|  |
| --- |
| **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 |

**Abstract**

An overlooked aspect of communication is the conceptual, where literacy corresponds conceptually to written language, and orality corresponds conceptually to spoken language. Non-standard French language data was obtained from eBay, SMS chats and Wikiconflits to explore how these two facets are realized in different internet domains. Training data was automatically developed using classification criteria that is typical of literacy and orality in French. This data was then used to train a naïve Bayes model to assign the most probable conceptual classification feature to a document. The results show that those who use the platform eBay tend to express themselves conceptually in a more literate fashion. SMS chats display a high level of conceptual orality and Wikiconflits show a tendency be conversative in their orality.

**Main Table of Contents:**

[1. Introduction 8](#_Toc79642329)

[2. Related Works 9](#_Toc79642330)

[2.1. Theoretical Linguistics 9](#_Toc79642331)

[2.2. Computational Linguistics 9](#_Toc79642332)

[3. Language as a Construct 10](#_Toc79642333)

[3.1. General Features of Language 10](#_Toc79642334)

[3.2. Medial Features 11](#_Toc79642335)

[3.3. Conceptual Features 12](#_Toc79642336)

[4. Diaphasic and Diastratic Registers 14](#_Toc79642337)

[4.1. Le Français 15](#_Toc79642338)

[4.2. Français Cultivé 15](#_Toc79642339)

[4.3. Français Familier 16](#_Toc79642340)

[4.4. Français Populaire 17](#_Toc79642341)

[4.5. Français Vulgaire 18](#_Toc79642342)

[4.6. Francais Argotique 18](#_Toc79642343)

[4.7. Français Technique 18](#_Toc79642344)

[5. The French Language Corpora 19](#_Toc79642345)

[5.1. Data Sets 20](#_Toc79642346)

[5.2. Pre-processing 21](#_Toc79642347)

[6. Methodology 22](#_Toc79642348)

[6.1. Naïve Bayes 22](#_Toc79642349)

[6.2. A Worked Example 25](#_Toc79642350)

[6.3. Combining Registers and Discourse 26](#_Toc79642351)

[7. System Evaluation 28](#_Toc79642352)

[7.1. Developmental Overhead 28](#_Toc79642353)

[7.2. Classification Sets and Naïve Bayes 28](#_Toc79642354)

[7.3. Sentence Tokenizer 31](#_Toc79642355)

[8. Results 32](#_Toc79642356)

[8.1. Development phase 32](#_Toc79642357)

[8.2. Training phase 34](#_Toc79642358)

[8.3. Testing phase 36](#_Toc79642359)

[9. Discussion 36](#_Toc79642360)

[10. Conclusion 40](#_Toc79642361)

[11. References 41](#_Toc79642362)

**List of Equations:**

[Equation 1. Bayes’ Theorem 23](#_Toc79609235)

[Equation 2. Bayes’ Theorem Reversed 24](#_Toc79609236)

[Equation 3. Normalizing Constant 24](#_Toc79609237)

[Equation 4. Naïve Bayes’ Classifier 24](#_Toc79609238)

[Equation 5. Argmax 24](#_Toc79609239)

[Equation 6. Argmax of Classification 24](#_Toc79609240)

[Equation 7. Model Probabilities 25](#_Toc79609241)

[Equation 8. Likelihood 25](#_Toc79609242)

[Equation 9. Composition of Likelihood 25](#_Toc79609243)

[Equation 10. Argmax of Likelihood 25](#_Toc79609244)

[Equation 11. Calculating Argmax 25](#_Toc79609245)

[Equation 12. MLE 26](#_Toc79609246)

[Equation 13. Calculating Prior Probability 26](#_Toc79609247)

[Equation 13. Ng Smoothing 26](#_Toc79609248)

**List of Figures:**

[Figure 1. Bühler Organ-Modell 11](#_Toc79608623)

[Figure 2. Medium and Concept 13](#_Toc79608624)

[Figure 3. Spoken and Written vs. Graphic and Phonic 14](#_Toc79608625)

[Figure 4. Nähesprache and Distanzsprache 14](#_Toc79608626)

[Figure 5. French Registers 16](#_Toc79608627)

[Figure 6. Registers According to Literacy and Orality 28](#_Toc79608628)

[Figure 7. Literacy and Orality 28](#_Toc79608629)

**List of Tables:**

[Table 1. Mini corpus 25](#_Toc79609256)

[Table 2. Classification Values 26](#_Toc79609257)

[Table 3. Classification Assignment 26](#_Toc79609258)

[Table 4. MLE Values 26](#_Toc79609259)

[Table 5. Evaluation of Training Classification Criteria for Literacy 29](#_Toc79609260)

[Table 6. Training Classification Criteria for Literacy 29](#_Toc79609261)

[Table 7. Evaluation of Classification of Orality 30](#_Toc79609262)

[Table 8. Classification for Orality 30](#_Toc79609263)

[Table 9. Naïve Bayes Evaluation 31](#_Toc79609264)

[Table 10. Development Results of the Classification Data 33](#_Toc79609265)

[Table 11. Top Development Classification Criteria for Wikiconflits 34](#_Toc79609266)

[Table 12. Top Development Classification Criteria for SMS 34](#_Toc79609267)

[Table 13. Naïve Bayes Development Results 34](#_Toc79609268)

[Table 14. Training Results of the Classification Data 35](#_Toc79609269)

[Table 15. Top Training Classification Criteria for Wikiconflits 35](#_Toc79609270)

[Table 16. Top Training Classification Criteria for SMS 35](#_Toc79609271)

[Table 17. Naïve Bayes Training Results 35](#_Toc79609272)

[Table 18. Naïve Bayes Testing Results 36](#_Toc79609273)

**List of Abbreviations:**

|  |  |
| --- | --- |
| CMRW | CMR-wikiconflits |
| CoMeRe | Corpora of Computer-Mediated  Communication in French |
| EPA | eBay petites annonces |
| FA | français argotique |
| FCO | Français courant |
| FCU | Français cultivé |
| FRÉ | Français écrit |
| FF | Français familier |
| FPA | Français parlé |
| FP | Français populaire |
| FV | Français vulgaire |
| LP | Langue parlé |
| LT | Langues techniques |
| MLE | Maximum Likelihood Expectation |
| NZ | Niveau zéro |
| NLP | Natural Language Processing |
| OOV | Out-of-Vocabulary |
| POS-Tagging | Part of Speech Tagging |

# Introduction

Excluding other modes by which human communication can be realized such as via sign language, body language, whistling, etc., human languages are generally expressed medially through either text or speech (Bader, 2002). Oral, i.e., spoken discourse, can be understood as a process which employs audible sounds to express meaning, whereas literate, i.e., written discourse, is the visual medium that uses written symbols (Bader, 2002). An aspect that is often overlooked is conceptual discourse. In other words, what is the actual conceptual intent that a speaker wishes to communicate with their message? Does the speaker intend for their message to be more in line with spoken or written language?

With these distinctions in mind, the concepts of written vs. spoken and literacy vs. orality arise. The former represents the medial aspect of language, whereas the latter represents the conceptual intent of a speaker. 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 (Koch & Oesterreicher, 1985).

Determining the conceptual intent of a speaker’s discourse is to be done using a multinomial naïve Bayes algorithm (Jurafsky & Martin, 2020). A simple, but effective smoothing algorithm as proposed by Ng (1997) will be used to solve the out-of-vocabulary problem. Multinomial naïve Bayes, which will be referred henceforth as naïve Bayes, requires training data for it to be able to properly determine the conceptual discourse type of a given document. Therefore, a classification system must first be developed that can automatically label documents according to their conceptual discourse type. This serves as the basis from which the naïve Bayes algorithm draws its training data and computes the discourse classification probabilities.

