* through a *deep semantic analysis* (syntactic, semantic, discursive and structural analysis) and *inter-conversation analysis* (definition of semantic and discursive similarity between conversations). It will propose innovative solutions in various use-cases including *analytics report generation*, *conversation success prediction* on the basis of criteria defined by operational units, and *online conversation solving*.

Résumé

Le but du projet DATCHA est de permettre l'extraction de connaissance à partir de très vastes corpus de conversation de type "chat" entre des clients et des opérateurs. Extraire des connaissances dans ce contexte demeure un défi pour les méthodes de traitement automatique des langues. Les méthodes de fouille de texte classique sont clairement sous-optimales en ne prenant en compte, ni la dimension interactive, ni les propriétés de ce nouveau type de de langage à l'intersection du langage écrit et parlé.

Le projet DATCHA va répondre à ces défis à travers des analyses profondes pour une conversation au niveau syntaxique, sémantique et discursif et pour des collections de conversations en définissant des mesures de similarités sémantiques et discursives. Les solutions innovantes développées seront évaluées dans le contexte des centres de relation clientèle d'Orange sur plusieurs cadres applicatifs tels que la génération de rapports, la prédiction de succès d'un dialogue et l'aide en ligne.

Table des matières

Tableau récapitulatif des personnes impliquées dans le projet

Partner	Last Name	First Name	Situation	M/M	Role
AMU	BECHET	FREDERIC	Professor	20	Project Coordinator Research activities: Natural Language Understanding
AMU	NASR	ALEXIS	Professor	10	Research activities: Syntactic/ Semantic Parsing
AMU	FAVRE	BENOIT	Associate Professor	10	Research activities: Machine Learning for NLP
AMU	PhD	PhD	PhD	36	PhD student on human-human conversation analysis
AMU	Internship	Internship	Internship	6	Prototype development
AMU	Internship	Internship	Internship	6	Prototype development
AMU	Internship	Internship	Internship	6	Prototype development
IRIT	ASHER	NICHOLAS	Researcher Director	22	Scientific supervisor for IRIT Research activities: Discourse Parsing
IRIT	MULLER	PHILIPPE	Associate Professor	22	Research activities: Discourse Parsing, Summarization
IRIT	AFANTENOS	STERGOS	Associate Professor	20	Research activities: Discourse Parsing, Summarization
IRIT	BENAMARA	FARAH	Associate Professor	20	Research activities: Opinion Mining, Sentiment Analysis
IRIT	PostDoc	PostDoc	Post Doc	36	Discourse parsing, joint semantic analysis, dialogue clustering
Orange	DAMNATI	GERALDINE	Research Engineer	12	Scientific supervisor for Orange Research activities: Natural Language Processing, Data Mining
Orange	CHARLET	DELPHINE	Research Engineer	4	Research activities: Natural Language Understanding, Applicative framework
Orange	GUERRAZ	ALEKSANDRA	Research Engineer	9	Research activities: Syntactic/ Semantic parsing, Linguistic Ressources

1 Contexte, positionnement et objectif de la proposition détaillée

1.1 Concept and Vision

The goal of the DATCHA project is to perform knowledge extraction from very large databases of WEB chat conversations between operators and clients in customer contact centers (technical and commercial).

Multichannel capability is a key component of today's Customer Relationship Management (CRM). If traditional telephone call-centers remain central in customer interactions with a company, the proportion of inputs from other channels (web, mail, social media...) is constantly growing. Among these new interaction patterns, *chat* is becoming more and more popular. It has the advantage of remaining a synchronous interaction mode while offering interesting properties: *for customers* who are proposed a live interaction at an appropriate moment (pro-active chat or ClickToChat), *for agents* who generally prefer synchronous interactions and who report that the distance induced by written interaction reduces conflicts and aggressive behavior, and *for companies* who can reduce their contact center costs while gathering via direct interaction with customers rich and varied information concerning "real" problems or issues of concern.

Accessing this rich knowledge however is a challenging research issue. Simply applying traditional text mining tools is clearly sub-optimal as it takes into account neither the interactive dimension nor the particular nature of this language which shares properties with both spoken and written language. While most research on synchronous contact center data has been carried out on telephone interactions, this project would represent a great opportunity to study a new type of interaction that has potential impact on many other fields beyond CRM. Unlike research on spoken conversations where manual transcriptions are always problematic, the availability of very large corpora collected on a daily basis is guaranteed for this project, thus offering opportunities for engaging research on relevant unsupervised analysis approaches.

From a societal point of view the DATCHA project is an opportunity to better understand the way people use these new interactive interfaces that are present in all WEB-related social media. WEB chat conversations represent massive corpora of such interactive data that can be mined to extract such knowledge.

The DATCHA project will address scientific issues including *intra-conversation analysis* through a *deep semantic analysis* (syntactic, semantic, discursive and structural analysis) and *inter-conversation analysis* (definition of semantic and discursive similarity between conversations). It will propose innovative solutions in various use-cases including *analytics report generation*, *automatic synthesis generation* of a (potentially dynamic) set of conversations, *conversation success prediction* on the basis of criteria defined by operational units, and *online conversation solving*.

1.2 Chat language

Chat language received attention in recent years as part of the general "social media" galaxy. More precisely it is often referred to as "*Computer-mediated communication (CMC)*". This term refers to any human communication that occurs through the use of two or more electronic devices such as instant messaging, email or chat rooms. According to a study by Jonsson [Jonsson-DCOURSE1998], who conducted an early work on data gathered through the Internet Relay Chat protocol and through emails: ''*electronic discourse is neither writing nor speech, but rather written speech or spoken writing, or something unique*".

Recent projects in Europe, such as the CoMeRe[1] or the STAC[2] project gathered collections of CMC data in several languages in order to study this new kind of language. Most of the effort has been dedicated to collect "chat room" data as it is the kind of data which is the most accessible on the WEB.

This kind of data falls under the "informal register" whereas we are interested in this project in understanding the mechanisms of a more formal kind of CMC: dialog chat in contact centers.

[1] https://corpuscomere.wordpress.com

[2] http://www.irit.fr/STAC

In the book entitled “Digital textuality” [Trimarco-PM2014] the author points out that “[…] *it would be more accurate to examine Computer Mediated Communication not so much by genre (such as e-mail, discussion forum, etc…) as in terms of communities*”. The importance of relation between partcipants is also pointed out in [Kucukyilmaz-IPM2008]. The authors insist on the fact that chat messages are targeted for a particular individual and that the writing style of a user not only varies with his personal traits, but also heavily depends on the identity of the receiver (corresponding to the notion of sociolinguistic awareness). Customer-agent chat conversations could be considered as being closer to customer-agent phone conversations than to chat-room informal conversations. However the media induces intrinsic differences between Digital talk and phone conversations. The two main differences related in [Trimarco-PM2014] are related to turn taking and synchronicity issues on one side and the use of semiotic resources such as punctuation or emoticons on the other side.

In the case of assistance contact centers, customers engage a chat conversation in order to solve a technical problem or to ask for information about their contract. In certain cases, the conversation follows a linear progress (as in the following example) and in some other cases, the agent can perform some actions (such as line tests) that take some time or the client can be asked to do some operations on his installation which also imply latencies in the conversation flow. In all cases, a chat conversation is logged as shown in the figure below (where _TC_ stands for *téléconseiller* or agent and _CLIENT_ stands for customer): the timestamps at the beginning of each line corresponds to the moment when the participant (agent or customer) presses the “Enter” key*, i.e.* the moment when the message becomes visible for the other participant.

```
[12:04:20] Vous êtes en relation avec _TC_.
[12:04:29] _TC_    : Bonjour, je suis _TC_, que puis-je pour vous ?
[12:05:05] _CLIENT_: mes enfant ont perdu la carte dans le modem et je nai plus de
                     tele comment dois je faire?
[12:05:27] _TC_    : Pouvez vous me confirmer votre numéro ligne fixe afin que je sois sûr
                     d'avoir le bon dossier ?
[12:05:56] _CLIENT_: _NUMTEL_
[12:07:04] _TC_    : Si je comprend bien vous avez perdu la carte d'accès de votre décodeur.
[12:07:27] _CLIENT_: oui ces bien sa
[12:07:47] _CLIENT_: code erreure S03
[12:09:09] _TC_    : Pas de souci, je vais vous envoyer une autre carte par voie postale à
                     votre domicile.
[12:09:38] _CLIENT_: est ce que je peux venir chez orange la chercher aujourdui
[12:10:36] _TC_    : Vous ne pouvez pas récupérer une carte depuis une boutique Orange
                     puisque vous n'avez pas une.
[12:11:02] _TC_    : Car dans une boutique Orange, ils peuvent seulement faire un échange.
[12:11:33] _CLIENT_: ok merci de me lenvoyer au plus vite vous avez bien mes coordonnée
[12:11:57] _TC_    : Oui je les bien sur votre dossier.
[12:12:51] _CLIENT_: ok tres bien dici 48h au plus tard 72h pour la carte
[12:14:06] _TC_    : Vous la recevrez selon les délais postaux à l'adresse figurant sur
                     votre dossier (entre 3 et 5 jours).
[12:14:25] _CLIENT_: ok tres bien en vous remerciant a bientot
[12:15:20] _TC_    : Je vous en prie.
[12:15:29] _TC_    : Avant de nous quitter avez-vous d'autres questions ?
[12:17:23] _CLIENT_: non merci
```

A conversation is a succession of messages, where several consecutive messages can be posted by the same participant. The temporal information only concerns the moment when the message is sent and there is no clear evidence on when the message was started to be written. There is no editing overlap in the Conversation Interface as the messages appear sequentially but it can happen that participants write simultaneously and that a message is written while the writer is not aware of the preceding message.

