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 Michael

Abgabedatum
23.08.2021

Prof. Dr. Stefanie Dipper
Prof. Dr. Ralf Klabunde

Abstract

An overlooked aspect of communication surrounds conceptual literacy and conceptual orality, where literacy corresponds conceptually to written language, and orality corresponds conceptually to spoken language. Non-standard French language data was obtained from eBay postings, SMS chats, and Wikipedia discussions to explore how literate and oral discourse are realized across different internet domains. Training data was automatically developed using features that are typical of literate and oral discourse. This was then used to train a naïve Bayes model to assign the most probable feature to a document. When creating the training data, eBay postings and Wikipedia discussions had high levels of literacy with SMS chats having a high level of orality. When testing with the naïve Bayes, eBay postings and Wikipedia discussions had levels of literacy like those of the classification phase. However, SMS chats had a low level of orality as the ambiguity of their non-standard nature made classification difficult.

Table of Contents:

List of Figures:

List of Equations:

List of Tables:

List of Abbreviations:

CMRW	CMR-wikiconflits
CoMeRe	Corpora of Computer-Mediated Communication in French
FA	Français argotique
FC	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
NLP	Natural Language Processing
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, and whistling, human languages are generally expressed medially by using graphic symbols or audible sound (Bader, 2002). Written language is mediated visually, using symbols, whereas spoken language can be understood as a process, which employs audible sounds to express meaning (Bader, 2002). However, an aspect that is often overlooked is literate vs. oral discourse.

With these distinctions in mind, the concepts of written vs. spoken and literal vs. oral arise. The former represents the medial aspect of language, whereas the latter represents conceptual literacy and orality. In other words, is the message of the speaker conceptually representative of written or spoken language irrespective of the medium? These two domains do not represent a natural dichotomy, as one might automatically assume, but rather, they are two domains of language that regularly correlate to one another (Koch & Oesterreicher, 1985).

Determining the literate and oral discourse is to be done using a multinomial naïve Bayes (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, referred henceforth to as naïve Bayes, requires training data for it to be able to determine if a document is more in line with conceptual orality or conceptual literacy.

As there is no training data available, features for classification sets must first be developed that can be used to automatically label documents according to conceptual literacy and orality. The naïve Bayes is a probabilistic algorithm that can be independently used from these classification sets. It thus allows for more freedom with respect to determining conceptual literacy and conceptual orality.

As the goal is not to only develop training data, but also to investigate how literate and oral discourse are realized, French language data from three internet domains is to be used: eBay postings, Wikipedia discussions, and SMS chats. SMS chats are the most likely candidate for realizing typically oral features due to their informal nature (Bader, 2002; Rehm, 2002). These are then to contrast with the Wikipedia discussions, as the content therein pertains to scientific and intellectual communication (*Overview of*

Wikiconflits-QI from CoMeRe, n.d.), which can realize features typical of conceptual orality (Koch & Oesterreicher, 1985). eBay postings are to be seen here as a "control" as it does not intrinsically contain features that are necessarily representative of literate or oral discourse. For this reason, it is to be assumed that this corpus contains mixed features, i.e., containing features typical of both conceptual literacy and orality.

2. Related Works

2.1. Theoretical Linguistics

Koch and Oesterreicher (1985) constructed the medial-conceptual paradigm of written vs. spoken and literate vs. oral by providing a situational context in which these two facets of language can occur. Koch and Oesterreicher (1985) also placed focus on sociolinguistic contexts regarding this paradigm by expounding upon the notions of *distanzsprache* and *nähesprache*, which are additional crucial factors in distinguishing between literate and oral discourse.

Distanzsprache represents contexts of conceptual literacy as there is a high degree of formal and physical distance between both the speaker and the listener. Typical situations include public events, official events, and newspaper articles (Koch & Oesterreicher, 1985). Nähesprache represents contexts that are conceptually oral in nature such as informal conversations or a phone call with a friend as these events are close in terms of proximity and informality (Koch & Oesterreicher, 1985). Using the criteria of distanzsprache and nähesprache, texts and documents can be initially assigned to conceptual literacy or orality. Koch and Oesterreicher (2007) offered a more detailed explanation regarding types by expanding their examples and explanations to include French, German, and English.

Even though Müller (1975) predates Koch and Oesterreicher (1985), Müller (1975) discusses the notions of medial and conceptual literacy and orality as it applies to French. In terms of medial literacy and orality, *message écrit* is seen as written language, whereas *message oral* is the phonic code that is directly analogous to spoken language (Müller, 1975, p. 57). In terms of conceptual literacy and conceptual orality, Müller (1975, p. 65) referred to them as *français écrit* and *français parlé* respectively.

Müller (1975) explored this distinction, and how it is realized chronologically, quantitatively, qualitatively, diatopically, and diastratically within the French language by consequently subscribing the various French sociolinguistic registers, *français cultivé*, *français familier*, *français populaire*, *français vulgaire, français argotique* and *français technique* to either français écrit or français parlé. Müller (1975) expounds directly and indirectly upon typical features of français écrit and français parlé, e.g., the frequency of word classes, sentence length, and word length. By extracting syntactical, part-of-speech, and morphological criteria such from the respective French registers and then assigning them to either français écrit or français parlé, features for the classification set according to these registers can be ascertained.

2.2. Computational Linguistics

Bader (2002) provided a rounded, general approach on how to properly assess conceptual literacy and orality in texts in the same vein as Müller (1975). However, Bader (2002) applied 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) offered 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. Features by both authors (Bader, 2002; Rehm, 2002) related to conceptual orality such as emoticons, repeated used of words, or of symbols were listed.

Ortmann and Dipper (2019) explored the ideas as proposed by various authors (Bader, 2002; Koch & Oesterreicher, 1985; Rehm, 2002) to be able to automatically identify conceptual literacy and orality in modern German texts. Features included, but not limited to, sentence length, word length, punctuation, and subordinating conjunctions (Ortmann & Dipper, 2019). The results of using these criteria showed many of these features are in some way related to oral conceptuality (Ortmann & Dipper, 2019, p. 1). Ortmann and Dipper (2020, p. 1) also assessed conceptual literacy and orality regarding historical texts by using a slightly altered feature set that was more fitting for historical texts as the non-standardized nature of historical documents could not be properly analyzed using modern criteria.

3. Language as a Construct

3.1. General Features of Language

Human language is first and foremost, the production of audible sounds, i.e., speech (Bader, 2002; Müller, 1975). 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, p. 10). If there is a distinction between these two elements, then opposition is present, but should they have the same function, then it would be necessary to refer to them as variants of one another (Stein, 2014, p. 10). This leads into the distinction of *langue* vs. *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, p. 10).

The Organon-Modell, 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 and sachverhalte are the messages being transmitted (Stein, 2014). 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 *appell*. 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 appell elicits a desired response from the listener that conforms to the sprachliches zeichen (Stein, 2014). All three are present in every message, but generally one will dominate over the others (Bader, 2002).

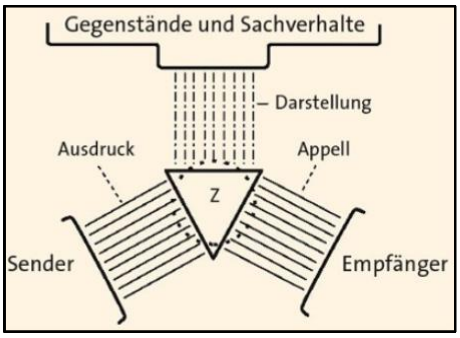

Figure 1.Bühler Organon-Modell (Stein, 2014, p. 1)

3.2. Medial Literacy and Orality

Spoken language can be understood as the the use of audible sounds to express meaning (Bader, 2002, p. 15). This is in line with structural linguists, who 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 & 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 (Bader, 2002; Stein, 2014).

