Radboud Universiteit Nijmegen

Discovering the Linguistic Landscape of Music: An Investigation into Topic Variety between Songs Written in First and Second Language

by Stan Vermeeren

A Bachelor's thesis

International Business Communication – Radboud University Nijmegen

Supervisors: dr. Iris Hendrickx & dr. Frank van Meurs

8 June 2023

ABSTRACT

The current study analysed the effect of first and second language use on topic variety in songs. In this way, it researches if an artist sings about different topics when writing in a second language. By creating a corpus of both Dutch and English songs in the genres of pop and hip-hop written by native and second language speaking artists, six groups of songs were formed. Subsequently, all six groups' 100 most used content words were labelled as belonging to a topic based on the 21 semantic themes of the UCREL semantic analysis system (USAS). These categorizations have been compared to make a statement about the topic variety. Results showed that the Dutch group differed significantly from both English groups (English music artists who write in English and Dutch artists who write in English) when comparing the categorizations of the USAS model. Therefore the conclusion can be made that Dutch songs written by Dutch speakers differ in topic variety from English songs in general, although further research is needed to investigate if this is caused by first or second language use. However, L1 and L2 speakers of English did not vary in terms of topic variety in English songs, which means first and second language use had no influence on songwriting here.

Keywords: first language, second language, topic variety, song lyrics, corpus study

INTRODUCTION

From the tracks of *Suzanne & Freek* and *Flemming* to the music of *Antoon* and *Goldband*. Songs written in the mother tongue, are totally in again in the Netherlands and are dominating the Dutch music charts. Six of Holland's ten most streamed songs on Spotify, which can be considered the most common streaming platform, in 2022, were written in Dutch. Although three of them were remakes of older songs (the most listened to track *Vluchtstrook* from *Kriss Kross Amsterdam*, *Sigourney K*, and *Antoon* is even based on an English song), all six contained newly written Dutch lyrics (Spotify, 2022). Taking the views of music videos in the Netherlands on YouTube into concern, the popularity of songs in the Dutch language is practically undeniable: no less than seven of the most watched videos connected to music are in that language (Trouw, 2022). However, the representativeness of this data can be questioned as this list of music clips is highly influenced by the platform's characteristics. The average age of people who listen to music on YouTube is not that high and therefore the

top ten most viewed music videos reasonably contain children's songs (*Kinderen voor Kinderen*) and music created by YouTube stars (*Bankzitters*). However, when considering the history of Dutch music and the recent cultural 'taste patterns', there can be established that songs written in Dutch are, again, more and more 'accepted' and even on the rise in the national music industry (Van der Hoeven et al., 2015).

Nonetheless, the (revived) popularity of Dutch-written music does certainly not imply a decreased reputation of English-written music in the Netherlands. When the most listened to songs on Spotify and YouTube are deliberated once again, one can examine that a clear majority of the songs that are not in Dutch are songs written in the English language, most of the time by artists that speak English as their L1. In this way, English music for a long time and even today dominates the music charts in Holland (Van der Hoeven et al., 2015). What is striking, is the smarter share of English-written music by Dutch artists, at least when we consider the charts. Still, enough well known Dutch-speaking artists write in English, which means enough songs are available in this group. This observation is therefore not evidently generalizable for the entire music industry, although artists with Dutch nationality progressively seem to prefer writing songs in their native language. This first 'musical language choice' of musicians regularly results from coincidences or general unwritten rules (Grijp, 2003). An e example that is given by Grijp (2003), is when someone likes the music of the British artist Elton John and therefore starts writing songs in English as well.

However, it is more interesting to analyse certain 'language switches', which usually do not happen too often in artists' careers. This is more relevant because, contrary to the first 'musical language choice', this switch is usually not based upon random coincidences. In fact, different factors can initiate a switch from writing and singing in English to Dutch, or vice versa (Grijp, 2003). The shift to another singing language can be influenced by the comprehensibility and desired (interaction with the) target group of songs. The (de)limitation of the expressiveness one experiences when writing a song, however, seems to be the main reason to perform a language switch in music. Grijp (2003) described in his paper how lead singers of famous Dutch bands like *The Scene* and *De Dijk* explained that they had an adequately higher capacity to express themselves in their L1 (Dutch) compared to their L2 (English). Therefore, writing in their L1 allowed them to access a broader range of topics because it provided more possibilities to verbalize their feelings and emotions, and to personalize their texts even more (Grijp, 2003). In addition, songs written in Dutch are more accessible for people who are less capable of understanding English as L2. Writing Dutch songs as an artist can, therefore, cause a more intense connection with their fan base (Van der

Hoeven et al., 2015). These utterances about expressiveness suggest that Dutch artists, in general, have a higher topic variety in their songs when they write in their first language (Dutch) compared to their second language (English). Whether this is actually the case, will be further examined in this study.

Language and Thought

On a general language level, the statements made by Grijp (2003) are in line with the findings of Caldwell-Harris and Acacetin-Dinn (2009), to the extent that they provide evidence that emotions are experienced stronger in one's native language. However, the paper by Caldwell-Harris and Acacetin-Dinn (2009) focuses especially on the other side of the story, namely the emotions of the listener instead of the writer. In their study, they established native speakers of Turkish experienced stronger emotions when listening to phrases with an emotional charge and lies in their L1, than in English (L2). Although this is not specifically representative of language use in music, it does not contradict the suggestion that songwriters are capable to express a wider variety of (emotional) topics in songs when writing in their L1. In addition, when the 'reading' of texts is considered, people who read texts that are high in emotional load, such as the Harry Potter books, experience more intense and separate emotions when the text is in their native language than when it is written in a second language (Hsu et al., 2015).

