

Syntactic parsing of chat language in contact center conversation corpus

Alexis Nasr, Geraldine Damnati, Aleksandra Guerraz, Frederic Bechet

▶ **To cite this version:**

Alexis Nasr, Geraldine Damnati, Aleksandra Guerraz, Frederic Bechet. Syntactic parsing of chat language in contact center conversation corpus. Annual SIGdial Meeting on Discourse and Dialogue, Sep 2016, Los Angeles, United States. pp.175 - 184, 10.18653/v1/W16-3621 . hal-01454768

HAL Id: hal-01454768
https://hal.science/hal-01454768

Submitted on 15 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Syntactic parsing of chat language in contact center conversation corpus

Alexis Nasr[1], **Geraldine Damnati**[2], **Aleksandra Guerraz**[2], **Frederic Bechet**[1]
(1) Aix Marseille Universite - CNRS-LIF, Marseille, France
(2) Orange Labs - Lannion, France

Abstract

Chat language is often referred to as *Computer-mediated communication (CMC)*. Most of the previous studies on chat language has been dedicated to collecting "*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 paper in understanding the mechanisms of a more formal kind of CMC: dialog chat in contact centers. The particularities of this type of dialogs and the type of language used by customers and agents is the focus of this paper towards understanding this new kind of CMC data. The challenges for processing chat data comes from the fact that Natural Language Processing tools such as syntactic parsers and part of speech taggers are typically trained on mismatched conditions, we describe in this study the impact of such a mismatch for a syntactic parsing task.

1 Introduction

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 (Jonsson, 1997), who conducted an early work on data gathered through the Internet Relay Chat protocol and through emails: *"eletronic discourse is neither writing nor speech, but rather written speech or spoken writing, or something unique"*.

Recent projects in Europe, such as the CoMeRe (Chanier et al., 2014) or the STAC (Asher, 2011) 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 "*chat room*" data as it is the kind of data which is the most accessible on the WEB. (Achille, 2005) constituted a corpus in French. (Forsyth and Martell, 2007) and (Shaikh et al., 2010) describe similar corpora in English. (Cadilhac et al., 2013) have studied the relational structure of such conversations through a deep discursive analysis of chat sessions in an online video game.

This kind of data falls under the *informal register* whereas we are interested in this paper in understanding the mechanisms of a more formal kind of CMC: dialog chat in contact centers. This study is realized in the context of the DATCHA project, a collaborative project funded by the French National Research Agency, which aims at performing unsupervised knowledge extraction from very large databases of WEB chat conversations between operators and clients in customer contact centers. As the proportion of online chat interaction is constantly growing in companies' Customer Relationship Management (CRM), it is important to study such data in order to increase the scope of Business Analytics. Furthermore, uch corpora can help us build automatic human-machine online dialog systems. Among the few works that have been published on contact center chat conversations, (Dickey et al., 2007) propose a study from the perspective of the strategies adopted by agents in favor of mutual comprehension, with a focus on discontinuity phenomena, trying to understand the reasons why miscomprehension can arise. (Wu et al., 2012) propose a typology of communication modes between customers and agents through a study on a conversa-

Proceedings of the SIGDIAL 2016 Conference, pages 175–184,
Los Angeles, USA, 13-15 September 2016. ©2016 Association for Computational Linguistics

tion interface. In this paper we are interested in evaluating syntactic parsing on such data, with a particular focus on the impact of language deviations.

After a description of the data and the domain in section 2, we introduce the issue of syntactic parsing in this particular context in section 3. Then a detailed analysis of language deviations observed in chat conversations is proposed in section 4. Finaly, experiments of part of speech (pos hereafter) tagging and syntactic parsing are presented in section 5.

2 Chat language in contact centers

In the book entitled "*Digital textuality*" (Trimarco, 2014), 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 participants is also pointed out in (Kucukyilmaz et al., 2008). 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 described in (Trimarco, 2014) are related to turn taking and synchronicity issues on the one side, and the use of semiotic resources such as punctuation or emoticons on the other.

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. The corpus used in this study has been collected from Orange (the main French telecom operator) online assistance for Orange TV customers who contact the assistance for technical problems or information on their offers. In certain cases, the conversation follows a linear progress (as the example given in Figure 1) 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: 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.