Spoken language is a spontaneous act that is directly coupled with transience (Nerius, 1987, as cited in 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, p. 66). This also has a direct impact on syntax meaning that the active voice and elliptical structures are preferable in spoken language as they are easier to process (Ortmann & Dipper, 2019, p. 66). This is evident in the lexical aspect as spoken language makes frequent use of various particles, e.g., answer and modal particles, vague expressions, and interjections (Ortmann & Dipper, 2019, p. 67).

If spoken language is the phonetic expression of thought, written language is then to be seen as a graphical depiction and recording of said thought (Bader, 2002; Stein, 2014). The reason as to why written language exists at all is because it is essential in transcribing thoughts and transporting messages over long temporal and physical distances (Nerius, 1987, as cited in Bader, 2002).

Written language often contrasts with spoken language due to the dichotomous nature of the language paradigm (Bader, 2002; Koch & Oesterreicher, 1985). Where spoken language is restricted to being less complex, written language can benefit from the 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 writing in general without having to take the speaker's ability to process information into consideration (Ortmann & Dipper, 2019).

An important property is that an author of written language can express features of conceptual orality via the omission of characters, i.e., incorrect spellings, word contractions, or use of ellipsis dots, em dashes, or apostrophes (Ortmann & Dipper, 2019, p. 67). The opposite is true in that written language can also express features of conceptual literacy by strictly adhering to orthographic norms, employing complex syntactical and lexical structures (Bader, 2002; Ortmann & Dipper, 2019).

3.3. Conceptual Literacy and Orality

Although it would be possible to see a dichotomy being present between conceptual literacy and orality, this is not strictly correct. The dichotomy does exist, but it only applies to the medial aspect of language, i.e., to written vs. spoken language specifically (Koch & Oesterreicher, 1985). The other question remains though: What is to be done with conceptual literacy and orality? As features regarding medial and conceptual literacy and orality contrast, they can be grouped together as seen in figure 2.

		Konzeption	
		gesprochen	geschrieben
Medium	graphischer Kode	Faut pas le dire	Il ne faut pas le dire
	phonischer Kode	[fopaldiʀ]	[ilnəfplalədiʀ]

Figure 2. Medium and Concept
(adapted from Koch & Oesterreicher, 1985, p. 17)

The medium is either the *phonischer kode*, i.e., phonic code or it is the *graphischer kode*, i.e., graphic code. This means that a message like *faut pas le dire* is medially literal, but conceptually oral. In this particular example, it is due to the omission of *il* and *ne*, which belong to standard French (Koch & Oesterreicher, 1985; Müller, 1975). The opposite of this applies as well where *il ne faut pas le dire* is conceptually and medially literal as it complies with the written norms set forth by the governing linguistic bodies of the French language (Müller, 1975).

The phonetic element plays an important role as well but is only relevant if the message is audible. As text is not a medium that can transport audio, the phonetic element does not necessarily apply here. It only applies indirectly if the speaker first starts with a phonetic message that is then transcribed graphically (Bader, 2002; Koch & Oesterreicher, 1985). Koch and Oesterreicher (1985) see spoken and written language as being on a continuum with different conceptual possibilities that have different levels, which are exemplified in figure 3.

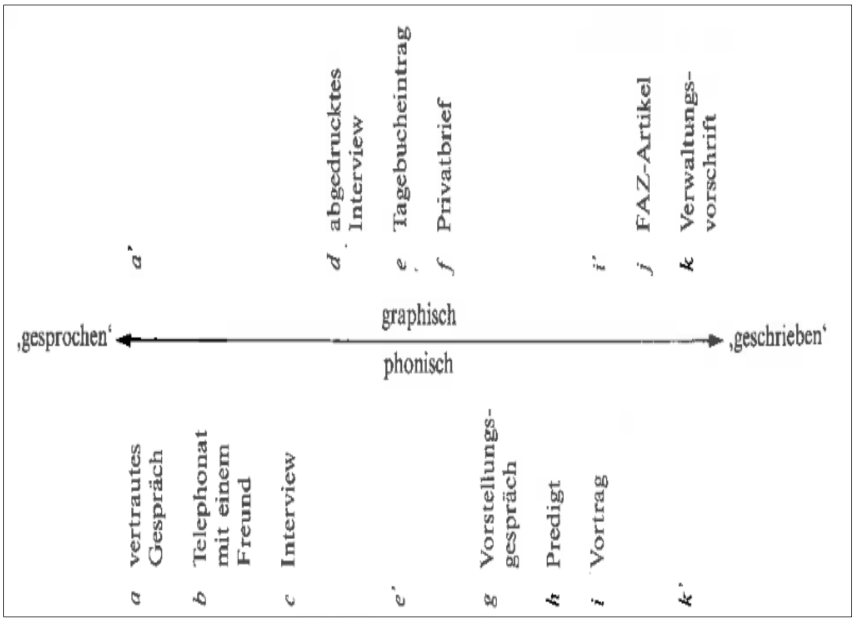

Figure 3. Spoken and Written vs. Graphic and Phonic
(Koch & Oesterreicher, 1985, p. 18)

Koch and Oesterreicher (1985) see spoken and written language as being on a continuum with conceptual possibilities that have different levels which are exemplified in figure 3.

On the *phonisch, i.e.,* phonic, portion of figure 3, all the texts are medially literal, but conceptually oral and gradually transition into conceptual literacy (Koch & Oesterreicher, 1985). When observing the two poles, *gesprochen*, i.e., spoken and *geschrieben*, i.e., written, there is a difference between a *vertrautes gespräch*, i.e., intimate conversation and a *vortrag*, i.e., presentation (Koch & Oesterreicher, 1985). The former represents spontaneous speech with the latter being prefabricated and then presented to an audience orally (Koch & Oesterreicher, 1985, p. 18).

On the *graphisch*, i.e., graphic, portion of the diagram, all documents are possible graphic representations of speech, with d, an *abgedrucktes Interview*, i.e., prepared interview, being conceptually the most oral with, k, v*erwaltungsvorschift*, i.e., an administrative regulation, being conceptually the most literal while still being medially oral (Koch & Oesterreicher, 1985, p. 18).

Figure 3 demonstrates, what was tabularly presented in figure 2, which is that conceptual literacy and orality exist on a continuum. An important element missing from figure 2 is how this relates to communication and discourse as was touched upon in figure 1. Figure 4 solves this dilemma by combining figure 2 and 3 together.

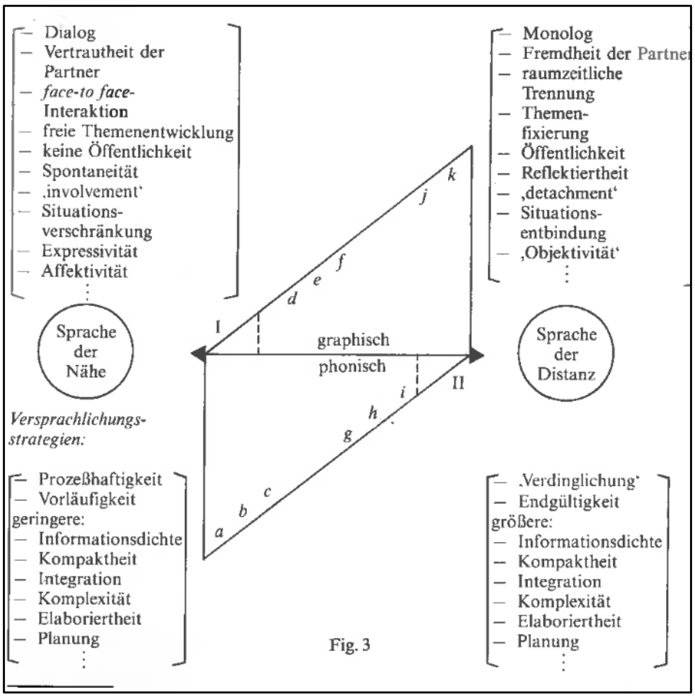

Figure 4. Nähesprache and Distanzsprache
(Koch & Oesterreicher, 1985, p. 23)

Figure 4 addresses how close in terms of proximity and familiarity the speakers are to one another. *Nähesprache*, or *sprache der nähe*, is reserved for situations that are physical and familiar in nature such as spontaneous, face-to-face and familiarity with the communication partner (Koch & Oesterreicher, 1985). *Distanzsprache*, or *sprache der distanz*, represents the opposite pole in that it depicts speech that includes communication that is detached, objective, unfamiliar, and fixed topic (Koch & Oesterreicher, 1985).

