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

For most human languages, a message can either be communicated through text or through speech. These two refer to the medium in which information can be delivered from one person to the next. However, an aspect of communication which is often overlooked is the conceptual. The conceptual here referring to the thought and intent behind the speaker. If the speaker's message is more in line with written text, then it is referred to as literacy, whereas a message being more in line with spoken language is referred to as orality.

Non-standard French language data was obtained from eBay, SMS chats and Wikiconflits was analyzed to explore the 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 as a probabilistic text classification identifier that assigned the most probable conceptual classification feature to a document.

The results showed that those who use the platform eBay tend to express themselves conceptually in a more literate fashion, but less so than those of wikiconflits authors. SMS chats display a high level of conceptual orality, but less so than initially assumed. The reasons behind this are that eBay sellers tend to use a mixture of both to attract customers or potential buyers. The orality of wikiconflits participants was confined to follow-up questions or short statements. Finally, SMS chat participants expressed themselves orally too high a degree due to the informal and close nature of the context.

Main Table of Contents:

List of Figures:

List of Tables:

List of Equations:

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
NZ	Niveau zéro
OOV	Out-of-Vocabulary
POS-Tagging	Part of Speech Tagging

1 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. Oral, i.e. spoken discourse, can be understood as process which employs audible sounds to express meaning, whereas literate, i.e. written discourse, is the visual medium that is uses visible, written symbols (Bader, 2002).

An aspect that is often overlook is the conceptual communication. In other words, what is the actual intent that a speaker wishes to communicate with their message? Should the speaker's message be more in line with written or spoken speech? The intent of this conceptual frame is often referred to as literacy and orality (Koch & Oesterreicher, 1985).

With these distinctions in mind, 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. 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 (Koch & Oesterreicher, 1985).

To explore the conceptual , French language data has been chosen from three main sectors: eBay, SMS chats and Wikiconflits chats. SMS chats have been chosen as they are the most likely candidate for orality (Bader, 2002). These are to contrast with the Wikiconflits chats as the content pertain to scientific and intellectual discourse. Finally, eBay postings are to be seen here as a control as they do not intrinsically represent one conceptual discourse style over another.

Determining the discourse will be done using a multinomial naïve bayes algorithm (Jurafsky & Martin, 2020) and a simple, but effect smoothing algorithm as proposed by Ng (1997) to address the OOV problem. Multinominal naïve bayes, which will be referred to from here on out simply as naïve bayes, requires training data for it be able to properly determine the conceptual discourse type of a given document. Therefore, a scoring system will first have to be developed that can automatically identify sentences according to their conceptual discourse type.

2 Related Works

2.1. Theoretical Linguistics

Koch and Oesterreicher (1985) were very influential in setting up the paradigm of literacy and orality. They did this by providing a distinction between the medial and the conceptual facets of language. 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. Furthermore, they also place focus on sociolinguistic aspects regarding this paradigm by mentioning *Distanzsprache* and *Nähesprache* which are additional factors that are crucial to identifying literacy and orality which are

Distanzsprache represents how far removed the speaker is mentally, conceptually and physically from other speakers. Situations that are more of a personal and physical nature will often be assigned to the category of Nähesprache. As the specific object language here in question is French, there are French-specific elements that are to be taken into and will be instrumental in determining the necessary features for language independent classification criteria.

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, diastratically within the French language.

2.2. Computational Linguistics

Using a naïve bayes classifier for text classification purposes is in of itself not a new process (Jurafsky & Martin, 2020). However, what has been seldom done is using this method to identify literacy and orality in any given text.

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. This was done 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). 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 (2002) 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. at the time of publication.

3 Language as a Construct

3.1. General Features of Language

To preserve the dichotomy of written versus spoken, literacy vs. orality, language here will be treated as being confined to these two dimensions. That is to say that all other forms of communication, .e.g., non-verbal, sign language, and the like will not be included when referring the concept 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 or written symbols, i.e., letters, characters, etc. (Bader, 2002). De Saussure makes the distinction of *parole* versus *langue*. Parole being the actual realization of language. Langue being the virtual construct of a given language that could be realized by a speaker of is said to be 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) .