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 writing starts. 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 one can see in the example in Figure 1, chat conversations are dissimilar from edited written text in that they contain typos, agrammaticalities and other informal writing phenomena. They are similar to speech in that a dialog 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, 2014).

All these properties along with the particular type of language used by customers and agents is the focus of this paper towards understanding this new kind of CMC data. The challenges for processing chat comes from the fact that analysis tools such as syntactic parsers and pos taggers are typically trained on mismatched conditions, we describe in this study the impact of such a mismatch for these two tasks.

3 Syntactic parsing of chat language

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

In the DATCHA project, we perform syntactic parsing as well as 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.

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

	You're in contact with AGENT
AGENT	Hello, I'm AGENT, how can I help you?
CUST	my children have lost the card in the modem and I don't have tv anymore what can I do?
AGENT	Can you confirm your line number?
CUST	NUMTEL
AGENT	If I understand correctly you lost your decoder card
CUST	Yes that's right
CUST	error code S03
AGENT	No problem, I will send you another card to your home address.
CUST	can I come and get it today
AGENT	You can't get a card from an Orange store because they can only proceed to exchanges.
CUST	ok thank you for sending it as soon as possible you have my coordinates
AGENT	Yes I have them in your record.
CUST	ok fine within 48h maximum 72h for the card
AGENT	You will receive it according to delivery time at the address in your record.
CUST	ok fine thank you
AGENT	You're welcome
AGENT	Before you go, do you any other question?
CUST	no thank you

Figure 1: Example of conversation in the TV assistance domain, in its original forme (above) and a translation without errors (below)

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 benchmark corpora and process text that comes from CRM chat. The experience we gathered on parsing speech transcriptions in the framework of the DECODA (Bazillon et al., 2012) and ORFEO (Nasr et al., 2014) projects showed that current parsing techniques can be successfully used to parse disfluent speech transcriptions.

Syntactic parsing of non canonical textual input in the context of human-human conversations has been mainly studied 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 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 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 a syntactic parser.

We present in this paper a detailed analysis of the impact of all these phenomena on syntactic parsing. Other types of social media data have been studied in the literature. In particular tweets have received lately more attention. (Ritter et al., 2011) for example provide a detailed evaluation of a pos tagger on tweets, with the final objective of performing Named Entity detection. They showed that the performances of a classical tagger trained on generic news data drop when applied to tweets and that adaptation with in-domain data helps increasing these performances. More recently (Kong et al., 2014) described a dependency parser for tweets. However, to the best of our knowledge, no such study has been published on social media data from formal on line web conversations.

4 A study on orthographic errors in agent/customer chat dialogs

Chat conversations are unique from several perspectives. In (Damnati et al., 2016), we conducted a study comparing contact center chat conversations and phone conversations, both in the domain of technical assistance for Orange customers. The comparative analysis showed significant differences in terms of interaction flow. If chat conversations were on average twice as long in terms of effective duration, phone conversations contain on average four times more turns than chat conversations. This can be explained by several factors: chat is not an exclusive activity and latencies are more easily accepted than in an oral conversation. Chat utterances are formulated in a more direct style. Additionally, the fact that an utterance is visible on the screen and remains visible, reduces misunderstanding and the need for reformulation turns in an interaction. Regarding the language itself, both media induce specific noise that make it difficult for automatic Natural Language Understanding systems to process them. Phone conversations are prone to spontaneous speech effects such as disfluencies, and the need to perform Automatic Speech Recognition generates additional noise. When processing online chat conversations, these issues disappear. However the written utterances themselves can contain errors, be it orthographic and grammatical errors or typographic deviations due to high speed typing, poor orthographic skills and inattention.

In this study we focus on a corpus of 91 chat conversations that have been fully annotated with correct orthographic form, lemma and pos tags. The annotator was advised to correct misspelled words but she/he was not allowed to modify the content of a message (adding a missing word or suppressing an irrelevant word). In order to compare the original chat conversations with

	Customer	Agent	Full
#words	11798	23073	34871
SER	10.5%	1.5%	4.5%
MER	41.3%	15.7%	27.2%

Table 1: Language deviation error rates

the corrected ones, punctuation, apostrophe and case have been normalized. The manually corrected messages have then been aligned with the original messages thanks to an automatic alignment tools using the classical Levenshtein distance, with all types of errors having the same weight. A post-processing step was added after applying the alignment tool, in order to detect agglutinations or splits. An *agglutination* is detected when a deletion follows a substitution (`[en->entrain] [train->]`) becomes (`[en train->entrain]`). Conversely, a *split* is detected when an insertion follows a substitution (`[télécommande ->télé] [->commande]`) becomes (`[télécommande ->télé commande]`). Instead of being counted as two errors, agglutinations and splits are counted as one substitution. The evaluation is given in terms of Substitution Error Rate (SER) which is the amount of substitutions related to the total amount of words, and the Message Error Rate (MER) which is the amount of messages which contain at least one Substitution related to the total number of messages. As we are interested in the impact of language deviations on syntactic parsing of the messages, the latter rate should also be looked at carefully.