In any case, thinking (which is a major factor in every listening or writing process, and the formation of emotions) is influenced by the language wherein thoughts occur (Akkermans et al., 2010; Chen & Bond, 2007; Ross et al., 2002). It has been proven that cultural associations, and therefore certain thinking patterns and one's morality, are heavily linked to the used language. In this way, Akkermans et al. (2010) found that Dutch students were confronted with this phenomenon called 'Language priming' when participating in a prisoner's dilemma experiment in English, which is their second language. During a prisoner's dilemma, two separate parties must simultaneously make a choice that affects the result of both. When both parties choose their 'dominant' option (the option with the highest possible outcome), for both sides the outcome will not be optimal. This dilemma, therefore, controls how cooperative-minded people are. Cooperative behaviour, in this case, will present the most acceptable results for both sides. However, the research found that L1 of Dutch made more competitive choices when participating in an English prisoner's dilemma than when deciding in their native language, even when these people had zero or less affinity with the English culture (Akkermans et al., 2010). This study, as well as the findings of Caldwell-

Harris and Ayçiçeği-Dinn (2009) and Hsu et al. (2015), once again support the suggestion that when music artists write a song in their L2, other (perhaps limited) topics and thinking patterns will be adopted than used in songs written in L1.

Alternative previous research, however, has provided information that is conceivably in conflict with the aforementioned assumption. That is, L1 speakers will be influenced by L2 or L3 languages even when performing tasks in a native context. Particular words, sentences, or phrases in their L1 will evoke associations in other languages. This implies that a person who has knowledge of more different languages is processing information without actively filtering it based on language (Van Hell & Dijkstra, 2002). These findings suggest that when an artist is writing a song, the difference in topic variety between their native and non-native language will not be that substantial, as producing text in L1 is always influenced by L2 and the other way around. In addition, when comparing Dutch as L1 to English as L2, the most (un)common words in both languages were similar (in an academic context). Although verbs in Dutch were more frequently used in the context of complex particular Dutch language phrases, these findings form a clear argument against the lack of expressivity in English as a second language (Tilstra & Smakman, 2017). The studies by Tilstra and Smakman (2017), and Van Hell and Dijkstra (2002) furthermore both especially focused on the comparison between expressiveness in Dutch compared to English. As this paper started with introducing the subject of the variety of Dutch music written by Dutch artists and English music written by Dutch artists, the prior studies are debatably even more relevant for this paper than earlier discussed research.

It must be admitted that the subjects in the discussed studies are not directly comparable. Some of the research covers the relationship between thinking and processing language, whereas other studies focus on the relationship between thinking and producing language. However, the comparison of L1 with L2 and the mutual influence of both languages is present in all elaborated studies. Therefore, this section demonstrates there are varied scientific perspectives that provide contrasting suggestions about the difference between L1 and L2 in the context of songwriting. On the one hand, previous research shows findings that could indicate the songs written in the first language of an artist, which is Dutch in this particular study, are marked by a different type of topic variety compared to songs written in the second language (English). Firstly, the studies of Caldwell-Harris and Acacetin-Dinn (2009) and Hsu et al. (2015) both display how emotions are experienced more intensely in L1. For Dutch songwriters, this could possibly mean they write their songs in Dutch about more emotional topics, with words that are more emotionally loaded. Next to this first hint of

variation in topic variety between songs written in L1 and L2, the study of Akkermans et al. (2010) adds the finding that people's thought processes are influenced by using English as a second language. The fact Dutch native speakers play the prisoner's dilemma in another way when it is presented in English could mean Dutch native songwriters write about different topics when writing in English. On the other hand, the studies of Tilstra & Smakman (2017), and Van Hell and Dijkstra (2002) provide another perspective, which is that first and second language use would have a minimal influence on topic variety within songs. This is because people seem to be influenced by their knowledge of a language regardless of whether they use that language at that moment (Van Hell & Dijkstra, 2002) and because a lack of expressivity when English is used as a second language, has proven to not be the case (Tilstra & Smakman, 2017).

The current study aims to give a definite answer about which of the two perspectives derived from previous research, corresponds best with reality.

Language in Music

When the topic of music is mentioned again, it is of high importance to keep in mind that the studies that have been specified until now, do not address the effect of 'language switching' on writing in general, let alone on writing songs in L1 versus L2. It, therefore, is of even higher importance to keep in mind the fact that musical texts differentiate from regular texts linguistically and stylistically. Kreyer and Mukherjee (2007) were confronted with this while analysing the Giessen-Bonn Corpus of Popular Music (GBoP). Although a pilot version of this music corpus was used, their study provided an outline of how corpus-linguistic research of language in songs can be performed. In addition, the study by Kreyer and Mukherjee (2007) presented an already clear image of how the general style of, in this case, pop music lyrics is slightly different from a general writing style. An internal variation of style within different pop songs was also indicated (Kreyer & Mukherjee, 2007). This illustrates one of the difficulties of researching music texts as they can barely be generalized. When style variation is already occurring between songs of an equal genre, the variation between songs of contrasting genres will be even broader. In addition, some genres seem to fit better into a certain language (which can be a foreign language) than others. While the so-called *levensliederen* (sentimental songs of life) are, because of their emotional content, much better fitting in Dutch, the English language has a higher artistic legitimacy for a genre like rock music. This 'language choice' based on genre depends on the style, topic, and target group

belonging to the genre concerned (Van der Hoeven et al., 2015).