Referring to figure 3 and figure 4, an intimate conversation is thus medially oral, that also has features of conceptual orality. The dynamic of the speakers is one of familiarity and closeness, and the speech can be assigned the label of nähesprache. The opposite can be said of administrative regulation texts. There is 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 (Koch & Oesterreicher, 1985). It can be assigned as being conceptually literal while being medially oral and thus belongs to distanzsprache. Using these parameters: medium, concept, and distance-proximity, a more detailed analysis of language is possible.

4. Registers and Styles

A speaker's linguistic choices often reveals information about their social and geographical background (Bieswanger & Becker, 2008, p. 182). 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 a specific communicative situation. (Bieswanger & Becker, 2008, p. 187)

Registers and styles are instrumental in determining conceptual literacy and orality since understanding how and when they are used allow for better identification of the literate and oral discourse in written language. Certain registers and styles are generally realized in specific situations and have similar properties akin to those presented in figure 3 and figure 4, e.g., belonging to distanzsprache or nähesprache, representing prototypical situations containing literate or oral discourse. The following French registers will therefore be seen in the light of having qualities of registers and styles as described by Bieswanger & Becker (2008).

4.1. Le Français

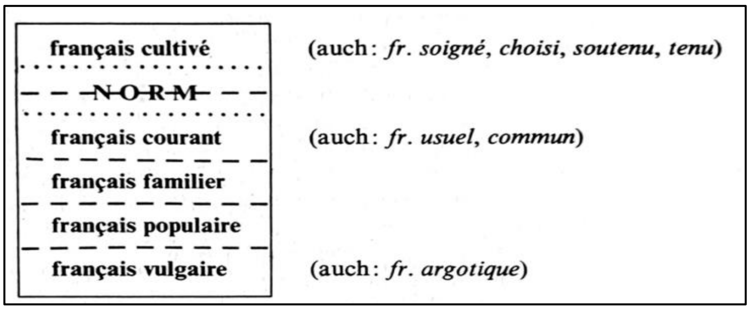

Figure 5. French Registers
(Müller, 1975, p. 184)

French was historically seen as having a single register (Müller, 1975). This is not in the sense that there was no variation, but rather, that there was one and only one correct way of using the French language referred to as *bon usage* (Müller, 1975). *Mauvais usage*, i.e., poor usage and *dites…ne dites pas,* i.e., say this, not that, dictated the correct usage of French for most of French language history (Müller, 1975). This was in part due to the academic body, Académie Française, who has been the governing body of the French language since its establishment in the 17th century (Müller, 1975).

Nevertheless, it is not necessarily feasible to entirely dictate what speakers of any given language do or say as this is directly antithetical to a defining characteristic of language, which is that languages are in a constant state of change (Müller, 1975; Stein, 2014). This led to the development of various French registers as seen in figure 5 (Müller, 1975; Stein, 2014).

4.2. Français Cultivé

Français cultivé, or FC, is often viewed in positive light and seen as the register that one should try to replicate due to its high prestige (Müller, 1975). It should not be used in banal or informal situations, otherwise the speaker risks being seen as pedantic and pretentious (Müller, 1975).

The most prominent feature of this register is the phonological component as it tends to consequently conserve sounds that are no longer used in the other registers (Müller, 1975). This includes phonetic opposition of certain sounds, the pronunciation of the schwa at the end of phonological words, and a more rigid syllable structure, e.g., *maître* vs. *mettre* (Müller, 1975). This has to do with the desire to retain the literary tradition, which is often dependent on such archaisms (Müller, 1975).

Certain verb tenses such as *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*, and *ne…guère* often appears with these verbal constructions (Müller, 1975).

Words, and thereby sentences, tend to be considerably longer as there is a high use of prefixes, suffixes, and negation particles (Müller, 1975). Furthermore, there is a high usage of conjunctions and particles of concessive subordination within this register (Müller, 1975).

Even though it is simultaneously considered français écrit and français parlé, it is often viewed as being conceptually literal as it retains the previously mentioned grammatical features, which are no longer used in contemporary speech either conceptually or medially (Müller, 1975). Whether written or spoken, it is considered artificial as it is a controlled process heavily reliant on proper word choice, intonation, and lengthy, detailed sentences (Müller, 1975).

4.3. Français Familier

Français 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, but it is used more frequently by those who have profited from a higher education (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*, and high use of topicalization, e.g., *cet homme, je l'ai vu* très souvent are typical of FF (Müller, 1975). It also makes high use of suffixes to denote agents and actors in speech context, e.g., *chançard*, *gueulard*, and *motard*. This also includes the diminutive suffixes such as *-et*, *ette*, *ot*. Reduplication is not only present among pronouns, but among nouns as well, e.g., *fla-fla*, *ronron*, *kif-kif* (Müller, 1975).

It is spontaneous in nature, and this is reflected in the fact that there is not much emphasis placed on pronunciation that is in lines with FC (Müller, 1975). This spontaneity is because FF, and français populaire by extension, are directly descended from Vulgar Latin, which itself was primarily an oral register of Latin, both medially as well as conceptually (Müller, 1975).

Due to its spontaneous nature, speakers tend to avoid overly complex and lengthy expressions when communicating strong feelings, which leads to a high number of simplified expressions, atypically using adverbs as intensifiers, shorter word length (Müller, 1975). Generally, what is not expressed either conceptually or medially is expressed through body language (Müller, 1975). The register is considered français parlé, and it is therefore conceptually oral as it signalizes a nonchalant attitude and, as the name implies, a familiar atmosphere.

4.4. Français Populaire

Français populaire, or FP, is considered neither proper nor good French as 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 (Müller, 1975). This is because it is not consistent with FC, and it 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, and whose speech was more commonly referred to as *lanuge du peuple* (Müller, 1975). In classifying it as such, FP is conceptually oral.

Verbal phrases are often formed without their corresponding grammatical subjects (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 (Müller, 1975). The *subjunctif* is only employed when a strong desire 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 preference for neglecting the spelling, especially when the message is clear from morphology. The most prominent example of this is the willingness to drop the ne of ne…pas (Müller, 1975). The lexicon does not differ in form from FC, but rather in usage, i.e., that speakers use the same words, but differently, which 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 *français vulgaire* (Müller, 1975).

4.5. 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 français argotique 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 (Müller, 1975). 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).

4.6. Français Argotique

Français argotique or *argot*, or FA, in its original form was meant to specify the speech patterns of marginal groups and that of professional jargon, which includes *français technique* (Müller, 1975). A defining feature of FA is that the speaker is intentionally trying to distance themselves socially from the outside group. At the same time, it is used as a way of identifying insiders and outsiders, which is usually the reason why argot is considered to be cryptic language (Müller, 1975).

Argot displays a high willingness to import loan words from dialects as well as other languages (Goudailler, 2002). Argot is conceptually oral as the need to record speech in a written form was completely secondary and 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 because it reflects the period in which the speakers live and not the continuing of a linguistic tradition (Müller, 1975).

4.7. Français Technique

Français technique, or FT, can be used to explain theoretical concepts to professionals the same field, or a reduction in complexity is introduced, i.e., vulgarization, making it more readily available to those who are not of a scientific background (Müller, 1975). A defining trait of it is the need to develop new terminology as the field of science is ever growing, which is done using complex use of morphological constructions (Müller, 1975).

