HAL
open science

Ubuntu-fr: a Large and Open Corpus for Supporting Multi-Modality and Online Written Conversation Studies

Nicolas Hernandez, Soufian Salim, Elizaveta Loginova Clouet

To cite this version:

Nicolas Hernandez, Soufian Salim, Elizaveta Loginova Clouet. Ubuntu-fr: a Large and Open Corpus for Supporting Multi-Modality and Online Written Conversation Studies. The Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 2016, Portorož, Slovenia. pp.1777-1783. hal-01503811

HAL Id: hal-01503811

https://hal.science/hal-01503811v1

Submitted on 7 Apr 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.

Ubuntu-fr: a Large and Open Corpus for Supporting Multi-Modality and Online Written Conversation Studies

Nicolas Hernandez, Soufian Salim, Elizaveta Loginova Clouet
LINA UMR 6241 - Université de Nantes
2 rue de la houssinière, 44322 Nantes Cedex 03
{firstname}.{lastname}@univ-nantes.fr

Abstract

We present a large, free, French corpus of online written conversations extracted from the Ubuntu platform's forums, mailing lists and IRC channels. The corpus is meant to support multi-modality and diachronic studies of online written conversations. We choose to build the corpus around a robust metadata model based upon strong principles, such as the "stand off" annotation principle. We detail the model, we explain how the data was collected and processed - in terms of meta-data, text and conversation - and we detail the corpus' contents through a series of meaningful statistics. A portion of the corpus - about 4,700 sentences from emails, forum posts and chat messages sent in November 2014 - is annotated in terms of dialogue acts and sentiment. We discuss how we adapted our dialogue act taxonomy from the DIT++ annotation scheme and how the data was annotated, before presenting our results as well as a brief qualitative analysis of the annotated data.

Keywords: corpus, conversation, dialogue, chat, forum, email

1. Introduction

Help desk and customer service are the most common areas of Customer Relationship Management (CRM) technology investment[1]. In order to offer new capabilities to CRM software, the ODISAE[2] project aims at developing a semantic analyser of written online conversations across several modalities (*i.e.* chat, forum, email). These capabilities are: multi-modal text information retrieval (*e.g.* finding the solution to a problem in a modality different from the one in which the request was formulated), automated FAQ and documentation management (*e.g.* automatic detection of the absence of a suitable solution to a recurring request), automated assistance generation (*e.g.* helping users to formulate problems, evaluating answers' exhaustivity), or conversation supervision (*e.g.* detecting attrition, irritation).

The project presents an opportunity to make the academic and the industrial worlds collaborate on real-world use-cases. Unfortunately, it also comes with heavy restrictions when it comes to disseminating data and results of scientific analysis. We consider that making our research reproducible and reusable in an open science perspective is an important issue (Nielsen, 2011).

In this paper, we present our efforts to build a large, free, French corpus of online written conversations from synchronous and asynchronous mediums and assorted explanatory texts. All items were gathered on the same period and pertain to the same type of discursive situations, namely user assistance in problem solving tasks. Our object of study is a specific domain of computer-mediated communication (CMC), computer-mediated conversations, and our scientific objectives are multiple:

1. Reaching a better understanding of the structure of problem-solving conversations;
2. Exploiting modality comparability in order to improve linguistic processing of less conventional user generated content (*e.g* by working on analysis engines' portability we may be able to use them to process data gathered on a medium they were not trained on);
3. Learning how to align or even "translate" content between different modalities (*i.e.* inter-conversations and between conversations and explanatory texts);
4. Allowing for the application of NLP techniques on the data (*e.g.* building statistical models dedicated to the identification of discursive structures in dialogue).

In order to support such applications, we consider the analysis of conversations in terms of dialogue acts as well as a few identified conceptual objects, such as submitted problems and their possible solutions. We are working on a taxonomical model intended to support the generic handling of online written conversations while allowing annotators to take into account the specific characteristics of each medium. This model is based on the taxonomy of dialogue acts and relations defined in the DIT++ scheme for the annotation of oral conversations (Bunt, 2009). In this paper we report our advancement, in particular: the data we gathered, our collection methodology, and the annotation scheme we used.