To explore the conceptual discourse in a more practical sense, French language data was chosen from three main internet domains: eBay, SMS and Wikiconflits. SMS chats were chosen as they are the most likely candidate for representing orality (Bader, 2002; Rehm 2002). These are then to contrast with the Wikiconflits chats, as the content therein pertains to scientific and intellectual communication, and therefore represents literacy (Koch & Oesterreicher, 1985). eBay postings are to be seen here as a control as they do not intrinsically represent one conceptual discourse style over another.

# Related Works

## Theoretical Linguistics

Koch and Oesterreicher (1985) constructed the medial-conceptual paradigm of written vs. spoken and literacy vs. orality by providing a situational context in which these two facets of language can occur. Koch and Oesterreicher (1985) also place focus on sociolinguistic contexts regarding this paradigm by expounding upon the notions of *Distanzsprache* and *Nähesprache*, which are additional factors crucial to identifying the correct discourse type. Koch and Oesterreicher (2007) offered a more detailed explanation regarding the medial and conceptual discourse types by expanding their examples and explanations to include German and English.

Even though Müller (1975) predates Koch and Oesterreicher (1985), the notion of literacy and orality was already known to Müller (1975) who refers to them as *français parlé, message oral, languée* and *français écrit, message écrit, langue écrite, langage écrit* respectively. Müller (1975) explores this distinction and how it is realized chronologically, quantitatively, qualitatively, diatopically, and diastratically within the French language.

## Computational Linguistics

Ortmann and Dipper (2019) explore the ideas as proposed by various authors (Bader, 2002; Koch & Oesterreicher, 1985; Rehm, 2002) to be able to automatically identify literate and oral discourse in modern German texts. Ortmann and Dipper (2020) applied the same methodology to assess the literacy and orality regarding historical texts. Ortmann and Dipper (2020) did this by using a slightly altered feature set that is more fitting for historical texts, as the non-standardized nature of historical documents cannot be properly analyzed using modern criteria (Ortmann & Dipper, 2020).

Bader (2002) provides a rounded, general approach to properly assess literacy and orality in texts in the same vein as Müller (1975). However, Bader (2002) applies the analysis to digital communication, e.g., e-mail, chat, newsgroups, forums, while also providing features to identify the precise nature of individual excerpts from said communication. Rehm (2002) offers a more restricted analysis by only detailing the nature, characteristics and features of conceptual orality in written language on the internet e.g., e-mail, chat data, websites, etc. at the time of publication.

# Language as a Construct

## General Features of Language

Language, as a mode of communication, is something of which humans have been capable for around 100,000 years (Stein, 2014). Human language is first and foremost, the production of audible sounds, i.e., speech (Bader, 2002). Furthermore, language is 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 (Stein, 2014). 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 (Stein, 2014). This leads into the distinction of *langue* versus *parole*, langue being the virtual construct of a given language that could be realized by a speaker and parole being the actual realization of langue (Stein, 2014).

|  |
| --- |
|  |
| Bühler Organon-Modell (Stein, 2014, p. 1) |

Independent of the medial and conceptual of the language is how exactly communication can work between speakers. The organon model, as seen in figure 1, models the way in which linguistic information is received and processed. Every communication process consists of the following: *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., *sprachliches zeichen* (Stein, 2014). The sprachliches zeichen is what is transmitted via language. It has three main functions: *Ausdruck*, *darstellung*, and *appel*. The ausdruck expresses the opinions and feelings of the speaker, which are the symptoms of the sprachliches zeichen (Stein, 2014). The darstellung is the symbol for the information while the appel elicits a desired response from the listener that is in line with the sprachliches Zeichen (Stein, 2014). All three are present in every message, but generally one will dominate over the others (Bader, 2002).

## Medial Features

Spoken language in the can be understood as the phonetic expression of thought (Bader, 2002). This is in line with structural linguists, saw spoken language superseding and being the precursor of written language (Stein, 2014). Due to the nature of spoken language being the primary factor chronologically speaking (Bader, 2002; Koch and Oesterreicher, 1985), it is the medium that is the most prominent and the one that has been object of great discussion, especially since the 20th century (Stein, 2014).

Spoken language is a spontaneous act that is directly coupled with the transience in that an oral statement is gone the moment it is expressed (Bader, 2002). This real-time process prevents spoken language from becoming overly complex as it would overload the listener’s ability to ascertain the meaning from the message (Ortmann & Dipper, 2019).

The speaker’s ability to be able to process the linguistic information in real-time also has a direct impact on syntax. Meaning that the active voice and elliptical structures are preferable in spoken language as they are most likely easier to process (Ortmann & Dipper, 2019). This is evident in the lexical as “spoken language is characterized by frequent use of various particles, e.g., answer and modal particles in German … and interjections” (Ortmann & Dipper, 2019, p. 4).

If spoken language is the phonetic expression of thought, written language is then to be seen as graphical depiction and recording of said thought (Bader, 2002). The reason as to why written spoken language exists at all is explained by the fact that it is essential in translating thoughts and transporting messages over long temporal and physical distances.

Written language has often been viewed as the true state of language as it allowed one to circumvent the transient nature of spontaneous speech. The prevailing assumption well into the 19th century was that language was synonymous with the written medium (Koch & Oesterreicher, 1985). This is the reason why text has often been the necessary default when examining human language.

Therefore, written language often contrasts with spoken language due the dichotomous nature of the language paradigm. Where spoken language is restricted to being less complex, written language can benefit from static properties of a textual medium (Ortmann & Dipper, 2019). This naturally carries over into the syntactical and lexical structure of any given written message. Syntactical and lexical properties can be expounded upon in general without having to take the speaker’s ability to process information into consideration (Ortmann & Dipper, 2019).

An important property is that “written language can express features of orality with specific graphical means, such as omission of characters, word contractions, or use of ellipsis dots, em dashes or apostrophes” (Ortmann & Dipper, 2019, p. 67). This can be exploited to identify markers that are proto-typical of spoken language (Bader, 2002; Ortmann & Dipper, 2019 ; Ortmann & Dipper, 2020).

## Conceptual Features

Koch and Oesterreicher (1985) have created a 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] |
|  | Medium and Concept | | |
|  | (Koch & Oesterreicher, 1985, p. 17) | | |

Although it would be possible to see a dichotomy being present between literacy and orality, this is not strictly correct. The dichotomy does exist, but it only applies to the dual nature of the discourse. 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?

|  |
| --- |
|  |
| Spoken and Written vs. Graphic and Phonic |
| (Koch & Oesterreicher, 1985, p. 18) |

It would be false to assume that spoken language can only represent spoken language and written language can only represent written language.

|  |
| --- |
|  |
| Nähesprache and Distanzsprache (Koch & Oesterreicher, 1985, p. 23) |

Koch and Oesterreicher (1985) see spoken and written as being on a continuum with conceptual possibilities that have different levels. On the phonic portion of figure 2, all the texts are medially spoken, but conceptually they start off being of informal and personal in nature and gradually become more formal and impersonal. In doing so, the language becomes more in line with the written medium. When observing the two poles, a and i, there is an obvious difference between an informal conversation and a presentation. The former represents spontaneous speech, while the latter is something that is refabricated and then presented to an audience in an oral form. On the graphic portion of the diagram, all documents represent 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. 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 are 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 to, communication that is detached, objective, unfamiliar. Using the parameters: medium, concept and distance-proximity, a more detailed analysis of language is possible

Referring to figure 4, an informal conversation is thus representative of spoken language, that is also conceptually representative of orality. The dynamic of the speakers is one familiarity and closeness, and the speech can be assigned the label of Distanzsprache. The opposite can be said of administrative regulation texts. There is great distance between the speakers, both in terms of familiarity and proximity. It is also not a message that can be communicated conceptually orally due to the very nature of the text. It can be assigned as being conceptually and medially written speech, while also belonging to Distanzsprache.