As can be seen in table 1, the overall proportion of misspelled words is not very high (4.5%). However, 27.2% of the turns contain at least one misspelled word. The number of words written by agents is almost twice as large as the number of words produced by Customers. In fact Agents have access to predefined utterances that they can use in various situations. They are also encouraged to formulate polite sentences that tend to increase the length of their messages, while Customers usually adopt a more direct and concise style. Consequently, Agents account for more in the overall SER and MER evaluation, artificially lowering these rates. In fact, as would be expected, Agents make much less mistakes and the distribution of their errors among conversations is quite balanced with a low standard deviation. The situation is different for Customers where both SER and MER have a high standard deviation (respectively 8.7% and 21.5%). The proportion of misspelled words depends on each Customer's linguistic skills and/or attention when typing.

In order to further study the impact of errors on Syntactic Analysis modules, we propose, as a preliminary study, to evaluate into more details the various types of substitutions encountered in the corpus. We make a distinction between the following types of deviations:

- `DIACR` *diacritic* errors are common in French as accents can be omitted, added or even substituted (`à ->a`, `très ->trés`, `énergie ->énérgie`).
- `APOST` for missing or misplaced *apostrophe*.
- `AGGLU` for *agglutinations* of two words into one.
- `SPLIT` for a word split into two words.
- `INFL` for *inflection* errors. Morpho-syntactic inflection in French is error prone as it is common that different inflected forms of a same word are homophones (`question ->questions`). Among these errors, it is very common (Véronis and Guimier de Neef, 2006) to find past participles replaced by infinitives for verbs that end with `er` (`j'ai changé -> j'ai changer`).
- `SWITCH` two letters are switched.
- `SUB1C` one character substituted.
- `DEL1C` one character missing.
- `INS1C` one character inserted.
- `OTHER` for all the other errors.

These types of errors are automatically evaluated in this order and are exclusive (*e.g.* `DEL1C` corresponds to words which have one missing character and are not of any preceding type).

Table 2 presents the proportion of each type of error observed in the corpus. As can be seen, diacritic deviations are predominant. On the overall, the second source of deviations is the use of erroneous inflection for a same word. It represents a higher proportion for Agents than for Customers.

Erroneous use of apostrophes is frequent for Customers but almost never occurs for Agents. Agglutinations are more frequent than splits, and constitue more than 11% of deviations for Agents.

	Customer	Agent	Full
DIACR	44.3%	34.5%	42.2%
APOST	12.0%	0.9%	9.6%
AGGLU	6.4%	11.2%	7.4%
SPLIT	1.7%	3.2%	2.0%
INFL	11.5%	25.0%	14.4%
SWITCH	0.7%	3.2%	1.3%
SUB1C	5.8%	4.3%	5.5%
DEL1C	7.4%	5.4%	6.9%
INS1C	3.4%	5.7%	3.9%
OTHER	6.8%	6.6%	6.8%

Table 2: Proportion (in %) of the different types of language deviations

Table 3 presents the repartition of language deviations by pos category. Observing this distribution can give hints on the problems that can be encountered for pos tagging and syntactic parsing. As one can see, function words are generally less error prone than content words. Apart from present participles that are always well written, only proper names and imperative verbs have an SER below the overall SER of 4.5%. But these categories are not highly represented in our data. All other content word categories have an SER above the overall SER. The most error prone category is past participle verbs, which are, as already mentioned, often confused with the infinitive form and which are also prone to inflection errors.