The NLP community has little available social media for French (Seddah et al., 2012; Falaise, 2014; Yun and Chanier, 2014). The CoMeRe[3] project attempts to fill that gap by releasing all efforts in that domain through the Ortolang[4] network, the French node of the CLARIN infrastructure (Chanier et al., 2014). Only the Simuligne corpus from the LETEC project covers several modalities (chat, forum, email, pedagogical support) for the French language (Reffay et al., 2008). But their corpus differs from ours in terms of object of study (collaborative distance learning

[1] Software Advice survey 2014 "CRM software users" www.softwareadvice.com/crm/userview/report-2014

[2] ODISAE is supported by the Unique Inter-ministerial Fund (FUI) no. 17 (wwww.odisae.com)

[3] corpuscomere.wordpress.com/apropos

[4] www.ortolang.fr/api/content/comere/latest/comere.html

and intercultural exchange), discursive scope (interactions between students learning French) and data volume (11,506 messages, 600,348 tokens and 67 participants). Falaise (2014) offers a general topic chat corpus made of five million turns. For the purpose of studying the automated parsing of raw user generated content, Seddah (2012) released a treebank containing 1,700 sentences from microblogging and web forums. Our initiative follows Uthus and Lowe whom proposed to exploit the Ubuntu open source community to build a large, public domain and technical chat corpus in English (Uthus and Aha, 2013; Lowe et al., 2015).

Firstly we present a general scheme for the unified modelling of written online conversation modalities, and we briefly discuss the data and metadata formats. Then we explain how the data was collected and prepared for NLP applications. We compare basic statistics for each modality. Finally, we report preliminary results based on a sizeable manual annotation effort for comparing various modalities in terms of dialogue acts and Opinion, Sentiment and Emotion (OSE).

2. Corpus construction

Discussions for encompassing multi-modal written interactions in a single model are still at a very early stage. A Text Encoding Initiative (TEI) special interest group[5] (SIG) considers CLARIN and CoMeRe proposals for adapting the TEI to represent genres of CMC. This denomination includes tweets, wiki discussions, blogs, SMS, audio channels... In comparison, our work focuses on conversational objects, more precisely on forum, mail and chat conversations. While we agree with the data models the SIG recommendations imply (in particular in terms of metadata), we disagree with an approach that fails to make a clear distinction between the model and its implementation. Indeed, the recommendation specifies how the content should be formatted and structured, ignoring the importance of the "stand off" annotation principle (Thompson and McKelvie, 1997). Bański (2010) discusses why the TEI's "stand-off" annotation is inadequate.

While this initiative is not part of the CoMeRe project (Chanier et al., 2014), we agree that it is necessary to think about the exploitation opportunities of a corpus ahead of its construction. Because of our NLP expertise we do not use the Text Encoding Initiative (TEI) or its extension to CMC, however we agree with the data models behind them.

2.1. Modelisation of computer-mediated written online conversations

We define a generic model to describe conversational metadata. The model results from a generalization of observed conversational modalities' structures. It takes into account recent evolutions of Internet message formats[6] as well as the current TEI-CMC meta data recommendations. It integrates common and specific attributes to include thematic categorisation of conversations, message view counts, participant role descriptions (*e.g.* ambassador, expert, customer), *etc*. The Conversation structures are available through some Message features: daytime and inReplyToMessageIds. Figure 1 shows the relations between the main conversational objects.

Forums topically structure Conversations which are made of Messages. A Forum is a recursive structure. It corresponds to the concept of Room for Chat. Forum, Conversations, Messages and Participants are uniquely identified. A Forum, a Conversation, and a Message can have a subject (*e.g.* "*SD card issue*") and can be categorized by Topics (*e.g.* "*Hardware*"). A Context informs about the medium type, private status, solved status, "likes" count, views, importance, and pinned state of a Conversation or a Message. A Message interconnects Participants with FROM, TO, CC, or BCC relations. It is made of Utterances, all of which are associated to an OSE tag and a dialogue act. The message's Body specifies the MIME type and character encoding of its contents. A Participant is characterized by a username, email address, role (*e.g.* customer), *etc*.