# Diaphasic and Diastratic Registers

Sociolinguistics is the scientific study of the relationship between language and society. It deals with the linguistic phenomena that occur within society (Bieswanger & Becker, 2008; Stein, 2014). A speaker’s linguistic choices often give information about their social and geographical background (Bieswanger & Becker, 2008). Registers, or styles, can be loosely defined as:

the function of language in a particular situation and the consideration of such factors as addressee, topic, location and the interactional goal rather than background of the speaker. The exact definition of style and register is difficult …. A common distinction is that style refers to the level of formality of an utterance or a text, whereas register refers to the choice of vocabulary in an utterance or a text (Bieswanger & Becker, 2008, p. 187).

Alongside style and register exist a whole host of other phenomena that are accounted for in sociolinguistics, such as: qualitive registers, quantitative registers, sociolects, diatopic view, diastratic view, gender, age, norms, etc. (Achim, 2014; Bieswanger & Becker, 2008; Müller, 1975).

These are instrumental in determining literacy and orality if there are textual identifiers for them. Certain registers, styles, etc. are usually only realized in a specific given situation. Therefore, if medium and concept do not align, it can be better identified in text.

## Le Français

|  |
| --- |
|  |
| French Registers (Müller, 1975, p. 184) |

French was historically seen as having a single register. This is not in the sense that it there was no variation, but rather, that there was one and only one correct way of using the French language, often referred to as *Bon usage* (Müller, 1975). *Mauvais usage* and *Dites …ne dites pas* dictated the correct usage of French for most of French language history.

This was in part due to the academic body, Académie Française, who sets norms for French (Müller, 1975). Nevertheless, it is not necessarily feasible to entirely dictate what speakers of any given language do or say as this is directly antithetically to a defining character of language construct which is that languages are in a constant state of change (Müller, 1975; Stein 2008).

French registers are classified as *français cultivé, français familier, français populaire, français vulgaire*, *français argotique* and *français technique.* Français cultivéis the mostformal and français vulgaire being the least formal (Müller, 1975; Stein 2014).As seen in figure 5, many of these registers have different referents, but denote the same speech patterns.

## Français Cultivé

Français cultivé, or FC, is often viewed in positive light and seen as the register that one should try to replicate (Müller, 1975), seeing as how this register is considered the highest register. It should not be used in banal or informal situations, otherwise the speaker risks being seen as pedantic and pretentious (Müller, 1975). It is used in official situations, special ceremonies or other special occasions (Müller, 1975).

The most prominent feature of this register is the phonological component. It tends to consequently conserve sounds that are no longer used in the other registers. This includes, but is not limited to, phonetic opposition of certain sounds, the pronunciation of the schwa at the end of phonological words and more rigid syllable structure. This has to do with the desire to retain the literary tradition, which is often dependent on such archaisms (Müller, 1975).

It is often viewed langage écrit retains certain grammatical features that have not been used in other contemporary registers for quite some time. Certain verb tenses such *passé simple, passé antérieur*, *subjonctif imparfait* or verbal constructs such as *inversion* are characteristic of this register. The strict adherence to proper negation e.g., *ne…pas*, *ne...point*, *ne…guère* often appears with these verbal constructions (Müller, 1975).

It is at its core medially and conceptually a textual register. Whether spoken or written it is artificial in the sense that it ss a controlled process heavily reliant on proper word choice, intonation and lengthy, detailed sentences (Müller, 1975).

## Français Familier

Francais familier, or FF, is a qualitative register that is often used in informal situations such as with family, job, daily routine, acquaintances and people from one’s inner social circle (Müller, 1975). It is a register that is indifferent to the social standing of the speaker. Nevertheless, it is used more frequently by those who have profited from a higher education than those who have not (Müller, 1975).

It is spontaneous in nature, and this is reflected in the fact that there is not a lot of emphasis placed on proper enunciation. This spontaneity is most likely since FF, and français populaire by extension, are directly descended from Vulgar Latin, which itself was primarily a spoken register of Latin, both medially as well as conceptually (Müller, 1975).

Statements and questions are generally formed through falling and rising intonation, respectively, even though questions using *est-ce que* are possible (Müller, 1975). The doubling of pronouns or referents e.g., moi je, ton père il, etc. is characteristic of FF and high use of topicalization e.g., Cet homme, je l'ai vu très souvent (Müller, 1975).

It makes use of a high level of suffixes to denote agents and actors in speech context e.g., chançard, gueulard, motard. This also includes the diminutive suffixes such as -et, ette, ot, etc. Reduplication is not only present among pronouns, but in nouns as well e.g., fla-fla, ronron, kif-kif, etc. (Müller, 1975).

Due to its spontaneous nature, speakers tend to avoid overly complex expressions when communicating strong feelings. This leads to a high number of simplified expressions, and using adverbs atypically as intensifiers (Müller, 1975). The register is often consigned to orality as it signalizes a nonchalant attitude and, as the name implies, a familiar atmosphere.

## Français Populaire

FP is considered neither proper nor good French it does not meet the requirements set by the norms or bon usage (Müller, 1975). Since it differs quite drastically from FC, it is often considered to be a language within a language congruence (Müller, 1975). This is because it is not consistent with FC, but rather within itself and presents grammar and orthography that while deviant, are internally consistent. It, along with FF, arose as a language of the people, meaning those who belonged to neither clergy nor nobility whose speech was more commonly referred to as *lanuge du peuple* (Müller, 1975).

Since communication is more important than grammatical correctness for speakers of this register , FP tends to forgo the linguistic norms. Verbal phrases are often formed without their corresponding personal pronouns (Müller, 1975). The appropriate auxiliary verbs, avoir and être, are used interchangeably. Nominal congruence with respect to gender and number are either ignored or forgotten all together. The subjunctif is only employed when a strong desire s expressed as would be the case with vouloir. Relative pronouns and conjunctions involving que tend to have a higher frequency for variability (Müller, 1975).

There is strong preference of neglecting the spelling, especially when the message is clear due to morphology. The most prominent example of this is the willingness to drop the ne of ne…pas. This is more noticeably in the phonetic realization as instable sounds such as /l/ and /e/(Müller, 1975).

The lexicon does not differ in form from FC, but rather in usage. That is to say that speakers use the same words, but differently. This leads to expressions being hyperbolic and suggestive (Müller, 1975). A great deal of the words that occur within FP are known to most speakers of French; they only make up a small portion of the language. Most of the words that appear in FP are from the 19th and 20th century, which mainly stem from dialects and FV (Müller, 1975). FP would therefore be representative of orality.

## Français Vulgaire

Français vulgaire, or FV, is the lowest register both in terms of prestige and formality, and therefore conceptually oral in nature, is often grouped together with français argotique (Müller, 1975). The difference is that it and its components are generally known to all speakers of French, whereas FA is restricted to certain milieus (Müller, 1975). Interjections, expressions of displeasure and expletives are present throughout FV. It is avoided whenever possible as it is in direct opposition to social norms regarding etiquette. It is notable for its lack of scientific jargon, Latin loanwords, and euphemisms, but it is also incredibly adept at coining new words that employ the method of directness (Müller, 1975).

## Francais Argotique

Argot in its original form was meant to specify the speech patterns of marginal groups and that of professional jargon. A defining feature of argot is that the speaker is intentionally trying to distance themselves socially. At the same time, it is used as a way of identifying insiders and outsiders (Müller, 1975). This is usually the reason why argot is considered to be cryptic language (Stein, 2014).

Argot employs metonym to a high degree by applying descriptions of food to refer to the body. It also displays a high willingness to import loan words from dialects as well as other languages (Goudailler, 2002). The high number of synonyms and polysems are as a byproduct of argot’s instability (Müller, 1975). Argot is highly representative of orality as the need to record speech in a written form was completely secondary. Due to the written aspect of language not being important, argot is relatively unstable (Müller, 1975). The extreme degree to which argot changes is also a defining feature (Goudailler, 2002). This is most likely since it reflects the period in which the speakers live and not the continuing of a linguistic tradition (Stein, 2014).

## Français Technique

Français technique can be used to explain theoretical concepts to those who are from the same field, or a reduction in complexity is introduced i.e., vulgarization (Müller, 1975). This makes it more readily available to those who are not from a specific scientific field.

