Code: https://github.com/Norton50/COMP551Project1/blob/master/__main__.py
XML: https://raw.githubusercontent.com/Norton50/COMP551Project1/master/mxdvmc_fre.xml

Reddit French Dialogue Corpus

A French conversational dataset for machine learning

Matthew Cooke | 260553365 | matthew.cooke2@mail.mcgill.ca
David Venuto | 260562974 | david.venuto@mail.mcgill.ca
Marley Xiong | 260739063 | marley.xiong@mail.mcgill.ca

Introduction

Machine learning research has been working to improve conversational AI for many years. Dialogue corpuses have become increasingly prevalent, however are dominated by those in English. In this report we outline our method for generating a French language, unstructured, dialogue corpus. We used the Python programming language to scrub through the online discussion board reddit/r/France[1] to data-mine French conversations.

Dataset description

The dataset contains conversations from 1000 posts in the r/France subreddit, an online community of over 160,000 users.

Avg. # Turns	Total # Utterance	Total # Dialogue	Total # Threads	Total # Words
3	65,651	20,913	1000	2M

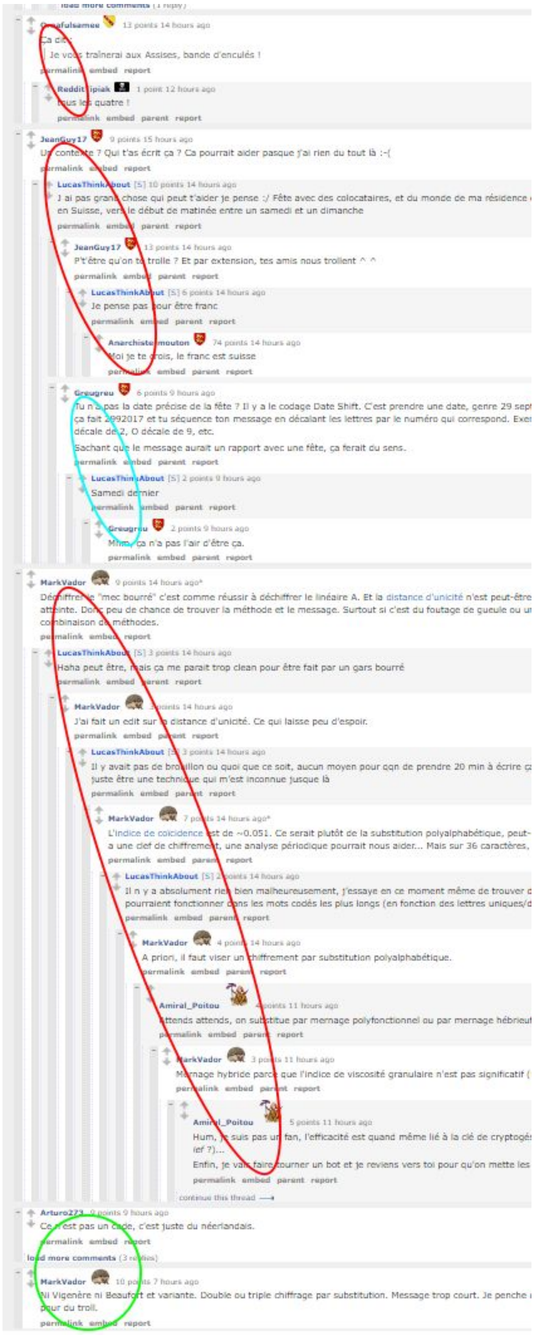
Figure 1. Overview of Reddit French dialogue corpus

The generated dataset had 20,913 conversations, 65,651 utterances and 2,175,274 words after removing all whitespace, numbers and stop characters. The average length of a conversation was therefore 104 words. The average number of utterances per conversation was 3.14, or approximately 3 turns per conversation. The average number of words per utterance was 33.13.

We used PRAW[2] to scrape comments from the top posts as determined by Reddit's native ranking algorithm. Since top posts tend to receive more comments and more conversational sequences, we opted to scrape the top 1000 threads for comments. Each sequence of comments, or chain where a comment is entered as a reply to another comment, was viewed conceptually as a separate conversation and represented as a series of utterances enclosed by <s></s>. The representation makes sense given that replies are produced one after another in time, much like a spoken conversation. Each comment, regardless of length, was represented as an utterance and enclosed with <utt></utt> tags, with the user ID numbered from 0 onwards. No special consideration was given to the relationship between different comment sequences in the same thread, as despite their topical similarities, logical continuation or sequential organization could not be assumed between lateral sets of comments.