As we can see, chat conversations are dissimilar from edited written text in that they contain typos, agrammaticalities and other informal writing phenomena (for example, the turn “`[12:07:27] _CLIENT_: oui ces bien sa`” contains homophonic errors which are frequent in spontaneous text writing). They are similar to speech in that a dialogue with a focused goal is taking place, and participants take turns for solving that goal, using dialogic idiomatic terms which are not found in typical written text. They differ from speech in

that there are no disfluencies, and that the text of a single turn can be repaired before being sent. We argue that these differences must be considered as relevant as the two differences pointed out by [Trimarco-PM2014].

All these properties along with the particular type of language used by customers and agents will guide our research towards understanding this new kind of CMC data. The challenges for processing chat data comes from the fact that analysis tools such as parsers are typically trained on mismatched conditions, which we plan to overcome by processing large quantities of chat and benefit from unsupervised training.

1.3 Relevance towards the objectives of the program - Compliance to the 7th challenge : « Société de l'information et de la communication »

DATCHA is directly related to « *Axe 7 : Interactions humain-machine, objets connectés, contenus numériques, données massives et connaissance* » of the 7th challenge where the following objectives are mentioned:

1. *«il s'agit ici des problématiques* […] *d'analyse sémantique, de modélisation et de représentation des connaissances», «interprétation sémantique des données non structurées»*
   →by exploring the nature of chat conversations with a deep semantic analysis approach at the semantic, discursive and dialogic level, the DATCHA project aims at improving the automatic analysis of human-human conversations beyond the "flat" analysis usually performed on text data for semantic classification or sentiment analysis.
2. *«automatisation de l'extraction des connaissances», «exploitation des données massives d'interaction»*
   → automatically adding structure to conversations and taking into account very large collections of them will open new opportunities for knowledge discovery from raw data collected in customer contact center. In particular having a rich description of conversation structures will allow us to consider the interactive nature of such data as an implicit supervision that can be used both for improving statistical models in an unsupervised way and acquiring new knowledge such as lists of Frequently Asked Questions with answers.
3. « *découverte dynamique et interactive de connaissances dans des données de plus en plus massives* », *« développement de techniques prédictives temps réel »*
   → the dynamic, interactive and real time aspects of the DATCHA project are illustrated by the "online conversation solver" use case that proposes to the operators, during an ongoing chat dialog, to mine the repository of conversation logs in order to find successful conversations similar to the current one. These retrieved conversations can help an operator to follow an efficient dialog strategy in order to perform the targeted task.

In addition to the 7th challenge, this research project also has an interdisciplinary dimension: the "online conversation solver" addresses the issue of helping operators to handle complex problems and contribute to the improvement of the working conditions in contact centers. This aspect corresponds to one of the objectives of « *Axe 1 : Travail- place de l'homme, organisation des écosystèmes, valeur sociétale* » in « *Défi 3 : Stimuler le renouveau industriel* ».

Moreover studying chat language at the semantic, discursive and dialogic levels is an opportunity firstly to describe and model this new kind of interactive language from a linguistic point of view; then to study the argumentative and negotiation strategies developed by chat users and more generally social media users.

These aspects correspond to objectives in « *Axe 6 - Révolution numérique et mutations sociales* » of « *Défi 8 - Sociétés innovantes, intégrantes et adaptatives* ».

1.4 Scientific and Technological objectives

The DATCHA project contains several innovative aspects that will yield technological breakthroughs.

- The first one is focused on the ***joint semantic/discursive analysis*** of human-human conversations for deep semantic analysis. Syntactic, semantic and discursive (or dialogic) analyses of textual data provide three distinct structuring of data. However, these structures are co-dependent and their interactions should be studied. The specific nature of the interactions studied in the DATCHA project (chat in a

restricted domain and in a specific interaction condition) yield textual data that is both realistic and reasonably simple from the syntactic, semantic and discursive perspective. The DATCHA data thus provides an ideal testbed to study, model and implement the interactions of these three processes.

- The second innovative aspect is linked to the ***nature*** of the corpus processed: *WEB chat conversation*. From the discourse parsing point of view, the corpus is a relatively well understood genre from a linguistic point of view featuring an expert and a tutee. However, we don't know of any work on corpora of a size comparable to the corpus for this project. Furthermore, we don't know of any attempts to analyze dialogues with full discourse structures other than our own efforts on a corpus of around 1000 negotiation dialogues in English from multiparty chat sessions in an online video game (STAC - http://www.irit.fr/STAC/). Most analyses of conversation involve dialogue act labeling, but these approaches do not capture the relational structure of conversation. Recovering a full discourse structure from a chat provides a strictly more expressive semantic representation of the conversation's content than dialogue act labeling does [cadilhac-emnlp2013]. In addition, with the advent of advanced discourse parsing methods in the last few years, the sort of genre of dialogue prevalent in our corpus is an ideal testbed for several hypotheses as answers to challenging questions:
    1. To what extent does the domain of discourse affect the structure of the dialogue? Will a discourse parser for the chat corpus of the current project output different types of structures from those for other chat corpora like that in the STAC corpus.
    2. Do different problems that users bring up yield different types of structures of dialogues or are they all of the same type? We hope to be able to answer "What is a typical way of responding to such a problem?" in an automated way. From the structural analysis we plan to look for regularities in the corpus in an unsupervised way, to build a typology of these situations.
- The third innovative aspect is the notion of ***similarity*** between conversations and parts of thereof: when are two discourse structures semantically similar? One can use a variety of tree metrics or metrics that take account of structural similarity such as those used in sentential parsing (e.g. kernels) as a first step. But these don't take account of the semantic contents of the leaves in the structure. We plan to combine methods for capturing structural similarities with the semantic analysis performed in work packages 1 & 2 to address similarity at several levels. Our similarity measures will enable us to classify dialogues but also types of customer issues and of solutions to these issues, which will enable us to test predictions about which conversations are successful and which are not. We believe that the large corpus with a relatively restricted domain of inquiry along with the idea that these dialogues will have several measures of evaluation will enable us to test various hypotheses about dialogue similarity. With a validated semantically sound similarity metric in hand for the dialogues in our corpora, we can see whether such a similarity metric applies to induced structures for other corpora.
- The fourth innovative aspect is ***unsupervised learning***. The interactive nature of chat conversations can be seen as an *implicit supervision* for statistical model adaptation and knowledge extraction from very large collection of interactive data. Having rich descriptions of dialogs structure and content as well as similarity measures between conversations can help discovering new knowledge about customers' issues in an unsupervised way as well as improving statistical model in charge of classifying conversations into categories linked to a task (topics, success/failure, satisfaction, etc.).

1.5 Progress beyond the state-of-the-art

1.5.1 Syntactic and semantic parsing of chat corpus

Robustness to noise In chat data

An accurate analysis of human-human conversation should have access to a representation of the text content that goes beyond coarse grain analyses such as keyword search.

We plan to perform syntactic parsing as well as a semantic analysis of the textual data in order to produce high-level features that will be used to evaluate human behaviors. Our target is not perfect and complete syntax and semantic analysis of the data, but rather to reach a level allowing to qualify and compare conversations.

We believe that the current models used in the fields of syntactic and semantic parsing are mature enough to go beyond normative data that we find in corpora and process text that come from CRM chat. The experience we gathered on parsing speech transcriptions in the framework of the DECODA [Bechet-LREC2012-2] and ORFEO [Nasr-LREC2014] projects showed that current parsing techniques can be successfully used to parse disfluent speech transcriptions with disfluencies.

Syntactic and semantic parsing of such textual genre is nevertheless challenging. We will focus, in the scope of this projects, on two aspects. The first one deals with syntactic parsing while the second concerns semantic parsing.

Syntactic parsing of non canonical textual input has been mainly studied [Nasr-LREC2014] in the context of textual transcription of spontaneous speech. In such data, the variation with respect to canonical written text comes mainly from syntactic structures that are specific to spontaneous speech, as well as disfluencies, such as filled pauses, repetitions and false starts. Our input has some of the specificities of spontaneous speech but adds new ones. More precisely, we find in our data syntactic structures found in speech (such as a loose integration of micro syntactic units into macro structures), and for obvious reasons we do not find other features that are characteristic to speech, such as repetitions and restarts. On the other hand, we find in our data many orthographic errors. The following example, taken in our corpus, illustrates the specific nature of our data:

```
ces deja se que j ai fait les pile je les est mit tou a l heure elle sont neuve
```

All words highlighted in yellow can be considered as erroneous either lexically or syntactically. This sentence could be paraphrased by:

```
c'est déjà ce que j'ai fait, les piles je les ai mises tout à l'heure,
elles sont neuves
```

Such an utterance features are an interesting mixture of oral and written characteristics: the syntax is close to oral, but there are no repetitions nor false starts. Orthographic errors are numerous and some of them are challenging for the syntactic parser.

We plan to explore two solutions in order to parse such input: the first one is quite standard and consists in correcting the input prior to parsing [Dahlmeier-EMNLP2012] [Han-HLT2011], using orthographic as well as phonetic transformation rules; the second one is based on word embeddings and will be described in the next sub-section on “Representation learning”.