A defining trait of it is the need to develop new terminology as the field of science is ever growing. This is done using complex use of morphological constructions. The high influx of new words also come from English, which is a point of contention with those working with français technique, but often French words are substituted to combat this (Stein, 2014).

The syntax and vocabulary are quite rigid, more so than that of FC, since precision in scientific fields is key. The syntactical structures are not per se complex. It also displays a high level of words that express causality which is to be expected as the goal is scientific in nature and conceptually literal.

# The French Language Corpora

French is not a monolith, but a language that is spoken across many domains, age groups, countries (Stein, 2014; Müller 1975). Whether a native speaker of metropolitan French, second-language speaker or speaker of given French dialect, this variation is present in France and outside as well (Stein, 2014). This poses a challenge of sorts since what is of the literal or oral discourse is to some extent dependent on the local and personal understanding of the language. Due to this, some concessions and compromises must be made for the subsequent chapters to be sound.

First and foremost, the object language here in question is that of Metropolitan French, which is French as it is spoken in France. The methods and reasoning will therefore apply to this variant of French. However, there is no feasible way to know if a speaker is completely in line with this standard. Seeing as how the internet is an open platform, and not bound to geographical constraints, it is plausible that speakers of other varieties have partaken in the conversations.

Finally, The data records stem from ca. 2000-2020 and can therefore only accurately encapsulate and illustrate the language at this stage. Of course, the age of each speaker would allow for an analysis that could potentially stretch further back into the past. However, no assumptions can or will be made about the language state before 2000 or projects about the language beyond 2020 as this would be purely conjecture.

## Data Sets

There are three primary data sets that will be the focus of the linguistic analysis: eBay petites annonces, or EPA, (Gerstenberg & Hewett, 2019), CMR-wikiconflits (Poudat et al., 2014) and 88milsms (Panckhurst, 2016).

The EPA corpus was compiled by the department of Romance studies at the University of Potsdam. It is a collection of around 1256 petites annonces which are split across four subcorpora. They online action listings from the online auction platform, eBay. The first three subcorpora deal with housing, vehicles, clothing, computer, telephones, children, collections and leisure, while the last corpus deals with professional activities e.g., stocks, shops, shipping, etc. (Gerstenberg & Hewett, 2019).

The first (e05p) is from 2005 and contains around 300 lists from private users. The second and third, collectively known as (e17p) are from 2017 which feature 300 listings from both private as well as professional. The final corpus is from 2018 (e18v) and has 365 listings from private users (Gerstenberg & Hewett, 2019).

Private users were those who had less than 200 reviews as of 2005, and over 200 were professional users. This process was replicated in 2017. The final corpus was gathered using a web scraping tool called ParseHub to facilitate the automation process. An upper bound of 1000 ratings and one listing per user was set to have a representative corpus (Gerstenberg & Hewett, 2019).

The next two corpora are distinct in nature but have provided and gathered by the CoMeRe Repository. The aim of CoMeRe is “to gather different corpora that represent the forms of communication in French on different networks (Internet, phone, etc.), all structured and informed in the same way, diffused in open access formats for research purposes.” (Poudat et al., 2015)

The first of the two, CMRW spans from 2004 to 2014 and contains discussions about the wikipedia article “Quotient intellectuel”. It contains around 52 participants, 170 contributions and 20 000 tokens. As is often the case with sites like Wikipedia, the information presented may not be factually correct (Poudat et al., 2015). This does not necessarily pose a problem as the accuracy of the information is irrelevant with respect to its literacy and orality

The second corpus is 88milsms which is a collection of more than 88,000 SMS messages that were collected from speakers in the Montpellier area in France. To comply with French data protection guidelines, the data has already been anonymized by Panckhurst et. al (2014). The SMS donors were asked to participate in a questionnaire, about the languages they speak, their telephone number, their profession, how they communicate through SMS, the frequency of their communication and what their opinions of SMS communication are (Panckhurst, 2016).

This French corpus was created as part of a greater project from sud4science, which sought to create many such corpora for various languages, such as German, English, Swiss German, etc. (Panckhurst et al., 2014).

The selection of the corpora was done in such a way as to provide three instances in which literacy and orality could appear in a data set. First and foremost, the SMS chats are generally forms of informal communication and because of this, they should contain data that is mostly representative of orality. Secondly, the wikiconflits documents contain discussions that generally relate to scientific and official matters. Therefore, it should fall more on the literal scale. Lastly, it predicted that the eBay texts should fall somewhere in between them.

## Pre-processing

All the corpora used in this research were created with the goal of computational linguistic analysis in mind (Gerstenberg & Hewett, 2019; Panckhurst et al., 2014; Poudat et al., 2015). Therefore, the data has been annotated and changed as little as possible by the respective institutions. This means that processes such as sentiment analysis, POS-Tagging, tokenization, etc. are possible without interference from foreign analysis. All the data sets are available in the .xml format, and do contains markers to identify author, date, time, title of the post.

The eBay corpus has been tagged with respect to typical features of ad postings such as abbreviations, misspellings, marketing language, slang, proper nouns, emoticons, etc. Before the individual entries could be properly processed, the corpus had to be first sub-divided.The other two data sets, Wikiconflits and SMS, were already in one homogenous corpus and sub-division was therefore not necessary. However, all three of the data sets were then equally divided into three parts: development, training and test data sets.

Since files were in an .xml format, it was not possible to directly access the text directly, but rather through their respective tags. This was done by parsing them .xml tags using the module *beautifulsoup*. A python function developed for accessing the tags of the eBay corpus and another function was developed for accessing the information of the other two corpora. Once the textual data was exposed, the respective entries were tokenized into their respective sentences using a custom tokenizer that uses regular expressions to identify the end of the sentences. Subsequently, information related to parts of speech, morphological and syntactical dependencies as well as tokens were ascertained from the sentences by using *Spacy*.

# Methodology

## Naïve Bayes

An efficient and well-known method of classifying a document is done using a group of classifiers known as naïve Bayes classifiers with multinomial and Bernoulli naïve Bayes classifiers being among the most common (Jurafsky & Martin, 2020). The main difference between the two is that Bernoulli naïve Bayes models the presence or absence of features, whereas multinomial naïve Bayes counts the number of times a given feature occur.

They work well with binary classification and are most often employed in sentiment analysis, spam detection, authenticating authorship (Jurafsky & Martin, 2020). The following explanation applies to the multinomial Bayes. The naïve Bayes’ algorithm is a conditional probabilistic algorithm that is first and foremost based on the Bayes’ theorem which is as seen in equation 1

|  |
| --- |
|  |
| Equation 1. Bayes’ Theorem |
| (Carstensen et al., 2010, p. 122) |

P represents the probability of an event with A and B representing two distinct events. P(A|B) is the probability of A given B (Carstensen et al., 2010). Since Bayes’ theorem is flexible, the events can be swapped as seen in equation 2.

|  |
| --- |
|  |
| Equation 2. Bayes’ Theorem Reversed |
| (Manning & Schütze, 1999) |

P(A) being the normalizing constant guarantees that the equation has a probabilistic aspect to it. P(A) is the combined probability of all events and is calculated as seen in equation 3.

|  |
| --- |
|  |
| *Equation 3. Normalizing Constant* |
| (Manning & Schütze, 1999, p. 43) |

Bayes’ theorem can then easily be converted into a classifier and takes on the form as seen in equation 4.

|  |
| --- |
|  |
| *Equation 4. Naïve Bayes’ Classifier* |
| (Jurafsky & Martin, 2020, p. 57) |

, the estimation of the correct class, represents the maximum posterior probability with d being the documents out of all classes *c ∈ C*. However, as is often the case with NLP tasks, only the maximum argument is relevant. The maximum argument consists of the product of the likelihood and prior probability must first be computed as presented in equation 5.

|  |
| --- |
|  |
| *Equation 5. Argmax* |
| (Manning & Schütze, 1999, p. 43) |

This also applies to the naïve Bayes’ classifier producing a simple, but effective model like that present in equation 6.

