|  |
| --- |
| **From T’es Qui to Qui Es-Tu*:* A Naïve Bayesian Approach to Assessing Literate and Oral Discourse in Nonstandard French Language Data**    Schriftliche Hausarbeit  für die Bachelorprüfung der Fakultät für Philologie  an der Ruhr-Universität Bochum  (Gemeinsame Prüfungsordnung für das Bachelor/Master-Studium  im Rahmen des 2-Fach-Modells an der RUB vom 03. November 2016)  Vorgelegt von  Chandler, Christopher  Abgabedatum  14.08.2021  Prof. Dr. Stefanie Dipper  Prof. Dr. Ralf Klabunde |

Table des matières

[Abstract 3](#_Toc75692638)

[List of Figures 3](#_Toc75692639)

[List of Tables 3](#_Toc75692640)

[List of Terms 3](#_Toc75692641)

[1 Introduction 3](#_Toc75692642)

[2 Related Works 3](#_Toc75692643)

[3 Features of Discourse 3](#_Toc75692644)

[3.1.Oral 3](#_Toc75692645)

[3.2.Literate 3](#_Toc75692646)

[4 Discourse Styles within French Language 3](#_Toc75692647)

[4.1.French Language Registers 3](#_Toc75692648)

[4.2.Typical Features of Oral French 3](#_Toc75692649)

[4.3.Typical Features of Literate French 3](#_Toc75692650)

[5 The French Language Corpora 3](#_Toc75692651)

[5.1.Nature of the Data Sets 3](#_Toc75692652)

[5.2.Data Pre-processing 3](#_Toc75692653)

[6 Methodology 3](#_Toc75692654)

[6.1.Discourse Classification with Naïve Bayes 3](#_Toc75692655)

[6.2.Feature Sets for Identifying Discourse Types 3](#_Toc75692656)

[6.3.Establishing Discourse Classification Baseline 3](#_Toc75692657)

[7 Evaluation 3](#_Toc75692658)

[7.1.Non-Statistical Evaluation Parameters 3](#_Toc75692659)

[7.2.Statistical Evaluation Parameters 3](#_Toc75692660)

[8 Discussion 3](#_Toc75692661)

[8.1.Developmental Phase 3](#_Toc75692662)

[8.2.Experimental and Training Phase 3](#_Toc75692663)

[8.3.Testing Phase 3](#_Toc75692664)

[9 Conclusion 3](#_Toc75692665)

[References 4](#_Toc75692666)

[Eigenständigkeitserklärung 6](#_Toc75692667)

# Abstract

# List of Figures

# List of Tables

# List of Terms

# Introduction

# Related Works

# Features of Discourse

## Oral

## Literate

# Discourse Styles within the French Language

## French Language Registers

## Typical Features of Oral French

## Typical Features of Literate French

# The French Language Corpora

## Nature of the Data Sets

## Data Pre-processing

# Methodology

## Discourse Classification with Naïve Bayes

## Feature Sets for Identifying Discourse Types

## Establishing Discourse Classification Baseline

# Evaluation

## Non-Statistical Evaluation Parameters

## Statistical Evaluation Parameters

# Discussion

## Developmental Phase

## Experimental and Training Phase

## Testing Phase

# Conclusion

# References

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>

# Eigenständigkeitserklärung

I hereby declare that the work submitted is my own and that all passages and ideas that are not mine have been fully and properly acknowledged. I am aware that I will fail the entire course should I include passages and ideas from other sources and present them as if they were my own.

Hiermit versichere ich, dass ich die Arbeit selbständig angefertigt, außer den im Quellen- und Literaturverzeichnis sowie in den Anmerkungen genannten Hilfsmitteln keine weiteren benutzt und alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, unter Angabe der Quellen als Entlehnung kenntlich gemacht habe.

|  |  |
| --- | --- |
| Ort/Place, Date/Datum | Name |
| Kamen, 14.08.2021 |  |
|  | Christopher Michael Chandler |