The variation within genres is something that is researched in the study of Waszink et al. (2018), which analysed the most 'popular' words in Dutch pop and hip-hop music. Although in this case the comparison was not performed with the goal of making general claims about Dutch music per se, the marked claim that one music genre is not representative of the entire music language of a certain language is clearly underlined by the findings. Most used terms in Dutch hip-hop lyrics are words in slang that are focused on flexing with money, expensive stuff, and girls, like *euro*, *gap* (which means something like *dude*), and *cake* (used as a metaphor for money or the ass of a girl). Pop songs written in Dutch, on the other hand, especially contain longer and more intense words that more often than not are describing the topic of love, like *voorgoed* (for good), *verlangen* (desire), and *afscheid* (farewell). The research of Waszink et al. (2018) furthermore already gives a sharp first impression of the topic variety in Dutch music, which can be rather indirectly derived from the most frequently used words in the lyrics of both genres. However, the adoption of only a list of popular words, without context or alternative supporting data, is not sufficient to state a case about topic variety in music.

Corpus Research

The (discussed) previous studies that researched certain patterns in music, or more specifically; music lyrics, have predominantly applied corpus research to do so. In general, this type of empirical research is suitable for linguistic analysis as tangible and factual material is operationalized. Although researchers usually still have to compose the corpus, working with already existing material allows the creation of a larger sample, with which the accuracy of the study increases. In this way, a corpus analysis was performed by Tilstra and Smakman (2018) to make a comparison of the language use of Dutch university lecturers in lectures spoken in Dutch and lectures spoken in English. The corpus in this study comprised transcriptions of discourses in the corresponding languages lectured by nine different speakers (Tilstra & Smakman, 2018). In another linguistic research, a corpus of English texts written by Dutch L1 speakers was created to establish a measurement tool for the evaluation of the texts of L2 speakers and their improvement in a specific language (Verspoor et al., 2012). However, both Tilstra and Smakman (2018) and Verspoor et al. (2012) had to put in some time and effort to single-handedly generate a corpus that was suitable for their research design. Therefore, the benefit of researching music (lyrics) is the availability of already

existing corpora. In such a manner, researchers are in no need to collect data units, in this case, the lyrics of songs, themselves. The Giessen-Bonn Corpus of Popular Music, consisting of pop songs, which was analysed in the study of Kreyer and Mukherjee (2007), is an adequate example of a similar ready-made corpus. Another appropriate and frequently used music collection is the lyrics website *Genius.com*, also used by Waszink et al. (2018), which contains hundreds of thousands of lyrics of songs in all different genres and languages. All semi-popular tracks are transcribed and often equipped with an explanation of the text. Therefore, this website is optimal for studies that require a high number of songs for a sufficient sample.

The study of Hanser et al. (2022), can possibly give even more information about how to specifically research topic variety in music. Although the focus of their paper is just on how funeral songs differ from normal songs, it is interesting to learn how they analysed the lyrical characteristics of these songs. Already existing software was used to investigate certain 'word categories', including topics, which is an interesting approach for the current research. The themes *friends*, *family*, *death,* and *religion* were utilized, after which the computer assigned the words in the lyrics of funeral and normal songs to one of the themes (Hanser et al., 2022).

Analysing Topics

To analyse topic variety, it is helpful to create, or use an already available, framework to classify topics on word level. In this way, words in songs can be labelled as belonging to a certain topical category, and after that get compared to each other. A well-known and commonly used framework to categorize words is the UCREL semantic analysis system (short: USAS) (Wilson & Rayson, 1993). As this framework will be used to conduct the current study it will be described more extensively in this section. The USAS makes use of a semantic set of 'tags' to (automatically) label words in 'early modern English'. The model, which is originally created by Paul Ryson of the Lancaster University based on the Longman Lexicon of Contemporary English by Tom McArthur (McArthur, 1981) was introduced for the first time in a paper by Wilson and Rayson (1993). The system currently counts 21 semantic categories that are referred to with letters of the alphabet and that contain, in their turn, sub- and sub-sub-categories (Archer et al., 2002; Archer et al., 2003). The 21 main categories and their features can be found in table 1.

Table 1. An overview of the categories of the UCREL semantic analysis system (USAS), with their names, labels, and descriptions (Archer et al., 2002).

Name of category	Label of category	Description of category
General and Abstract terms	A	This category contains words with a more general or abstract meaning and therefore not really fit in one of the other categories.
The body and the individual	B	This category contains words that are related to everything around the (human) body, such as health and anatomy but also individual properties.
Arts and crafts	C	This category contains words that are related to an artistic process or to the crafting of objects.
Emotion	E	This category contains words that are related to different states of emotion, both positive and negative.
Food and farming	F	This category contains words that are related to consumer goods and how these are obtained. By

		consumer goods, food and drinks are meant, but also cigarettes and drugs.
Government and public	G	This category contains words that are related to the state, politics, law, and other governmental processes.
Architecture, housing, and the home	H	This category contains words that are related to everything with buildings, living spaces and architecture.
Money and commerce in industry	I	This category contains words that are related to everything with money, business and the industry that belongs to it.
Entertainment, sport, and games	K	This category contains words that are related to entertaining activities, such as sports, music, theatre, and other games.
Life and living things	L	This category contains words that are related to life and death and non-human beings.

Movement, location, travel, and transport	M	This category contains words that are related to (not) moving or travelling to another place, and certain locations and directions.
Numbers and measurement	N	This category contains words that are related to terms that be used to measure things.
Substances, materials, objects, and equipment	O	This category contains words that are related to all the features an object can have, such as shape, texture, material, and colour.
Education	P	This category contains words that are related to the general process of education.
Language and communication	Q	This category contains words that are related to the communication forms of language and speech, and the media through which those are communicated.
Social actions, states, and processes	S	This category contains words that are related to social actions people can perform (for example participation), and traits,

		relationships, and (religious) beliefs they have.
Time	T	This category contains words that are related to everything with the process of time, such as the beginning and end, and the age of things.
World and environment	W	This category contains words that are related to the world as a planet in the universe and its environment.
Psychological actions, states, and processes	X	This category contains words that are related to mental processes, such as knowing, feeling, and deciding.
Science and technology	Y	This category contains words that are related to science and (information) technology, and computing.
Names and grammar	Z	This category contains words that are related to names of for example persons and geographical locations, but also words that are related to grammar,

such as prepositions and pronouns.

