|  |
| --- |
| **From T’es Qui to Qui Es-Tu*:* A Naïve Bayesian Approach to Assessing Literate and Oral Discourse in Non-standard 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  31.08.2021  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

# Terminology

# Introduction

Excluding other modes in which human communication can be realized such as via sign language, body language, whistling, etc., human languages are generally expressed using at least one of the two discourse types: oral or literate discourse (Bader, 2002). In the most literal sense, oral discourse can be understood as process that employs audible sounds to express meaning, whereas literate discourse is visual medium that is expressed through visible, written symbols (Bader, 2002).

Furthermore, despite the somewhat trivial natural of this discourse classification, these two domains do not represent a natural dichotomy, as one might automatically assume, but rather, they are two sectors of language that regularly overlap.

Oral and literate discourse are often realized by employing non-standard language and standard language, respectively. In the strictest sense, standard language is language which conforms to norms set forth by an official linguistic institution or work e.g., Académie Française, Real Academia *Española*, Oxford English Dictionary, Duden, etc. This contrasts with non-standard language which is not beholden to any norms whatsoever. It is therefore often characterized by the heavy usage of dialectal and vernacular speech (Bader, 2002).

With respect to the aforementioned definitions, data that stems from scientific or professional sources such as newspapers, scientific journals, political speeches, etc. tends to exhibit a style that is more akin to literate discourse. This contrasts with SMS, chat messages, online ad postings, forum discourse, etc. as they generally exhibit a style that is more representative of oral discourse. (Bader, 2002).

It is however possible to discern to the two irrespective of the domain. However, the underlying problem in identifying an oral or a literate style within a text is twofold. First, there is the issue that written speech inherently lacks several paralinguistic cues that are associated with spoken speech like intonation, prosody, speed, accent, and so on (Bader, 2002). This is further compounded by the fact that written speech can also represent language that can realized be orally but is done so artificially. This paradigm applies to documents such as presentations, speeches, sermons and news reports.

One naïve solution to solving this problem would involve native speakers sentences by hand and classify according However, this approach is cumbersome and length. It would thus only be suitable if the corpus were reasonably small i.e., a couple dozen sentences at most. Therefore, the most efficient way in solving this dilemma would have to involve a method which allows for the automatic disambiguation of oral and literate discourse.

While this method can theoretically be applied to any language, the object language of this thesis is modern French from c. 2005 – 2019. The language data has been sourced from reputable sectors that are generally considered to be representative of standard as well as non-standard language. These includes Wiki-discussions, eBay postings, and SMS chats.

A common method of disambiguation that I will employ involves the supervised usage of a naïve bayes classifier with a specific feature and classification sets (Jurafsky & Martin, 2009). First, the feature sets which are to be typical of the respective discourse styles and how they can occur in a written medium must be identified. Then the naïve bayes will be trained according to these features thereby allowing the classifier to determine to a relative degree of probability if a given text is representative of oral or literate discourse. The methodology as well as the results of this process will be presented in detail at the end of this face.

# Related Works

When examining the concept of oral and literate language within the discipline of computational linguistics, it is paramount to include not only the theoretical aspect of the dynamic, but the computational one as well.

## Theoretical linguistics

Koch and Oesterreicher (1985) were very influential in setting up the paradigm of literate and oral discourse. They did this by providing a distinction between the medial and the conceptual facets of language. The medial aspect of the language refers to whether a given message is expressed in either a graphic or a phonetic medium. The conceptual aspect is then the either spoken or written. Koch and Oesterreicher (2007) reprised their roles in offering a more detailed explanation regarding literate and oral discourse. They also place focus on sociolinguistic aspects regarding this paradigm.

Another factor that plays a role in identifying orality and literacy in a text is that of *Distanzsprache* and *Nähesprache[[1]](#footnote-1)*. Distanzsprache represents how removed the speaker is mentally, conceptually and physically. On the other hand, Nähesprache depicts how close one is to any given situation. Situations that are more of a personal and physical nature will often be assigned to the category of Nähesprache.

As the specific object language in question is French, there are French-specific elements that should be taken into and will be instrumental in determining the necessary features. While Müller (1975) predates Koch and Oesterreicher (1985), the notion of orality and literacy was already known to Müller (1975) and refers to them as *français parlé, message oral, languée* and *français écrit, message écrit, langue écrite, langage écrit* respectively. Alongside other aspects, Müller (1975) explores this distinction and how it is realized chronologically, quantitatively, qualitatively, diatopically, diastratically with special focus placed on the French language.