This model is implemented in W3C XML schemas and a UIMA type system[7]. The former is used to store and exchange raw data after extraction from their source format. The latter is used as soon as the data is taken as input in our NLP pipelines. They are based on Apache UIMA (Ferrucci and Lally, 2004) since it provides the level of abstraction necessary for the manipulation of annotated data. Our annotated data is serialized as XML Metadata Interchange (XMI) files, a standard for the exchange of UML metadata information. These choices take us away from the TEI but not necessarily from its conceptual model.

The manner in which annotation results are serialized, whether they are the product of manual or automatic annotation, is a secondary matter. The primary functions of this format are storage and exchange. The main imperative is to allow any user to edit and apply annotations on data in its original form whithout undermining it. This implies the adoption of certain principles (*e.g.* the "stand off" annotation) and the use of certain tools, such as Apache UIMA (Ferrucci and Lally, 2004)

2.2. Collection methodology

Our data was gathered from public resources hosted by the French Ubuntu community[8]. Besides a web-based documentation[9], the community offers several communication tools: forums[10], mailing lists[11] and IRC channels[12]. Their contents are publicly accessible online.

The documentation is supplied under the license CC BY-SA v3.0. For other resources, users delegate the usage of their messages to the editor, Ubuntu-fr; therefore if a user expresses his or her refusal to participate in this corpus, we will have to remove his or her messages.

The documentation is not static and must be version controlled. It can only be retrieved by scraping its web pages.

[5] wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

[6] tools.ietf.org/html/rfc6854

[7] docs.oasis-open.org/uima/v1.0/os/uima-spec-os.html#_Toc205201040

[8] ubuntu-fr.org

[9] doc.ubuntu-fr.org

[10] forum.ubuntu-fr.org

[11] lists.ubuntu.com/mailman/listinfo/ubuntu-fr

[12] irc.freenode.net/ubuntu-fr

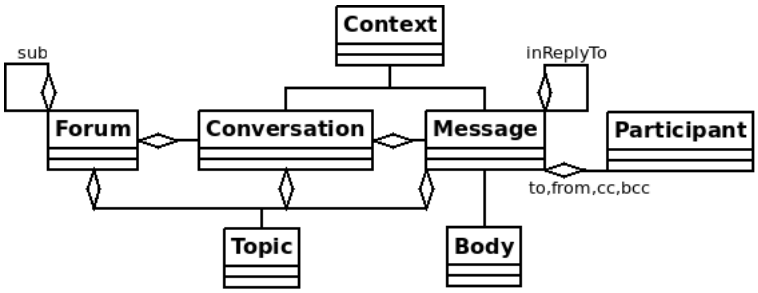

Figure 1: Metadata model for written online conversations.

We are working with the community to systematize its archiving and versioning. In the meantime, we have been scraping and storing it daily since November 2014. Emails are incrementally archived and are publicly available. Forums also grow incrementally but no public archive is available, we had to use web scraping techniques to collect their data. Chat messages are not saved at all, however we have been logging them since November 2014.

We have all forum messages and emails since the platform creation in 2004, and all data for documentation and IRC chats since October 2014. These subcorpora are perpetually growing, are representative of recent style of online writing and bear witness to the evolution of online communication style over a period of ten years. Moreover, this data can be easily extended with similar English language data already available (Lowe et al., 2015).

2.3. Data pre-processing

For the purpose of discourse annotation tasks and further specific NLP development, we prepared our data by performing the following processing:

1. Meta-data extraction. We developed *ad hoc* parsers to instantiate the model depicted in section 2.1. for each medium. For chats, we developed a specific analysis engine in order to identify the addressee of each message.
2. Text extraction. We handled character encoding and MIME type issues for each message. Email contents were also parsed to recognize quoted parts.
3. Conversation recovering and disentangling. The forum structure gives a direct access to the thread. The mail conversation structures were recovered by using the *inReplyTo* field when present and subject similarity (in terms of character edit distance) for mails occurring in a same period. However, for chats, multiple conversations can occur simultaneously in the same channel. In (Riou et al., 2015), we study the portability for French of a state-of-the-art disentanglement method (Elsner and Charniak, 2010) as well as the contribution of discourse information for this purpose.
4. Text segmentation in word tokens and sentences. We developed a custom rule-based approach exploiting punctuation, typography and text layout information. The segmenters were tuned to process HTML and CMC phenomena.