Interactive text

Processing interactive text brings another level of complexity beyond non-canonical language. The nature of the interactions in CRM chats yields short chatter turns with usually a simple semantic content. In term of frame semantics, the structure of the turns are generally simple, organized around one to three semantic frames, such as in the following examples:

```
_TC_: le défaut viendrait de la livebox
_TC_: je viens de lancer une réinitialisation de votre décodeur à distance
_TC_: Je lance une action afin de corriger ce souci.
```

However, our data features several interesting and challenging cases where the frames are not triggered by lexical elements but by the dialog act that corresponds to the turn as well as the semantic material that has been introduced by the preceding turns, such as in the following example:

```
_TC_: Comment avez-vous branché votre tv au décodeur? Avec un cable péritel ou hdmi?
_CLIENT_: hdmi
```

The second turn could be paraphrased as:

```
J'ai branché ma tv au décodeur avec un câble hdmi
```

This paraphrase is clearly impossible to predict based on the turn itself which is reduced to a single noun: hdmi.

This type of utterances is an example of a more general characteristic that can be observed on chat data: efficiency. One particularity of chat conversation is the remanence of the previous utterances in the chat interface on each participant's screen.

We plan to develop new semantic parsing models, inspired by dynamic discourse semantics [Asher-NLP-CUP2003], which predict the semantic structure of the utterances based on the lexical and syntactic cues that are present in the utterance as well as the semantic structure of the preceding turns and the dialog acts.

Reference resolution

As in any dialogue, CRM chat features a large number of referring expressions such as pronouns, demonstratives and definite noun phrases. Such as in the following example where the pronoun *il*, in the second turn refers to the noun phrase *le decodeur*.

```
_TC_ : Pouvez vous vérifier si [le décodeur] chauffe s'il vous plaît ?
_CLIENT : Non mais de temps en temps [il] émet un gros bruit comme de lancement de relais
```

In order to get access to the meaning of such utterances one should, in theory, resolve all such referring expressions. Although reference resolution is not the main topic of DATCHA, we plan to integrate in our system a reference resolution module. We will use existing systems and data, more specifically, we will adapt the BART system [Versley-LREC 2008] to French, using the ANCOR corpus [Muzerelle-CMLF2012].

The system produced will be evaluated on our data and adapted in order to reach reasonable performances.

1.5.2 Representation learning

Large scale semantic and structural analysis of text is currently dominated by machine learning approaches. Performing such learning entails the use of annotated corpora for training classifiers to predict the targeted structures from an input represented by carefully crafted features. Recent advances in NLP [Collobert-JMLR2011], speech recognition [Deng-ICASSP2013] and computer vision [Taigman-CVPR2014] have looked into relieving practitioners from feature design, through representation learning [Bengio-PAMI2012], which consists in learning from raw data the right set of features which should be extracted. Given very large datasets and deep learning architectures, more relevant representations of the input can be derived, which have the nice property of being dense, low-dimensional, adapted to co-variance hypotheses made by classifiers, and offer better generalization as well as being task-agnostic.

Word embeddings are a kind of representation learning for words which are starting to prove useful for a range of NLP tasks. Contemporary extraction methods originate from the analysis of the hidden state of neural language models [Bengio-JMLR2003] and more recent recurrent neural network language models [Mikolov-ASRU2011]. Representations are typically extracted by learning a model which predicts a word from its context (or vice versa) and using parameters from that model as representation [Mikolov-NIPS2013]. This has been proved to be related to dimensionality reduction of the word co-occurrence matrix [Goldberg-ARXIV2014]. Word embeddings trained from sufficiently large datasets show interesting properties where linear transforms correspond to linguistic and semantic regularities such as inflexion, meronymy, or analogy [Mikolov-NIPS2013].

Using word embeddings in place of 1-of-n word representations as input of NLP systems has been successfully applied to a range of tasks from part-of-speech tagging to semantic role labeling [Collobert-JMLR2011], spoken language understanding [Mesnil-IS2013], or translation [Zou-EMNLP2013]. They enable better generalization and allow for adaptation by projecting various embeddings from domain to domain or even cross-language training of topic classifiers [Lauly-NIPS2014] by enforcing bilingual alignment constraints when learning the embedding. It has been proved by that refining word embeddings and processing out-of-embedding words can lead to significant improvements in part-of-speech tagger adaptation to the speech domain when little on-domain data is available [Tafforeau-IS2015].

The way word representations are trained has an impact of their utility for a given task. For example, following dependency trees instead of linear order when extracting context yields syntactic embeddings which work better on related tasks [Levy-ACL2014]; in addition, a task layer can be added to the classifier that learns the embedding in order to bias the generation towards this task [Tang-ACL2014]. Refinement, which consists in initializing the word representation in a neural network with generic embeddings and training it for a target task, also performs an adaptation process which leads to better end-task performance [Tafforeau-IS2015]. There have also been efforts to overcome the monosemy assumption of word embeddings by taking into account the fact that words can have more than one dominant sense through the introduction of latent variables [Huang-ACL2012].

Recent efforts have focused on extending embeddings to smaller and larger structures than words. For instance, characters can be leveraged as input to reconsider the notion of words for text segmentation [Chrupala-ARXIV2013] with Recurrent Neural Networks which can model variable-length n-grams. Learning sentence or document-level representations has been explored by either composing word representations along the parse tree [Socher-EMNLP12] or using dependent words as operators [Pennington-EMNLP2014], or gathering paragraph-level residuals from the prediction of a word sequence [Le-ARXIV2014]. Long short-term memory (LSTM) cells in recurrent neural networks (RNN), with their ability to retain information much longer than RNNs, have proved useful for modeling larger structures and enable to learn fixed-size representations for variable length inputs [Sutskever-NIPS2014] with successful applications in machine translation [Cho-ARXIV2014] and image caption generation [Vinyals-ARXIV2014].

In Datcha, we will work on two open questions linked to mining large interaction text corpora which are sensibly different from available copora such as Wikipedia and Gigaword. The first question is "how can we create meaningful representations relevant for deep semantic, which can absorb the domain difference between generic text and chat interactions?" This question will be tackled at the word and sentence level in order to cope with typos, specific constructs of interactive text and context modeling. The second question is "how can we create conversation-level representations that help comparing conversations and identify success factors in interactions?" The main lead for answering this question is to look into variable size representations and take into account not only the textual content but the results of semantic and structure analysis for devising the representation. We envision that such representation will allow to explore similar phenomena as linguistic regularities found in word embeddings, in order to uncover for example, if linear transforms can locate conversations of similar structure in different domains.

1.5.3 Discourse parsing of human-human conversation

Discourse analysis is the study of the aspects of language that go beyond the level of isolated sentences. When applied to dialogues, it includes the phenomena that explains the linguistic interactions between speakers, how they understand each other and share information for instance. One of the key objects of study in this field are the factors that make a discourse coherent. Coherence can be accounted for by positing relations between clauses, sentences, or speech acts, that organize the writer's intentions (with explanations, elaborations, contrasts for instance), or explain speakers' turns (e.g. answer to a question, acknowledgment of a proposal or an assertion, correction of an assertion). A coherent text is one in which each constituent is connected via some discourse relation. A number of theories of relational coherence have been proposed, for written text and dialogue, which make different assumptions about the kinds of relations (thus yielding different taxonomies of discourse relations), or the resulting structure (a chain, a tree, diversely constrained types of graphs that influence the interpretation process).

If the analysis of discourse structure for traditionally written text is now well established, there is much less work on applying these theories to conversation. Discourse structure in dialogue s differs in a number of respects from that of text; it features different relations between more varied types of speech acts---e.g.

questions and directives, which are rare in monologue. Additionally dialogue can feature crossing dependencies and different sorts of structures than monologue.

In the context of the Datcha project, these notions are important for various reasons:

- The semantic analysis of speech turns is usually highly dependent on from each other, with a high presence of fragments, ellipsis and implicit references to the context, even more than in the case of a written document. Some specific aspects of the analysesis DATCHA wants towill perform, such as sentiment analysis, (to evaluate customer satisfaction for instance) rely on aspects that have been shown to benefit from discourse analysis [Asher-ECAI2008].
- In order to analyze the kind of interactions that take place in CRM, we need features describing groups of dialogs in a unsupervised way (via clustering). Lexical aspects will play an important role, both for problem-related identification in customer support and for describing the quality of the provided service, but it can only be a crude approach regarding the latter point. Customer-agent interaction needs to be accounted for with more elaborate description of the sequence of turns, and we expect that dialog structure analysis will provide a complementary way of characterizing conversations in this context.

Our experience from our work in ANNODIS and STAC that we have described above brings to DATCHA a theoretical competence concerning discourse structure and dialogue analysis and a computational competence in terms of discourse parsing.

However, DATCHA poses new challenges and opportunities that we will explore. First, we will investigate unsupervised learning of discourse structure. Our experience with annotation has shown us that while they are extremely important for learning about discourse structure, they are extremely time-consuming, expensive and difficult to do. Experts have to revise annotators' best efforts. DATCHA's overall goals are to increase the efficiency of on-line chat communication, and for this it is not necessary to have a completely accurate and detailed analysis of the dialogue structure. The moral we draw from DATCHA's downstream-directed tasks is that accuracy is less critical than it would be for a theoretical project investigating the nature of dialogue or text structure. We want to use unsupervised methods to cluster chat dialogues according to topics such as the challenges that users face, user's goals and user's preferences and sentiments. DATCHA also will provide a vast and in principle unlimited corpus on which to perform experiments in this paradigm.