## Computational linguistics

Using a naïve bayes classifier for text classification purposes is on itself not a new process. However, what is new and has been seldom done is using it to identify oral and literate nature of a give text. Ortmann and Dipper (2019) asses the ideas as proposed by Koch and Oesterreicher(1985) to be able to automatically identify literate and oral discourse in modern German texts.

Ortmann and Dipper (2020) also applied the same methodology, also referencing Koch and Oesterreicher (1985), to assess the oral and literate nature of historical texts. This was done by using a slightly altered feature set that is more adept to historical texts as the non-standardized nature of historical documents cannot be properly analyzed using modern criteria.

To better identify literacy and orality in a text Ortmann and Dipper (2019), drew from Bader (2002) and Rehm (2001) who both provide such criteria, albeit from slightly different angels. Bader (2002) provides a rounded, general approach to properly assess orality and literacy in texts in the same vein as Müller (1975). Bader (2002) applies the analyses to digital communication e.g. e-mail, chat, newsgroups, etc., while also providing features to identify the precise nature of individual excerpts from said communication. Rehm (2001) offers a more restricted analysis by only detailing the nature, characteristics and features of written language on the internet e.g. e-mail, chat data, websites, etc.

The internet is an everchanging dimension and one could therefore rightfully question the validity of using such articles from nearly two decades ago. However, despite both articles having been published at the turn of the 21st century, where the internet and digital communication were still in their infancy, both works comment and detail phenomena that are still relevant to modern internet communication.

# General Features of Language and Discourse

## Language

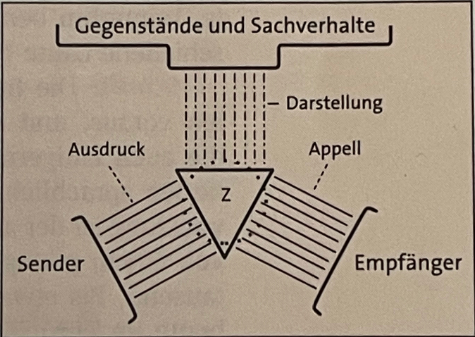
As a mode of communication, it is something that humans have been capable of for around 100,000 years (Stein, 2014). Human Language, is first and foremost, the production of audible sounds i.e., speech or written symbols i.e., letters, characters, etc. (Bader, 2002).

It is tempting to analyze and define language by contrasting it with the concept of dialects. While this can be done, the goal here is to provide a general definition of what exactly is meant by language as a construct. Furthermore, to preserve the dichotomy of written versus spoken, language here will be treated as being confined to these two dimensions.

De Saussure makes the distinction of *Parole* versus *Langue*. Parole being the actual realization of language system. Langue being the virtual construct of a given language that could be realized by a speaker of said language (Stein, 2014).

The language system is simply the aggregation of conventions, norms, value and opposition. The value of a given word, be it phonetic or graphic, is that it can be distinguished from another element. If there is a distinction between these two elements, then opposition is present (Stein, 2014). Should they have the same function, then it would be necessary to refer to them as variants of one another.

A final important aspect of language is the relationship that speakers have to one another. More specifically, how communication can work between speakers. There exist at least two main models Jakob and Bühler for explaining the communication aspect of language, but they serve the same purpose, which is to present the function of language (Stein, 2014) .



Bühler Organ-Modell (Stein, 2014, p. 1)

The organ model is a communication model that models the way linguistic information is received and processed. Every communication process consists of three essential parts: *Sender*, *Empfänger* and Gegenstände und Sachverhalte. *Sender* is the speaker, with *Empfänger* being the listener. Gegenstände und Sachverhalte are the messages being transmitted. All three of these are connected through Z which represents the language i.e., *das sprachliches Zeichen* (Stein, 2014).

The sprachliches Zeichen is simply what every is transmitted via language. It has three main functions: Ausdruck, Darstellung, Appel. The Ausdruck expresses the opinions and feelings of the speaker. These are the symptoms of the sprachliches Zeichen. (Stein, 2014). The Darstellung is the symbol for the information. The Appel has the function of eliciting a desired response from the listener that is in line with the the sprachliches Zeichen. (Stein, 2014). All three of these functions are present in every message, but general one message will dominate over the others (Bader, 2002).