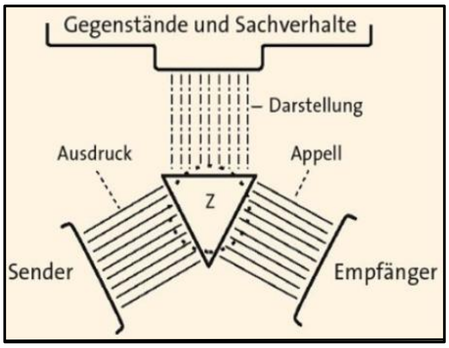


Figure 1. Bühler Organ-Modell
(Stein, 2014, p. 1)

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 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 elicits a desired response from the listener that is in line with the the sprachliches Zeichen (Stein, 2014). All three are present in every message, but general one message will dominate over the others (Bader, 2002).

3.2. Medial Features

Spoken language 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 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. That is to say that the active voice and elliptical structures are preferable in spoken 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, etc. 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 a 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).

3.3. Conceptual Features

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

Konzeption

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

Table 1. Medium and Concept

(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. 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?

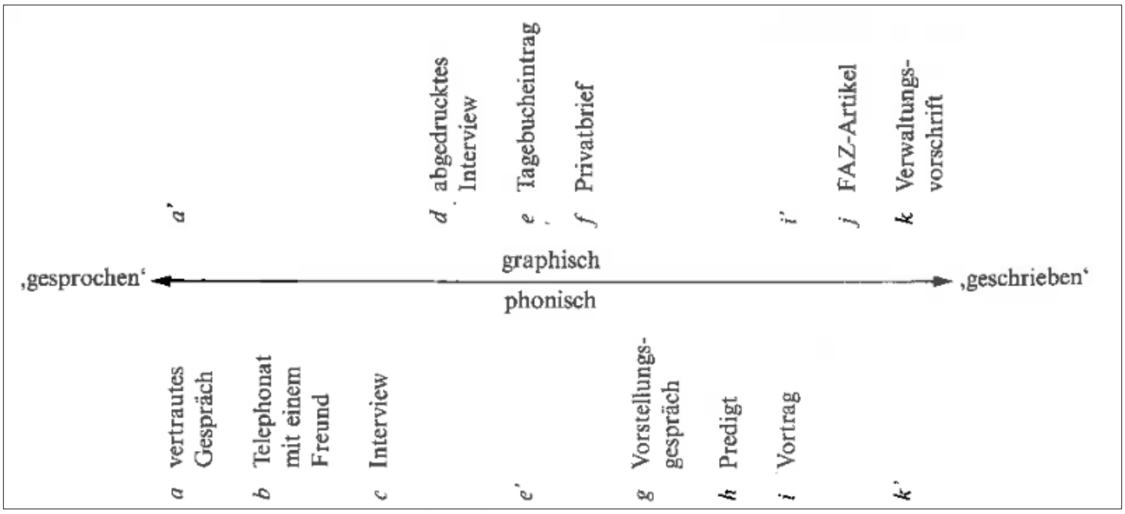


Figure 2. Written and Spoken Language

(Koch & Oesterreicher, 1985, p. 17)