We plan to leverage the features that we have exploited on the English STAC corpus to French. These features, like basic speech act type (question, assertion, demand), the speaker and characteristics of the speaker's turn are general enough to adapt to French and they have proven useful in determining certain crucial discourse relations for dialogue, such as for example when an assertion is an answer to a question, when a speaker acknowledges the contribution of another and when a speaker asks a follow up question that relies on a previous question or assertion. We also plan to transfer and adapt our model of dialogue acts to the information seeking task, and we will make heavy use of DAMSL's typology of acts in this respect. We will also adapt the preference lexicon developed for STAC by [Cadilhac-EMNLP2013] to DATCHA's framework, as well as exploit the sentiment analysis lexicon for French already developed by Farah Benamara, which also includes information useful to the detection of speaker attitudes. We also plan to make use of the templates that Orange provides to its technical staff that engage customers in chat sessions as a rough guideline for the discourse structure of a dialogue in our corpus. It will be interesting to see when dialogue participants adhere to a particular template and when they depart from it. We will rely on the structural aspects described above and the lexical representations presented in 1.6.2.

1.6 Relevance and Relation to other research programs

This section presents some research projects in which members of the DATCHA consortium have participated and which have a link with the current proposal. We show the complementarity of these previous or current works with DATCHA while highlighting the novelty of this proposal.

ANNODIS (ANR)

ANNODIS was a three year ANR funded project (2007-2010) in which members of the IRIT Datcha team acquired an expertise in the supervised learning of discourse parsing of texts. We built a small annotated corpus on which we carried out our parsing experiments [Muller-COLING2012]. We verified that our Segmented Discourse Representation Theory (SDRT) approach was robust, providing the basis of an understanding of the problems in mind as well as a theoretical basis for our annotation manual. DATCHA differs from ANNODIS in several ways. First, ANNODIS only dealt with single-authored text extracted from newspapers Wikipedia articles. The data in DATCHA is dialogue in a chat format, which will require adapting the annotation model to this domain.

STAC (ERC)

STAC is a 5 year multiparty and interdisciplinary project focused on strategic conversation funded by the ERC. It will end June 30, 2016. Of interest to DATCHA is the expertise we have gained in STAC concerning the analysis of chat dialogues. We have built a large corpus of multi-party chat dialogues (over 1130 dialogues) involving negotiations during the play of an on-line version of the well-known game *Settlers of Catan*. We have adapted SDRT and its accompanying model of annotation to the analysis of chat dialogue, integrating an analysis and annotation of dialogue in terms of dialogue acts specifically adapted to a task with an analysis and annotation in terms of discourse structure.

STAC like DATCHA investigates dialogue in a chat format. However, DATCHA differs from STAC since STAC employed only supervised learning techniques for discourse parsing, and the type of dialogues considered (negotiation within a specific game) are very different from contact-center chats. In DATCHA we will be investigating unsupervised techniques, and use discourse for specific applications, with already available supervision.

DECODA (ANR)

The DECODA project (2009-2012) was targeting speech analytics methods in telephone call-centre. The applicative framework was the RATP call centre service in Paris. An important part of this project was on the link between speech processing and linguistic processing. The Automatic Speech Recognition accuracy was quite low considering the difficulty of processing speech recorded from mobile phones in buses, metro and busy streets of Paris. Therefore the kind of linguistic analysis that could be performed was limited by the noise in the automatic transcriptions.

DATCHA differs from DECODA because the kind of noise that we will deal with in chat corpus is completely different from the one found in erroneous speech transcriptions. Moreover the interactive and dialogic dimensions were ignored in DECODA, and they will play a central role in DATCHA.

SENSEI (European Project - FP7)

The DATCHA project will build on the methodologies and project outcomes of the FP7 SENSEI European funded project (2013-2016) in which AMU is one of the partners. SENSEI focused on language understanding for human-human interactions in spoken conversations (telephone call-centre) and computer-mediated communication (in social media). Chat corpus are not part of SENSEI use cases, however the telephone call-centre use case and the analysis of social media data such as short messages share several common characteristics with DATCHA.

DATCHA differs from SENSEI since we will focus on the automatic adaptation of syntactic and semantic models to the particular problem of chat conversation processing. Unlike speech, chat data are not affected by the same kind of “noise”: automatic speech recognition errors for speech; typos, acronyms, out-of-vocabulary words for chat. Moreover SENSEI applicative framework is summarization. In DATCHA our focus will be on dialog strategy evaluation, conversation similarity estimation and success prediction.

ASFALDA

ASFALDA is a three year multiparty project, of which both IRIT and LIF are partners. It is funded by the ANR and will end in 31 march 2016. ASFALDA aims to provide both a French corpus with semantic annotations and

automatic tools for semantic analysis. The target semantic annotations can be characterized roughly as an explicitation of “who does what when and where”. This abstract representation is rather independent from word order and syntactic variation, and to some extend to the lexical variation found in natural language. The formal framewok chosen to represent semantic structure is Framenet. ASFALDA already produced a semantic lexicon, that maps lexical predicates to frames, a partially annotated corpus and a parser that predicts the semantic structure of a sentence.

DATCHA differs from ASFALDA mainly on the textual genre studied. ASFALDA focuses on canonical written text (the French newspaper Le Monde) while DATCHA aims to process dialogue in a chat format and will require adapting the annotation model to this domain.

1.7 Relevance for the industrial context of “Customer Relationship Management”

Improving Customer Relationship Management is a strategic issue for a company. Chat conversations between operators and customers are a unique opportunity for a company to understand the real needs and issues encountered by its customers. However current state-of-art in text analytics of such logs is limited to flat analysis about their semantic content and the estimation of customers’ satisfaction.

Improving Customer Relationship is at the heart of the strategy (Essentiels2020) of the Orange company for the next 5 years. “Reinventing Customer Relationship” is one of the five main priorities and this implies confidence, simplicity, quality and personalization. Understanding customers’ needs and points of view is crucial in order to improve Customer Experience in general.

The Orange contributors that are directly committed into the Datcha project have strong interactions with analysts from the Orange Customer Relation Branch, who are in charge of optimizing CRM from work organization or economic performance points of views. They specify together the indicators that are interesting to automatically extract from textual data. More specifically, Orange Labs contributors have strong interactions with colleagues in charge of the organization of CRM through the chat canal. This will allow us to have access to data along with relevant information about the scripts agents can follow, banks of utterances they can use. It will also guarantee that the conversation success indicators will be consistent with what is expected on-field. Finally, the use-case proposed in the Datcha project correspond to what is expected by operational units.

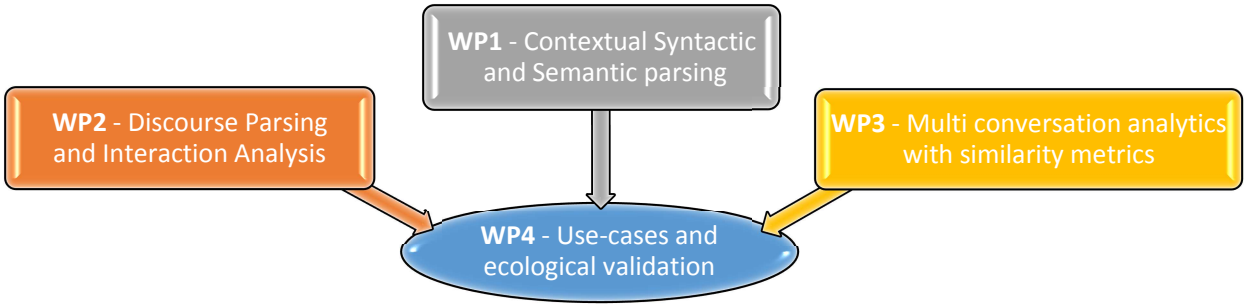
1.8 Evolution with respect to the pre-proposal

This detailed proposal is mostly similar to the pre-proposal. The consortium is unchanged and the management plan is conform to the initial pre-proposal. The only difference in the scientific objectives is related to the increased investment that will be devoted to representation learning. This is justified by very recent promising work published in the literature and conducted in the partners’ laboratories.

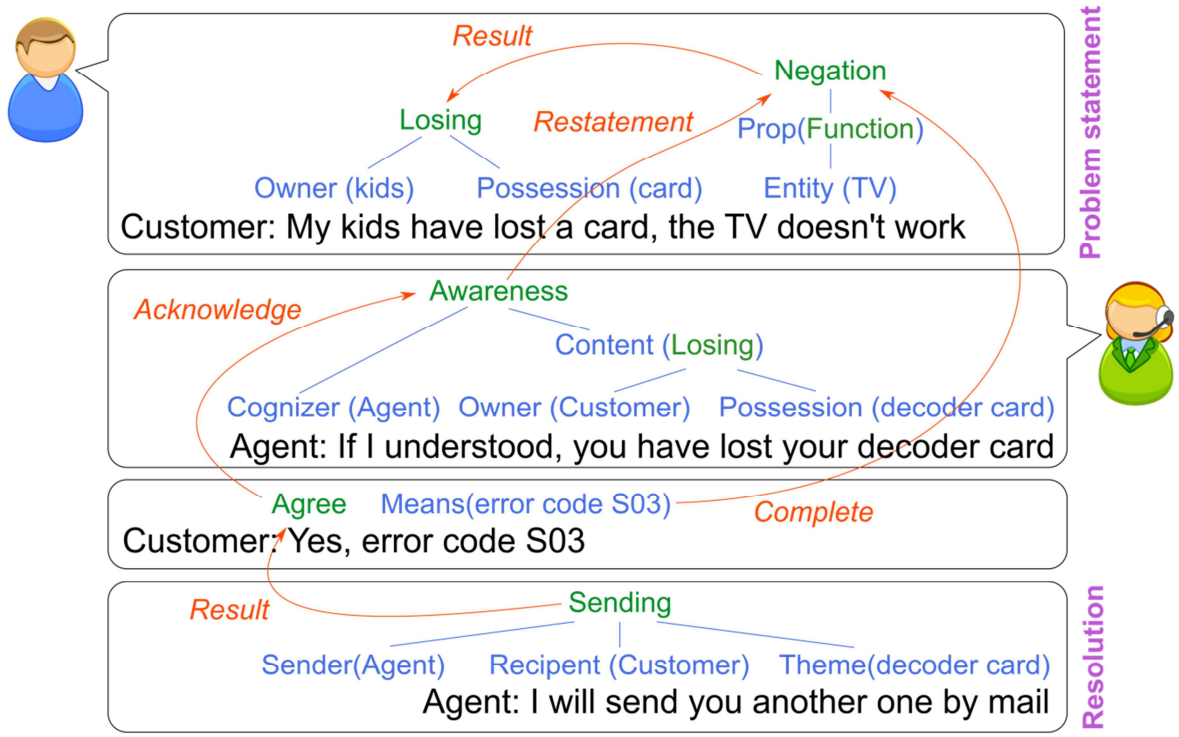
2 Programme scientifique et technique, organisation du projet