A high influx of new words also comes 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 (Müller, 1975). The syntactical structures of français technique are not per se complex, but FC displays a high level of words that express causality which is to be expected as the goal is scientific in nature and conceptually literal (Müller, 1975).

4.8. Combining Registers and Discourse

Conceptual literacy and orality represent the binary feature set that is to be assessed with in this paper. As the corpora that are to be analyzed are in already in a textual medium, they medially represent written language. Therefore, it must be decided if the textual data contains features pertaining to conceptual literacy or orality. Grouping the registers in a manner akin to figure 2 can be seen in figure 6:

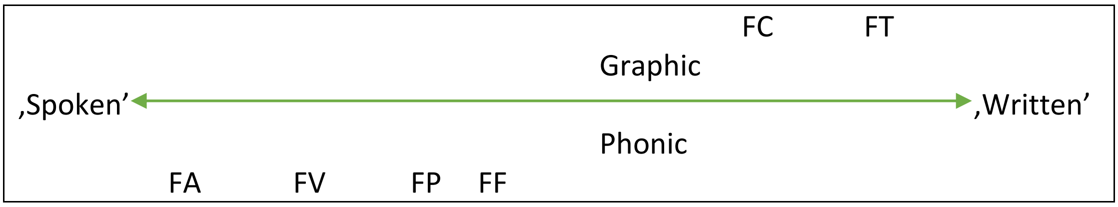

Figure 6. Registers According to Conceptual Literacy and Orality

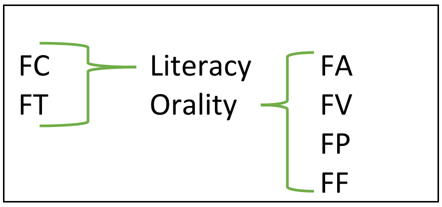

Figure 7. Conceptual Literacy and Orality

Figure 6 shows how the French styles and registers can be potentially mapped to written and oral discourse. Figure 6 can be further refined to allow them to be mapped to either conceptual literacy or conceptual orality as seen in figure 7. It would be neither reasonable nor feasible to train a model to recognize the individual registers due to the high overlap between them since some of the registers can belong to both français écrit and français parlé or neither. However, they all can express features pertaining to conceptual literacy and orality. By extracting characteristics and criteria from each register and grouping them accordingly, it is possible to create features for classification sets that allow for the automatic identification of conceptual literacy and orality.

5. The French Language Corpora

French is spoken across many domains, age groups, and countries (Müller, 1975; Stein, 2014). Whether a native speaker of French, a second-language speaker, or speaker of a given French dialect, this variation is present in France as well as outside (Stein, 2014). This poses a challenge since how orality and literacy are expressed is to some extent dependent on the local and personal understanding of the language (Bader, 2002). The object language here in question is that of Metropolitan French, which is contemporary French as spoken in Mainland 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 French standards, registers, and styles. Since the internet is an open platform, and not bound to geographical constraints, it is plausible that speakers of other varieties or languages have partaken in the conversations.

5.1. Data Sets

The three primary data sets that are the focus of the linguistic analysis are: eBay petites annonces (Gerstenberg & Hewett, 2019), CMR-wikiconflits (Poudat et al., 2015), and 88milsms (Panckhurst et al., 2014), which will be referred to as eBay or eBay postings, Wikiconflits or Wikipedia discussions, and SMS, or SMS chats respectively.

The eBay corpus contains listings from the online auction platform, eBay, and it was compiled by the department of Romance studies at the University of Potsdam with a collection of around 1256 online auction listings, which are split across four sub-corpora (*eBay petites annonces*, 2020). The first three sub-corpora deal with housing, vehicles, clothing, computer, telephones, children, collections, and leisure, while the last sub-corpus deals with professional activities, e.g., stocks, shops, and shipping (*eBay petites annonces*, 2020).

Wikipedia corpus contains discussions about IQ consisting of around 52 participants, 170 contributions, and 20,000 tokens (*Overview of Wikiconflits-QI from CoMeRe, n.d.*). As is often the case with sites like Wikipedia, the information presented may not be factually correct or in line with the terms of Wikipedia (*Overview of Wikiconflits-QI from CoMeRe, n.d.*). This does not necessarily pose a problem as the accuracy of the information is irrelevant with respect to its literacy and orality.

The SMS corpus is a collection of more than 88,000 SMS messages that were collected from speakers in the Montpellier area in France (Panckhurst, 2017). To comply with French data protection guidelines, the data had already been anonymized (Panckhurst, 2017). 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 their opinions on SMS communication (Panckhurst, 2017).

5.2. Pre-processing

The corpora (Gerstenberg & Hewett, 2019; Panckhurst et al., 2014; Poudat et al., 2015) were created with the goal of individual linguistic analysis in mind and so the data had been annotated and changed as little as possible by the respective institutions. This means that tokenization, sentence tokenization, POS-tagging, syntactical and morphological analysis were possible without interference from foreign analysis. They are available in the .xml format, and contain markers to identify author, date, time and title of the post. The eBay corpus was tagged with respect to typical features of ad postings such as abbreviations, misspellings, marketing language, slang, proper nouns, and emoticons (*eBay petites annonces*, 2020).

Before the individual entries could be properly processed, the eBay corpus had to be first sub-divided. Wikipedia and SMS were each already in one homogenous corpus and sub-division was therefore not necessary (Panckhurst et al., 2014; Poudat et al., 2015). However, all three of the corpora were then equally divided into three parts: development, training, and test data sets.

Since the corpora files were in an .xml format, it was not possible to directly access the text, but rather through their respective tags. This was done by parsing them using the Python module Beautiful Soup (*Beautiful Soup Documentation — Beautiful Soup 4.9.0 Documentation*, n.d.). Once the textual data was exposed, the respective entries were tokenized into their respective sentences using a custom sentence tokenizer that used regular expressions. Information related to tokens, parts-of-speech, morphological and syntactical dependencies were subsequently ascertained from the sentences by using French spaCy modules (*French · spaCy models documentation*, n.d.).

6. Methodology

The methodology involved using a probabilistic algorithm to recognize conceptual literacy and orality in texts. However, before this could be done, training data had to be ascertained. Due to the lack of known or adequate training data, another classification system had to be employed by which a training database could be built. From this database, probabilities could be calculated, and the conceptual orality or literacy of a given text could be made known.

Originally, a French-based classification set was meant to gauge the reliability of the language-independent classification sets as seen in table 1 and table 2. This relied upon using words to determine conceptual literacy and orality of a sentence. These words would then be deleted from the sentence so that they would not be made available to the other classification sets.

The validity of the language-independent classification criteria would be weighed against the language-dependent criteria set. This proved to be extremely ineffective since there were not enough unique words and criteria 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.

Another problem present throughout the eBay and SMS corpora was that the data was more non-standard than that of the Wikipedia corpus, 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 on a national, and not always a local scale (Gerstenberg & Hewett, 2019; Panckhurst et al., 2014).

6.1. Classification Sets

Various researchers (Bader, 2002; Ortmann & Dipper, 2019; Rehm, 2002) provided a plethora of criteria by which one can automatically identify literate and oral discourse. These criteria focused on creating a system which is to be linguistically and chronologically independent. However, since French data was being classified, characteristics of the French registers were taken into consideration when developing the classification criteria. Based on these researchers (Bader, 2002; Müller, 1975; Ortmann & Dipper, 2019; Rehm, 2002), two distinct classification sets were created as seen in table 1 and in table 2.