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

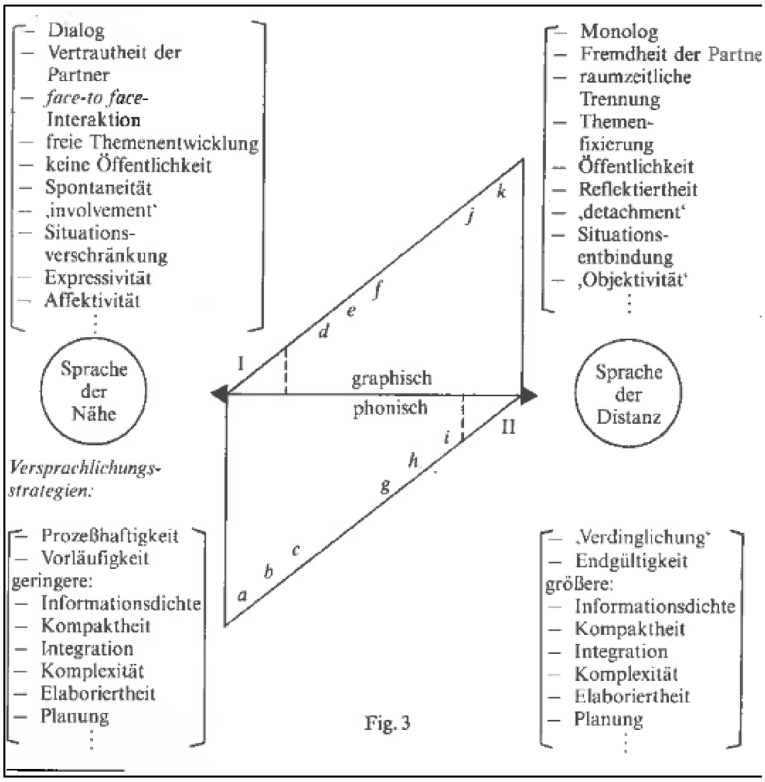


Figure 3. Communication diagram
(Koch & Oesterreicher, 1985, p. 23)

On the phonic portion of figure 2, a,b,c,g,h,i represent spoken speech that starts off being of informal and personal nature and gradually becomes less informal and personal. In doing so, the language becomes more in line 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 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 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.

Using all, three of these parameters: Medium, Conception and Distance-Proximity, a more detail analysis of language is possible. Referring to figure 3, 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 therefore 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 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.

4 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). By employing sociolinguistics, it is possible to investigate the effects of extra linguistic factors on society. Furthermore, a speaker's linguistic choices often give information about their social and geographical background (Bieswanger & Becker, 2008).

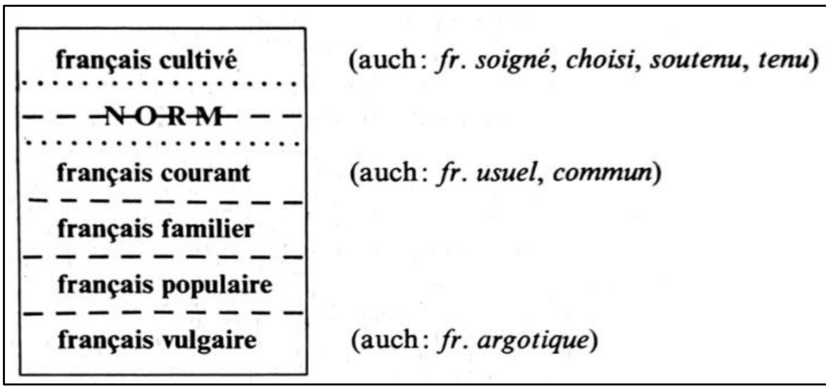
Registers are such linguistic phenomena that are general points of interest for linguistics involved in sociolinguists(Bieswanger & Becker, 2008). Registers, or styles, can be loosely defined as:

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 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).

All these elements can be 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.

4.1. Le Français



français cultivé	(auch: *fr. soigné, choisi, soutenu, tenu*)
~~NORM~~	
français courant	(auch: *fr. usuel, commun*)
français familier	
français populaire	
français vulgaire	(auch: *fr. argotique*)

Figure 4. French Registers (Müller, 1975, p. 184)

Historically speaking, French was 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 the usage of French for most of French language history.

This was in part due to the academic body, Académie Française who is instrumental in setting 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 says that languages are in a constant state of change (Müller, 1975; Stein 2008).

At the most fundamental level, French registers are usually classified as *français cultivé, français familier, français populaire, français vulgaire, français argotique* and *français technique*. Français cultivé being the most formal and français vulgaire being the least formal (Müller, 1975; Stein 2014). As seen in figure 4, many of these registers have different referents, but denote the same speech patterns. For the sake of simplicity, the French registers will only be referred to terms previously mentioned.

4.2. Français Cultivé

FC is often referred to as *français soigné, français choisi, langue recherché, langue tenue, langage soutentue, style noble*. This register often viewed in positive light and seen as the register that one should try to replicate. (Müller, 1975). Seeing as how this register considered the highest register. It is should not be used in banal or informal situation otherwise the speaker risks be seen as being pedantic and pretentious (Müller, 1975). It is used in official situations, special ceremonies or other special occasions.