On a closing note, one of the most prominent features of human language, is that it is possible to transmit abstract information in a phonetic or graphical form (Stein, 2014). This is an ability that connects all humans, while at the same distancing them from other animals (Bader, 2002).

## Oral Speech, Oral, Spoken, Speaking and Orality

Oral speech in the most simplest can be understood as the phonetic expression of thought (Bader, 2002). This is in line with De Saussure, who along with other structural linguists, saw spoken language superseding and therefore being the precursor of written language (Stein, 2014). This in of itself not surprising seeing as how one e.g., a child learns how to speak before one can write. With this in mind, speaking is a spontaneous process that is directly coupled with the transience (Bader, 2002).

Due to the nature of orality being a primary factor chronologically speaking (Bader, 2002), it is the feature that is the most prominent and the one that has been object of great discussion, especially since the 20th century (Stein, 2014),

## Written Speech, Written, Literate, Literacy

Seeing as orality is the phonetic expression of thought, literacy is then to be seen as graphical depiction and recording of said thought (Bader, 2002). This. In that case,

## Medium, Conception and Distance-Proximity

Koch and Oesterreicher (1985) have created an elegant, but simple paradigm of addressing the conceptual and medial nature of discourse types.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Konzeption | | |
| Medium |  | Gesprochen | Geschrieben |
| Graphischer  Kode | Faut pas le dire | Il ne faut pas le dire |
| Phonischer  Kode | [fopaldiʀ] | [ilnəfplalədiʀ] |

Fig. 1 (Koch & Oesterreicher, 1985, p. 17)

Although it would be wrong to see a dichotomy being present between orality and literacy, this is not strictly correct. The dichotomy does exist, but it only applies to the dual nature of the discourse. Excluding all other modes, either a message is communicated through spoken speech, or it written speech. Regarding the medial representation, i.e., the graphic code and the phonetic code , a dichotomy is present. The other question remains though: What is to be done with the conceptual aspect of language?

Here, it would be false to assume that spoken speech can only represent spoken speech and written speech can only represents speech. Koch and Oesterreicher(1985) see ,spoken’ and ,written’ as being a continuum with conceptual possibilities that have different levels. They exemplify this in the following diagram.



Fig. 2 (Koch & Oesterreicher, 1985, p. 17)

On the phonic portion of the diagram, a,b,c,g,h,i represent spoken speech that starts of being of a informal and personal nature and gradually becomes less informal and person. In doing so, it is more in lines with written speech. When observing the two poles, a and i, there is an obvious difference between an informal conversation and a presentation. The former most likely represents spontaneous speech, while the latter is something that prefabricated and then presented to an audience in an oral form. On the graphic portion of the diagram, d,e,f,j,k all represents possible graphic representations of speech, with a prepared interview being the most oral and an administrative regulation being the most written and least spoken realization.

Another dynamic presented by Koch and Oesterreicher(1985) is that of *Nähe* and *Distanz*. It is not enough to simply address the written or spoken nature of any given speech, but also address how close in terms of proximity and familiarity the speakers are to one another. Nähesprache is reserved for situations that physical and familiar in nature. This includes, but is not limited to, communication that is spontaneous, face-to-face and familiar. Distanzsprache represents the opposite pole in that it depicts speech that includes, but is not limited, communication that is detached, objective, unfamiliar.



Fig. 3 (Koch & Oesterreicher, 1985, p. 17)

Using all, three of these parameters: Medium, Conception and Distance-Proximity, a more detail analysis of language is possible. An informal conversation is thus representative of spoken speech, that is also conceptual representative of spoken speech. The dynamic of the speakers is one familiarity and closeness, and the speech can therefore be assigned the label of Distanzsprache. The opposite can be said of administrative regulation. There is great distance between the speakers, both in terms of familiarity and proximity. It is also not a message that can be communicated orally due to the very nature of the text. Therefore, it can be assigned as being conceptually and medially written speech, while also belonging to Distanzsprache.

# Registeres 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.

Koch, P., & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift Für Germanistische Linguistik*, *35*, 246–275.

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.

# 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 |

1. These are also sometimes referred to as Sprache der Distanz and Sprache der Nähe respectively without any change meaning. [↑](#footnote-ref-1)