A sentence was automatically analyzed according to both classification sets. If a given criterion for a sentence was true, then it received points equal to the respective category as specified in table 1 and table 2. At the end of the analysis, two scores will have been calculated. The sums of the respective scores were then compared. The feature of the higher score was assigned to a sentence of a document. This means that if a sentence received more point with respect to conceptual orality, then the sentence will be classified as such and vice-versa. Table 1 was created partially by extracting specific features present in français cultivé and français technique and partially by those mentioned by other researchers (Bader, 2002; Müller, 1975; Ortmann & Dipper, 2019; Rehm, 2002).

Criterion	Description	Point Amount
ABBR_NO_VOWEL	Abbreviations without vowels	Number of abbreviations without vowels
AVG_WORD_LEN	Average word length	The length of the average word
CCONJ_VB_RATIO	More coordinating conjunctions than verbs	Number of coordinating conjunctions and verb
LOW_VERB_HIGH_ADJ	Low number of numbers, but high number of adjectives	Number of verbs and adjectives
NOM_SUBJ	Sentence Length	The number of nominal subjects
NP_VB_RATIO	Noun to verb ration	The number of nouns and verbs
PRES_TENSE	Present tense verbs	The number of present tense verbs
SEN_LEN	Sentence Length	The length of the sentence in character length (only considered if sentence is longer than ca. 9 – 10 words)
SHORT_SEN_LENGTH_PRESENCE_OF_NUMBERS	Short sentences that consist of only numbers	Only one point
THIRD_PERSON_EXPL	Dummy Subjects	The number of dummy subjects

Table 1. Classification Criteria for Literacy

Table 2 was created in the same fashion as table 1, but also by extracting features present in français argotique, français vulgaire, français populaire, and français familier.

Criterion	Description	Point Amount
ABBR	Abbreviations and acronyms	The number of abbreviations and acronyms
ALL_CAPS	All caps	The number of words written in all caps
AVG_WORD_LEN	Average word length	The length of the average word
EMOTIOCONS	The usage of emoticons in a sentence	The number of emoticons used in a sentence
HIGH_PUNCTION	High use of punctuation	The number of punctation symbols
ISOLATED_VERBS	Only verbs in a sentence	The length of the sentence
MULTI_CHAR_ REDUPLICATION	Using the same character multiple times	The number of symbols that occur more than once
PRES_TENSE	Present tense verbs	The number of present tense verbs
SEN_LEN	Sentence Length	The length of the sentence in character length (only considered if the sentence is less than ca. 5 words)
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
WORD_ WORD_REDUPLICATI ON	Using the same word back-to-back	The number of times a word is used more than once back-to-back

Table 2. Classification Criteria for Orality

6.2. Bayes' Theorem: Basis of Naïve Bayes

An efficient and well-known method of classifying a document is 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 occurs (Jurafsky & Martin, 2020).

They work well with binary classification and are most often employed in sentiment analysis, spam detection, and 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:

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

Equation 1. Bayes' Theorem (adapted from Carstensen et al., 2010, p. 122)

P represents the probability of an event with *A* and *B* representing two distinct events, while *P(A|B)* is the probability of event *A* given event *B* (Carstensen et al., 2010). Since Bayes' theorem is flexible, the order of dependency between the events can be swapped around (Manning & Schütze, 1999). This is demonstrated in equation 2:

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

Equation 2. Bayes' Theorem Reversed (adapted from Manning & Schütze, 1999, p. 43)

P(A), as seen in equation 2, is the normalizing constant that guarantees that the equation has a probabilistic aspect to it (Manning & Schütze, 1999). *P(A)* can be broken down into its individual elements as it is the combined probability of all events and is calculated as seen in equation 3:

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) \text{ [additivity]}$$
$$= P(A|B) \cdot P(B) + P(A|\overline{B}) \cdot P(\overline{B})$$

Equation 3. Normalizing Constant
(adapted from Manning & Schütze, 1999, p. 43)

$\overline{B}$ represents not *B*, and both serve to split *A* into two disjoint parts with the possibility of $\overline{B}$ being empty, whereas ∩ represents the intersect between two respective events (Manning & Schütze, 1999, p. 44).

6.3. Naïve Bayes as a Classifier

A document classifier can be created by using Bayes' theorem as a basis. To make the explanation more suitable for text classifications, the variables have been changed, as seen in equation 4, but the base form of Bayes' theorem remains intact.

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

Equation 4. Naïve Bayes Classifier
(adapted from Jurafsky & Martin, 2020, p. 57)

The naïve Bayes returns $\hat{c}$[1], which represents the maximum posterior probability given a document with *d* being the documents out of all classes, *c∈ C* (Jurafsky & Martin, 2020). However, as is often the case with NLP tasks, only the maximum argument, or *argmax*, is of importance. *Argmax* consists of the product of the likelihood and prior probability and this means that the denominator, in this case *P(d)* can simply be ignored as it remains the same for each class (Jurafsky & Martin, 2020; Manning & Schütze, 1999).

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

Equation 5. Argmax
(adapted from Manning & Schütze, 1999, p. 43)

Equation 5 represents how this can be computed, but it can be converted to be more in line with the variable labels of naïve Bayes classifier as seen in equation 4, which produces the following as presented in equation 6:

[1] '^' is the estimation of the correct class.

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

Equation 6. Argmax of Classification
(Jurafsky & Martin, 2020, p. 58)

In equation 6, and by extension, equation 5, equation 4, there are two main probabilities after having dropped the denominator, which are seen in equation 7:

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

Equation 7. Model Probabilities
(adapted from Jurafsky & Martin, 2020, p. 58)

Using equation 7, it is possible to determine the classification of a given document. $\hat{c}$ is the most probable class, which is computed using *P(c)*, the prior probability of a given class, and *P(d|c),* the likelihood of the document (Jurafsky & Martin, 2020, p. 58). The likelihood of a given document can be expounded upon as seen in equation 8:

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

Equation 8. Model Probabilities Expanded
(adapted from Jurafsky & Martin, 2020, p. 58)

Equation 8 is still too difficult to calculate as it forces one to calculate every given possibility, and therefore, it does not consider the simplifying assumptions of the naïve Bayes (Jurafsky & Martin, 2020, p. 58). The first assumption is the *bag-of-words* principle, which states that the position of the words within a given document is irrelevant as only number of times a word occurs, i.e., its frequency is important (Jurafsky & Martin, 2020; Manning & Schütze, 1999). The second assumption, sometimes referred to as the *naïve Bayes assumption*, is that probabilities $P(f_i|c)$ are independent of a given class and can be computed naively (Jurafsky & Martin, 2020, p. 58). Equation 9 considers the said assumptions and by applying equation 9, equation 10 results:

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

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

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

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

To apply equation 10 to a text, it is only necessary to traverse all words in each document in the order in which they occur within the document as detailed in equation 11:

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

Equation 11. Calculating Argmax
(adapted from Jurafsky & Martin, 2020, p. 58)

6.4. Naïve Bayes Classification Probabilities

To use the naïve Bayes classifier, it is first necessary to ascertain the probabilities of *P(c)* and *P*(f_i*|c)* (Jurafsky & Martin, 2020). This is done by computing the frequencies in the training data as presented in equation 12:

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

Equation 12. Probability of $P\ (c)$
(adapted from Jurafsky & Martin, 2020, p. 59)

Equation 12 states for the class prior *P(c)*, what is the frequency of a given class as it occurs within this document (Jurafsky & Martin, 2020). Finally, to compute *P(*f_i*|c)* as *P(*w_i*|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.