On the matter of data anonymization, we consider several alternatives: we can avoid it entirely and delete messages on demand, we can restrict it to metadata, or we can refrain from releasing any data at all and merely provide our collection and processing tools.

2.4. Corpus statistics

Table 1 details the contents a sample of one year of data. While the table details recent samples only for better comparison, in the case of emails the size of such a sample would have been much bigger in the early years of the platform. For example, the 2005 email sample contains 7,034 messages. This reflects the deep evolutions of user behaviour observed in CMC over the past ten years. Such data can be used to perform diachronic studies. However this level of analysis is out of the scope of this paper and will not be detailed here.

The statistics provided in Table 1 show significant differences between modalities. A thread is twice as likely to be answered in forums than in emails or chats, and they stay active longer too. Email messages are longer, and are much more likely to reference external resources and especially the Ubuntu documentation. Forums are mostly referenced in other forums, and almost nobody is ever redirected to the mailing list these days.

3. Annotating a subpart of the corpus with dialogue acts and OSE information

For the purpose of studying the interactions over various modalities, we annotated a part of the corpus in terms of dialogue acts and OSE information. From November 2014, 29 forum threads, 45 mail threads, and 6 days of IRC activity were annotated. To support future supervised machine learning approaches, we arbitrary decided to start by annotating at least 1,200 utterances for each modality.

3.1. Annotation taxonomy and procedure

In many works related to conversational phenomena, dialogue annotation and dialogue systems, conversation interactions between the participants are modeled in terms of dialogue acts. Dialogue act theory is an extension of speech act theory (Austin, 1975) which attempts to describe utterances in terms of communicative function (*e.g.* question, answer, thanks). While until the 1990s speech act theory was mostly concerned with the analysis of isolated utterances, it later incorporated notions of context and common ground (*i.e.*, information that needs to be synchronized between participants for the conversation to move forward)

	emails	forums	chats	doc.
pages (total)	-	-	-	4 600
conversations (total)	100	23K	4.4K‡	-
unanswered threads (percentage)	23%	12%	21%‡	-
messages (total)	448	189K	114K	-
messages per conversation (median)	3	5	4‡	-
words (total)	40K†	25M	1M	4M
words per message (median)	59†	46	7	-
participants (total)	75	12K	2.4K	-
references to documentation (per conv.)	2.71†	0.78	0.02	-
references to forums (per conv.)	0.44†	2.13	0.11	-
references to mailing list (per conv.)	0.17†	0.00	0.00	-
references to external resources (per conv.)	5.08†	3.82	2.58	-
conversation duration (median)	4h33	5h51	-	-

Table 1: One year of data: January to December 2014 for forums and emails, October 2014 to October 2015 for chats. Numbers marked by † do not take quoted text into account. Numbers marked by ‡ are estimates based on chat disentanglement results on a sample of 1,229 messages (Riou et al., 2015).

(Traum and Hinkelman, 1992). A number of conversation annotation schemes sprouted from this theoretical foundation, such as DAMSL (Core and Allen, 1997) and DIT++ (Bunt, 2009).

DAMSL is a *de facto* standard in dialogue analysis, due to its theoretical foundation (acts are annotated as context update operations), its genericity (high level classes allow for the annotation of a wide range of conversations types) and multidimensionality (each utterance can be annotated with several labels). However, its dimensions are not discussed and lack conceptual significance. DIT++ builds upon DAMSL and attempts to improve these aspects, which is why we chose it as a basis for our taxonomy.

DIT++ proposes to annotate dialogue acts in terms of semantic content and communicative functions. Each utterance can contain several dialogue acts. The taxonomy presents 10 dimensions defined as independent classes of conversational behaviours (*e.g.* TIME MANAGEMENT, SOCIAL OBLIGATIONS MANAGEMENT). Each dialogue act falls within one of these dimensions and is further annotated with its communicative function. Communicative functions capture the intended effect on addressees (*e.g.* THANKING, REQUEST FOR INFORMATION). DIT++ provides the annotator with over 50 communicative functions, some of which are general-purpose, meaning they can be applied with any dimension, while the others are dimension-specific. Moreover, DIT++ provides a number of qualifiers used to further qualify utterances in terms of sentiment, partiality, conditionality, and certainty.