2.1 Overall strategy of the work plan

In order to address these technological breakthroughs, the work plan of the DATCHA project is divided into 5 work packages, 1 related to management, and 4 technical ones presented in the following figure:



Work packages 0 is related to the management of the project. WP1 and WP2 are dedicated to the analysis of single conversations, at the syntactic, semantic and discourse level. An example of the possible output of this single conversation parsing process is given in the following figure:



Example of possible parsing output from WP1 and WP2: a dialog is analysed as a common structure, including frames (in green), frame arguments (in blue) and discourse relations (in orange). Domain-specific concepts (in purple) can be extracted from that structure.

WP3 is about the processing of a collection of conversations in order to compare and evaluate dialog strategies and success. Finally WP4 is in charge of evaluating the impact of features provided by downstream Work Packages within the context of "ecological" use-cases defined with Orange Labs.
All these work packages are described in the tables of the next section.

2.2 Work package descriptions

Work package number	0	**Start date or starting event:**	M1
Work package title	Project Management		
Participant short name	**AMU**	IRIT	Orange
Person-months per participant	**4**	4	1

Objectives

This work package deals with general administrative responsibilities and tasks for coordinating the financial and technical aspects of the project, in order to ensure the successful fulfilment of the projects objectives. This includes ensuring effective implementation of the project according to the Contract and the Consortium Agreement.

Description of work (possibly broken down into tasks) and **role of partners**

Management activities must be adapted to the needs of the project as it evolves, but will include at least the following objectives:

- To ensure the overall general strategic and operational management and steering of the project, ensuring the accuracy, quality and timeliness of deliverables.
- To ensure the seamless integration of the activities (i.e. manage the time and result dependencies).
- To conduct the financial and administrative management of the project.
- To establish and manage effective collaboration and communications systems between partners.
- To set up and run a web-based shared information space for effective handling of documents and communications within the Consortium.
- To manage liaison with ANR, and the production of periodic reports.
- To monitor the progress of individual work packages in terms of production of deliverables according to schedule and other key indicators of progress.
- To manage the Intellectual Property and its exploitation, including any legal agreements such as the Consortium Agreement and other agreements required for the use, access and exploitation of foreground and background knowledge.
- To continuously monitor significant project risks: identify, assess probability and consequences, and devise mitigation strategies.
- To deal with any conflicts which may arise between project participants.

Deliverables

- **D0.1 First periodic report (M6 – all partners)**
- **D0.2 Second periodic report (M21 – all partners)**
- **D0.3 Final periodic report (M42 – all partners)**

Work package number	1	Start date or starting event:	M1
Work package title	Contextual Syntactic and Semantic parsing of human-human dialogs		
Participant short name	**AMU**	IRIT	Orange
Person-months per participant	**29**	14	3

Objectives

Discourse analysis of dialog relies on the syntactic and the semantic structure of the utterances produced during the dialog. While syntactic parsing of written text is a well-studied domain for which high accuracy parsers has been developed, syntactic parsing of WEB data has not received as much attention and performances are much lower. In this work package, we will develop robust and accurate parsers for chat data based on existing parsers already developed for written data and speech transcription, which handle phenomena that are specific to chat dialogs.

Description of work (possibly broken down into tasks) and **role of partners**

The semantic parsing approaches developed in this work package will learn multiple granularities of semantic analysis from labelled and/or unlabeled chat corpora. We will have two tasks, one focusing on lexical representation robust to the "noise" occurring in chat data and the second one on syntactic and semantic parsing integrating dialog context.

- Task 1.1 – Robust lexical representation (AMU, Orange)

  The first task in this WP will focus on dealing with lexical noise in chat data. Two tracks will be followed: the first one will focus on "*normalizing*" the input data in order to correct spelling mistakes, typos and chat-specific phenomenon; the second one is based on *representation learning* with embeddings in order to answer to the following question: "how can we create meaningful representations relevant for deep semantic, which can absorb the domain difference between generic text and chat interactions?" This question will be tackled at the word and sentence level in order to cope with typos, specific constructs of interactive text and context modeling.

- Task 1.2 – Contextual parsing (AMU, IRIT)

  Semantic content useful for the application purpose can be expressed at phrase, sentence, dialogue segment and whole conversation level. One particularity in the semantic analysis of dialogs is the necessity to take the dialog context into account when estimating the semantic content of a turn. Links to previous turns as well as dialog structure are necessary in order to perform semantic parsing of such data. Co-reference between several dialog turns will be studied in this task.

Deliverables

- **D1.1 – Report on robust lexical representation for chat language – M12**
- **D1.2 – Report on syntactic and semantic contextual parsing – M24**

Work package number	2	Start date or starting event:	M1
Work package title	Discourse Parsing and Interaction Analysis		
Participant short name	AMU	**IRIT**	Orange
Person-months per participant	19	**48**	4

Objectives

Discourse parsing is the automatic analysis of dialog structure, providing links between conversation turns during a chat, and labelling them with their function in the conversation. From this, Interaction analysis will consist in studying typical sequences of moves (requests, answers, acceptances, ...) and global features of the exchange, such as the temporal flow of interactions. One of the issues in this perspective is for instance to detect disruptions in the flow, when customers and agents are typing messages simultaneously, possibly leading to miscomprehensions.

Description of work (possibly broken down into tasks) and **role of partners**

This WP will have an important interactive component with WP1, and provide input for WP3/4. As a first step, we will consider the tasks as independent from WP1, in a pipe-line approach, with local models at the semantic level, and then we will consider joint models, something we are already familiar within the context of discourse parsing.

- Task 2.1– Model adaptation and discourse parsing (IRIT, AMU)

  We will adapt annotation schemes developed for journalistic text in the ANR ANNODIS and for bargaining chat conversations in the STAC project to the chat data of DATCHA. We plan to leverage the features exploited on the English STAC corpus to French. These features, like basic speech act type (question, assertion, demand), the speaker and characteristics of the speaker's turn are general enough to adapt to French and they have proven useful in determining certain crucial discourse relations for dialogue. For the structural part we will investigate how the Stac models behave on the DATCHA chat data, to decide if they are sufficient for the downstream tasks, and/or if they can be used

  As a weak supervision to be combined with more corpus-specific scripted approaches, we plan to make use of the templates that Orange provides to its technical staff that engage customers in chat sessions as a rough guideline for automatically building the discourse structure of a dialogue in our corpus.

  We also plan to transfer and adapt our model of dialogue acts to the information seeking task. We will make use of DAMSL's typology of acts in this respect. We will also adapt the preference lexicon developed for STAC by Cadillac (2013 thesis 2013 EMNLP) to DATCHA's framework, as well as exploit the sentiment analysis lexicon for French already developed by Farah Benamara, which also includes information useful to the detection of speaker attitudes.

- Task 2.2 – Dialogue structure and temporal flow (Orange, IRIT)

  We will conduct a deep analysis of the temporal flow of a conversation. Preliminary studies revealed that message length in terms of number of characters and message duration are not necessarily correlated and cross observation of these features, combined with discourse analysis will lead to a better understanding of the conversation flow, with the possibility to detect disruptions or inconsistencies that can lead to misunderstandings between customer and agents.

  We also plan to use session templates in this task, as it will be interesting to see when dialogue participants adhere to a particular template and when they depart from it. These features should allow us to structure a dialogue at different levels of granularity with different degrees of accuracy. These outcomes can provide relevant features for further tasks studied in WP3 and WP4.

Deliverables

- **D2.1 Report on model adaptation and dialogue structure– M18**
- **D2.2 Dialogue structure parser – M18**
- **D2.3 Report on interaction analysis – M24**

Work package number	3	Start date or starting event:	M12
Work package title	Multi conversation analytics with similarity metrics		
Participant short name	AMU	**IRIT**	Orange
Person-months per participant	20	**36**	5

Objectives

Downstream applications considered for this WP all rely on the availability of punctual information describing certain aspects of the interaction: lexical signals of specific features, sentiment expression, but they would also benefit from more global information, some of which are easily collected (number of turns for instance), some of which need a more elaborate analysis of the interaction: question/answer pairs for instance, feedback turns and the level of agreement they reflect, phenomenon raising various difficult problems. Beyond that, it remains an open question as to what structural properties can reflect different kind of interactions that are relevant to predict, e.g., customer satisfaction.