|  |
| --- |
|  |
| *Equation 6. Argmax of Classification* |
| (Jurafsky & Martin, 2020, p. 57) |

To determine the most fitting class, the two probabilities must first be computed as in equation 7.

|  |
| --- |
|  |
| *Equation 7. Model Probabilities* |
| (Jurafsky & Martin, 2020, p. 58) |

P(c) is the prior probability of a given class. The likelihood, as in equation 8, is assumed to be in line with the bag-of-words principle, which states that the position of the words is irrelevant.

|  |
| --- |
|  |
| *Equation 8. Likelihood* |
| (Jurafsky & Martin, 2020, p. 58) |

Thus, the naïve Bayes assumes that occurrence of the features, but not their position as laid out in equation 9:

|  |
| --- |
|  |
| *Equation 9. Composition of Likelihood* |
| (Jurafsky & Martin, 2020, p. 58) |

By applying equation 9, the production in equation 10 results.

|  |
| --- |
|  |
| *Equation 10. Argmax of Likelihood* |
| (Jurafsky & Martin, 2020, p. 58) |

To apply this formula, it is only necessary to traverse all words in each document as detailed in equation 11

|  |
| --- |
|  |
| *Equation 11. Calculating Argmax* |
| (Jurafsky & Martin, 2020, p. 58) |

To apply the formula, it is first necessary to train the model by calculating the probabilities of P(c) and P(|c). This is done by using the frequencies in the data to ascertain the MLE as presented in equation 12.

|  |
| --- |
|  |
| *Equation 12. MLE* |
| (Jurafsky & Martin, 2020, p. 59) |

This states that for a given number of documents, how many times does a given class occur within this document. Finally, to compute P(|c) as P(|c), the frequency of a given word occurring within a given class is calculated, then divided by the sum of how often words within a given class occur as presented in equation 13.

|  |
| --- |
|  |
| *Equation 13. Calculating Prior Probability* |
| (Jurafsky & Martin, 2020, p. 59) |

The problem here comes when a given word does not occur within a certain class, this means that the effective frequency is zero. To remedy this problem, a smoothing algorithm must be applied. There are many methods to choose from such as La-Place, Good Turing, held-out, etc. (Jurafsky & Martin, 2020). However, the one used in this paper is based on that of Ng (1997):

|  |
| --- |
|  |
| *Equation 13. Ng Smoothing* |

With all other parameters being equal, *N* here represents the amount of training data from a given corpus, the amount of which must be squared.

## A Worked Example

|  |  |  |
| --- | --- | --- |
|  | Feature | Document |
| Training |  |  |
|  | ORAL | Vous dites quoi ? |
|  | ORAL | Faut partir parce qu’ il pleut . |
|  | ORAL | Je n’ sais pas . |
|  | LIT | Il faut partir, car il pleut . |
|  | LIT | Elle m' a dit que j' étais une imbécile . |
| Test | ? | Vous dites imbécile |
| Mini corpus Examples taken from Müller (1975, p.185) | | |

|  |  |  |
| --- | --- | --- |
| Tokens | LIT ORAL | |
| . | 0.66 | 1.0 |
| , | 0.12 | 0.5 |
| ? | 0.33 | 0.08 |
| Elle | 0.12 | 0.5 |
| Faut | 0.33 | 0.08 |
| Il | 0.12 | 0.5 |
| Je | 0.33 | 0.08 |
| Vous | 0.33 | 0.08 |
| a | 0.12 | 0.5 |
| car | 0.12 | 0.5 |
| dit | 0.12 | 0.5 |
| dites | 0.33 | 0.08 |
| faut | 0.12 | 0.5 |
| il | 0.33 | 0.5 |
| imbécilee | 0.12 | 0.5 |
| j' | 0.12 | 0.5 |
| m' | 0.12 | 0.5 |
| n’ | 0.33 | 0.08 |
| parce | 0.33 | 0.08 |
| partir | 0.33 | 0.5 |
| pas | 0.33 | 0.08 |
| pleut | 0.33 | 0.5 |
| que | 0.12 | 0.5 |
| quoi | 0.33 | 0.08 |
| qu’ | 0.33 | 0.08 |
| sais | 0.33 | 0.08 |
| une | 0.12 | 0.5 |
| étais | 0.12 | 0.5 |
| MLE Values | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | ORAL | LIT | |  |  |  | | Prior Probability | .60 | .40 | | Smoothing | .12 | .08 |  Classification Values |
| |  |  |  | | --- | --- | --- | |  | ORAL | LIT | | Vous | 0.3 | 0.08 | | Dites | 0.3 | 0.08 | | Imbéciile | .12 | 0.05 | | Prior Probabiltiy | 0.60 | .40 | | Product | 0.00798 | 0.00128 | |
| Classification Assignment |

It is assumed in the following corpus , as seen in table 1, that the sentences have the following features as listed in the feature column. Using these sentences as a training corpus, it is possible to ascertain the most probabilistic classification of a sentence the *vous dites imbécile.* The prior probabilities and smoothing probabilities for the respective classes must first be calculated and the results of which are present in table 2. Combined with the values in table 2, the MLE values can be calculated as seen in table 3.

The final step is simply to traverse the sentence, retrieve the respective values from table 3 and multiply the respective products by their respective prior probabilities. The result, as seen in table 4, shows that the sentence is most likely ORAL based on the corpus as presented above.

## Combining Registers and Discourse

Literacy and orality represent the binary feature set that is to be assessed by the naïve Bayes. As the medium is apparent from the textual nature of the data set, it is assumed then that when the textual and medial discourse overlap, they represent literacy. If they are to diverge, then they represent orality. Therefore, it is possible to to group the registers in a manner akin to the diagram as presented by Koch and Oesterreicher (1985):

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | FC | TL |  |
|  |  |  |  |  | Graphic |  |  |  |
| ,Spoken’ |  |  |  |  |  |  |  | ,Written’ |
|  |  |  |  |  | Phonic |  |  |  |
|  | FA | FV | FP | FF |  |  |  |  |

##### Registers According to Literacy and Orality

|  |  |  |
| --- | --- | --- |
|  |  |  |
| FC | Literacy | FA |
| TL | Orality | FV |
|  |  | FP |
|  |  | FF |
| Literacy and Orality | | |

By grouping the registers in this manner, it is easier to ascertain where FRP and FRÉ overlap medially and conceptually. This graph can be further refined to allow them to be mapped to the conceptual as seen in figure 7. Registers by their very nature represent different discourse types and situations. In this case, registers can be assigned to FPA or FRÉ, which do line up with orality and literacy.

There is a lot of variation and overlap between the respective registers. So, it would not be reasonable or feasible to train a model to recognize the individual registers. However, by extracting characteristics and criteria from each class and grouping them according to their discourse type, it was possible to fit a model with criteria that allowed it automatically recognize literacy and orality and thus proving that the principal discourse types apply to French as well.

# System Evaluation

## Developmental Overhead

As was the case with the corpora used in this project, most linguistic data is typically stored in an .xml format. The training files created by the program were saved as .csv files. Finally, the program had to also be able to accept .txt files as well as strings as these would be the most common way of training and inputting data into the system. The program is dynamic and allows for user input which required the implementation of error correction and prevention.

The optimization of the program was done in two main steps: development, and training, with testing being done in the last phase. The training of the program varies depending on the amount of data being input into the system and the system resources.

The classification criteria used to create training data could theoretically be retrained to recognize any language supported by Spacy and that uses the Latin script. As for applying the algorithm to a domain other than literacy and orality, this would also heavily depend on the training data being supplied to the naïve Bayes.

Naïve Bayes is in of itself a relatively flexible algorithm that can be applied to a whole host of classification tasks. Therefore, if the program were supplied with slightly different parameters and training data, it could be restructured to recognize data with other binary classifications in mind e.g., positive vs. negative, spam vs. not spam, detection between two languages, positive vs impolite, etc. The limitation does not lie necessarily within the program, but rather within the training data made available to the naïve Bayes.