While we accept without reservation the conceptual framework of DIT++, we had to adapt its taxonomy for our corpus. Since the DIT++ scheme was developed around the study of oral dialogue corpora, it is reasonable to expect that some aspects of it may not be relevant or even applicable to online conversations, and that it could fail to capture some important information that is not typically observed in oral dialogues. Moreover, we had to ensure that the scheme we use supports our stated scientific and applicative goals, which differ from those of Bunt.

In order to do this, for each dimension, communicative function and qualifier, we had to ask two questions: "is this useful?" and "is this present in the data?". While we looked for the answer to the first question by confronting the taxonomy to various use cases, the answer to the second one required a data-driven approach. And for each utterance, we also asked whether wa have an appropriate function in the scheme. Therefore, three annotators (with a computational linguistics background) performed several exploratory annotation sessions on a sample of conversations from November 2014.

These experiments resulted in some modifications in the annotation scheme. While adapting the scheme to online conversations, we had to add the EXTRA-DISCOURSE dimension in order to capture all non-discursive text that participants can include in their messages (such as quoted text or automatic formatting artifacts). We also added the PSYCHOLOGICAL STATE dimension for utterances that aim at discussing the psychological states of the conversation's participants (*e.g.* "I'm getting angry"[13], ":D"). Inversely, some of DIT++'s dimensions may not be necessary for written data. Thus, the TIME MANAGEMENT class is irrelevant in the context of online communication (except some very rare utterances in chat). Other classes, such as CONTACT MANAGEMENT, COMMUNICATION MANAGEMENT and ATTENTION-PERCEPTION-INTERPRETATION have only a marginal resence in the data. The scheme was simplified to make it usable to support a task of supervised machine learning classification of DA. The dimensions ALLO FEEDBACK and AUTO FEEDBACK

[13]Since we do not expect that all of our readers understand French, examples presented in this paper are not taken from the corpus.

were unified in one dimension FEEDBACK, instead a qualifier "allo" or "auto" is used to determine the subject. We also did away with the separation between OWN COMMUNICATION MANAGEMENT and PARTNER COMMUNICATION MANAGEMENT since the distinction is captured by this newly introduced qualifier.

Communicative functions were simplified in many ways while preserving the same overall structure. For example, the taxonomy does not distinguish between various types of questions, commissives and directives; instead we use REQUEST FOR INFORMATION, REQUEST FOR ACTION and REQUEST FOR DIRECTIVES. Annotation results showed that dimension-specific functions related to rarely used dimensions (*e.g.* STALL, SHIFT TOPIC) were almost inexistent in the data.

Sentiment analysis would be useful for evaluating user satisfaction as well as customer support performance. We also felt it would be interesting to be able to capture satisfaction feedback independently of the effective answer evaluation (*e.g.* "thanks for the quick answer, however unfortunatly this solution doesn't work for me"). Thus, we decided to opt for a more detailed annotation of sentiment and we based our OSE annotations on the fine-grained semantic classes defined in the Ucomp[14] project (Fraisse and Paroubek, 2014). OPINION covers DEPRECIATE and PROMOTE tags, SENTIMENT covers SATISFIED and UNSATISFIED, and EMOTION covers seven tags including ANGER, HAPPINESS and others. Emotion categories were generalized from the original tagset to prevent some ambiguities. Sentiment-less utterances would be classified under the generic label INFORMATION. However, after examining the results of the annotation we found that only a small subset of all utterances, about 10% of them, bore any one of those sentiments. Nonetheless, we found that answer satisfaction can be evaluated by other means, most notably by functions in the FEEDBACK and SOCIAL OBLIGATIONS dimensions.

On other qualifiers, we chose to keep them all. Partiality can be useful to identify incomplete answers, even though their automatic detection would be a difficult task. Certainty is also useful as a participant's own confidence in his or her answers is useful to evaluate them. Conditionality is also useful as it may often indicate that a user is about to give up or that a customer engages in threatening behaviour ("If I don't get a refund, I won't shop here ever again"). Unfortunately, annotation results were vastly underwhelming. Clear marks of partiality, uncertainty and conditionality are near absent in the data: less than one utterance out of twenty was annotated with any of them.

Tables 3 and 4 detail the functions and the dimensions we used for the annotation task.

