From T'es Qui to Qui Es-Tu*:*
A Naïve Bayesian Approach to Assessing Literate and Oral Discourse in Nonstandard French Language Data

Exposé

1 Computational Linguistic Aspect

While there are many ways in which language can be realized such as non-verbal communication, e.g., sign language, hand gestures or even whistled languages, for all intents and purposes, most languages fall into two main domains: oral and literate. This in of itself is nothing particularly profound. What is more interesting is that these two domains do not represent a natural dichotomy, as one might automatically assume, but rather, they represent two sectors of language that regularly overlap.

If one were to analyze standard language data that is typical of a specific domain, one would get unsurprising results. Data from scientific or professional sources such as newspapers, scientific journals, etc. tend to exhibit a more literate style, whereas SMS and ad postings generally exhibit a style that is more representative of oral discourse.

1.1. Computational Linguistic Problem

The underlying problem of identifying an oral or a literate style within a text lies in the fact that texts such as SMS inherently lack many of the elements that are associated with spoken speech like intonation, prosody, speed, accent, etc. This problem can be overcome by identifying other features which are typical of the respective discourse style and that can occur in a written medium.

For example, a prominent feature of oral discourse would be the extensive use of contractions. If a lot of contractions occur within a text, one could naively assume that this text string represents oral discourse. By combining this and other features together and analyzing the texts in this manner, it is possible to determine if a given a text string is of the oral or literate discourse.

The realization of the overlap of the oral and literate styles and to what degree will be exemplified on non-standard French language data from three main areas: Wiki-discussions, eBay postings, and SMS chats.

2 Related Works

2.1. Computational Linguistics

Ortmann, K., & Dipper, S. (2019). Variation between different discourse types: Literate vs. oral. *Proceedings of the sixth workshop on NLP for similar languages, varieties and dialects*, 64–79. https://doi.org/10.18653/v1/W19-1407

Ortmann, K., & Dipper, S. (2020). Automatic orality identification in historical texts. *Proceedings of the 12th language resources and evaluation conference*, 1293–1302. https://www.aclweb.org/anthology/2020.lrec-1.162

2.2. Theoretical Linguistics

Bader, J. (2002). Schriftlichkeit und Mündlichkeit in der Chat-Kommunikation. *Network*, *29*, https://doi.org/10.15488/2920

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe—Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, *36*, 15–43.

Rehm, G. (2002). Schriftliche Mündlichkeit in der Sprache des World Wide Web. In A. Ziegler & C. Dürscheid (Eds.), *Kommunikationsform E-Mail* (pp. 263–308). Tübingen. http://www.georg-re.hm/pdf/Rehm-Muendlichkeit.pdf

Vilmos, Á., & Mathilde, H. (2012). Überlegungen zur Theorie und Praxis des Nähe- und Distanzsprechens. *Zeitschrift Für Rezensionen Zur Germanistischen Sprachwissenschaft*, *4*(2), 156–161. https://doi.org/doi:10.1515/zrs-2012-0032

3 Methodology

3.1. Proposed Solution

One naïve solution that would mostly likely produce the highest accuracy would involve having native speakers, in this case native French speakers, identify sentences by hand and tag them as either belonging to oral or literate discourse. However, this approach would only be applicable if the corpus were reasonably small, e.g., a couple dozen sentences at best.

The solution that I find to be better suited for the task at hand involves using another naïve solution in the form of a naïve bayes classifier to classify sentences according to the discourse nature of the texts. For this to be feasible, an accurate feature set must be first be present.

3.2. Feature Set

A modified feature set as presented by Ortmann, K., & Dipper, S. (2019) will serve as the foundation for identifying oral features:

Feature	**Description**
ABR	The number of abbreviations
COORDINIT	Proportion of sentences beginning with a coordinating conjunction.
DEM	Ratio of demonstrative pronouns (tagged as PDS) to all words.
DEMSHORT	Proportion of demonstrative pronouns (tagged as PDS) with lemmas dies 'this/that' or der 'the' which are realized as the short form (lemma der 'the').
EXCLAM	Proportion of exclamative sentences, based on the last punctuation mark of the sentence.
INTERJ	Proportion of primary, i.e. one-word interjections (e.g. ach, oh, hallo) to all words.