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

Equation 13. Probability of P(f_i*|c)*
(adapted from Jurafsky & Martin, 2020, p. 59)

However, when a given word does not occur within a certain class, the effective frequency is zero as seen in equation 14:

$$\boxed{\hat{P}(\text{"}meuf\text{"}\,|LIT) = \frac{count\ (\text{"meuf"}, \mathrm{LIT})}{\sum_{w \in V} count(w, \mathrm{LIT})} = 0}$$

Equation 14. Null Frequency
(adapted from Jurafsky & Martin, 2020, p. 59)

This is a problem as the naïve Bayes multiplies all values together, and a frequency of 0 causes the whole product to be 0. To solve this problem, a smoothing algorithm must be applied. A popular smoothing algorithm is LaPlace, add one-smoothing, since it is a simple method, that involves adding 1 to corpus frequencies (Carstensen et al., 2010; Jurafsky & Martin, 2020; Manning & Schütze, 1999). However, its simplicity is the exact reason as to why it is an ineffective method, and it is used best only for exemplary purposes regarding smoothing (Jurafsky & Martin, 2020).

Ng (1997) offers a simple smoothing algorithm that works well with naïve Bayes classifiers, while achieving a relatively high accuracy compared to other smoothing algorithms.

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

Equation 15. Ng Smoothing
(adapted from Ng, 1997, p. 209)

As stated in equation 15, with all other parameters being equal, *N* here represents the amount of training data from a given corpus squared. This equation must be applied for each respective class in the training data.

6.5. A Worked Example

For the sake of simplicity, it is assumed in the following corpus, as seen in table 3, that the sentences have the following features as listed in the feature column. They have not necessarily been analyzed using the classification criteria as specified in table 1 and table 2, but rather were taken from Müller (1975, p. 185) who assigned them specific registers. The classification features where then assigned according to their sociolinguistic register as seen in figure 7.

	Feature	Document [2]
Training		
	LIT	Il faut partir , car il pleut .
	LIT	Elle m' a dit que j' étais une imbécile .
	ORAL	Vous dites quoi ?
	ORAL	Faut partir parce qu' il pleut .
	ORAL	Je n' sais pas .
Test	?	Vous dites imbécile

Table 3. Example Corpus
Examples adapted from Müller (1975,p.185)

	LIT	ORAL
Feature Count	2	3
Prior Probability	0.40	0.60
Smoothing	0.08	0.12

Table 4.Classification Values

	LIT	ORAL
Vous	0.08	0.33
Dites	0.08	0.33
Imbécile	0.5	0.12
Prior probability	0.40	0.60
Total probability	0.00128	0.0079

Table 5.Classification Assignment

Using these sentences as a small training corpus, as seen in table 3, it is possible to ascertain the most probabilistic classification of a sentence the *vous dites imbécile.*

The prior probability and smoothed values of the respective features must be ascertained from the corpus in table 3 using equation 12 and equation 15 respectively. There are five documents in total, with two being LIT and three being ORAL. With this information, equation 12 and 15 can be applied. The following results of the preliminary calculations can be seen in table 4.

[2] The spaces between the words are intentional as they indicate that the text has already been tokenized.

Combined with the values in table 4, the MLE values for the respective tokens can be calculated according to equation 13. If a given word does not occur in a specific class, then the respective smoothing probability from table 4 must be added. The MLE results are present in table 6.

The final step is simply to traverse the test sentence, as specified in table 3 and apply equation 11 by retrieving the respective values from table 6 and then to multiply the respective products by their respective prior probabilities. The result, as seen in table 5, shows that the sentence is most likely ORAL based on the training data from corpus in table 3.

Token	LIT	ORAL
,	0.5	0.12
.	1.0	0.67
?	0.08	0.33
Elle	0.5	0.12
Faut	0.08	0.33
Il	0.5	0.12
Je	0.08	0.33
Vous	0.08	0.33
a	0.5	0.12
car	0.5	0.12
dit	0.5	0.12
dites	0.08	0.33
faut	0.5	0.12
il	0.5	0.33
imbécile	0.5	0.12
j'	0.5	0.12
m'	0.5	0.12
n’	0.08	0.33
parce	0.08	0.33
partir	0.5	0.33
pas	0.08	0.33
pleut	0.5	0.33
que	0.5	0.12
quoi	0.08	0.33
qu’	0.08	0.33
sais	0.08	0.33
une	0.5	0.12
étais	0.5	0.12

Table 6. MLE Values

7. System Evaluation

7.1. Developmental Overhead

As was the case with the corpora (Gerstenberg & Hewett, 2019; Panckhurst et al., 2014; Poudat et al., 2015) 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.

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 system used to create training data could theoretically be retrained to recognize any language supported by Spacy (*French · spaCy models documentation*, n.d.). As for applying the algorithm to a domain other than conceptual literacy and orality, this would also heavily depend on the training data being supplied to the naïve Bayes.

Naïve Bayes is a relatively flexible algorithm that can be applied to a whole host of classification tasks, and the limitation does not lie necessarily within the program, but rather within the training data made available (Jurafsky & Martin, 2020). 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 (Jurafsky & Martin, 2020).

	Values
Accuracy	0.94
Error Rate	0.05
Precision	1.0
Recall	0.69
F-Score	0.82

Table 7. Evaluation of Training Classification Criteria for Literacy

	Values
Accuracy	0.91
Error Rate	0.08
Precision	0.77
Recall	0.31
F-Score	0.45

Table 8. Evaluation of Classification of Orality

	Values
Accuracy	0.94
Error Rate	0.05
Precision	0.89
Recall	0.50
F-Score	0.64
10-Fold Cross Validation	0.69

Table 9. Naïve Bayes Evaluation

7.2. Classification Sets and Naïve Bayes

The original classification sets were 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 in the data set.

The first classification set, as seen in table 1, considered features that were prevalent throughout texts which often exhibited conceptual literacy. These were weighted according to their prevalence and importance. Using these criteria, training data was created, labeled, and then evaluated. The results of this evaluation can be seen in table 7.

A second classification set, table 2, considered factors that often occurred in French texts expressing orality. This classification set was then tested and evaluated, the results of which can be seen in table 8. Using a separate sub data set within the development corpus, a training database was created. This database was then made available to the naïve Bayes algorithm. The results of this process can be seen in table 9. The results of table 7, table 8 and table 9 were ascertained by manually creating a gold reference file for the respective systems.

7.3. Sentence Tokenizer

	Accuracy
eBay	1.0
SMS	.95
Wikipedia	.94

Table 10. Sentence Tokenization Accuracy

Since the data is non-standard, it was not always clear which sentences should be parsed, or where they should be parsed. Data from all three corpora (Gerstenberg & Hewett, 2019; Panckhurst et al., 2014; Poudat et al., 2015) often lacked any meaningful punctuation, or punctuation was used incorrectly in that there was often reduplication of certain symbols to create an emphatic impression. This was especially true of the SMS corpus, where conservative definitions of sentences do not necessarily apply (Panckhurst et al., 2014). This includes beginning a sentence with capital letters or ending a sentence with punction such as a period, exclamation mark, or question mark (Bader, 2002; Rehm, 2002).

This resulted in sentences that were sometimes too long or too short, which skewed the results. Due to this, some sentences were added together that should have been split by the author. It was apparent from the data, such as eBay postings, that bullet points, rather than sentences were the intent of the author (Gerstenberg & Hewett, 2019). The decision was made to include bullet points as sentence markers as well. Dates and times were also seen as marking the end of sentences as many entries only contained such information (Gerstenberg & Hewett, 2019). There was no explicit regex expression that split sentences containing only numbers, but this was a result of the way the authors formulated their sentences.

7.4. spaCy Module

	Projected Accuracy	System Accuracy
Tokenization	1.0	1.0
POS	0.93	0.93
Dependency	0.96	0.99
Morphology	0.90	0.93

Table 11. spaCy Accuracy

The spaCy module was used for tokenization, part-of-speech tagging, syntactical dependencies, and assessing morphology (*French · spaCy models documentation*, n.d.). Tokens included punctuation and non-letter symbols as they were often essential in emoticons and reduplication. 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.

French · spaCy models documentation (n.d.) states that tokenization, part-of-speech, syntactical dependency, and morphology have an accuracy of 100%, 93%, 96%, and 90% respectively. These values align with the actual values obtained from a small data set of data from each development corpus set with a small deviation, the results of which can be seen in table 11.