pos	prop.	SER
VER:ppre pres. participle	0.3%	0.0%
DET determiner	13.2%	1.3%
NAM proper name	1.7%	1.5%
INT interjection	2.1%	1.5%
PRO:REL relative pronoun	0.8%	1.6%
KON conjunction	4.6%	1.8%
NUM numeral	2.0%	2.4%
VER:imp verb imperative	0.9%	3.1%
PRP preposition	11.9%	3.5%
VER:inf verb infinitive	5.1%	4.6%
PRO pronoun	13.7%	5.2%
ADV adverb	6.9%	5.6%
VER verb	10.9%	5.8%
ADJ adjective	3.9%	6.7%
NOM name	19.6%	6.7%
ABR abbreviation	0.2%	10.0%
VER:pper past participle	2.2%	16.9%

Table 3: Language deviation by pos: proportion of each pos in the corpus and corresponding Substitution Error Rate

5 Evaluation and Results

5.1 Corpus description

In order to evaluate the impact of errors on pos tagging and parsing, the corpus has been split into two sub-corpora (DEV and TEST]) of similar sizes.

Conversations have been extracted from logs in a chronological way, meaning that they are representative of real conditions, with a variety of call motives and situations. Hence splitting the corpus into two parts by following the chronological order reduces the risk of over-fitting between the DEV corpus and the TEST corpus.

Table 4 illustrates the lexical composition of the DEV corpus, with a comparison between the original forms and the corresponding manually corrected version. All conversations have been anonymized and personal information has been replaced by a specific label (one label for Customer names, one for Agent names, one for phone numbers and another one for addresses). Hence, the entities concerned by this anonymization step do not account for lexical variety. It is interesting to notice that the number of different words on the Full corpus drops from 2381 when computed on the raw corpus to 2173 (15.3% relative) when computed on the corrected corpus. The proportion of words occurring just once is also reduced when computed over the manually corrected tokens. The statistics of the TEST corpus are comparable. However, the lexical intersection of both corpora is not very high as 10.3% of word occurrences in the TEST corpus are not observed in the DEV corpus (9.1% for Agents and 19.8% for Customers). When computing these rates over the manually corrected tokens, the overall percentage goes down to 9.0% (8.6% for Agents and 17.3% for Customers). These last figures remain high and show that the lexical diversity, if enhanced by scripting errors is already inherent to the data and the domain, with a variety of situations encountered by Customers. Adapting our pos tagger on the DEV corpus is a reasonable experimental approach as the preceding observations exclude the

	DEV original			DEV corrected		
	Customer	Agent	Full	Customer	Agent	Full
#words	5439	11328	16767	5425	11325	17338
diff. words	1431	1468	2381	1301	1414	2173
1 occ. words	879 (61.4%)	652 (44.4%)	1205 (50.6%)	764 (58.7%)	599 (42.4%)	1020 (46.9%)

Table 4: Description of the DEV corpus in terms of number of words, different words and words occurring only once. Figures vary because of splits and agglutinations.

risk of over-fitting bias at the lexical level.

5.2 Tagging

The pos tagger used for our experiments is a standard Conditional Random Fields (CRF) (Lafferty et al., 2001) tagger which obtains state-of-the-art results on traditional benchmarks. We use a coarse tagset made of 18 different parts of speech.

Three different taggers based on the same architecture are evaluated, the first one, T_F, is trained on the French Treebank (Abeillé et al., 2003), which is composed of newspaper articles. The second one, T_D, is trained on our DEV corpus and the third one, T_{FD} on the union of the French Treebank and our DEV corpus.

Taggers are usually evaluated with an accuracy metric, which is based on the comparison, for every token, of its tag in the output of the tagger (the hypothesis) and its tag in the human annotated corpus (the reference). In our case, the number of tokens in the reference and the hypothesis is not the same, due to agglutinations and splits. In order to account for these phenomena in the evaluation metric, we define conventions that are depicted in Table 5: in case of an agglutination, the tag of the agglutinated token t in the hypothesis is compared to the tag of the first token in the reference (see left part of table 5, where the two tags compared are in bold face). In case of a split, the tag of the first token in the hypothesis is compared to the tag of the token in the reference (see right part of the table).

agglutination				split			
REF		HYP		REF		HYP	
tok	tag	tok	tag	tok	tag	tok	tag
A	$\mathbf{T_A}$	AB	$\mathbf{T_{AB}}$	AB	$\mathbf{T_{AB}}$	A	$\mathbf{T_A}$
B	T_B					B	T_B

Table 5: Conventions defined when computing the accuracy of the tagger for a token. Tags in bold face are compared

	tok.	T_F	T_{FD}	T_D
Cust.	Corr.	91.13	93.26	94.36
	Orig.	86.59	88.83	90.38
Agent	Corr.	91.01	96.60	97.30
	Orig.	90.23	95.51	96.50

Table 6: Pos accuracy of the three taggers computed on the original (Orig.) and the corrected (Corr.) versions of the TEST corpus, for Customers and Agents parts of the corpus.