This is where a notion of similarity of two conversations is crucial (1) to define global features of conversations that can be useful to discriminate between them in a supervised setting (where one want to predict the success of an interaction), (2) to define a criterion on which to perform clustering of conversations, as an analytical tool, which in turn is useful : (3) for online customer support, where finding similar interactions (both from the content and the interaction point of view) can help an agent interacting with the customer, or help predict when an interaction is going in an undesirable direction and address the situation as soon as possible.

Description of work (possibly broken down into tasks) and **role of partners**

WP3 consists in two tasks related to dialog representation and clustering

- Task 3.1 – Learning of conversation level representations (AMU)

  This task consists in learning conversation-level representations from deep semantic and dialog structure analyses extracted from chat corpora in WP1 and WP2. In the same way word embeddings extracted from large corpora exhibit the property that words with similar embeddings are used in similar context, leading to the emergence of synonyms as well as various kind of cohyponyms (such as countries, car models...) or subdomains (foreign words, slang...), we expect conversation representations to place conversations with similar semantic or structural properties in similar locations. This task will consist in finding the right regime for learning conversation-level embeddings and form a space where conversations can be compared with simple metrics, leveraging current work on representation learning. An analysis of the output space and the different features that can be used as input will be performed in this task.

- Task 3.2 – Defining distances and similarity metrics for conversation clustering (AMU, IRIT, ORANGE)

  Representations learned as embeddings have the problem of being opaque, preventing from analyzing dialogs in regard to particular features. Therefore, in addition to representations explored in task 3.1, different similarity metrics will be defined for conversation clustering to compare dialogs according to various features. For instance, dialogues may share a common topic, resolution strategy or customer attitude. They may also be similar in that they follow a certain pattern of questions and answers. The sort of information sources that will be exploited for dialog analysis and parsing in WP2 will provide natural clusterings of the conversations. This task will study different structural distances (some preliminary work on this topic was done in [Venant-SIGDIAL2013]) with attention paid to robustness, given noisy discourse parsing output. A combination with more topically-oriented information and sentiment analysis will also be studied.

Deliverables

- **D3.1 Report on conversation-level representation learning – M24**
- **D3.2 Report on metric design for conversation clustering – M36**

Work package number	4	Start date or starting event:	M21
Work package title	Use-cases and ecological validation		
Participant short name	AMU	IRIT	**Orange**
Person-months per participant	22	18	**12**

Objectives

An important technical problem in all areas of NLP that deal with the analysis and retrieval of deep semantic information is evaluation (see e.g [Lewis-TAVL2013]). While [Cadilhac-EMNLP2013] shows that some aspects of discourse structure can substantially help tasks like the retrieval of preferences from chat, the problem remains a difficult one. The use-cases defined in this work package, however, provide us with ecological evaluation protocols.

Very large chat conversation corpora are available at Orange Labs for various domains (including TV, Internet or Mobile), both for technical assistance and commercial services. An anonymization process is already in place allowing data to be shared with partners in a contractual context such as a PRCE, without infringing customers' or operators' privacy rights. A selected corpus of chat conversation has already been manually annotated in terms of problem category and solution provided.

Models and algorithms developed in the first three work packages will be applied to specific tasks (*use-cases*) evaluated in *ecological conditions* thanks to the direct access to CRM professional provided through Orange Labs. These use-cases will include *conversation analytics* at the conversation and collection levels, *conversation success prediction* on the basis of criteria defined by operational units, and *online conversation solving*.

Description of work (possibly broken down into tasks) and **role of partners**

- Task 4.1 Conversation analytics (Orange)

  The objective of this task is to measure the impact of downstream analyses performed in WP1-2-3 for classical analytics tasks such as call reason classification. Annotated corpora already exists to evaluate this classification task and we will measure the improvement yielded by advanced semantic, discursive and structural features with respect to baseline statistical classification approaches that only rely on words to train classification models. Beyond these objective measures, the resulting analytics result will be integrated in the Orange VisualCRM platform in order for the operational CRM analysts to validate the improvement.

  CRM Analytics also implies providing efficient search capacities in order for analysts to be able to retrieve particular conversations depending on their needs. Based on an already available implementation of a search engine based on ElasticSearch, which as a baseline only performs full-text search, we will integrate the higher level characterizations provided by WP1-2-3 into the search process and evaluate the impact of these new functionalities with CRM analysts.

- Task 4.2 Dialogue Success prediction (Orange, IRIT, AMU)

  Additionally to conversation logs, we will have access to customer surveys performed just after the conversation, including answers to questions related to customer satisfaction. In this context, we will be able to study the correlation between the clusters provided by WP3, and to another extent, we will be able to perform customer satisfaction prediction, with a direct way to evaluate the impact of all the features provided by downstream work packages.A recent study [Dixon-HBR2010] emphasizes that customers' satisfaction alone is not sufficient to create customer loyalty: solving customer problems and reducing customer effort, expressed by a "Customer Effort Score" (CES) are a complement to the widely used "Net Promoter Score" or "Customer Satisfaction Score". [DeHaan-IJRM2015] conducted a comparative study of customer feedback metrics. As an attempt to cross automatic knowledge discovery with marketing trends in CRM, we propose as an exploratory task, to model Customer Effort from the vast collection of logs available through DATCHA thanks to dialog success and dialog policy/strategy estimation. We believe that the joint analysis of conversations from both semantic and

discourse parsing angles is a key component towards this objective.

- Task 4.3 Online conversation solving (Orange; IRIT, AMU)

  The objective of this task is to perform online prediction of the problem's possible solutions. If coupled with the features developed throughout the project (conversation similarity, dialogue success prediction, and dialogue structure analysis) we can design an efficient prototype where agents, could be proposed online some patterns that yielded successful conversations for a similar solution. Corpora are already available with a manual annotation of solutions proposed to customers. Hence an objective evaluation can be performed for this task, with an interesting question being: when are we able to efficiently predict the solution? (after how many turns?, from which dialogue phase?...). Additionally we will develop an online conversation solving prototype to demonstrate our approach.

Deliverables

- **D4.1 Report on conversation analytics M28**
- **D4.2 Report on dialogue success prediction M36**
- **D4.3 Report on online conversation solving M42**

2.3 Gantt Chart

		Année 1												Année 2											
	mois	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
WP0	**Management**						◆						◆												◆
WP1	**Contextual syntactic and semantic analysis**																								
T1.1	Robust lexical representation												◆												
T1.2	Contextual parsing																								◆
WP2	**Discourse Parsing and interaction analysis**																								
T2.1	Model adaptation and discourse parsing																		◆						
T2.2	Dialog structure and temporal flow																		◆						◆
WP3	**Multi conversation analytics with similarity metrics**																								
T3.1	Learning of conversation level representations																								◆
T3.2	Defining distances and similarity metrics for conversation clustering																								
WP4	**Ecological validation and evaluation**																								
T4.1	Conversation analytics																								
T4.2	Dialogue Success prediction																								
T4.3	Online conversation solving																								
	◆ Livrables																								

		Année 3												Année 4					
	mois	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
WP0	**Management**												◆						◆
WP1	**Contextual syntactic and semantic analysis**																		
T1.1	Robust lexical representation																		
T1.2	Contextual parsing																		
WP2	**Discourse Parsing and interaction analysis**																		
T2.1	Model adaptation and discourse parsing																		
T2.2	Dialog structure and temporal flow																		
WP3	**Multi conversation analytics with similarity metrics**																		
T3.1	Learning of conversation level representations																		
T3.2	Defining distances and similarity metrics for conversation clustering												◆						
WP4	**Ecological validation and evaluation**																		
T4.1	Conversation analytics				◆														
T4.2	Dialogue Success prediction												◆						
T4.3	Online conversation solving																		◆
	◆ Livrables																		

2.4 Summary of effort

DATCHA	Partners			
	1	2	3	
	AMU	IRIT	Orange	TOTAL
WP0 - Management	4	4	1	**9**
WP1 – Syntactic/ Semantic parsing	29	14	3	**46**
WP2 – Discourse parsing/Interaction	19	48	4	**71**
WP3 – Dialog similarity/success	20	36	5	**61**
WP4 – Use-case/Validation	22	18	12	**52**
TOTAL	**94**	**120**	**25**	**239**

2.5 Budget

The requested funding for DATCHA is 380K euros (90% personnel, 3% equipment, 7% functioning) with the funding of 1 PhD and 1 PostDoc, each for 3 years. For the functioning budget, we plan to have 2 internal meetings every year, for 2 people per site (500€ each), as well as funding for the attendance to 2 scientific conferences for 2 people every year for each partner (1000€ each).
Equipment will involve work stations for personnel hired for the project as well as computing power with GPU cards for neural network experiments as well as storage capacity for processing large corpus of chat.

For personnel, we have the following repartition:

- AMU requests funding for a PhD (36 months) as well as 3 students for a period of 6 months each (internship). The PhD will work on robust conversation parsing, more specifically on joint syntax/ semantic/ discourse analysis and representation learning. The internship students will come from AMU Master Degree in Computer Science and will work on prototype development for WP4. In addition to requested men/months, AMU will dedicate 40 men/months of permanent staff for the project, on management and all scientific WPs.