The most prominent feature of this register in speech 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 appear with these verbal constructions (Müller, 1975).

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

4.3. Français Familier

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 that there is not a lot of emphasis placed on proper enunciation. This spontaneity is most likely due to the fact that FF, and FP by extension directly descended from Vulgar Latin, which itself was the primary spoken register of Latin, both in terms of medium and concept. (Müller, 1975).

Statements and questions are generally formed through falling and rising intonation, respectively, even though question 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. This construction is a left-over of bon usage and free syntax rules (Müller, 1975).

Furthermore, 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 expression when communicating strong feelings. This leads to a high number of simplified expressions, animal-inspired metaphors and using adverbs atypically as intensifiers (Müller, 1975). Therefore, the register is often consigned to orality as it signalizes a nonchalant attitude and, as the name implies, familiar atmosphere.

4.4. Français Populaire

FP is not considered to be proper and good French. This means that is 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. Historically speaking, this 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, FP tends to forgo the linguistic norms. Verbal phrases are often formed without their corresponding personal pronouns. 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 desired is 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 of course 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 they 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.

4.5. Français Vulgaire

FV is lowest register both in terms of prestige and formality (Müller, 1975). It is often grouped together with FA. The difference being that it and its components are generally known to all speakers of French, whereas FA is restricted to certain milieus. Interjections, expressions of displeasure and expletives are present throughout FV.

FA is therefore avoided whenever possible as it is in direct opposition to social norms regarding etiquette due to how it can be used to described things in an indecent manner. It is notable for its lack of scientific jargon, Latin loanwords, euphemisms. It is also incredibly adept at coining new words that employ the method of directness. It is also conceptually oral in nature.

4.6. 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 is 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 usually the reason why argot is considered to a 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. The high number of Synonyms and polysems are als a byproduct of argot's instability (Müller, 1975).

Argot is for all intents and purposes 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. This is most likely due to the fact that it reflects the time period in which the speakers live and not the continuing of a linguistic tradition (Stein, 2014).

4.7. Français Technique

LT is often grouped together with argot and were historically seen as being one and the same (Müller, 1975)..LT can be viewed as a microcosmos of sorts as there exist two poles within LT. It 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. 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 LT. 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 of most LT is scientific in nature and therefore conceptually literal.

5 The French Language Corpora

French, as with any natural human language, is not a monolith, but a language that is spoken across, many domains, age groups, countries, etc. (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. The methods and reasoning will therefore apply to this variant of French. While it might very well be possible that the methods and reasoning are applicable to other varieties such as Swiss French, Belgian French, Canadian French, etc., that is not necessarily goal, but be an unintended byproduct.

Secondly, 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.

Lastly, even though Metropolitan French is the object language, 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 therefore not bound to geographical constraints, it is plausible that speakers of other varieties have partaken in the conversations.

5.1. Data Sets

There are three primary data sets that will be the focus of the linguistic analysis: eBay petites annonces (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, online action listings from the online auction platform eBay which are split across four subcorpora. 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 only 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.

The goal of using this information is to analyze how people would react knowing that their information can be viewed by others. However, 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 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.

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 wiki chats 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.

5.2. 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, do contains markers to identify author, date, time, title of the post, etc.

The eBay corpus has been tagged to with respect to typical features of ad postings such as abbreviations, misspellings, marketing language, slang, emoticons, etc. The remaining data sets have been tagged for emoticons, and personal pronouns. As previously mentioned, this data set is comprised of 4 sub-corpora. Before the individual entries could be properly processed, the corpora had due to be sub-divided into their respective components. The other two data sets 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 to prevent accidentally training and testing on the same data.

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. Subsequently, information related to parts of speech, morphological and syntactical dependencies as well as tokens were ascertained from the sentences by using *Spacy*.

6 Methodology

6.1. Classification with 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 multinominal and Bernoulli naïve bayes classifiers being among the most common (Jurafsky & Martin, 2020). The main difference between the two is that Bernoulli naive bayes models the presence or absence of feature, whereas multinominal bayes counts the number of times a given feature occur.