The taggers have been evaluated on the TEST corpus. The results are displayed in Table 6 which shows several interesting phenomena.

First, the three taggers obtain significantly different results. T_F, which is trained on the French Treebank, obtains the lowest results: 86.59% accuracy on the customer part of the corpus and 90.23% on the agent part. Adding to the French Treebank the DEV corpus has a benefic impact on the results, accuracy reaches respectively 88.83% and 95.51%. The best results are obtained by T_D with 90.38% and 96.50% accuracy, despites the small size of the DEV corpus, on which it is trained.

Second, as could be expected, the results are systematically higher on the corrected versions of the corpora. The results are around 4.5 points higher on the customer side and around 1 point higher on the agent side. These figures constitute the upper bound of the tagging accuracy that can be expected if the corpus is automatically corrected prior to tagging.

Third, the results are higher on the agent side, this was also expected from the analysis of the errors in both parts of the corpus (see Table 1).

Tables 7 and 8 give a finer view of the influence of errors on the pos tagging accuracy for tagger T_D. Each line of the table corresponds to the status of a token. If the token is correct, the status is CORR, otherwise it corresponds to one label of the

status	occ.	corr.	acc.	contrib.
CORR	5916	5547	93.76	59.23
DIACR	201	120	59.70	13.00
AGGLU	76	23	30.26	8.51
SUB1C	46	13	28.26	5.30
INFL	67	45	67.16	3.53
DEL1C	43	22	51.16	3.37
OTHER	40	23	57.50	2.73
INS1C	20	12	60.00	1.28
APOST	47	40	85.11	1.12
SPLIT	6	3	50.00	0.48
SWITCH	2	2	100.00	0.00

Table 7: Influence of token errors on pos tagging, computed on the customer side of the TEST corpus.

status	occ.	corr.	acc.	contrib.
CORR	12883	12517	97.16	79.91
DIACR	61	36	59.02	5.46
INFL	46	25	54.35	4.59
AGGLU	32	18	56.25	3.06
OTHER	11	3	27.27	1.75
SPLIT	8	4	50.00	0.87
DEL1C	10	6	60.00	0.87
SUB1C	8	4	50.00	0.87
INS1C	9	8	88.89	0.22
SWITCH	4	4	100.00	0.00

Table 8: Influence of token errors on pos tagging, computed on the agent side of the TEST corpus.

error types of Table 2. The second column corresponds to the number of occurrences of tokens that fall under this category. The third column is the number of tokens of this status that were correctly tagged, column four is the accuracy for this status and column five, the contribution to the error rate.

Table 7 shows that misspelled tokens are responsible for roughly 40% of the tagging errors. Among errors, the DIACR type has the highest influence on the pos accuracy, it corresponds to 13% of the errors, followed by agglutination. Table 8 shows that erroneous tokens account for 20% of the errors on the agent side. And the first cause of token deviation that provokes tagging errors is DIACR.

5.3 Parsing

The parser used in our experiment is a transition based parser (Yamada and Matsumoto, 2003; Nivre, 2003). It is a dependency parser that takes as input tokens with their pos tag and selects for every token a syntactic governor (which is another token of the sentence) and a syntactic label. The prediction is based on several features that combine lexical information and pos tags. Orthographic errors have therefore a double impact on the parsing process: through the errors they provoke on the pos tagging process and the errors they provoke directly on the parsing process. The parser was trained on the French Treebank. Contrary to taggers, a single parser was used for our experiments since we do not have hand corrected syntactic annotation of the DATCHA corpus.

In order to evaluate the parser, we have parsed our DEV corpus with corrected tokens and gold pos tags and considered the syntactic structures produced to be our reference. The results that are given below should therefore be taken with caution. Their absolute value is not reliable (it is probably over estimated) but they can be compared with one another.

The metric used to evaluate the output of the parser is the Labeled Attachement Score (LAS) which is the ratio of tokens for which the correct governor along with the correct syntactic label have been predicted. The conventions of Table 5 defined for the tagger were also used for evaluating the parser.