The entire tagset can be found in part A of the appendix of this paper again.

The USAS framework was originally created to automatically analyse the semantic (which means the meaning on a word or sentence level) features of a text. This means an online tool, which can be used for scientific purposes, is available that automatically tags words. In the meantime, various online 'semantic taggers' are developed for other languages such as Finnish (Löfberg et al., 2003) and Russian (Mudraya et al., 2006). The USAS model is yet available in twelve different languages, including Dutch (Piao et al., 2016). However, this Dutch tool is less advanced and much more complex to use in comparison to its English counterpart and therefore will not be used in this study.

Current Research

Making use of a corpus analysis, which as discussed is a widely used approach in music research, this study will focus on the difference in topic variety between songs written in L1 and L2. With topic variety, the range of most used themes in a certain group, genre, or language is meant. This can be measured on the level of an entire song, but also on word level as will be done in this study. In doing so, certain overarching themes will be introduced to divide words, a bit similar to the approach of Hanser et al. (2022) in which words were labelled under funeral-linked themes. However, in this study, the framework of the earlier discussed UCREL semantic analysis system (USAS) will be used to categorize words. This model namely provides a clear and accessible categorizing of word topics, by grouping words on the base of semantic themes (Wilson & Rayson, 1993; Archer et al., 2002). These themes can be used as certain overarching topics and thereby the difference in topic variety between songs, genres, and languages becomes measurable. In addition, the USAS model is equipped with a handy and easy-to-use tool that can provide relevant results for this study in a fast and organized way.

With the help of the USAS framework and its corresponding, tool, the current research will respond to the theme of the revival of Dutch music and the remaining popularity of English songs in the Netherlands, and the motivations behind writing songs in one's first language rather than their second language. This will be done by analysing the difference in

topic variety between songs written by L1 and L2 speakers. In addition, insights into the influence of language on thought patterns and writing processes in a musical context will be provided. This approach will be carried out by examining the following research question(s):

RQ: What is the difference in topic variety between songs written by L1 speakers and songs written by L2 speakers?

(SQ1) What is the difference in topic variety between Dutch songs written by Dutch artists and English songs written by Dutch artists?

(SQ2) What is the difference in topic variety between Dutch songs written by Dutch artists and English songs written by English-speaking artists?

(SQ3) What is the difference in topic variety between English songs written by Dutch artists and English songs written by English-speaking artists?

These research questions are based on the research gap that derives from previous research. When reflecting on this previous research it can be concluded that many studies have investigated the relation between communicating and thinking in L1 and L2. It has been proven that emotions are experienced as stronger in one's native language, independent of the difference between sending (writing or speaking) and receiving information (reading or listening) (Caldwell-Harris & Acacetin-Dinn, 2009; Grijp, 2003; Hsu et al., 2015). In addition, the language in which thoughts are produced appeared to influence thinking patterns, as shown in the papers of Akkermans et al. (2010), Chen and Bond (2007), and Ross et al. (2002). When those findings are combined, one can conclude native and non-native communication certainly differ in terms of mental perception. Linking this to the introduced subject of 'language choice' in music, the suggestion can be made that songwriters will have different thought processes when writing in L1 or L2. This could result in the application of different topics in lyrics when writing in their second language.

On the other hand, previous studies that are in contrast with this claim were discussed. It turns out people barely filter information from various languages and are thus influenced by their L2 when performing communicative activities in an L1-only context, and the other way around (Van Hell & Dijkstra, 2002). Additionally, the study by Tilstra and Smakman (2018) shows that Dutch (L1) and English (L2) almost share the same commonly and uncommonly

used words, which could indicate that the difference in produced thoughts between both languages will not be that large. Again, this can relate to musical 'language choice', but this time it will provide a conflicting suggestion. When an assumption has to be made based on Tilstra and Smakman (2018) and Van Hell and Dijkstra (2002), a negligible difference between the range of topic variety in lyrics in L1 and L2 songs can be suggested.

Previous research thus could not give a definitive answer about whether there is a presence or absence of a difference in topic variety between writing in L1 our L2. Therefore, the current study will focus on this issue because in general, the effect of language on topic choice in writing seems to be an underexposed research subject. A lack of information especially occurs within the writing of song lyrics, which has been proven to be yet another style of writing than that of normal texts (Kreyer & Mukherjee, 2007).

By examining and comparing the most used words in Dutch hip-hop and pop lyrics, Waszink et al. (2018) studied a subject that is already closer to the core of the current paper. In that study, however, only the difference within one language, thus between genres, was analysed. Although it could be interesting to compare the findings with the results of their English counterpart study carried out by Daniels (2017), the most frequently used words in song lyrics do not give a clue about topic variety per se. The selected artists in this paper, moreover, only write in their L1, which makes the comparison less relevant for the current study.

METHOD

Materials

To assess the research questions a corpus analysis was performed. The corpus consisted of Dutch and English songs by Dutch and English artists in the genres of pop and hip-hop that were at least active between the years 2010 and 2019. This period was chosen, firstly because usually decades (60s, 70s, 80s, etc.) are used when referring to music in general and secondly this study is aiming to concentrate on the current (or in any case most recent) patterns in music. The *10s* is the decade that was most recently concluded and therefore a corpus of songs coming from this era was applied.