## Classification Sets and Naïve Bayes

Various researchers (Bader, 2002; Ortmann & Dipper, 2019; Rehm, 2002) have provided a plethora of criteria by which one can automatically identify literacy and orality in discourse. These criteria focus on creating a system which is to be linguistically and chronologically independent. However, since French data is being classified, characteristics of the French registers were taken into consideration when developing the classification criteria.

The original scoring system was to assign one point if a criterion in any given classification was met. However, this proved to be ineffective, as it treated all criteria equally. This often caused the sentences to be either assigned to the wrong category or all of them to be assigned to only one category. The solution to this entailed weighting the criteria according to the importance and prevalence of the data set as in table 5.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Description | | Point Amount | |
| SEN\_LEN | | Sentence Length | | The length of the sentence in character length |
| AVG\_WORD\_LEN | | Average word length | | The length of the average word length |
| THIRD\_PERSON\_EXPL | | Dummy Subjects | | The number of dummy subjects |
| NOM\_SUBJ | | Sentence Length | | How often nominal subjects occur |
| PRES\_TENSE | | Present tense verbs | | The number of present tense verbs |
| ABBR\_NO\_VOWEL | | Abbreviations without vowels | | Count of abbreviations without vowels |
| NP\_VB\_RATIO | | Noun to verb ration | | Noun count plus verb count |
| LOW\_VERB\_HIGH\_ADJ | | Low number of numbs, but high number of adjectives | | Verb and adjective count |
| CCONJ\_VB\_RATIO | | More coordinating conjunctions than verbs | | Coordinating conjuct plus verb count |
| SHORT\_SEN\_LENGTH\_PRESENCE\_OF\_NUMBERS | | Short sentences that consist of only numbers | | Only one point |

|  |  |
| --- | --- |
|  | Values  (in Percent) |
| Accuracy | 94 |
| Error Rate | 28 |
| Precision | 91 |
| Recall | 69 |
| F-Score | 79 |
| Evaluation of Training Classification Criteria for Literacy | |

###### Training Classification Criteria for Literacy

The first classification criteria considered features that were prevalent throughout texts which often expressed a high degree of literacy. These were weighted according to their prevalence and importance. Using these criteria, training data was created and then evaluated which produced the following results in table 6. A second classification, table 7, was created that mirrored the first classification set to a certain extent but considered factors that often occurred in French texts expressing orality.

|  |  |  |
| --- | --- | --- |
| Variable | Description | Point Amount |
| SEN\_LEN | Sentence Length | The length of the sentence in character length |
| AVG\_WORD\_LEN | Average word length | The length of the average word length |
| VERB\_SEN\_LEN\_RATIO | Short sentences without verbs, high number of pronouns | The number of verbs and pronouns that occur within the sentences |
| WORD\_REDUPLICATION | Occurrence of a word more than once in a text. | The number of words that occur more than once |
| PRES\_TENSE | Present tense verbs | The number of present tense verbs |
| HIGH\_PUNCTION | High use of punctuation | The number of punctation symbols |
| MULTI\_CHAR\_REDUPLICATION | Using the same character multiple times | The number of symbols that occur more than once |
| WORD\_WORD\_REDUPLICATION | Using the same word back-to-back | The number of times a word is used more than once back-to-back |
| ALL\_CAPS | All caps | Words written in all caps |
| ISOLATED\_VERBS | Only verbs in a sentence | The length of the sentence |
| EMOTIOCONS | The usage of emoticons in a sentence | The number of emoticons used in a sentence |
| ABBR | Abbreviations and acronyms | The number of abbreviations and acronyms as they occur in the text. |

|  |  |
| --- | --- |
|  | Values  (in Percent) |
| Accuracy | 91 |
| Error Rate | 29 |
| Precision | 1 |
| Recall | 68 |
| F-Score | 0,81 |
| Evaluation of Classification of Orality | |

###### Classification for Orality

These classification set was then tested and evaluated, the results of which can be seen in table 8. The sentence is analyzed according to both criteria and the highest score determines the feature of the document. Throughout all of the corpora, word length, sentence length, and reduplication of symbols played the biggest role in determining the feature of the sentence. This lines up with the sources (Bader, 2002; Koch & Oesterreicher 1985; Ortmann & Dipper 2019; Rehm 2002) that also show that expressions of literacy tend to

|  |  |
| --- | --- |
|  | Values  (in Percent) |
| Accuracy | 94 |
| Error Rate | 0,46 |
| False positive rate | 0,0045 |
| Precision | 1,0 |
| Recall | 0,88 |
| F-Score | 0,936 |
| Cross Validation | 78% |
| Naïve Bayes Evaluation | |

have longer sentences and longer words, whereas expressions of orality tend to show the opposite.

Abbreviations, acronyms, were statistically insignificant. The problem in identifying these features is that users, especially in non-standard communication, often use abbreviations and acronyms that might be non-standard as well. Thus, there is no clear way to always identify acronyms properly. After training data was gathered using the scoring system, it was given to naïve Bayes as proper training data which produced the results as seen in table 9.

The error rate in table 9 is relatively high due to the way the system was evaluated. The true negatives in this case were the ORAL tags, with LIT being the true positives. Due to the way the error rate is generally calculated, it appears as if the system has more errors than in actuality. Therefore, the false positive rate should be interpreted as the error rate in this case as both true positives and true negatives are the desired results.

## Sentence Tokenizer

Since the data is often non-standard, it was not always clear which sentences should be parsed and where they should be parsed. Data from all three domains often lacked any meaningful punctuation, punctation was used incorrectly in that there was often reduplication of certain symbols to create an emphatic impression. This resulted in sentences that were sometimes too long or too short, which skewed the results. Sentences that were generally short i.e., less than a couple of words, were generally representative of orality. The reverse of that being that the longer sentences were often representative literality.

Long sentences could not be parsed without syntactically and semantically analyzing the sentence. Due to this, some sentence sentences were added together that should have been split by the author. The reverse, however, cannot necessarily be said. It was apparent from the data, such eBay online postings, that bullet points, rather than sentences were the intent of the author. Therefore, the decision was made to use this bullet points as sentence markers.

7.4 Spacy Module

The spacy module was used for tokenization, part-of-speech tagging, syntactical dependencies and assessing morphology. Tokens included punctuation and non-letter symbols as they were often essential in emoticons and expressive thought. Therefore, there was no changes were made to the data to make it easier to be processed by Spacy as the linguistic nature of the data was to remain as unaltered as possible.

Fr\_core\_news\_sm (n.d.) states that tokenization, part-of-speech tagging, syntactical dependency tagging, and morphology tagging have an accuracy of 100% , 93%, 96% and 90% respectively. These values align with the actual values pulled from a small data section of data from each development corpus set with a deviation of ∓ 5% .

The Wikiconflits and eBay data were easily processed by spacy with minimal errors. This was due in part to the authors following orthographic norms and not using non-standard language excessively. A challenge posed to Spacy was that users of the SMS chats often had incorrect spellings, made high use of emoticons or created new unknown abbreviations. However, emoticons were classified as punctuation, rather than as emoticons. This caused Spacy to perform poorly compared to the other data sets, but the values were still with an acceptable range.

# Results[[1]](#footnote-1)

## Development phase

The number of sentences and tokens were not distributed equally among all three of the original corpora with the SMS corpus being the largest and the wikiconflits being the smallest corpora. Therefore, it was ensured that the development and training would only entail a small portion of each data set to ensure that there was an even distribution of quality. The maximum number of documents extracted from corpus was kept to a minimum.

Originally, a separate classification set was meant to evaluate the first classification set. A process that was akin to a two-fold cross validation. The validity of the first classification would be weighed against the second classification set. However, this proved to be extremely ineffective since there were not enough unique words to push a sentence into one category over another. The result of this was that sentences were either wrongly classified or the number of unknown sentences was extremely high.