LEXDENS	Ratio of lexical items (tagged as ADJ.*, ADV, N.*, VV.*) to all words.
MEAN_SENT	Mean sentence length, without punctuation marks.
MEAN_WORD	Mean word length.
MED_SENT	Median sentence length, without punctuation marks.
MED_WORD	Median word length.
ORTH	Orthographical mistakes such as a spelling errors
PRON1ST	Ratio of 1st person sg. and pl. pronouns with lemmas ich 'I' and wir 'we' to all words
PRONSUBJ	Proportion of subjects which are realized as personal pronouns, based on the head of the subject.
PTC	Proportion of answer particles (ja 'yes', nein 'no', bitte 'please', danke 'thanks') to all words.
QUESTION	Proportion of interrogative sentences, based on the last punctuation mark of the sentence.
SUBORD	Ratio of subordinating conjunctions (tagged as KOUS or KOUI) to full verbs.
V:N	Ratio of full verbs to nouns.

3.3. Modules

The modules that will be used to identify the features include, but are not limited to:

Standard	**Pip**
Re - Assisting spacy in identifying non-standard elements of language such as abbreviations, misspellings, etc. Os - Detecting if the program has all the necessary files it needs to function. Shutil - Moving and copying files Tkinter - Dynamically selecting files Datetime/time - Time stamps for the data files	Beautifulsoup - Parsing the XML files Spacy - Tagging the text NLTK - Tokenizing the text Matplotlib - Creating graphical representation of the data pyspellchecker - Identify misspelled words in a text

3.4. Sparse Data Problem

As with any classifier, there is the general problem of sparse data. To resolve this, the data must be adequately smoothed. A simple, but efficient smoothing method that can be used with a naïve bayes classifier is that of Ng (1997):

- if $C(C_j, S_n) = 0$
  - $P(c_j|s_n)\frac{P(s_n)}{N} = \frac{C(s_n)}{n^2}$
  - N = the amount of training data

3.5. Evaluation

3.5.1. Non-Statistical Evaluation Criteria

- Programming Overhead
    - Training required
    - Additional resources
    - Time needed for creating rules and templates.
- Language dependency
    - Does this work for all languages?
    - Can it be adjusted for a new language?
- Domain dependency
    - Would this approach work in other domains?
- Efficiency
    - How fast is the program?
    - How much memory does it require?

3.5.2. Spacy Evaluation

Seeing as how Spacy has been mainly trained on standard language data and not non-standard language data, its accuracy and error rate must first be properly assessed.

A set number of words from each corpus will be set aside and tagged by Spacy. The results will then be compared to a gold standard that I create by hand. This will demonstrate the overall reliability of the tagger, the results of which will appear as part of the documentation. This data set will then be part of the development corpus so as not to influence the results.

3.5.3. Bayes Evaluation

To be able to evaluate the accuracy/error rate of the naïve bayes tagger and the usefulness of the feature set from 3.2, language data from the development corpus will be used to develop a second feature set. Features akin to Ortmann, K., & Dipper, S. (2019) will be used, wherever applicable. These include, but are not limited to:

- **Participants**: Number of participants
- **Interactiveness**: Monolog, dialog
- **Production circumstances**: synchronous, quasisynchronous, asynchronous

- **Reception circumstances:** synchronous, quasisynchronous, asynchronous

In addition to the Ortmann, K., & Dipper, S. (2019) feature set, separate linguistic features will be implemented to further asses the nature of the sentences.

- A fix set list on words/expressions that are deemed to be a part of oral speech.
- A set list of words/expressions that are deemed to be a part of literate speech.