The Wikipedia discussions and eBay postings were easily processed by spaCy with minimal errors. This was due in part to the authors in the texts following orthographic norms and not using non-standard language excessively (Gerstenberg & Hewett, 2019; Poudat et al., 2015). SMS chats often had incorrect spellings, made high use of emoticons, or created new unknown abbreviations (Panckhurst et al., 2014). However, emoticons were classified as punctuation, rather than as emoticons, which caused spaCy to perform poorly compared to the other data sets, but the values were still with an acceptable range.

8. Results

Using table 12 as an example, the corpus id represents the part of the corpus that was extracted from the respective corpus. 0 – 29507 is the span of possible documents within this corpus section. The number of documents pulled is 150, which consists of 3444 tokens. Finally, there 349 sentences within 150 documents, with 129 sentences being classified as conceptually literal, 218 being classified as conceptually oral, and 2 unknown sentences.

8.1. Development phase

Using the Wikipedia and SMS corpora (Panckhurst et al., 2014; Poudat et al., 2015), sentences were labeled according to the classification features mentioned in table 1 and table 2.

	Corpus ID	Documents	Tokens	Sentences	LIT	ORAL	UNK
SMS	sms_0_29507	150	3444	349	129	218	2
Wiki	wikiconflits_0_53	53	6766	345	234	110	1

Table 12. Development Results of the Classification Data

While creating the training data, the most relevant classification criteria were retrieved for Wikipedia, table 13, and for SMS, table 14 respectively.

Feature	Classification Criteria
LIT	SEN_LEN
LIT	PRES_TENSE
LIT	NP_VB_RATIO
ORAL	AVG_WORD_LENGTH
ORAL	ALL_CAPS
ORAL	SEN_LEN

Table 13. 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

Table 14. Top Development Classification Criteria for SMS

Sentence length, noun-to-verb-ratio, and average word length were decisive in determining the feature for the training data set for both corpora. After having acquired the training data using the classification set, it was entered into the naïve Bayes algorithm as training data. All four of the eBay sub-corpora were used as testing corpora. The results in table 15 show that all four of the eBay sub-corpora contain documents that exhibit a high level of conceptual literacy with a low rate of conceptual orality (Gerstenberg & Hewett, 2019).

	Corpus Id	Documents	Tokens	Sentences	LIT	ORAL	UNK
eBay	ebayfr-e05p_0_100	100	5800	380	361	8	11
eBay	ebayfr-e17p_0_100	100	6195	317	312	3	2
eBay	ebayfr-e17xp_0_100	100	21184	1028	995	32	1
eBay	ebayfr-e18v_0_100	100	9321	563	551	9	3

Table 15. Naïve Bayes Development Results

Even though all the corpora contained 100 documents (see table 15), the number of sentences and tokens contained within vary significantly. Despite this, they are uniform in the way conceptual literacy and orality are distributed across the data.

8.2. Training phase

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

The results of which mirror those of the development phase to a certain degree and can be seen in table 16. The documents in the Wikipedia corpus again displays a high level of conceptual literacy while SMS displays a high level of conceptual orality (Gerstenberg & Hewett, 2019; Poudat et al., 2015). As during the development phase, the top classification criteria were retrieved from and can be seen in table 17 and table 18.

	Corpus Id	Documents	Tokens	Sentences	LIT	ORAL	UNK
SMS	sms_29508_59014	255	4138	458	140	317	1
WIKI	Wikiconflits_54_106	52	8226	463	303	160	0

Table 16. Training Results of the Classification Data

Feature	Classification Criteria
LIT	SEN_LEN
LIT	AVG_WORD_LEN
LIT	NP_VB_RATIO
ORAL	SEN_LEN
ORAL	ALL_CAPS
ORAL	AVG_WORD_LENGTH

Table 17. Top Training Classification Criteria for Wikiconflits

Feature	Classification Criteria
LIT	SEN_LEN
LIT	NP_VB_RATIO
LIT	NOM_SUBJ
ORAL	SEN_LEN
ORAL	ALL_CAPS
ORAL	AVG_WORD_LENGTH

Table 18. Top Training Classification Criteria for SMS

These results do not differ drastically from those of the development phase. The process from the development phase was then repeated by retraining a new database with new training data created from the classification set as seen in table 1 and table 2. After that, the naïve Bayes was then tested again on the eBay corpus (Gerstenberg & Hewett, 2019).

	Corpus Id	Documents	Tokens	Sentences	LIT	ORAL	UNK
eBay	ebayfr-e05p_101_200	100	5225	315	283	32	0
eBay	ebayfr-e17p_101_200	100	6242	373	337	36	0
eBay	ebayfr-e17x_101_200	100	24477	1202	1112	89	1
eBay	ebayfr-e18v_0_100	100	9784	542	503	39	0

Table 19. Naïve Bayes Training Results

The results of the training phase, as seen in table 19, mirror those of the development phase as well. The documents in the eBay corpora display a high level of conceptual literacy with a low level of conceptual orality (Gerstenberg & Hewett, 2019).

8.3. Testing phase

	Corpus Id	Documents	Tokens	Sentences	LIT	ORAL	UNK
eBay	ebayfr-e05p_201_300	100	4063	249	229	20	0
eBay	ebayfr-e17p_201_300	100	4680	275	254	21	0
eBay	ebayfr-e17x_201_300	100	17155	922	830	92	0
eBay	ebayfr-e18v_201_300	100	9824	558	515	43	0
SMS	sms_59015_88522	250	3523	342	293	49	0
Wiki	wikiconflits_79_159	53	9172	487	441	46	0

Table 20. Naïve Bayes Testing Results

Using the training data created during the training phases as described in 8.2, the naïve Bayes was trained to assess conceptual literacy and orality of each corpus (Gerstenberg & Hewett, 2019; Panckhurst et al., 2014; Poudat et al., 2015). The results of which can be seen in table 20.

9. Discussion

9.1. Results of Classification Sets and Naïve Bayes

An earnest attempt was made at ascertaining reliable French examples of conceptual literacy and orality. 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 of the training data for the naïve Bayes. Müller (1975) offers readers prototypical texts of the respective French registers that can be mapped to conceptual orality and literacy. Despite this, it was the quantity, and not the quality of the texts, that proved to be a hindrance as Müller (1975) did not have enough training data for the naïve Bayes. Had more information been readily available by Müller (1975) or other similar sources, then less emphasis and time would have been placed on developing classification sets.

The classification sets relied heavily on naïve assumptions that often proved to be correct (see table 13, table 14, table 17, and table 18). More points were given to sentences that were longer, and fewer to sentences that were shorter. It was not uncommon for sentence length to be the decisive factor in determining conceptual literacy and orality. Sentences that were long tended to represent conceptual literacy as opposed to conceptual orality (see table 1). Upon manual inspection of the results, this turned out to be correct in most instances. However, sentence length was also highly dependent upon the user correctly using punctuation (Bader, 2002). If the author of the text incorrectly used punctuation, the sentence would be split prematurely and thus skewing the results.

The data between the development and the training phase was also relatively consistent. The documents in the Wikipedia corpus had high level of conceptual literacy as a lot of the discussions revolved around topics that were highly scientific and intellectual in nature (*Overview of Wikiconflits-QI from CoMeRe, n.d.*). This entails high word length and high sentence length as seen in table 17. When conceptual orality did occur, then it was only in short bursts or small statements.

The documents in the SMS corpus during classification were highly representative of orality for various reasons. The authors of the documents were very familiar with one another, and this was reflected in the language used by them. Intimate conversations as specified in figure 3 are representative of orality and nähesprache as specified in figure 4. Furthermore, there were a high number of pronouns, nouns, proper nouns, and redacted names[3].