Owing to the diverse media (pictures, videos) of the main posts on Reddit, we ignored the main posts and focused on comments instead. The format of Reddit's discussion board is such that multiple branches can form in a conversation. In order to maximize the wealth of data obtained without skewing the representation of comments, we treated each branch as a separate conversation starting at the base of the branch. In this way, top-level comments are featured only once in the corpus instead of in all conversations spawned from the comment, which would artificially inflate the frequency at which top level comments are featured in the corpus.

Code: https://github.com/Norton50/COMP551Project1/blob/master/__main__.py
XML: https://raw.githubusercontent.com/Norton50/COMP551Project1/master/mxdvmc_fre.xml



load more comments (1 reply)

[illegible]afulsamee 13 points 14 hours ago
Ça dit :
Je vous trainerai aux Assises, bande d'enculés !
permalink embed report

Reddit ipiak 1 point 12 hours ago
Tous les quatre !
permalink embed parent report

JeanGuy17 9 points 15 hours ago
Un contexte ? Qui t'as écrit ça ? Ca pourrait aider pasque j'ai rien du tout là :-(
permalink embed report

LucasThinkAbout [S] 10 points 14 hours ago
J ai pas grand chose qui peut t'aider je pense :/ Fête avec des colocataires, et du monde de ma résidence
en Suisse, vers le début de matinée entre un samedi et un dimanche
permalink embed parent report

JeanGuy17 13 points 14 hours ago
P't'être qu'on te trolle ? Et par extension, tes amis nous trollent ^ ^
permalink embed parent report

LucasThinkAbout [S] 6 points 14 hours ago
Je pense pas pour être franc
permalink embed parent report

Anarchiste mouton 74 points 14 hours ago
Moi je te crois, le franc est suisse
permalink embed parent report

Greugreu 6 points 9 hours ago
Tu n'a pas la date précise de la fête ? Il y a le codage Date Shift. C'est prendre une date, genre 29 sep
ça fait 2992017 et tu séquence ton message en décalant les lettres par le numéro qui correspond. Exe
décale de 2, O décale de 9, etc.
Sachant que le message aurait un rapport avec une fête, ça ferait du sens.
permalink embed parent report

LucasThinkAbout [S] 2 points 9 hours ago
Samedi dernier
permalink embed parent report

Greugreu 2 points 9 hours ago
Mhm, ça n'a pas l'air d'être ça.
permalink embed parent report

MarkVador 9 points 14 hours ago*
Déchiffrer le "mec bourré" c'est comme réussir à déchiffrer le linéaire A. Et la distance d'unicité n'est peut-être
atteinte. Donc peu de chance de trouver la méthode et le message. Surtout si c'est du foutage de gueule ou u
combinaison de méthodes.
permalink embed report

LucasThinkAbout [S] 3 points 14 hours ago
Haha peut être, mais ça me parait trop clean pour être fait par un gars bourré
permalink embed parent report

MarkVador 3 points 14 hours ago
J'ai fait un edit sur la distance d'unicité. Ce qui laisse peu d'espoir.
permalink embed parent report

LucasThinkAbout [S] 3 points 14 hours ago
Il y avait pas de brouillon ou quoi que ce soit, aucun moyen pour qqn de prendre 20 min à écrire ç
juste être une technique qui m'est inconnue jusque là
permalink embed parent report

MarkVador 7 points 14 hours ago*
L'indice de coïcidence est de ~0.051. Ce serait plutôt de la substitution polyalphabétique, peut-
a une clef de chiffrement, une analyse périodique pourrait nous aider... Mais sur 36 caractères,
permalink embed parent report

LucasThinkAbout [S] 2 points 14 hours ago
Il n y a absolument rien bien malheureusement, j'essaye en ce moment même de trouver
pourraient fonctionner dans les mots codés les plus longs (en fonction des lettres uniques/
permalink embed parent report

MarkVador 4 points 14 hours ago
A priori, il faut viser un chiffrement par substitution polyalphabétique.
permalink embed parent report

Amiral_Poitou 4 points 11 hours ago
Attends attends, on substitue par mernage polyfonctionnel ou par mernage hébrieu
permalink embed parent report