They work well with binary classification and are most often employing in sentiment analysis, spam detection, authenticating authorship (Jurafsky & Martin, 2020). The following explanation applies to the multinominal bayes. The naïve bayes' algorithm is at its core a conditional probabilistic algorithm that is first and foremost based on the Bayes' theorem which is as follows:

$$P(A|\mathrm{B}) = \frac{P(A|\mathrm{B}) \cdot \mathrm{P(A)}}{P(B)}$$

Equation 1. Bayes' Theorem

(Carstensen et al., 2010):

P represents the probability of the given even with A and B representing two distinct events. Therefore , P(A|B) is the probability of A given B (Carstensen et al., 2010). Since Bayes' theorem is flexible, the events can be swapped, which produces the following formula:

$$P(B|\mathrm{A}) = \frac{P(B|\mathrm{A}) \cdot \mathrm{P(B)}}{P(A)}$$

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 follows (Manning & Schütze, 1999):

$$P(A \cap B_1) + P(A \cap \overline{B}_1\ )$$

$$P(A \mid B)\ \cdot\ P(B)\ + P(A \mid \overline{B})\ \cdot\ P(B)$$

Equation 3. P(A)

(Manning & Schütze, 1999)

When converting this theorem into a classifier, it results in the following formula (Jurafsky & Martin, 2020):

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|\mathrm{d}) = \frac{P(d|\mathrm{c}) \cdot \mathrm{P(c)}}{P(d)}$$

Equation 4. Naïve bayes' classifier

(Jurafsky & Martin, 2020)

$\hat{c}$, the estimation of the correct class, represents the maximum posterior probability with d being the documents out of all classes $c \in C$. However, as is often the case with NLP, natural language processing, tasks, only the maximum argument is relevant:

$$argmax_B P(B|\mathrm{A}) = \frac{P(A|\mathrm{B}) \cdot \mathrm{P(B)}}{P(A)} = argmax_B (A|\mathrm{B}) \cdot P(B)$$

Equation 5. Argmax

(Jurafsky & Martin, 2020)

This also applies to the naive bayes' classifier producing a simple, but effective model:

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|\mathrm{d}) = \underset{c \in C}{argmax}\, P(d|\mathrm{c}) \cdot \mathrm{P(c)}$$

Equation 5. Argmax of Classification

(Jurafsky & Martin, 2020):

To determine the most fitting class, the two probabilities must first be computed

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|\mathrm{d}) = \overbrace{P(d|\mathrm{c})}^{Liklihood} \cdot \overbrace{\mathrm{P(c)}}^{prior}$$

Equation 6. Model Probabilities

(Jurafsky & Martin, 2020)

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

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|\mathrm{d}) = \overbrace{P(f_1, f_2 \ldots., f_n|\mathrm{c})}^{Liklihood} \cdot \overbrace{\mathrm{P(c)}}^{prior}$$

Equation 7. likelihood

(Jurafsky & Martin, 2020)

Thus, the naïve bayes assumes that occurrence of the features, but not their position:

$$P(f_1, f_2, \dots, f_n|c) = P(f_1|c) \cdot P(f_2|c) \cdot \dots \cdot P(f_n|c)$$

Equation 8. Composition of likelihood

(Jurafsky & Martin, 2020)

This results in the formula:

$$C_{NB} = \underset{c \in C}{argmax}\, P(c) \prod_{f \in F} P(f\,|c)$$

Equation 9. argmax of likelihood

(Jurafsky & Martin, 2020)

To apply this formula, it is only necessary to traverse all words in each document:

$$C_{NB} = \underset{c \in C}{argmax}\, P(c) \prod_{i \in positions} P(\,w_i|c)$$

Equation 10. Calculating argmax

(Jurafsky & Martin, 2020)

To apply the formula, it is first necessary to train the model by calculating the probabilities of P(c) and P(f_i|c). This is done by using the frequencies in the data to ascertain MLE, or maximum likelihood estimate.

$$\hat{P}_{(c)} = \frac{N_c}{N_{doc}}$$

Equation 11. MLE

(Jurafsky & Martin, 2020)

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

$$\hat{P}(W_i|c) = \frac{count\ (wi, c)}{\sum_{w \in V} count(w, c)}$$

Equation 12. Calculating prior

(Jurafsky & Martin, 2020)

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 turning, held-out, etc. (Jurafsky & Martin, 2020). However, the one used in this paper is based on that of Ng(1997):

$$P\ (W_i|C_n) = \frac{C(w_n)}{N^2}$$

Equation 13. Ng Smoothing

Ng(1997)

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.