Should one of these words occur in a sentence, then the sentence is automatically marked as being either literate or oral.

3.5.4. Simplified Worked Example

The features that are listed here have not been selected according to any empirical criterial. Therefore, let us naively assume for demonstrative purposes that a sentence must have **at least one of** the following properties to be recognized by our naïve bayes tagger as either literate or oral:

Naïve Bayes Set

1. Oral
   a. ABR >2
   b. INTERJ >2
2. Literate
   a. MEAN_SENT >7
   b. VERB > 1

Simplified Tagger Set

1. Oral
   a. Ouais
   b. Mdr
   c. Bises
2. Literate
   a. S'agit-il
   b. Échantillon

1. Meme pas de texto le gars
   a. The simplified tagger would mark this sentence as oral because it includes the word "le gars"
   b. The naïve tagger would mark this sentence as oral because it lacks a verb.
2. Mdr ouai mais bon à se point c'est nul
   a. The simplified tagger would mark this sentence as oral because it includes the words "Mdr" and "Ouai"
   b. The naïve bayes tagger would mark this sentence as **literate** because it has a verb, and the sentence has at least 7 words.

3 S'agit-il d'un pays à une époque donnée

a. The simplified tagger would make this as literate because of S'agit-il.
b. The naïve bayes would mark this as literate because it is longer than 7 words.

This simplified tagger will be used to generate a gold standard against which the naïve bayes will compared. This means that there will then be two sets of results: the results from the naïve bayes (main program) and the gold standard (the simplified tagger/recognizer). By comparing these two results, we can assess accuracy and the error rate. The goal here is for the naïve bayes to perform better than the simplified tagger.

4 Corpora

4.1. Data Sets

The first data set was compiled by the department of romance studies of the University of Potsdam. The other two data sets were obtained from Corpora of Computer-Mediated Communication in French, also known as the CoMeRe Repository.

1. eBay Ad Postings (La-bank: Resources for Research and Teaching)

Gerstenberg, A., & Hewett, F. (2019). *A collection of online auction listings from 2005 to 2018 (anonymised)* [Data set]. La-bank: Resources for Research and Teaching. https://www.uni-potsdam.de/langage/la-bank/ebay.php

2. Wiki discussions (TEI-CMC version of Wikipedia discussions associated with the article "Quotient intellectuel")

Poudat, C., Grabar, N., Kun, J., & Paloque-Berges, C. (2015). *TEI-CMC version of wikipedia discussions associated to the article "Quotient intellectuel"* (Cmr-wikiconflits-qi_discu-tei-v1) [Data set]. CoMeRe Corpora Repository. https://hdl.handle.net/11403/comere/cmr-wikiconflits/cmr-wikiconflits-qi_discu-tei-v1

3. French SMSs (88milSMS. A corpus of authentic text messages in French.)

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., & Verine, B. (2016). *88milSMS. A corpus of authentic text messages in French (nouvelle version du*

corpus ISLRN : 024-713-187-947-8) (Cmr-88milsms-tei-v1) [Data set]. Banque de Corpus CoMeRe. https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1

4.2. Pre-processing

All the data sets are available in the .xml format. The eBay data has been tagged to with respect to typical features of ad postings such as abbreviations, misspellings, marketing language, slang, etc. The remaining data sets have been tagged for emoticons, and personal pronouns. All the data sets have markers to identify author, date, time, title of the post, etc.

Before the individual entries can be properly classified, the data sets must first be extracted from the .xml files. This will be done using the library *beautifulsoup*. Then, they will have to be tokenized, which will be done using the *nltk tokenizer* library. Finally, they will have to be POS-tagged which will using the *Spacy* Tagger.

To account for the non-standard nature of the data set, regular expressions will be implemented to detect things such as misspellings, abbreviations, proper nouns, expressive reduplication of letters or symbols.

4.3. Development, Training and Test Corpus

1 Development

A portion of each data set will be involved in creating, refining the algorithms, and addressing other unforeseen. The data will be visible to both to me and the program. The statistical results will not be included in the documentation, but only in the parts of the methodology.