Three series of parsing experiments were conducted, the first one takes as input the tokens as they appear in the raw corpus and the pos tags predicted with our best tagger (T_D). These experiments correspond to the most realistic situation, with original tokens and predicted pos tags. The second series of experiments takes as input the corrected tokens and the predicted pos tags. Its purpose is to estimate an upper bound of the parsing accuracy when using an orthographic corrector prior to tagging and parsing. The third experiment takes as input raw tokens and gold pos tags. It corresponds to an artificial situation, its purpose is to evaluate the influence of orthographic errors on parsing, independently of tagging errors.

Table 9 shows that the influence of orthographic errors on parsing is limited, most parsing errors are due to pos tagging errors.

The table also shows that the difference in parsing accuracy between the customer part of the corpus and the agent part is higher than what it was for tagging. This can be explained by the fact that,

	tok.	pos acc.	LAS
Cust.	O	90.38	73.47
	C	94.36	81.30
	O	100	94.68
Agent	O	96.50	82.12
	C	97.30	86.43
	O	100	95.74

Table 9: LAS of the parser output for three types of input: original tokens (O) and predicted pos tags, corrected tokens (C) and predicted pos tags and original tokens and gold pos tags, computed on the TEST corpus for the customer and the agent parts of the corpus.

from the syntactic point of view, agent utterances are probably closer to the data on which the parser has been trained (journalistic data) than customer utterances.

6 Conclusion

We study in this paper orthographic mistakes that occur in data collected in contact centers. A typology of mistakes is proposed and their influence on part of speech tagging and syntactic parsing is studied. We also show that taggers and parsers trained on standard journalistic corpora yield poor results on such data and that the addition of a limited amount of annotated data can significantly improve the performances of such tools.

Acknowledgments

This work has been funded by the French Agence Nationale pour la Recherche, through the project DATCHA (ANR-15-CE23-0003)

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In *Treebanks*, pages 165–187. Springer.

Falaise Achille. 2005. Constitution d'un corpus de français tchaté. In *Rencontre des tudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, Dourdan, France.

Nicholas Asher, 2011. *Strategic Conversation*. Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, France. `https://www.irit.fr/STAC/`.

Thierry Bazillon, Melanie Deplano, Frederic Bechet, Alexis Nasr, and Benoit Favre. 2012. Syntactic annotation of spontaneous speech: application to call-center conversation data. In *Proceedings of LREC*, Istambul.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *EMNLP*, pages 357–368.

Thierry Chanier, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The comere corpus for french: structuring and annotating heterogeneous cmc genres. *JLCL-Journal for Language Technology and Computational Linguistics*, 29(2):1–30.

Géraldine Damnati, Aleksandra Guerraz, and Delphine Charlet. 2016. Web chat conversations from contact centers: a descriptive study. In *International Conference on Language Resources and Evaluation (LREC)*.

Michael H Dickey, Gary Burnett, Katherine M Chudoba, and Michelle M Kazmer. 2007. Do you read me? perspective making and perspective taking in chat communities. *Journal of the Association for Information Systems*, 8(1):47.

Eric N Forsyth and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 19–26. IEEE.

Ewa Jonsson. 1997. Electronic discourse: On speech and writing on the internet.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *Proceedings of EMNLP*.

T. Kucukyilmaz, Cambazoglu B. B., C. Aykanat, and F. Can. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management*, 44:1448–1466.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Alexis Nasr, Frederic Bechet, Benoit Favre, Thierry Bazillon, Jose Deulofeu, and Andre Valli. 2014. Automatically enriching spoken corpora with syntactic information for linguistic studies. In *International Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*. Citeseer.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Sarah M Taylor, and Nick Webb. 2010. Mpc: A multi-party chat corpus for modeling social phenomena in discourse. In *LREC*.

Paola Trimarco. 2014. *Digital Textuality*. Palgrave Macmillan.

Jean Véronis and Emilie Guimier de Neef. 2006. Le traitement des nouvelles formes de communication écrite. *Compréhension automatique des langues et interaction*, pages 227–248.

Min Wu, Arin Bhowmick, and Joseph Goldberg. 2012. Adding structured data in unstructured web chat conversation. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 75–82. ACM.

Hiroyasu Yamada and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of IWPT*, volume 3, pages 195–206.