Using the training data gathered using the classification set, the naïve Bayes was tested in multiple phases. It was initially only trained on the SMS and Wikipedia corpora, which were thought to have documents that displayed conceptual orality and literacy respectively (Panckhurst et al., 2014; Poudat et al., 2015). Upon analyzing the eBay corpus with naïve Bayes (Gerstenberg & Hewett, 2019), it was found to indeed have a high level of literacy, but a lower-than-expected level of orality (see table 15).

This process was repeated in the training phase (see table 19) and produced the same level of results. The unexpected high conceptual literacy in eBay data can be attributed to buyers and sellers using an imbalanced combination of both (Gerstenberg & Hewett, 2019). The eBay documents had features pertaining to distanzsprache and therefore features of conceptual literacy (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 by presenting part of their postings using conceptual literacy, and a blend of the two was thus inevitable (Gerstenberg & Hewett, 2019).

In the final portion, training data that was created by the classification system (see table 16) was used on all corpora portions (see table 20). The naïve Bayes showed that all the texts had a high level of conceptual literacy. While this does line up with most of the corpora, there were some deviations. The biggest deviation in the testing results those of the SMS data which shows a high level of conceptual literacy as opposed to conceptual orality (see table 20).

Typical punctuation such as periods, exclamation marks, and question marks were used emphatically rather than syntactically. That is to say that they were more often employed to express conceptual orality, rather than to mark the end of a sentence. Finally, many sentences lacked any coherent or predictable endings. This had the side-

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

effect of the program classifying sentences as being literal when they were not, as long sentence length is a sign of conceptual literacy in the texts.

9.2. Classification Set vs. Naïve Bayes

The use of classification sets was essential as it provided more control and more speed with respect to building a necessary training data. The naïve Bayes was then trained using this data and probabilistically assigned the conceptual literacy or orality feature to a given sentence in a document. This approach provided objective criteria by which a training database could automatically be built and then given to a probabilistic classifier.

The biggest advantage that a classification set has over the naïve Bayes is that the results do not become diluted as the training data grows. If the training data does not contain enough of a certain classification feature, then it logically follows that the naïve Bayes cannot assign a feature to a given document as the probabilities of doing so would be too low. This advantaged was visible in how sentences were correctly assigned their feature with respect to conceptual literacy and conceptual orality (see table 12, table 16). However, this did not prevent the training data from becoming slightly skewed and it developed a bias towards assigning conceptual literacy instead of orality and literacy equally (see table 20).

There are a few reasons as to why this bias exists. The first and foremost being that the training data was small and somewhat imbalanced. While every precaution was taken to ensure that the corpus was as balanced as possible, e.g., not testing and training on the same documents, setting aside a portion of each corpus, using the same number of documents, it was not possible to balance the training data in a way so that the naïve Bayes classifier can properly analyze documents of SMS corpus despite their non-standard nature.

To solve this imbalance, it would be worthwhile to employ a multinomial binary naïve Bayes, as this places more emphasis on the presence or absence of a term as opposed to its frequency (Jurafsky & Martin, 2020). The classification set does not suffer from this problem as it only considers what the qualities of the sentence being analyzed. Thus, it has nothing from which to remember probabilities from and can therefore not be influenced by imbalanced properties.

10. Conclusion

Since the aspects of conceptual literacy and orality can exist on a continuum, the French registers and their features were able to be grouped accordingly. This information was applied to create a scoring system to classify sentences automatically and prototypically according to their literate and oral discourse. This was the training data for the naïve Bayes classifier, which assigned the most probable feature to documents from the eBay, Wikipedia, and SMS corpora.

The results of the preliminary classification results showed that conceptual literacy is prominent throughout the documents of the Wikipedia and eBay corpora with conceptual orality being most prevalent in the SMS corpus. These results did not transfer over to the naïve Bayes classifier, which showed that there was a higher-than-expected bias towards conceptual literacy in the SMS corpus as opposed to the predicted orality.

Authors of the documents in the Wikipedia corpus limited themselves conceptually orality. This was due in part to the precision of intellectual and scientific discourse as set forth by FC and FL. However, rebuttals and follow-up questions were often expressed in terms of orality. In the documents of the eBay corpus, conceptual literacy was fairly dominate, not due to the FT or FC, but rather employing a blend of FF and FC to sell their wares to potential customers. Conceptual orality was often used to the give the potential buyer the feeling that they were being addressed by the author through the frequent use of capital letters and short descriptions.

While the texts of the authors in the other two corpora were often of a standard nature, the SMS corpus often lacked consistency with respect to orthographic conventions, e.g., improper use of punctuation, improper spelling, and neologisms. Furthermore, English and German were sometimes found within the documents of this corpora, which further skewed the results. Authors of the SMS documents often switched between conceptual literacy and conceptual orality within the same sentence, and thus making it hard to determine what the most appropriate feature of a given sentence should be. Initial results showed that the SMS authors preferred conceptual orality over literacy with the results of the naïve Bayes showing that there was preference for conceptual literacy. The bias of the system could be resolved by either introducing an algorithm less prone to bias, refining a non-probabilistic algorithm to recognize conceptual discourse or having native speakers build an appropriate training database.

If the goal is to assess conceptual literacy and orality independent of a language, then it would be worthwhile to further develop a non-probabilistic classification system. However, if the gain insight on the nature of discourse within a specific language, then a refined probabilistic algorithm with proper training data would be advantageous as it eliminates the need of a classification set. The bias present in the naïve Bayes system shows that conceptual literacy and orality are often much more difficult to define and determine than the medial orality and literacy of language due to the correlation of written vs. spoken and literate and oral discourse.

11. References

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

Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation. (n.d.). Crummy.Com. Retrieved August 17, 2021, from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Bieswanger, M., & Becker, A. (2008). *Introduction to English linguistics* (2nd ed.). UTB.

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S. J., Klabunde, R., & Langer, H. (2010). *Computerlinguistik und Sprachtechnologie* (3rd ed.). Spektrum Akademischer Verlag.

eBay petites annonces. (2020). La-Bank: Resources for Research and Teaching. https://www.uni-potsdam.de/langage/la-bank/ebay.php

French · spaCy models documentation. (n.d.). Spacy.io. Retrieved August 17, 2021, from https://spacy.io/models/frx

Gerstenberg, A., & Hewett, F. (2019). *A collection of online auction listings from 2005 to 2018 (anonymised)* [Data set]. University of Potsdam: LA-bank. 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*, 5–24. 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.). [Unpublished]. 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 (1985)*, *36*(1985), 15–43. http://dx.doi.org/10.15496/publikation-20410

Koch, P., & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik*, *35*(3), 346–375. https://doi.org/10.1515/zgl.2007.024

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*. Carl Winter Universitätsverlag.

Ng, H. T. (1997). Exemplar-based word sense disambiguation: Some recent improvements. *Second Conference on Empirical Methods in Natural Language Processing*, 208–213. 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

Overview of Wikiconflits-QI from CoMeRe. (n.d.). Ortolang.Fr. Retrieved August 17, 2021, from https://repository.ortolang.fr/api/content/comere/v3.3/cmr-wikiconflits/cmr-wikiconflits-qi_discu-tei-v1.html

Panckhurst, R. (2017). A digital corpus resource of authentic anonymized French text messages: 88milSMS—What about transcoding and linguistic annotation? *Digital Scholarship in the Humanities*, *32* (Suppl. 1), 92–102. https://doi.org/10.1093/llc/fqw049

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., & Verine, B. (2014). *88milSMS. A corpus of authentic text messages in French* (ISLRN 024-713-187-947-8; cmr-88milsms-tei-v1) [Data set]. ISLRN . https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1

Poudat, C., Grabar, N., Jin, K., & 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

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.

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.

Ort/Place, Date/Datum	Name
Kamen, 23.08.2021	Christopher Michael Chandler