The corpus that was used consisted of six different groups based on three conditions: first language, lyrics, and genre. With 'first language', the mother tongue and/or most spoken language of an artist was meant. This indicates whether the musician speaks Dutch or English

as their L1. The second factor, ‘lyrics’, was intended to label the language a song was written in. Again, the options were Dutch or English, as Dutch artists are generally capable to write songs in both languages. The third and last factor, ‘genre’, focused on the music genre the respective song belonged to. The genres that were used were pop and hip-hop as these are both popular as contrasting and various music styles. This makes it possible to draw on a larger sample of songs. In fact, a fourth (sub)factor was created within ‘nationality’ to differentiate UK English from American English, as they differ regarding lexicon and spelling. However, during the final study, these two pieces of the corpus were blended into one. Dutch L1 speakers namely have been shown to use a mix of UK and American English when writing songs in the English language and therefore it is not necessary to adopt yet another factor. This means, nevertheless, that artists from other English-speaking countries, like Canada and Australia are also not included in the corpus. In fact, the music lists on Wikipedia, which were used as an encyclopaedia to select artists, did not contain British hip-hop artists at all. Because of that, only USA hip-hop artists were included in that group. In total, the three described factors resulted in six groups of songs: 1. Dutch hip-hop songs written by Dutch artists (nl_nl_hiphop), 2. Dutch pop songs written by Dutch artists (nl_nl-_pop), 3. English hip-hop songs written by Dutch artists (nl_en_hiphop), 4. English pop songs written by Dutch artists (nl_en_pop), 5. English hip-hop songs written by English-speaking artists (en_en_hiphop) and 6. English pop songs were written by English-speaking artists (en_en_pop). Although the application of two extra groups of Dutch hip-hop and pop songs written by English artists would definitely provide an even better answer to the research question, it is not conducted in this study. Indeed, the range of English L1 artists that write Dutch songs is simply too small to create a sufficient corpus. Artists that meet those requirements were in any case not found in the Wikipedia music encyclopaedia that was used for this study.

Per group, a representative simple random sample of 20 popular artists was composed, based on the Wikipedia music corpora, and complying with the above-elaborated requirements. This was computed out of a list of artists that included only the artists in the Wikipedia corpora causes that met the requirements. Singer-songwriters or bands that write their own music were selected in this respect, to exclude artists that perform songs written by another (song)writer, who could possibly have other topic preferences or a higher level of the L2 (English). However, on the Wikipedia lists, only three Dutch artists that write English hip-hop songs came to the surface. Therefore, this group of the corpus was considerably taller than the other groups.

For each artist, a certain number of songs were by means of a tool derived from the music lyrics website *Genius.com*. This study aimed to apply a number of 20 songs per artist for each corpus group to ensure comparable representativity. Unfortunately, that number of songs could not be achieved for every artist, as not every artist had lyrics of minimum 20 tracks available on Genius.

After these songs were derived, the corpus was manually reviewed to filter out errors. In this way, it was again checked if an artist genuinely wrote their songs and if a native speaker of either Dutch or English is the case. In addition, songs with featured artists and covers, reperformed songs that were originally written and performed by another artist, were excluded from selection as well. The features and complete sizes of those sample groups can be found in table 2.

Table 2. An overview of the corpus with the number of artists, songs, and words per group

	Artists *n*	Songs *n*	Words *n*
nl_nl_hiphop	17	198	89,140
nl_nl_pop	19	318	88,196
nl_en_hiphop	3	24	46,581
nl_en_pop	19	366	164,098
en_en_hiphop	17	265	46,581
en_en_pop	40	767	205,703
Total	115	1938	640,299

Procedure

To derive information from the corpus about topic variety in lyrics, several steps were taken. Firstly, a general picture of the most frequent words used in the different parts of the corpus had to be painted. This was done by using a word analyse tool created by the communication office Debatrix (Debatrix, 2022) that counted and created a list of the most common words for each of the six groups. The next step was to filter these lists to make sure only content words, meaning words with explicit meaning, were included. This was necessary as this paper focuses on topic variety, which can only be analysed if words have a certain topical meaning. With the use of highly regarded dictionaries (*Van Dale* for Dutch and *Cambridge Dictionary* for English), the nouns, verbs, and adjectives were manually selected out of the lists until a new list of the 100 most used content words in the relevant group was created. This is done for each of the six groups. Besides that, terms like *we're* (we are) and *I'm* (I am) were broken up and added to the total of the original words. This is about types rather than tokens. The wordlists can be found in Part B of the appendix of this paper.

To assign the coded words in the wordlist to a fitting topic category, the model of the earlier discussed UCREL semantic analysis system was used. This model, which contains 21 individual semantic categories, can be found in Part A of the appendix of this paper. For the English corpora, the lists that count 100 words were put into a tool corresponding with the USAS model. This tool automatically 'tagged' these words in a certain category, which gave an impression of how often a category occurred in the list. It is hereby about tokens rather than types. Unfortunately, the same tool was not capable of tagging Dutch words. Therefore, the two lists with Dutch words were labelled semi-manually. This means the Dutch words were tagged by comparing them with the tag of the corresponding English term. To check if the words were correctly judged in this way, a 'first' coder tagged 10 percent of both wordlists (20 in total: 10% of 100 + 10% of 100) without using the tool. The interrater reliability of the coding of the semi-manual tagged nl_nl_hiphop corpus between the tool and the first coder was satisfactory: $\kappa = .73$, $p < .001$, and the interrater reliability of the coding of the semi-manual tagged nl_nl_pop corpus between the tool and the first coder was also satisfactory: $\kappa = .80$, $p < .001$. In addition, a second 'external' coder, who was not involved in the study, coded 10 percent of the Dutch lists to test the subjectivity. The interrater reliability of the coding of the semi-manual tagged nl_nl_hiphop corpus between the tool and the second coder was moderate $\kappa = .47$, $p < .001$, as well as between the first and second coder $\kappa = .60$, $p < .001$. For the nl_nl_pop list, the interrater reliability between the second coder and both the

tool and the first coder was moderate as well $\kappa = .48$, $p < .001$, $\kappa = .48$, $p < .001$.