2 Training (10-Fold Validation)

After the developmental stage, experimental testing will be done using a portion from each dataset. This will be visible to both to me and the program. The results of which will be included in the discussion portion of the documentation.

3 Test (10 K-Fold Validation)

Once the developmental and testing phases have been successfully completed, a portion from each data set will be used for the finally testing. During the testing phase, the data set will only be visible to the program. The results of which will be included in the discussion portion of the documentation.

5 Documentation Draft Layout

Abstract

List of Figures

List of Tables

List of Terms

1. Introduction
2. Related Works
3. Features of Discourse
   - 3.1. Oral
   - 3.2. Literate
4. Discourse Styles within the French Language
   - 4.1. French Language Registers
   - 4.2. Typical Features of Oral French
   - 4.3. Typical Features of Literate French
5. The French Language Corpora
   - 5.1. Nature of the Data Sets
   - 5.2. Data Pre-processing
6. Methodology
   - 6.1. Discourse Classification with Naïve Bayes
   - 6.2. Feature Sets for Identifying Discourse Types
   - 6.3. Establishing Discourse Classification Baseline
7. Evaluation
   - 7.1. Non-Statistical Evaluation Parameters
   - 7.2. Statistical Evaluation Parameters
8. Discussion
   - 8.1. Developmental Phase
   - 8.2. Experimental and Training Phase
   - 8.3. Testing Phase
9. Conclusion

References

Eigenständigkeitserklärung

6 README

1 License

2 Requirements Text

   a. Python libraries

3 Description of Program Functionality

7 Typical Examples of the Corpora

The following texts represent typical examples from each of the corpora to be used. The examples have not been sorted by any algorithm, but rather hand-picked by me as indicators of what a particular style could look like.

7.1. Oral

- **88SMS**
  - Ben ouais mais des fois ils trouvent jamais ou ils sont pas d'acord ou ils peuvent faire que des hypotheses. [1]
  - Mdr ouai mais bon à se point c'est nul[2]
- **Wiki**
  - Oh, on a beaucoup ri de ce ministère et de son ministre ![3]
- **eBay**
  - Je vends une caméra de surveillance et un écran de 13 cm Ce matériel se monte très facilement. Il suffit de brancher la caméra et l'écran à une prise électrique et le tour est joué[4]

7.2. Literate

- **88SMS**
  - Bonjour. Peux-tu me confirmer ton horaire d'arrivée ? Bises et bon voyage[5]
- **Wiki**
  - C'est le rapport entre l'âge « mental » que donne le résultat du test sur l'âge réel, multiplié par 100. [6]
  - S'agit-il d'un pays à une époque donnée ? [7]
- **eBay**
  - 10 pelotes Phildar Coton Phil 51 Coloris n° 423 (noir), même bain. 51% coton, 49% acrylique. Aiguilles 2 - 3, 116 m. Échantillon 10 x 10 cm: 25 mailles, 34 rangs.[8]

[1] <post xml:id="cmr-88milsms-a10211" when-iso="2011-09-24T12:06:27" who="#cmr-88milsms-p399" type="sms">
[2] <post xml:id="cmr-88milsms-a10217" when-iso="2011-09-24T12:07:48" who="#cmr-88milsms-p504" type="sms">
[3] <text xml:id="cmr-wiki-c002-rev5008107" prev="#cmr-wiki-c002-rev5008019">
[4]
[5] <post xml:id="cmr-88milsms-a42378" when-iso="2011-10-16T23:22:58" who="#cmr-88milsms-p248" type="sms">
[6] <text xml:id="cmr-wiki-c002-rev4846889" prev="#cmr-wiki-c002-rev4845720">
[7] <text xml:id="cmr-wiki-c002-rev741072" prev="#cmr-wiki-c002-rev740972">
[8]