Token	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
imbécile	0.12	0.5
j'	0.12	0.5

6.1. A worked example with naïve bayes

	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

Table 1 worked example

Examples take from Müller (1975, p.185)

P(ORAL) = .60

P(LIT) = .40

Smoothing (ORAL) = .12

Smoothing (LIT) = .08

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

Document	**Feature**	**Prob**
Vous dites imbécile		
P(vous|ORAL) * P(dites |ORAL)* P(imbécile |ORAL) * P(ORAL)		
0.3 * 0.3 * 0.12 * 0.60	ORAL	0.00798
0.08 * 0.08* 0.5 * .40	LIT	0.00128

$$\frac{C(word)}{C(Classification)}$$

6.2. 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 & Oesterreicher (1985):

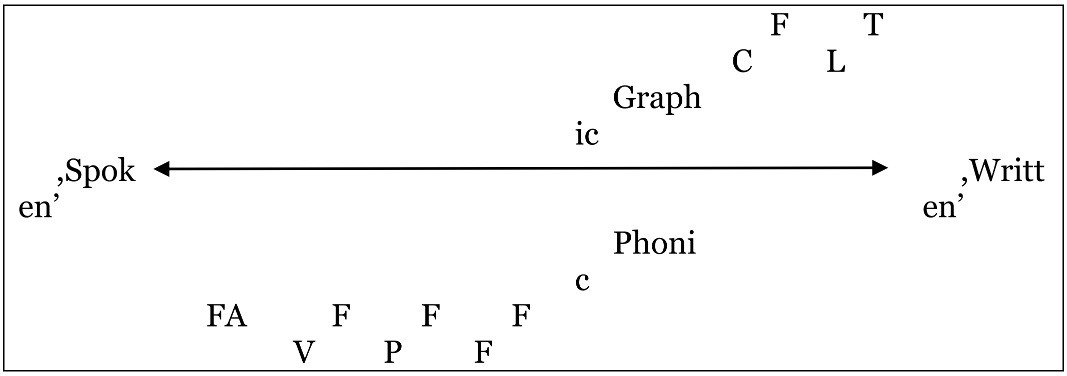


Figure 5. Registers on Orality and Literate Scale

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

:

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 orality and literacy.

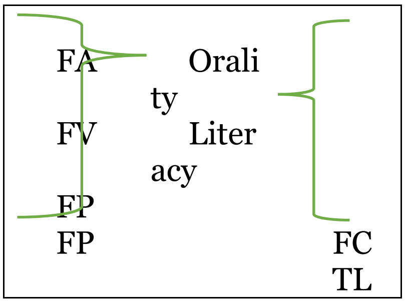


Figure 6. literacy and orality

7 System Evaluation

7.1. Developmental Overhead

As was the case with the corpora used in this project, most of the linguistic data is typically saved in an .xml format. Furthermore, 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. For these reasons, the writing of the program took around two weeks since it was necessary to write multiple functions that could accept .xml, .csv and .txt data files. Furthermore, 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, training, with testing being done in the last phase. This required the data sets to be split accordingly and evenly along all three phrases. The training of the program varies depending on the amount of data being input into the system and the system resources. Starting the program, the reading in of a single document and creating a training document from it typically did not take longer than a couple of minutes.

The classification criteria used to create training data could theoretically be retrained to recognize any language, more specifically, 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.

7.2. Classification Sets and Naïve Bayes

Various researchers (Bader, 2002; Ortmann & Dipper, 2019; Rehm, 2002;) have provided various 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 intent of the scoring system used was meant to assign one point if a criterion met the parameters set forth. However, this proved to be extremely 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. To remedy this problem, another option was chosen which entailed weighting weighing the criteria according to the importance and prevalence of the data set.