Although the validity and subjectivity of the categorization by the tool can be questioned, the decision was made to stick with the results of the tool for all the lists. In this way, one clear guideline was maintained regarding the labelling of the words. The last step was to compare the ranking of the nl_nl group with the nl_en group (to find an answer on SQ1), the nl_nl group with the en_en group (to find an answer on SQ2), and that of the nl_en group with the en_en group (to find an answer on SQ3). Firstly, separated based on the genres hip-hop and pop, after which the separation was lifted, and the songs were combined.

STATISTICAL TREATMENT

In total nine Chi-square tests were performed to analyse a difference between the categories of the USAS model and therefore a variation in topic variety between the corpora. Three tests were performed for the genre of hip-hop, three for the genre of pop, and three for a sum of these two genres to give an image of the songs in total. The results of these Chi-square tests are presented in the results section after which an answer could be formulated to the research question.

As this study works with relatively small sample sizes, six additional Fisher's Exact tests were performed as well to analyse a difference between the categories of the USAS model and therefore a variation in topic variety between the corpora in a more suitable way. However, these additional Fisher's Exact tests were not performed for the three groups that summed up the genres of hip-hop and pop. This was, because these sample sizes turned out to be not small enough for the Fisher's Exact test to compute a *p*-value. The results of the six executed Fisher's Exact tests are presented in the results section.

RESULTS

In this results section, the results of the performed research are presented in three separate parts: that of the songs with the genre hip-hop, that of the songs with the genre pop, and one part where these two genres are put together. This allows to only make a separation between language groups. For each part, a table with the frequency of the 21 topics of the USAS-model is shown, supported by the results of Chi-square and Fisher's Exact tests that provide a

picture of the similarity between the frequencies of the groups.

Hip-hop Songs

Table 3 shows the frequencies of how the 100 most common words in the music corpus for the genre hip-hop were labelled, on the base of the 21 topics of the USAS-model. In this way table 3 shows that, for example, 29 of the 100 words in the nl_nl_hiphop list were labelled as belonging to the category 'General/Abstract.' The three groups this corpus consists of were Dutch hip-hop songs written by Dutch artists (nl_nl_hiphop), English hip-hop songs written by Dutch artists (nl_en_hiphop), and English hip-hop songs written by English artists (en_en_hiphop). In general, the frequencies do not differ that much between the three groups. This is underlined by the conducted chi-square tests as well (which can be found below table 3).

Table 3. The occurrence frequency of the topics of the USAS-model in the wordlists of of the three corpora of the genre hip-hop

	Frequency of Topic		
Topic	In nl_nl_hiphop list	In nl_en_hiphop list	In en_en_hiphop list
General/Abstract	29	25	28
Body/Individual	5	6	8
Arts and Crafts	0	0	0
Emotion	2	4	4
Food and Farming	0	0	0
Government/Public	2	1	0

Housing/Home	0	0	0
Money/Commerce	4	0	1
Entertainment	0	4	4
Life and Living	1	1	2
Movement/Transport	13	13	7
Numbers/Measure	5	3	3
Substances/Materials	1	3	1
Education	0	0	0
Language/Comm.	2	4	4
Social actions	14	6	11
Time	6	5	8
World/Environment	1	2	1
Psychol. actions	13	14	9
Science/Technology	0	0	0
Names/Grammar	1	9	9
Total	100	100	100

A first Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (15) = 21.86, p = .112) when comparing the nl_nl_hiphop group with the nl_en_hiphop group. An additional Fisher's Exact test showed no significant relationship (p = .083) between the nl_nl_hiphop group and the nl_en_hiphop group as well. This indicates that there is no significant difference in topic variety between the 100 most common words of the Dutch hip-hop songs written by Dutch artists and the English hip-hop songs written by Dutch artists.

A second Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (15) = 20.03, p = .171) when comparing the nl_nl_hiphop group with the en_en_hiphop group. An additional Fisher's Exact test showed no significant relationship (p = .136) between the nl_nl_hiphop group and the en_en_hiphop group as well. This indicates that there is no significant difference in topic variety between the 100 most common words of the Dutch hip-hop songs written by Dutch artists and the English hip-hop songs written by English speaking artists.

A third Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (15) = 9.17, p = .868) when comparing the nl_en_hiphop group with the en_en_hiphop group. An additional Fisher's Exact test showed no significant relationship (p = .903) between the nl_en_hiphop group and the en_en_hiphop group as well. This indicates that there is no significant difference in topic variety between the 100 most common words of the English hip-hop songs written by Dutch artists and the English hip-hop songs written by English speaking artists.

In general, the determination can be made that the categorizing of the wordlists of the different groups within the genre of hip-hop do not significantly differ. This suggests that there is no variation in topic variety, for at least hip-hop songs, between a first and second language. However, this issue will be discussed in more detail in the conclusion and discussion sections of this paper.

Pop Songs

Table 4 shows the frequencies of how the 100 most common words in the music corpus for the genre pop were labelled, on the base of the 21 topics of the USAS-model. In this way

table 4 shows that, for example, 32 of the 100 words in the nl_nl_pop list were labelled as belonging to the category 'General/Abstract.' The three groups this corpus consists of were Dutch pop songs written by Dutch artists (nl_nl_pop), English pop songs written by Dutch artists (nl_en_pop), and English pop songs written by English artists (en_en_pop). In general, the frequencies do not differ that much between the three groups. This is underlined by the conducted chi-square tests as well (which can be found below table 4).