MarkVador 3 points 11 hours ago
Mernage hybride parce que l'indice de viscosité granulaire n'est pas significatif (
permalink embed parent report

Amiral_Poitou 5 points 11 hours ago
Hum, je suis pas un fan, l'efficacité est quand même lié à la clé de cryptogé
ief ?)...
Enfin, je vais faire tourner un bot et je reviens vers toi pour qu'on mette les
permalink embed parent report
continue this thread ⟶

Arturo273 9 points 9 hours ago
Ce n'est pas un code, c'est juste du néerlandais.
permalink embed report
load more comments (3 replies)

MarkVador 10 points 7 hours ago
Ni Vigenère ni Beaufort et variante. Double ou triple chiffrage par substitution. Message trop court. Je penche
pour du troll.
permalink embed report

Figure 2. Illustration of a comment sequence for a typical Reddit post. Circled in red and blue are four comment sequences that would form separate conversations in the corpus. In red are the conversations that start from top-level comments. In blue is an auxiliary conversation, and in green are single-utterance comments that are ignored during the creation of the dataset.

We chose to use a package called langdetect[3], to help us remove any non-French artifacts from our dataset. After preliminary analysis of 12,820 comment sequences, we found 778 French conversations (about 6% of the corpus), substantial enough to make it worth preprocessing non-French utterances. In the case where a conversation had both French and non-French utterances, we removed the non-French utterance and all subsequent utterances but preserved the preceding French utterances if there were any. Where a conversation consisted entirely of French utterances, no comment string or <s>...</s> tags were written altogether.

After processing for non-French comments, we checked for single-utterance conversations i.e. top-level comments without replies and branches that were only one comment long. The 20,913 conversations in the final dataset represent fully French sequences that were found to have at least one turn in dialog from a total of 52,344 comment branches that were parsed.

Analysis

The distribution of utterance lengths (by number of words) is shown (Figure 3). It follows a left skewed normal distribution. The average utterance would have the length of 1-2 sentences.

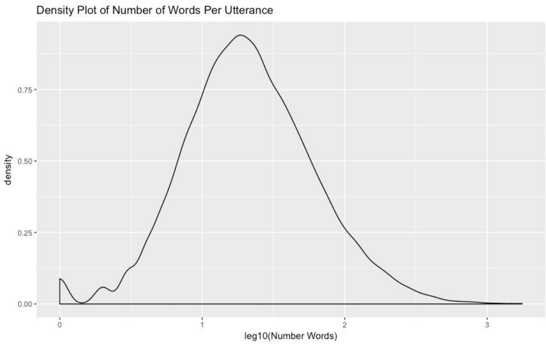


Figure 3. Density plot of utterance lengths by number of words. Lengths are $\log_{10}$ transformed.

Also, the number of utterances per conversation was examined. A histogram of the frequencies of utterances per conversation is plotted and a maximum of 10 utterances per conversation is observed.

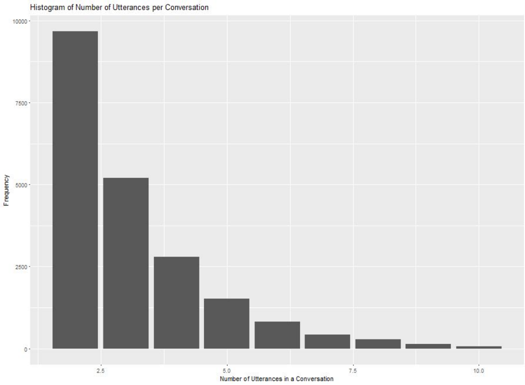


Figure 4. Histogram of number of utterance per conversation frequencies.

We analyzed the most common words and bigrams in the dataset (Figure 5). “Que” (English Translation: that, but, than, as or use a conjunction) was the most commonly written word at 39,194 occurrences. “Que” is the 9th most used French word[5] obtained from an analysis of all French movie subtitles. “Pas” and “les” held the second and third most common words at 37,710 and 36,825 occurrences. The most common bigarm was “de la”, this is consistent with the Google French corpus. [6]

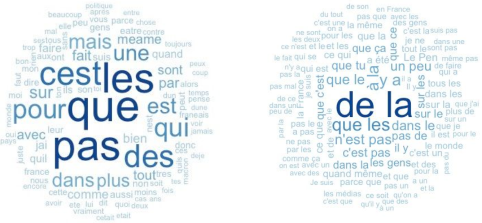


Figure 5. Illustration of word cloud of most common words in the dataset along with a wordcloud of most common bigrams.