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

Table 2. 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 by hand. This produced the following results.

	Values (in

	Percent)
Accuracy	94
Error Rate	28
Precision	91
Recall	69
F-Score	79

Table 2.1 Evaluation for Literacy Classification

A second classification set was created that mirrored the first classification set to a certain extent but considered factors that often occurred in French texts expressing orality. These classification criteria set was then tested and evaluated.

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.

Table 3. Classification for Orality

	Values (in Percent)

Accuracy	91
Error Rate	29
Precision	1
Recall	68
F-Score	0,81

Table 3.1. Evaluation of Classification of Orality

The sentence is analyzed according to both criteria and the highest score determines the feature of the document. Throughout all the corpora, word length, sentence length, reduplication of symbols played the biggest role in determining the feature of the sentence. This lines up with the sources (add sources) that also show that expressions of literacy tend to have longer sentences and longer words, whereas expressions of orality tend to show the opposite. Abbreviations, acronyms, while important, 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 the database was trained, sentences were tagged according to their highest probability.

	Values (in Percent)
Accuracy	97
Error Rate	0,1
Precision	1,0
Recall	0,88
F-Score	0,936
Cross Validation	78%

Table 4. Naïve Bayes Evaluation

7.3. Sentence Tokenizer

Two of the most popular NLP libraries, NLTK and Spacy, both provide sentence and word tokenizers that can be used in NLP tasks. However, the overhead with respect training them to recognize non-standard data and importing them slowed down the run time of the program. Furthermore, they did not provide any additional benefit over using a custom regex expression to parse the sentences.

Since the data is often non-standard i.e., does not follow the norms of the French language, it was not always clear which sentences should be parsed and where they should be parsed. A naïve approach might involve simply splitting texts using punctuation. This initial approach proved to be extremely effective because sentence punctuation was often used correctly in sentences that were standard.

Nevertheless, 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 result 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.

The points did not to seem to affect the accuracy of the sentence tagger and worked well across all three domains

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. It should be noted that the definition of sentence is being somewhat expanded to encompass such thought.

7.4 Spacy Module

The spacy module was used for tokenization, part-of-speech tagging, syntactical dependencies and assessing morphology. Using Spacy was preferred over a custom function was preferred even though it added to the duration of the run time of the program. Furthermore, it provided to be more reliable and up to date than NLTK It was initially thought that spacy would not be able to provide reliable and accurate information due the fact that most of the data is non-standard and therefore particularly difficult to analyze. This turned out to be incorrect and spacy provided relatively high accuracy in all three areas.

There was no correct made to the data to make it easier to be processed by spacy as the linguistic nature of the data should as unaltered as possible. A challenged posed by spacy was that users of the SMS chats often had incorrect spellings or made high use of emoticons. These did not impact Spacy's performance. However, emoticons were classified as being punctuation, rather than as emoticons. This could have been remedied by training a pipeline to recognize such symbols. This minor setback did not affect the creation of the training data. It was therefore not deemed necessary to train such a pipeline.

8 Results[1]

8.1. Development phase

As was mentioned in chapter 5, the data was split equally into three sets: development, training and testing sets. However, the number of sentences and tokens were not distributed equally among all three of the original corpora. With the SMS corpus being the biggest 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 to 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. This could be remedied by having more data to train a French-specific identifier

The second problem, however, defeats this solution as too many features were deleted from a sentence which caused it to be unable to be recognized 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

[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

other system were generalized and incorporated into the second classification system.

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 orality and literality in datasets. Using a combination of criteria proposed 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 eBay set, there were traces of German and English since postings were most likely on a national, and not a local scale.

Using the scoring system as specified in chapter, Wiki and SMS as training data, data was labeled according to the classification sets mentioned above.

	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

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

Table 7. Results of development training data results

Table 7.1 Most important development classification 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

Table 7.2 Most important development classification for SMS

	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

Table. 7.3 Naïve bayes development results

8.2. 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 error in the code that would assign incorrect scores to the ratios. These results of which mirrored those of the development phase to a certain degree.