Table 4. The occurrence frequencies of the topics of the USAS-model in the wordlists of of the three corpora of the genre pop

	Frequency of Topic		
Topic	In nl_nl_pop corpus	In nl_en_pop corpus	In en_en_pop corpus
General/Abstract	32	33	30
Body/Individual	3	3	6
Arts and Crafts	0	0	0
Emotion	3	1	0
Food and Farming	0	0	0
Government/Public	0	0	0
Housing/Home	0	1	2
Money/Commerce	1	0	0
Entertainment	1	0	2

Life and Living	1	1	0
Movement/Transport	13	7	8
Numbers/Measure	3	4	3
Substances/Materials	2	4	2
Education	0	0	0
Language/Comm.	2	4	5
Social actions	12	6	10
Time	9	8	7
World/Environment	2	4	3
Psychol. actions	16	12	10
Science/Technology	0	0	0
Names/Grammar	0	8	12
Total	100	100	100

A first Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (15) = 16.96, p = .322) when comparing the nl_nl_pop group with the nl_en_pop group. An additional Fisher's Exact test showed no significant relationship (p = .234) between the nl_nl_pop group and the nl_en_pop group as well. This indicates that there is no significant difference in topic variety between the 100 most common words of the Dutch pop

songs written by Dutch artists and the English pop songs written by Dutch artists.

A second Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (15) = 24.89, p = .051) when comparing the nl_nl_pop group with the en_en_pop group. An additional Fisher's Exact test, however, showed a significant relationship (p = .016) between the nl_nl_pop group and the en_en_pop group. This indicates that there certainly is a significant difference in topic variety between the 100 most common words of the Dutch pop songs written by Dutch artists and the English pop songs written by English speaking artists. Although the frequencies are comparable in places, some striking differences are definitely present. In this way, the categories of 'Movement/transport' (13 versus 8) and 'Psychologic actions' (16 versus 10) occur more often in the list of the 100 most common words of the Dutch pop songs written by Dutch artists in comparison with that of English pop songs written by English artists. On the other hand the category of 'Body/Individual' appears double as much (6 versus 3) in the list of the en_en_pop group comparted to the nl_nl_pop group.

A third Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (14) = 9.06, p = .827) when comparing the nl_en_pop group with the en_en_pop group. An additional Fisher's Exact test showed no significant relationship (p = .895) between the nl_en_pop group and the en_en_pop group as well. This indicates that there is no significant difference in topic variety between the 100 most common words of the English pop songs written by Dutch artists and the English pop songs written by English speaking artists.

The results show a significant difference between, at least, two of the groups of the genre pop, which are those of nl_nl_pop and nl_en_pop. This suggests there is a variation in topic variety between Dutch pop songs written by native Dutch speakers and English pop songs written by native English speakers. However, because both speakers in this condition are first language speakers, this not necessarily indicates an influence of first and second language competency on topic variety. This assumption is enhanced by the fact the other two conditions (nl_nl_pop versus nl_en_pop and nl_en_pop versus en_en_pop) do not significantly differ. This issue will be discussed in more detail in the conclusion and discussion sections of this paper.

Songs in General

When the genre-related barriers are raised, and the corpora of hip-hop and pop songs are combined, an utterly different set of results is the outcome. Table 5 shows how often the 21 topics of the USAS-model occurred in the lists of the 200 most common words in the three language groups. Table 5 shows the frequencies of how the 200 most common words for the three language groups were labelled, on the base of the 21 topics of the USAS-model. In this way table 5 shows that, for example, 61 of the 200 words in the nl_nl list were labelled as belonging to the category 'General/Abstract.' This groups, based on the language of the song and the original language of the performing artist, are as follows: Dutch songs written by Dutch artists (nl_nl), English songs written by Dutch artist (nl_en), and English songs written by English artists (en_en). In contrast with when the genres of hip-hop and pop were separated, these three groups show quite some difference regarding the occurrences of the topics within the word lists. This is stressed and clarified by the conducted chi-square tests as well (which can be found below table 5).

Table 5. The occurrence frequencies of the topics of the USAS-model in the wordlists of of the three language groups

	Frequency of Topic		
Topic	In nl_nl corpus	In nl_en corpus	In en_en corpus
General/Abstract	61	58	58
Body/Individual	8	9	14
Arts and Crafts	0	0	0
Emotion	5	5	4
Food and Farming	0	0	0

Government/Public	2	1	0
Housing/Home	0	1	2
Money/Commerce	5	0	1
Entertainment	1	4	6
Life and Living	2	2	2
Movement/Transport	26	24	15
Numbers/Measure	8	7	6
Substances/Materials	3	4	2
Education	0	0	0
Language/Comm.	4	8	9
Social actions	26	12	21
Time	16	13	15
World/Environment	3	6	4
Psychol. actions	29	26	19
Science/Technology	0	0	0
Names/Grammar	1	17	21

Total	200	200	200

A first Chi-square test showed a significant relation between music group and word tag ($\chi 2$ (16) = 32.20, p = .009) when comparing the nl_nl group with the nl_en group. This indicates that there is a significant difference in topic variety between the 200 most common words of the Dutch songs written by Dutch artists and the English songs written by Dutch artists. The most striking differences are for the topic categories 'Social actions' and 'Names/Grammar'. No less than 26 words in the word list of Dutch songs were labelled as belonging to the category 'Social Actions', whereas the list of English songs written by Dutch artists does not contain even half of that amount (12). On the other hand, the Dutch corpus includes only one word that was labelled as a name or grammatical word. Very little in comparison with the 17 'Names/Grammar' words in the corpus of English songs by Dutch artists. In addition, the distribution within the category Money/Commerce (5 – 0) is seen as a significant difference.

A second Chi-square test showed a significant relation between music group and word tag ($\chi 2$ (16) = 38.19, p = .001) when comparing the nl_nl group with the en_en group. This indicates that there is a significant difference in topic variety between the 200 most common words of the Dutch songs written by Dutch artists and the English songs written by English-speaking artists. Although the frequencies are slightly comparable, the largest difference again occurs in the category 'Names/Grammar'. As indicated, the corpus of Dutch songs only contained one word in that category compared to the 21 words of the English-English corpus.