Discussion

The conversations span a plethora of topics and offer valuable insight on colloquial, written French conversation.

Our corpus is comparable in size to existing French language corpora. Interview-based datasets such as Corp-Aix-2 contain around 1.7 million words, most of which are unavailable to other researchers and the public[8]. One of the largest French datasets, the *corpus de référence du francais contemporain* (CRFC) presents 60 million words from discussion forums[8]. An advantage of using our corpus compared to other datasets of online discussion is a homogenous comment sequence structure. Whereas many online forums are organized by date and time only, with many conversations being broken up by other conversations, our Reddit corpus has a clear conversational structure without sacrificing breadth. The corpus also has a fair representation of conversations of different lengths, with greater than 1700 conversations that consist of more than 5 turns (Figure 4).

When considering English datasets of a similar nature, we find many of the same characteristics as our French corpus. The average number of turns for the Twitter Corpus, which also describes online conversations, is 2, while our French corpus features around 3 turns per conversation[7]. Among non-chat internet corpora such as Agreement in Wikipedia Talk Pages and Agreement by Create Debaters, 2 seems to be the standard number of turns per conversation. The French corpus would appear to be largely reflective of typical conversation structures on the Internet, but provide greater utility in that single-comment conversations are filtered out. The French corpus is particularly useful for conversation analysis since each dialogue has at least two utterances; if our interest is in the generation of responses using data-driven approaches, the corpus provides suitable targets since it is structured to present relationships between utterances. Unlike in Twitter, responses in Reddit may be nested indefinitely, and we can see 58 conversations of length 10 that take advantage of this feature.

Some qualitative characteristics of the data to note are that conversations are natural (unscripted) and unrestricted in the number of contributing users. The limit on the number of characters per comment is generous at 10,000 characters, meaning the length of utterances are determined naturalistically, by the users themselves.

Overall, the Reddit corpus is a sizable, openly available repository of French dialogues reflecting humanistic conversation content and structure.

Statement of Contributions

Throughout the project, the entire group communicated about all major decisions and worked together to determine the direction we would take. Matthew worked heavily designed the python code to visit reddit, scrape for comments[2] and verify that they were in French[3,4]. He also helped write the report, focusing on the introduction and references. Marley wrote the logic for parsing and representing comment branches, conducted preliminary analysis of the corpus to guide decisions on how to handle non-French comments and comment branching, and wrote the dataset description and discussion. David preformed analysis of the dataset, wrote the analysis section and formatted the xml files for analysis in R. We hereby state that all the work presented in this report is that of the authors.

Code: https://github.com/Norton50/COMP551Project1/blob/master/__main__.py
XML: https://raw.githubusercontent.com/Norton50/COMP551Project1/master/mxdvmc_fre.xml

References

[1]“reddit: the front page of the internet", Reddit.com, 2017. [Online]. Available: http://reddit.com.

[2]B. Boe, "PRAW: The Python Reddit API Wrapper", Praw.readthedocs.io, 2017. [Online]. Available: https://praw.readthedocs.io/.

[3]M. Danilak, "langdetect 1.0.7 : Python Package Index", Pypi.python.org, 2017. [Online]. Available: https://pypi.python.org/pypi/langdetect?.

[4]Cybozu Labs, "shuyo/language-detection", GitHub, 2014. [Online]. Available: https://github.com/shuyo/language-detection/blob/wiki/ProjectHome.md.

[5]Open Subtitles,”Open Subtitles - movie subtitle database”, open-subtitles, 2017, [Online].www.opensubtitles.org

[6]Google ,”Google Ngram Viewer - Google Books", Books.Google.com, May 2012, [Online]. books.google.com/ngrams.

[7]Serban, I. (2017). A Survey of Available Corpora for Building Data-Driven Dialogue Systems. [online] Available at: https://arxiv.org/pdf/1512.05742.pdf

[8]Siepmann, D. (2017). Dictionaries and Spoken Language: A Corpus-Based Review of French Dictionaries. [online] Available at: https://academic.oup.com/ijl/article/28/2/139/2413042/Dictionaries-and-Spoken-Language-A-Corpus-Based