		Tokens	Sentences	Documents	LIT	ORAL
Wiki	sms_29508_59014	8226	463	52	303	160
SMS	Wikiconflits_0_54_106	4138	458	255	140	317

Table 9. Classification training data results

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

Table 9.1 Most important training classification criteria

Feature	Classification Criteria

LIT	NP_VB_RATIO
LIT	SEN_LEN
LIT	NOM_SUBJ
ORAL	SEN_LEN
ORAL	ALL_CAPS
ORAL	AVG_WORD_LENGTH

Table 9.2. Most important training classification criteria (SMS)

	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

Table 12. Naïve bayes training results

8.3. Testing phase

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

	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
Muller Lit	Mueller_LIT	699	20	20	20	0
Muller oral	Muellerr_ORAl	1971	59	59	50	9

Table 13. Analyzing all corpora using training dataset

9 Discussion

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 naïve training algorithm would suffice as opposed to having a naïve bayes and training algorithm in one program. One would be right in raising such concerns. However, this problem could not be avoided as there existed no reliable or accurate training data the for the program.

The idea of a scoring system was inspired by various authors (Bader 2002; Koch & Oesterreicher 1985; Ortmann and Dipper 2019; Rehm 2002) and is not to be assigned strictly to one author. The scoring system was to incorporate elements that are prototypical of the respective discourse types. The idea of using criteria was put forth by the likes Koch & 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 agnostic 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 source of much of contextual French information for the training data as well as the naïve bayes. Surprisingly, 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.

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

That is why more attention and thought was put into continuing with a universal classification set as opposed to French language classification set.

The scoring system relies heavily on naïve assumptions that often prove to be correct. More points were given to sentence 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 etc.

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 punction, 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 scoring system and scoring system that is more finely tuned to the French language. With more time and resources, this would be a possibility

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.

In the development phase, the wiki document had approximately 30/60 split across all domains regarding the expected orality. This means that 30% of the documents represent orality, whereas 60% represent literacy.

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 literal.

The unexpected high literacy in eBay data can be attributed to buyers and sellers using 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 high scientific and intellectual in nature. If orality did occur, then it was only in short burst or uttering small statements.

Finally, the SMS chats were of a high orality quality, this was to be expected and extracting literality from these texts proved to the be 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]. Second, the end of sentences were more often marked by capitalized words, in particular capitalized pronouns.

Third, 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.

Overall, the system delivers results that are reliable and reflect of the discourse types. 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 the author's intent. This is the inherent issue in using a naïve bayes to calculate such features. It simply takes what it is has gathered from a training data set and then presents a result based on said data. However, it cannot give one

[2] The names being redacted was part of the pre-processing done the respective institutions and was thus not part of this project.

insight into the thoughts of the author as to why one discourse type was preferred over the other.

In addition to that, it has been mentioned 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). Lastly, eBay was discontinuous, asynchronous communication in that buyer and seller were not necessarily in permanent contact with one another. This expectedness 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 French registers were mentioned in chapter 4, but were 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 and their properties.

As previously mentioned, various authors (Bader 2002; Koch & Oesterreicher 1985; Ortmann and Dipper 2019; Rehm 2002) have proposed methods and ideas that are directly related to literacy and orality. Before these classification criteria were to be used, they were checked against the French registers to determine overlap. This overlap did appear in much of the data, but these registers are to be seen as an orientation and not a strict guideline by which one must fervently abide.

The classification of the registers, as seen in figure 2 and figure 3, 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. Therefore, they were indirectly incorporated into this project with respect to the selection of criteria and the selection of the corpora.

10 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 wanting to use non-standard French language data. The challenges posed by non-standard are many, but the most noticeable one is that the data is often in a state that makes it difficult to process directly due to the authors of such texts not always adhering to the norms set forth by the respective linguistic institutions.

Therefore, it was necessary to pre-process the data in a such way as to be useful. After successfully developing a system to read in the non-standard data, another challenge cropped that had to be addressed.

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.

The most important point of which being as to why the author of a document chose one discourse style over another. This can only be answered through speculation and inference. 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.

11 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

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

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.

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

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

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