- Orange requests a 25 men-months funding for three permanent research engineers. 6% of the effort will be devoted to management and dissemination, 24% will be affected for the core analysis workpackages WP2 and WP3, 24% to WP3 on multi conversation analytics and 46% will be devoted to the use-case and ecological validation workpackage (WP5).

- IRIT requests funding for 36 months of post-docs for the following tasks: adaptation of discourse parsing to the Orange data and coordination of joint prediction of dialogue structure with semantic parsing (1.5y), similarity for conversation clustering and task-oriented validation (1.5y).

  In addition to the requested funding, IRIT permanent staff will contribute to the various work packages for 84 men.months.

2.6 Risks and Associated contingency plans

Risk	Impact	Probability	Contingency plan
No access to Orange chat data	High	Low	Orange already collected and annotated large corpus of chat conversation from their customer contact platform. New logs can be collected every day.
No access to Orange CRM for ecological validation	Medium	Low	The Orange researchers involved in DATCHA have already a strong link with the visual CRM team. Even if internal change in Orange make these ecological experimentation impossible, simulation based on previous CRM experience will allow us to validate our models on realistic scenarios.
Difficulty in parsing chat data due to lexical/syntactic "noise"	High	Low	DATCHA partners have experience in parsing "noisy" data such as spontaneous speech (AMU, Orange) and chat room data (IRIT). All the methods developed can generate and use partial parse rather than full parse of every turn of a conversation.

2.7 Presentation of the consortium

We propose a PRCE partnership between Orange and two academic laboratories in order to carry on breakthrough research on large scale realistic data. The availability of large « real-life » corpora is crucial for research in today's information society. If it is not realistic to consider making publicly available such strategic data as contact center conversations, collaborative research projects are an excellent opportunity for academic laboratories to access such data, and for companies to increase the value of these data by collaborating with internationally recognized researchers.

The main strength of the DATCHA consortium is to group together key players in France in the fields of automatic language processing (**LIF, IRIT**) and an industrial research lab (**Orange Labs**). **LIF** and **IRIT** have internationally recognized skills in complementary domains (respectively semantic and discursive analysis) that are essential to reach the objectives of the project. **Orange Labs** researchers have experience in collaborative research in the field of language processing and are can bridge the gap between applicative needs expressed by operational teams and high level formalisms explored in the academic research. **Orange** will bring realistic use-cases for contact-centre optimization and diagnostic to the project to help designing and evaluating new models and systems to process human-human dialog strategies.

The small size of the DATCHA consortium will guarantee optimal collaboration between partners as all the scientific WPs are shared by everybody. This will be a real collaborative project in the sense that the scientific complementarity of each partner matches the description of each work package: no WP can be realized by 1 partner alone. Moreover the three partners have experience in working together: AMU researchers and Orange have been partners in several previous ANR and EU projects (LUNA, PERCOL); AMU and IRIT researchers have collaborated in the past through PhD programs or exchange of postdocs

Partner n°1	**Aix Marseille Université – AMU**	**Academic**
Description	The TALEP team of the LIF laboratory of AMU develops symbolic and statistical methods for the automatic processing of textual and speech data. LIF will bring to the project its expertise in the linguistic analysis of non-canonical language.	
Role in the project	Developing deep semantic analysis methods on chat data	
Involved scientists	**Prof. Frédéric Béchet** (coordinator; former coordinator of the ANR DECODA and ANR PERCOL - DEFI REPERE) [Bazillon-LREC2012,Bechet-LREC2012,Camelin-TASLP2010]; **Prof. Alexis Nasr,** (research interests on automatic processing of written language and speech, with a focus on probabilistic syntactic models) [Bazillon-LREC2012]; **Dr. Benoit Favre** (research interests on speech and text summarization and understanding with a focus on machine learning methods) [Bazillon-LREC2012].	

Requested funding	1 PhD (3 years), **120 k€** (88% non-permanent personnel, 4% equipment, 8% functioning)	
Partner n°2	**IRIT (Institut de Recherche en Informatique de Toulouse)**	**Academic**
Description	The MELODI team of IRIT is a subgroup of 14 permanent researchers focusing on natural language processing, and knowledge representation. The group has an internationally recognized expertise in formal semantics, formal pragmatics and discourse analysis.	
Role in the project	Development of discourse models	
Involved scientists	**Nicholas Asher** (DR CNRS), **Stergos Afantenos**, and **Philippe Muller** (both University of Toulouse associate professors) for discourse parsing and summarization [Muller-COLING2012,Venant-SIGDIAL2013]; **Farah Benamara** (associate-professor at the University of Toulouse) for opinion mining and sentiment analysis [Cadilhac-EMNLP2013].	
Requested funding	3 years PostDoc, **165 k€** (91% non-permanent personnel, 3% equipment, 6% functioning)	
Partner n°3	**Orange Labs**	**Group**
Description	The Orange Labs network is the key contributor for the Orange Group's innovation. The FAST team (Future Architecture and Textual Analysis) gathers researchers, developers and architects that provide Natural Language Processing solutions towards the Group.	
Role in the project	Development of solutions for the foreseen use-cases; Integration in existing CRM tools; Interface with operational units for providing corpora and gathering feedback on results.	
Involved scientists	**Dr Géraldine Damnati** (research interests in natural language understanding and data mining for CRM, involved in internal projects with operational units) [Camelin-TASLP2010, Bouchekif-ICASSP2014]; **Dr Delphine Charlet** (research interests on language understanding with special attention paid to applicative constraints) [Bouchekif-ICASSP2014]; **Dr Aleksandra Guerraz** (research interest in syntactic and semantic analysis with a focus on linguistic resource construction and maintenance) [Collin-DEFT2013]	
Requested funding	25 men-month, 92.8 k€ (94% permanent personnel, 6% functioning)	

2.8 Curriculum Vitae of the coordinator of the project

Frédéric Béchet is a researcher in the field of Speech and Natural Language Processing. His research activities are mainly focused on Spoken Language Understanding for both Spoken Dialogue Systems and Speech Mining applications. After studying Computer Science at the University of Marseille, he obtained his PhD in Computer Science in 1994 from the University of Avignon, France. Since then he worked at the Ludwig Maximilian University in Munich, as a Professor Assistant at the University of Avignon, as an invited professor at ATT Research Shannon Lab in Florham Park, New Jersey. Frédéric Béchet is currently a full Professor of Computer Science at the Aix Marseille University, and a member of the Natural Language Processing research group of the Laboratoire d'Informatique de Marseille (LIF-CNRS).

Frédéric Béchet has been involved in many French and European research programs in the fields of Speech Processing and Spoken Dialog Systems: FP5 SMADA STREP, FP6 LUNA STREP, FP6 PASCAL NoE, FP7 SENSEI, ANR EPAC, ANR SEQUOIA, ANR EDYLEX, ANR ORFEO. He was the coordinator of two French research programs funded by the French ANR agency: DECODA - speech mining in call-center data (ANR CONTINT 2009) and PERCOL - people identification in broadcast videos (REPERE - CONTINT 2010).

3 Stratégie de valorisation, de protection et d'exploitation des résultats, impact global de la proposition

The major objective of the dissemination plan is to bring the intermediate and final results of the project to the attention of the Natural Language Processing research community as well as the technology and service providers.

Dissemination will take place within the whole duration of the project as-a-whole and individually by partners in their own professional and business network.

3.1 Dissemination toward scientific community

Scientific dissemination of DATCHA's results will use the traditional channels of scientific journals, international conferences and workshops. In these events intermediate results and demos will be presented. Beyond the Natural Language Processing community, we also plan to disseminate our work towards other communities linked to Customer relationship Management that can use our results (eg. Interface community or Marketing community).

To engage the scientific community to the research issues dealt by DATCHA we plan on two separate activities:

- Firstly we will annotate with DATCHA's technology a corpus of publically available chat messages such as [Martell-ICSC2007] for English and the corpus for French initially described in [Falaise-RECITAL2005] and completed in [Chanier-JLCL2014]. We will build links with the CoMeRe project and other project dedicated to study this new form of textual data.
- Secondly we plan on launching shared evaluation tracks sponsored by DATCHA in evaluation programs such as DEFT in France. This type of events are very effective into both sharing resources back to the community and bringing in more researchers and evaluate externally DATCHA's approach to the chat conversation understanding research lines.

3.2 Expected exploitable results for Orange

The Datcha project results are directly exploitable for Orange, where a better understanding of customer-agent interactions is at the heart of the company's strategy plan.

Orange Labs has developed an analytics tool with Textmining functionalities (classification, keyword extraction) which is already deployed towards Orange operational units.

This tool, called VisualCRM, deals with regular flows of text inputs ranging from verbatims of customer surveys to written reports about customer complaints. It performs automatic analysis and provides the results through an interactive dashboard web interface.

So far the integration of chat conversations into the VisualCRM tool is limited because the available Textmining analysis approaches can only provide global classification of the call motive, relying on simple text classification approaches.

The Orange contributors involved in the Datcha project have developed the Textmining aspect of the VisualCRM tools and have contributed to the specifications of the interface. The Datcha project outcomes will be directly transferred to improve the exploitation of chat conversation analytics within the VisualCRM platform.

This constitutes a good opportunity for our research results to be evaluated by end users from operational units. Additionaly, the use-cases defined in WP4 correspond to concerns that are regularly expressed by operational teams and we will be able to gather feedback on the relevance of our solutions.

Beyond Orange, the results of the project will be useful for all the companies who wish to develop this communication canal with their customers.