8 Bibliography

Bader, J. (2002). Schriftlichkeit und Mündlichkeit in der Chat-Kommunikation. *Network*, *29*. https://doi.org/10.15488/2920

Barme, S. (2012). *Gesprochenes Französisch*. De Gruyter. https://doi.org/10.1515/9783110279832

Cook, J. (2012). Les marques lexicales du français familier dans la traduction polonaise des dialogues romanesques. *Traduire*, *226*, 93–107. https://doi.org/10.4000/traduire.162

Gerstenberg, A., & Hewett, F. (2019). *A collection of online auction listings from 2005 to 2018 (anonymised)* [Data set]. La-bank: Resources for Research and Teaching. https://www.uni-potsdam.de/langage/la-bank/ebay.php

Goudailler, J.-P. (2002). De l'argot traditionnel au français contemporain des cités. *La linguistique*, *38*(1), 5–24. Cairn.info. https://doi.org/10.3917/ling.381.0005

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing,computational linguistics, and speech recognition* (2nd ed.). Prentice-Hall, Inc.

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe—Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, *36*, 15–43.

Müller, B. (1975). *Das Französische der Gegenwart: Varietäten, Strukturen, Tendenzen*. Winter.

Ng, H. T. (1997). Exemplar-Based Word Sense Disambiguation" Some Recent Improvements. *Second Conference on Empirical Methods in Natural Language Processing*, 208–2013. https://www.aclweb.org/anthology/W97-0323

Ortmann, K., & Dipper, S. (2019). Variation between different discourse types: Literate vs. oral. *Proceedings of the sixth workshop on NLP for similar languages, varieties and dialects*, 64–79. https://doi.org/10.18653/v1/W19-1407

Ortmann, K., & Dipper, S. (2020). Automatic orality identification in historical texts. *Proceedings of the 12th language resources and evaluation conference*, 1293–1302. https://www.aclweb.org/anthology/2020.lrec-1.162

Panckhurst, R. (2016). A digital corpus resource of authentic anonymized French text messages: 88milSMS—What about transcoding and linguistic annotation? *Digital Scholarship in the Humanities*, *21*, 92–102. https://doi.org/10.1093/llc/fqw049

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., & Verine, B. (2016). *88milSMS. A corpus of authentic text messages in French (nouvelle version du corpus ISLRN : 024-713-187-947-8)* (Cmr-88milsms-tei-v1) [Data set]. Banque de

Corpus CoMeRe. https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1

Poudat, C., Grabar, N., Kun, J., & Paloque-Berges, C. (2015). *TEI-CMC version of wikipedia discussions associated to the article "Quotient intellectuel"* (Cmr-wikiconflits-qi_discu-tei-v1) [Data set]. CoMeRe Corpora Repository. https://hdl.handle.net/11403/comere/cmr-wikiconflits/cmr-wikiconflits-qi_discu-tei-v1

Prüßmann-Zempher, H. (2010). 337. Varietätenlinguistik des Französischen / Linguistique des variétés. In G. Holtus, M. Metzeltin, & C. Schmitt (Eds.), *Band V/1 Französisch* (pp. 830–843). Max Niemeyer Verlag. https://doi.org/10.1515/9783110966091.830

Rehm, G. (2002). Schriftliche Mündlichkeit in der Sprache des World Wide Web. In A. Ziegler & C. Dürscheid (Eds.), *Kommunikationsform E-Mail* (pp. 263–308). Tübingen. http://www.georg-re.hm/pdf/Rehm-Muendlichkeit.pdf

Stein, A. (2014). *Einführung in Die Französische Sprachwissenschaft* (4th ed.). J.B. Metzler.

Vilmos, Á., & Mathilde, H. (2012). Überlegungen zur Theorie und Praxis des Nähe- und Distanzsprechens. *Zeitschrift Für Rezensionen Zur Germanistischen Sprachwissenschaft*, *4*(2), 156–161. https://doi.org/doi:10.1515/zrs-2012-0032