3.2. Inter-annotator agreement

The annotation process was performed iteratively and incrementally. At each iteration, new data was annotated mainly by a single annotator. The annotator was a postdoctoral researcher with a computational linguistics background but no specific annotation skills. New phenomena and uncertain cases were the subject of intense discussion between three researcher fellows working on the domain. Label definitions and guidelines were consequently strengthened and previously annotated data was revised.

When the guidelines were stable and our quantitative objective was reached, a third-party annotator with background in computational linguistics was involved. This annotator was provided with the guidelines, various annotated examples and a corpus sample, and was further trained through correction of test data and free discussion. This third-party annotator was then tasked with the annotation of a different corpus made of CRM conversations. This second corpus is not detailed here due to licence restriction. The same amount of data was annotated.

To measure the homogeneity of the annotated data, two annotators were requested to annotate three new conversations (about 80 sentences and 1,200 words). Both annotated 110 utterances. Since the annotators did not always segment the sentences into the same utterances, we decided to calculate the agreement coefficient at the token level. We obtain the following Kappa values: 0.69 and 0.70 for the dimension and the communicative function features, respectively. These values indicate a substantial agreement. Only the dimension feature and the communicative function feature were compared. Further work should evaluate the divergence for each dimension and function values.

3.3. Annotated corpus overview

	# conv.	# msg	# token	# DA
chat		2,320	17,448	1,989
forum	29	258	25,205	1,338
mail	45	200	19,798	1,382

Table 2: Characteristics of each modality.

Over 4,700 dialogue acts were annotated. Table 2 provides more details on the size of the annotated corpus for each modality in terms of conversations, messages, tokens and dialogue acts. The forum, email and chat parts of the corpus are of sufficiently similar size to be compared.

Table 3 highlights some variation in conversation participants' discursive behaviours. Social obligations management is much more present in emails. In chat, discourse management (*e.g.* "here's my question") is rare, unlike in forums and especially emails. However, the EVALUATION ("OK let me see"), ATTENTION-PERCEPTION-INTERPRETATION ("I understand") and PSYCHOLOGICAL STATE ("I'm feeling good") dimensions are much more prevalent there, which shows that grounding as well as informational and emotional synchronization between participants is more important in synchronous conversations than in asynchronous ones. Utterances that deal with time (*e.g.* stalling), contact ("is anyone here?") and communication ("sorry I meant Ubuntu") management are only found in chat, and absent from asynchronous modalities.

Table 4 confirms that in emails greeting, valediction, thanking, and such social acts seem to be expected and serve as typical protocol for framing a message. We also observe

[14] www.ucomp.eu

Dimension	chat	forum	mail
domainActivities	82,35	80,1	67
socialObligationManagement	9,25	12,85	30,35
discourseManagement	0,85	4,8	2
evaluation	3,95	1,65	0,15
psychologicalState	1,45	0,45	0,35
attentionPerceptionInterpretation	0,6	0,15	0,05
communicationManagement	1,35	0	0
contactManagement	0,9	0	0
timeManagement	0,2	0	0

Table 3: Percentage distribution of dimensions in dialogue acts.

Function	chat	forum	mail
inform	26,95	31,3	33,35
answer	17,65	20,05	11,2
requestForInformation	16,35	9,2	6,95
answerPositively	9,1	5,45	3,05
requestForAction	8,85	8,45	6,15
greetings	5,65	5,1	6,8
correct	4,75	3	1,3
answerNegatively	3,4	2,85	1,9
thanking	2,05	2,85	3,4
commit	1,95	1,05	1
requestForDirective	0,85	0,9	0,35
valediction	0,5	1,35	6,15
apologizing	0,45	0,65	0,6
anticipateThanking	0,4	1,95	2,95
finalSelfIntroduction	0	0,9	9,2
announce	0,35	0,8	1,25

Table 4: Percentage distribution communicative functions in dialogue acts.

that chat conversations are more to the point and action-driven: compared to participants in other modalities, chat participants are twice as likely to perform a commissive act and dedicate more utterances to requesting actions, instructions or information.

Figure 2 shows the distribution of OSE[15] tags. While the number of utterances bearing a clear OSE mark is relatively small - less than 10% of all utterances - we can already see a pattern. Chats appear to be more emotional, while forums are more opinionated (the opinion tagset contains the DEPRECIATE and PROMOTE tags).