In fact, even if the definition of use-cases is driven by Orange particular needs, they can easily be generalized to other companies.

3.3 Management of knowledge and intellectual property

The DATCHA Consortium partners are ready to bring in their individual expertise and knowledge to make the project a success and implement an appropriate IP management process. Knowledge created during the project will be distributed within the consortium to enable a targeted and coordinated development towards the project goals, which also requires active knowledge exchange between the project partners and WPs.

Formal issues of Management of Knowledge will be addressed in the Consortium Agreement (CA).
This includes the coordination at consortium level of new knowledge generated by the project, i.e. the results, including information, whether or not they can be protected.
The idea is to regulate rights and obligations and detailed access rules to knowledge, whenever such is necessary to create a fair and comfortable basis for all partners to work together in the project.

All partners have already participated in ANR projects and have a substantial experience from CA negotiations.

4 References

[Asher-ECAI2008] N. Asher, F. Benamara, Y. Mathieu, 'Categorizing Opinion in Dis- course', Proceedings ECAI 2008.

[Asher-Strategic2013] N. Asher, S. Paul, `Infinite games with uncertain moves', First International Conference on Strategic Reasoning, ETAPS workshop, Rome, 2013.

[Asher-SALT2015] N. Asher, A. Venant, 'OK or not OK?', Semantics and Linguistic Theory 25, Stanford, 2015

[Asher-NLP-CUP2003] N. Asher, A. Lascarides, "Logics of Conversation", studies in NLP, Cambridge University Press, 2003

[Bazillon-LREC2012] T. Bazillon, M. Delplano, F. Bechet, A. Nasr, B. Favre, "Syntactic Annotation of Spontaneous Speech: Application to Call-Center Conversation Data", *Proc. LREC*, 2012

[Bechet-LREC2012-1] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, E. Arbillot, "DECODA: a call-centre human-human spoken conversation corpus", *Proc. LREC*, 2012

[Bechet-LREC2012-2] Thierry Bazillon and Mélanie Deplano and Frédéric Bechet and Alexis Nasr and Benoît Favre "Syntactic annotation of spontaneous speech: application to call-center conversation data" International Conference on Language Resources and Evaluation (LREC) 2012

[Bengio-JMLR2003] Bengio, Yoshua, et al. "A neural probabilistic language model." The Journal of Machine Learning Research 3 (2003): 1137-1155.

[Bengio-PAMI2012] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.8 (2013): 1798-1828.

[Bouchekif-ICASSP2014] A. Bouchekif, G. Damnati, D. Charlet, "Intra-content term weighting for topic segmentation", *Proc. ICASSP*, 2014.

[Cadilhac-EMNLP2013] A. Cadilhac, N. Asher, F. Benamara, & A. Lascarides, "Grounding Strategic Conversation: Using Negotiation Dialogues to Predict Trades in a Win-Lose Game". *Proc. EMNLP (pp. 357-368),* 2013.

[Camelin-TASLP2010] N. Camelin, F. Béchet, G. Damnati, R. de Mori, "Detection and Interpretation of Opinion Expressions in Spoken Surveys", IEEE Trans. on Audio, Speech, and Language Processing, Vol. 18, N° 2, February 2010, Pages 369-381.

[Chanier-JLCL2014] Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi J. & Seddah, D. "The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres.", Journal for Language Technology and Computational Linguistics, Vol. 29, N° 2, 2014.

[Cho-ARXIV-14] Cho, Kyunghyun, et al. "Learning phrase representations using rnn encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[Chrupala-ARXIV2013] Chrupała, Grzegorz. "Text segmentation with character-level text embeddings." arXiv preprint arXiv:1309.4628 (2013).

[Collin-DEFT2013] O. Collin, A. Guerraz, A. Hiou, N. Voisine, "Participation de Orange Labs à DEFT 2013", *in Actes DEFT*, 2013.

[Collobert-JMLR2011] Collobert, Ronan, et al. "Natural language processing (almost) from scratch." The Journal of Machine Learning Research 12 (2011): 2493-2537.

[Dahlmeier-EMNLP2012] Daniel Dahlmeier, Hwee Tou Ng, A beam-search decoder for grammatical error correction, Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012

[Deerwester-JASIS-1990] Deerwester, Scott C., et al. "Indexing by latent semantic analysis." JAsIs 41.6 (1990): 391-407.

[DeHaan-IJRM2015] De Haan, E., Verhoef, P. C., & Wiesel, T., "The predictive ability of different customer feedback metrics for retention". International Journal of Research in Marketing, 2015.

[Deng-ICASSP2013] Deng, Li, et al. "Recent advances in deep learning for speech research at Microsoft." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[Dixon-HBR2010] M. Dixon, K. Freeman, N. Toman "Stop trying to delight your customers", *Harvard Business Review, 88 (7/8)*, 2010.

[Falaise-RECITAL2005] Falaise, A. (2005). Constitution d'un corpus de franç}ais tchaté. RECITAL. Dourdan.

[Forsyth-ICSC2007] Eric N. Forsyth and Craig H. Martell, "Lexical and Discourse Analysis of Online Chat Dialog," Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 19-26, September 2007.

[Golberg-ARXIV2014] Goldberg, Yoav, and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." arXiv preprint arXiv:1402.3722 (2014).

[Han-HLT2011] Bo Han, Timothy Baldwin, "Lexical normalisation of short text messages: Makn sens a\# twitter" Human Language Technologies, 2011

[Hochreiter-NC1997] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[Huang-ACL2012] Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

[Jonsson-DCOURSE1998] "Electronic discourse: On speech and writing on the Internet." Retrieved from http://www.ludd.luth.se/~jonsson/D-essay/ElectronicDiscourse.html

[Kucukyilmaz-IPM2008] Kucukyilmaz, T., Cambazoglu B. B., Aykanat, C. And Can, F., "Chat mining: Predicting user and message attributes in computer-mediated communication", Information Processing and Management, Vol. 44, pp. 1448-1466, 2008.

[Lauly-NIPS2014] Lauly, Stanislas, et al. "An Autoencoder Approach to Learning Bilingual Word Representations." Advances in Neural Information Processing Systems. 2014.

[Le-ARXIV2014] Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

[Levy-ACL2014] Levy, Omer, and Yoav Goldberg. "Dependency based word embeddings." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Vol. 2. 2014.

[Lewis-TACL2013] M. Lewis & M. Steedman. "Combined Distributional and Logical Semantics". *TACL*, 1, 179-192, 2013.

[Martell-ICSC2007] Martell, E., Forsyth , N., & Craig, H. (September 2007). Lexical and Discourse Analysis of Online Chat Dialog. Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), (pp. pp. 19-26,)

[Mesnil-IS2013] Mesnil, Grégoire, et al. "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding." INTERSPEECH. 2013.

[Mikolov-ASRU2011] Mikolov, Tomas, et al. "RNNLM-Recurrent neural network language modeling toolkit." Proc. of the 2011 ASRU Workshop. 2011.

[Mikolov-NIPS2013] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems. 2013.

[Muller-COLING2012] P. Muller, S. Afantenos, P. Denis, and N. Asher. "Constrained decoding for text-level discourse parsing", *Proc. COLING* 2012

[Muzerelle-CMLF2012] Muzerelle J., E. Schang E., Antoine J.-Y., Eshkol I., Maurel D., Boyer A., Nouvel D. (2012). Annotations en chaînes de coréférences et anaphores dans un corpus de discours oral spontané en français. Actes 3ème Congrès Mondial de Linguistique Français

[Nasr-LREC2014] Alexis Nasr, Frederic Bechet, Benoit Favre, Thierry Bazillon, Jose Deulofeu, Andre Valli, "Automatically Enriching Spoken Corpora with Syntactic Information for Linguistic Studies", International Conference on Language Resources and Evaluation (LREC), 2014

[Pennington-EMNLP2014] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014) 12 (2014).

[Socher-EMNLP2012] Socher, Richard, et al. "Semantic compositionality through recursive matrix-vector spaces." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.

[Sutskever-NIPS2014] Sutskever, Ilya, Oriol Vinyals, and Quoc VV Le. "Sequence to sequence learning with neural networks." Advances in Neural Information Processing Systems. 2014.

[Taigman-CVPR-2014] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014.

[Tafforeau-IS2015] Jeremie Tafforeau, Thierry Artières, Benoit Favre and Frederic Bechet. "Adapting lexical representation and OOV handling from Written to Spoken language with Word embedding", Interspeech 2015, Submitted.

[Tang-ACL2014] Tang, Duyu, et al. "Learning sentiment-specific word embedding for twitter sentiment classification." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014.

[Trimarco-PM2014] Trimarco, P., “Digital Textuality”. Palgrave Macmillan, 2014.

[Turian-ACL2010] Turian, Joseph, Lev Ratinov, and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010.

[Venant-SIGDIAL2013] A. Venant, N. Asher, P. Muller, P. Denis, S. Afantenos. “Expressivity and comparison of models of discourse structure”, *Proc. SIGDIAL*, 2013.

[Venant-IWCS2015] A. Venant & N. Asher, ’Dynamics of Public Commitments in Dialogue’, Proceedings of IWCS, London, 2015.

[Versley-LREC2008] Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A. (2008) *BART: A Modular Toolkit for Coreference Resolution*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)

[Vinyals-ARXIV-2014] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." arXiv preprint arXiv:1411.4555 (2014).

[Zou-EMNLP2013] Zou, Will Y., et al. "Bilingual Word Embeddings for Phrase-Based Machine Translation." EMNLP. 2013.