The second problem voids this solution as too many features were being deleted from a sentence which caused it to be unrecognizable by the other classification. The first classification relied heavily on sentence, word length, reduplication and emoticons, which are crucially for determining literacy and orality. Therefore, the features that would have been present in the other system were generalized and incorporated into the second classification system. This ultimately led to having two files that were imbalanced and not reliable for evaluation.

The developmental phase of this project was therefore crucial since there were no French training data and criteria available by which it was possible to ascertain literacy and orality in datasets. Using a combination of criteria proposed by various authors (Bader, 2002; Koch & Oesterreicher, 1985; Ortmann and Dipper, 2019; Rehm 2002) it was possible to develop and refine a system by which literacy and orality could automatically be assigned to sentences.

Problems that were touched upon earlier were present throughout the eBay and SMS corpora which was that the data was non-standard, this made the classification quite difficult as there was no way to guarantee uniformity. This was compounded by the fact that French was not exclusively used in all the data sets. In the SMS and eBay corpus, there were traces of German and English since postings and conversation were most likely on a national, and not always a local scale.

Using the scoring system with Wiki and SMS as training data, data was labeled according to the classification sets mentioned in table 5 and table 7.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Corpus ID | Sentences | Tokens | Documents | LIT | ORAL |
| Wiki | wikiconflits\_0\_53 | 345 | 6766 | 53 | 234 | 110 |
| SMS | sms\_0\_29507 | 349 | 34454 | 150 | 129 | 218 |

###### Development Results of the Classification Data

While creating the data, the most relevant classification criteria were retrieved for Wikiconflits, table 11, and for SMS, table 12 respectively.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Feature | Classification Criteria | | LIT | SEN\_LEN | | LIT | AVG\_WORD\_LEN | | LIT | NP\_VB\_RATIO | | ORAL | AVG\_WORD\_LENGTH | | ORAL | MULTI\_CHAR\_REDUPLICATION | | ORAL | SEN\_LEN | | Top Development Classification Criteria for Wikiconflits | | | |  |  | | --- | --- | | Feature | Classification Criteria | | LIT | SEN\_LEN | | LIT | NP\_VB\_RATIO | | LIT | PRES\_TENSE | | ORAL | SEN\_LEN | | ORAL | ALL\_CAPS | | ORAL | AVG\_WORD\_LENGTH | | Top Development Classification Criteria for SMS | | |

Sentence length, noun-to-verb-ratio, average word length were decisive in determining the feat for the training data set for both corpora. After having acquired the training data using the scoring system, it was inputted into the naïve bayes and trained on it. Then, all four of the eBay corpora were used as testing material. The results of table 13 show that all four of the eBay sub-corpora displayed a high rate of literacy with a low rate of orality.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Corpus Id | Tokens | Sentences | Documents | LIT | ORAL |
| eBay | ebayfr-e05p\_0\_100 | | 4929 | 380 | 100 | 361 | 8 |
| eBay | ebayfr-e17p\_0\_100 | | 6195 | 317 | 100 | 312 | 3 |
| eBay | ebayfr-e17xp\_0\_100 | | 21184 | 1028 | 100 | 995 | 32 |
| eBay | ebayfr-e18v\_0\_100 | | 9321 | 563 | 100 | 551 | 9 |

###### Naïve Bayes Development Results

Even though all contained 100 documents, the number of sentences and tokens contained within varied significantly. Despite this, they were uniform in the way literacy and orality was distributed across the data.

## Training phase

After the development phase and with only slight modification to the data and criteria set, the model was then retrained using the same process on the second portion of the data without incorporating the results from the first phase. The modification included correcting errors in the code that would assign incorrect scores to the ratios.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Corpus Id | Tokens | Sentences | Documents | LIT | ORAL |
| Wiki | sms\_29508\_59014 | 8226 | 463 | 52 | 303 | 160 |
| SMS | Wikiconflits\_0\_54\_106 | 4138 | 458 | 255 | 140 | 317 |

###### Training Results of the Classification Data

These results of which mirrored those of the development phase to a certain degree and can be seen in table 14. Wiki again displays a high level of literacy while SMS displays a high level of orality.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Feature | Classification Criteria | | LIT | SEN\_LEN | | LIT | AVG\_WORD\_LEN | | LIT | NP\_VB\_RATIO | | ORAL | AVG\_WORD\_LENGTH | | ORAL | MULTI\_CHAR\_REDUPLICATION | | ORAL | SEN\_LEN | | |  |  | | --- | --- | | Feature | Classification Criteria | | LIT | NP\_VB\_RATIO | | LIT | SEN\_LEN | | LIT | NOM\_SUBJ | | ORAL | SEN\_LEN | | ORAL | ALL\_CAPS | | ORAL | AVG\_WORD\_LENGTH | |
| Top Training Classification Criteria for Wikiconflits | Top Training Classification Criteria for SMS |

As during the development phase, the top classification criteria were retrieved from and can be seen in table 15 and table 16. These results do not differ from those of the development corpus. The process from the development phase was then repeated by retraining a new database with new training data created from the classification criteria. After that, the naïve bayes was then tested again on the eBay corpus data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Corpus Id | Tokens | Sentences | Documents | LIT | ORAL |
| eBay | ebayfr-e05p\_101\_200 | 5225 | 315 | 100 | 283 | 32 |
| eBay | ebayfr-e17p\_101\_200 | 6242 | 373 | 100 | 337 | 36 |
| eBay | ebayfr-e17x\_101\_200 | 24477 | 1202 | 100 | 1112 | 89 |
| eBay | ebayfr-e18v\_0\_100 | 9784 | 542 | 100 | 503 | 39 |

###### Naïve Bayes Training Results

The results of this mirror those of the development phase as well.

## Testing phase

The testing phase of the system was implemented differently. Using the training data from the training phases, a training database was built. This was then used to train the naïve Bayes and data from each of the corpora was used.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Corpus Id | Tokens | | Sentences | Documents | LIT | ORAL |
| eBay | ebayfr-e05p\_201\_  300 | | 4063 | 249 | 100 | 229 | 20 |
| eBay | ebayfr-e17p\_201\_300 | | 4680 | 275 | 100 | 254 | 21 |
| eBay | ebayfr-e17x\_201\_300 | | 17155 | 922 | 100 | 830 | 922 |
| eBay | ebayfr-e18v\_201\_300 | | 9824 | 588 | 100 | 5151 | 43 |
| Wiki | wikiconflits\_79\_159 | | 9172 | 487 | 53 | 441 | 46 |
| SMS | sms\_59015\_88522 | | 3523 | 342 | 250 | 293 | 49 |

###### Naïve Bayes Testing Results

The biggest deviation in the testing results those of the SMS data which shows a high literacy and opposed to orality. Another anomaly is the ebay 17x eBay corpus as it shows orality and literacy equally. The other values in the chart are in line with the other training phases.

# Discussion

Various authors (Bader, 2002; Koch & Oesterreicher 1985; Ortmann & Dipper, 2019; Rehm 2002) have proposed methods and ideas that are directly related to literacy and orality. The use of a scoring system was essential as it provided more control and more speed with respect to building up a necessary training data set. It might seem somewhat redundant to have a training data algorithm and a naïve Bayes in the same program. It could be rightfully said that having a naïve training algorithm would suffice as opposed to having a naïve Bayes and training algorithm in one program. However, this problem could not be avoided as no reliable or accurate training data existed for the program. Thus, a program to create training data and then train the naïve bayes was a necessity.

The scoring system was to incorporate elements that are prototypical of the respective discourse types. The idea of using such criteria was put forth by the likes Koch and Oesterreicher(1985) who supposed regarding syntax, sentence length, lexical property with respect to literacy and orality. This was partially the basis for the research as done by by Ortmann & Dipper (2019; 2020). Where they used German data as the object language in their research, French language data was used here. This theoretically did not pose any limits on the creation of a training dataset and using language independent classification criteria.