Overall, it seems that forums are a balanced form of online communication, positioned between the more formal email style of discourse and more informal chats.

4. Conclusion

We presented a large, free, French corpus of online written conversations extracted from forums, emails and chats gathered on the Ubuntu platform. The corpus is meant to support multi-modality and diachronic studies of online written conversations. The corpus is built around a robust metadata model based upon strong principles, such as the "stand off" annotation principle. We showed how the data was collected and processed, and we detailed the corpus' contents. A subpart of the corpus was annotated in terms of dialogue acts and OSE. We had to modify the DIT++ annotation scheme to adapt it to the written texts. Through the analysis of distributions of communicative functions and dimensions in a corpus sample, we showed their relevance according to modalities.

5. Acknowledgements

This material is based upon work supported by the Unique Interministerial Fund (FUI) 17. It is part of the ODISAE[16] project. The authors would like to thank the anonymous reviewers for their valuable comments.

6. Bibliographical References

Austin, J. L. (1975). *How to Do Things With Words*. Oxford University Press, second edition.

Bański, Piotr. (2010). *Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless*.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for*

[15] Opinion, Sentiment, Emotion

[16] http://www.odisae.com

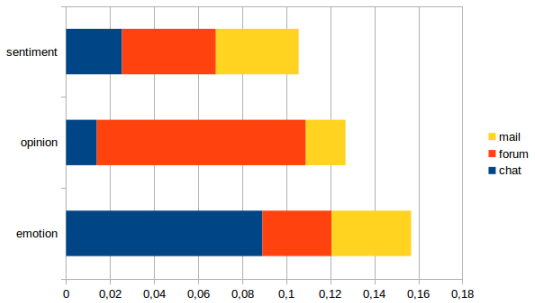

Figure 2: OSE distribution in each modality.

Embodied Dialogue Acts" (EDAML 2009), pages 13–24, Budapest, Hungary.

Core, M. and Allen, J. (1997). Coding Dialogs with the DAMSL Annotation Scheme. In *Workin g Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Boston, MA, USA.

Elsner, M. and Charniak, E. (2010). Disentangling chat. volume 36, pages 389–409. MIT Press.

Ferrucci, David and Lally, Adam. (2004). *UIMA: an architectural approach to unstructured information processing in the corporate research environment*. Cambridge Univ Press.

Fraisse, A. and Paroubek, P. (2014). Toward a unifying model for opinion, sentiment and emotion information extraction. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3881–3886, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1009.

Ryan Lowe and Nissan Pow and Iulian Serban and Joelle Pineau. (2015). *The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems*.

Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton, N.J. Princeton University Press.

Reffay, C., Chanier, T., Noras, M., and Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. volume 15.

Riou, M., Hernandez, N., and Salim, S. (2015). Using discursive information to disentangle french language chat. In *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media workshop at GSCL Conference 2015*, pages 23–27, Essen, Germany, September.

Thompson, H. S. and McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97: the next decade – pushing the envelope*, pages 227–229.

Traum, D. R. and Hinkelman, E. A. (1992). Conversation Acts in Task-Oriented Spoken Dialogue. volume 8, pages 575–599. Wiley Online Library.

7. Language Resource References

Chanier, Thierry and Poudat, Céline and Sagot, Benoit and Antoniadis, Georges and Wigham, Ciara R and Hriba, Linda and Longhi, Julien and Seddah, Djamé. (2014). *The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres*.

Falaise, A. (2014). *Corpus de français tchaté getalp_org*. Ortolang.fr.

Ryan Lowe and Nissan Pow and Iulian Serban and Joelle Pineau. (2015). *The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems*.

Seddah, Djamé and Sagot, Benoît and Candito, Marie and Mouilleron, Virginie and Combet, Vanessa. (2012). *The French Social Media Bank: a Treebank of Noisy User Generated Content*.

Uthus, David C. and Aha, David W. (2013). *The Ubuntu Chat Corpus for Multiparticipant Chat Analysis*. AAAI.

Yun, H. and Chanier, T. (2014). *Corpus d'apprentissage FAVI (Français académique virtuel international)*. Ortolang.fr.