Nevertheless, there was an earnest attempt at ascertaining reliable French examples of literacy and orality by developing a separate French classification set. One of the most reliable and well-known sources of information regarding French philology comes from Müller (1975). This was initially set to be the source of much contextual French information for the training data as well as the naïve Bayes. Despite the age of this work, much of the information contained within is still relevant to the French language. Many of the descriptions about literacy and orality that appeared within were essential in refining and rechecking the algorithms, defining sentence length and even developing a scoring system purely based on French.

Müller (1975) offers readers prototypical texts of the respective French registers that can be graphed to respective discourse types. Despite all of this, it is the quantity, and not the quality of the texts, that proved to be a hindrance with respect to training a naïve Bayes to recognize literacy and orality in French discourse data. Had more information been readily available by Müller (1975) or other similar sources, then less emphasis and time would have been placed on developing a scoring system.

The scoring system relies heavily on naïve assumptions that often prove to be correct. More points were given to sentences that are longer, and less to sentences that are shorter. This often created an imbalance and drowned out the other classification criteria such as, but not limited to, adverbs, pronouns, adjectives.

It was not uncommon for sentence length to be the decisive factor in determining literacy and orality. Sentences that were long tended to represent literacy as opposed to orality upon manual inspection of the results, this turned out to be correct in most instances. This was not necessarily universally correct as sentence length was also highly dependent upon the user correctly using punctuation, which in turn was to be recognized by the sentence tokenizer. If the author of text incorrectly used punctuation, the sentence would be split prematurely and thus skewing the results.

This naïve approach poses a problem as it prevents the system from having a precise reason as to why a particular sentence is representative of orality and opposed to literacy. With that being said, the scoring system would benefit from having a more evenly distributed point and scoring system that is more fine-tuned to the French language.

It was initially hypothesized that Wiki documents would show the highest amount of literacy, and the lowest amount of orality. The SMS chats would be on the opposite end of this spectrum and display the highest amount of orality, whilst having the lowest amount of literacy. To strike a theoretical balance between the two corpora, the eBay corpus was chosen to serve as a control to be between the two corpora.

Using the eBay corpora as a control, the naïve Bayes was trained so that it could recognize sentences that it had not seen before. More documents were classified as literal than expected. This was due to the criteria seeing more of the data as being literal than oral and was thus these results transferred over to the naïve Bayes. Even taking this into account, eBay data is more along the lines of being representative of literacy and orality.

The unexpected high literacy in eBay data can be attributed to buyers and sellers using an imbalanced mixture of both. The postings had to be of a literal quality to attract buys as literacy used in such business situations (Koch & Oesterreicher, 1985). That is to say that using it lends credence to the belief that one is being more serious and professional (Koch & Oesterreicher, 1985). However, some buyers did not want to exaggerate this and offset this discourse type by presenting part of their postings. A blend of the two was thus inevitable.

The wiki data showed a high level of orality, but this was to be expected as a lot of the discussions revolved around topics that were highly scientific and intellectual in nature. If orality did occur, then it was only in short bursts or uttering small statements.

The SMS chats were of a high orality quality, this was to be expected and extracting literality from these texts proved to be the most difficult. First, the authors of the documents were very familiar with one another, and this was reflected in the language used by them. There were a high number of pronouns, nouns, proper nouns and redacted names[[2]](#footnote-2). The end of sentences was more often marked by capitalized words, in particular capitalized pronouns.

Typical punctuation such as periods, exclamation marks, question marks were used emphatically rather than syntactically. That is to say that there were more often employed to express orality, rather than to mark the end of a sentence. Finally, a lot of sentences lacked any coherent or predictable endings. This had the unfortunate side effect of the program classifying sentences as being literal when they were not, as long sentence length, as previously mentioned is a sign of literacy in the texts.

The main issue with the system is that it cannot necessarily tell one why exactly a sentence is representative of literacy or orality with respect to the author’s intent. This is the inherent issue in using a naïve Bayes to calculate such features. It simply takes what it has gathered from a training data set and then presents a result based on said data. However, it cannot give one insight into the thoughts of the author as to why one discourse type was preferred over the other.

In addition to that, that the results were different than had initially been expected. This expectation was not based on a scoring system or probabilistic reasoning as was the case with Ortmann and Dipper (2019; 2020), but rather on the notion that these data sets must be inherently different due to the nature of their data.

SMS is interpersonal communication which generally brings along the characteristics of being representative of orality (Koch & Oesterreicher, 1985). Wikiconflits discussions are group discussions of a scientific quality and thus must most likely represent literacy (Koch & Oesterreicher, 1985). As for the eBay corpus, the buyer and seller were not necessarily in permanent and constant contact with one another and this indicates discontinuous, asynchronous communication which is a sign of literacy (Ortmann & Dipper, 2019). The expectation was based on the notion that eBay posters creating their postings would have more time to prepare and rehearse them and this preparation is often reflective of literacy (Koch & Oesterreicher, 1985; Ortmann & Dipper, 2019).

The classification of the registers, as seen in figure 5 and figure 6, were the catalyst for reaffirming the notions of Koch and Oesterreicher (1985) and Ortmann and Dipper (2019). Unfortunately, there was not enough data provided by Müller (1975) to strictly rely on such sociolinguistic parameters. However, automatically recognizing the French registers was not the direct goal of the program. The ideal situation would have entailed having the program classify an utterance according to its register, which then could be graphed on a discourse type. This was indirectly done by having examined the registers, their properties and how they map to literacy and orality, but not individually.

# Conclusion

With all things being equal, the internet, and by extension digital communication, are still in their infancy. They provide a wealth of information that can be useful for linguistic analysis among other things. This was the reason for wanting to use non-standard French language data in an electronic form.

The goal of this project was to assess literacy and orality in non-standard data. This was to be done using a naïve Bayes classifier which only works if it has training data from which it can learn. As there was no such data available, it was necessary to create a scoring to automatically and prototypically tag sentences that represented literacy and orality in French. Creating an algorithm that creates training data and then successively training a naïve Bayes allowed for insight into the nature of discourse information in non-standard French data.

The initial thought behind using the three corpora was that eBay would serve as the midway point SMS and Wikiconflict chats. However, this proved to be false as the conceptual is often much more difficult to define and determine than the medial representation of language. Despite this initial set back, the nature of literacy and orality in non-standard data could be determined to a certain degree. The data shows that there is indeed a spectrum of literacy and orality within data and that the data, while it can represent a discourse, it cannot be entirely dictated by that. The most interesting point here was that the discourse type can be determined in a text using universal and general classification features as well as a naïve Bayes classifier.

The domains of the respective texts can offer up plausible reasons as to why certain discourse types were chosen as opposed to others. SMS being interpersonal communication, Wikiconflits being scientific in nature and eBay representing asynchronous communication. Despite only having speculative answers and having minor setbacks, it is worth noting that the results line up with previous research and the assumed domains of the research types. More research and devotion to this topic would allow linguistic analysis to show as to why authors prefer one discourse type over another.

# References

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

Bieswanger, M., & Becker, A. (2008). *Introduction to English Linguistics* (2nd ed.). Narr Franke Attempto Verlag.

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Klabunde, R., & Langer, H. (Eds.). (2010). *Computerlinguistik und Sprachtechnologie* (3rd ed.). Spektrum. <https://doi.org/10.1007/978-3-8274-2224-8>

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>

*Fr\_core\_news\_sm*. (n.d.). Spacy. <https://spacy.io/models/fr#fr_core_news_sm>

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. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf>

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. <http://dx.doi.org/10.15496/publikation-20410>

Koch, P., & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift Für Germanistische Linguistik*, *35*, 246–275. <http://dx.doi.org/10.15496/publikation-20391>

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press

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>

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. The sentences that do not appear in the following results are not accounted for since they were classified as being unknown. This means that it could not be determined if they were representative of literacy or orality [↑](#footnote-ref-1)
2. The names being redacted was part of the pre-processing done the respective institutions and was thus not part of this project. [↑](#footnote-ref-2)