A third Chi-square test showed no significant relation between music group and word tag ($\chi 2$ (16) = 12.25, p = .726) when comparing the nl_en group with the en_en group. This indicates that there was no significant difference in topic variety between the 200 most common words in the English songs written by Dutch artists and the English songs written by English-speaking artists.

The results show a difference between, at least, some of the corpora when the genre factor is eliminated. The wordlist of Dutch songs written by Dutch artists differs significantly from both the lists of English songs by Dutch artists and English songs by English-speaking artists. The two groups of English songs do not differ significantly. This, however, does not

detract from the fact the suggestion can be made that a variation in topic variety exists between the first and second languages. That could indeed be the consequence of a difference in language competence but could be influenced by a distinction of specific language features. This issue will be discussed more extensively in the conclusion and discussion section of this paper.

CONCLUSION

The results that are derived from the research contribute to the formulation of an answer to the following research question:

> **RQ: What is the difference in topic variety between songs written by L1 speakers and songs written by L2 speakers?**

With topic variety, the distinctive range of commonly used themes is meant. It is thereby important to emphasize that this study is limited to the popular genres of hip-hop and pop and not L1 and L2 songs in general, as the research question suggests. Before a statement can be made about this question, a more in-depth examination of the sub-questions is necessary. When the answers to those questions are put together, a general conclusion about this paper can be made.

(SQ1) What is the difference in topic variety between Dutch songs written by Dutch artists and English songs written by Dutch artists?

When the genre of hip-hop is first taken into consideration, no (significant) difference in topic variety is found between Dutch songs written by Dutch artists and English songs written by Dutch artists. The wordlists of both groups are too similar, to speak of a variation in topic variety. Secondly, the genre of pop was analysed. Here too, a (significant) difference in topic variation between Dutch songs written by Dutch artists and English songs written by Dutch artists variety was not found. The wordlists of both groups are again too similar, to speak of a variation in topic variety.

However, when the corpus is no longer divided into the genres of hip-hop and pop, an effectively (significant) difference in topic variety is found. This means that, based on this study, it can be concluded that Dutch songs written by Dutch artists and English songs written by Dutch artists differ regarding topic variety. The groups mainly vary in the use of words that describe social actions, states, and processes, which are used more often in Dutch songs,

and words that are labelled as names or have a grammatical function. This last category was present more frequently within the English songs written by Dutch artists.

(SQ2) What is the difference in topic variety between Dutch songs written by Dutch artists and English songs written by English-speaking artists?

When here the genre of hip-hop is first taken into consideration, no (significant) difference in topic variety is found between Dutch songs written by Dutch artists and English songs written by English-speaking artists. The wordlists of both groups are too similar, to speak of a variation in topic variety. Secondly, the genre of pop was analysed. Here, however, a (significant) difference in topic variation between Dutch songs written by Dutch artists and English songs written by English-speaking artists variety is present. This means that, based on this study, it can be concluded that Dutch pop songs written by Dutch artists and English songs written by English artists differ regarding topic variety. The Dutch pop songs in general seem to contain more words that are related to movement and travelling, and social actions, states, and processes, which are amongst others linked to relations and religion.

In addition, when the corpus is no longer divided into the genres of hip-hop and pop, again an effectively (significant) difference in topic variety is found as well. This means that, based on this study, it can be concluded that Dutch songs written by Dutch artists and English songs written by English-speaking artists differ regarding topic variety. The groups, again mainly vary in the use of words that are labelled as names or have a grammatical function. This category was present more frequently within the English songs written by English-speaking artists. In addition, English-speaking artists seem to write more often about the body and the human as an individual. On the other hand, Dutch artists are more likely to write about the topic of movement and transport in Dutch songs.

(SQ3) What is the difference in topic variety between English songs written by Dutch artists and English songs written by English-speaking artists?

The answer to this third and last sub-question can be formulated shortly. Both for the music genres of hip-hop and pop, as well as for the combined corpus of these two genres, no (significant) difference in topic variety is found.

Now the sub-questions are discussed an answer to the research question of this paper can be formulated. This research question concerns the issue of whether songs written by an L1 speaker differ regarding the topic variety from songs written by L2 speakers. However, after the research that was performed in this study, it is not certain if topic variety is per

definition influenced by first and second language use.

The categorizing of the wordlists of the nl_nl and nl_en groups is significantly different in any case. This could suggest a difference in topic variety, because of second language competency. The significant difference between the wordlists of the nl_nl and en_en corpora rather emphasise, than contradict this. That is because it, in the same way, suggests a distinction in topic variety between Dutch and English songs, which could be caused by the fact the artists in both groups are native speakers.

However, the fact the nl_en (of second language speakers) and en_en (of first language speakers) corpora do not differ, questions that suggestion about topic variety. When the Dutch group differs from the two English groups, but the songs of non-native speakers of English do not differ compared to those of native speakers of English, it could be that the difference in topic variety between the Dutch and the English corpora is only caused by a variation in the features of both languages, rather than whether it is spoken as a first or second language. Dutch and English are simply unique languages that both have their own characteristics. Therefore, it is not crazy to assume some topics are more likely to occur in Dutch than in English and vice versa, regardless of whether the spoken language is the native language of the speaker. If this is really the case could not yet be confirmed nor invalidated, because this paper only detected a difference without researching its cause.

In summary, the research question could be answered by saying that based on this study songs by written Dutch L1 speakers differ from Dutch L2 speakers of English in terms of topic variety, although it is not certain if this is really influenced by first and second language speaking. What, based on this study, can indeed be said with certainty, is that, at least in songs written the English language, no difference in topic variety between L1 and L2 speakers is present.

DISCUSSION

As expected, this paper has given answers but also raises questions. The main answer of the research performed in this study is that a difference in topic variety between English L1 and L2 speakers in songwriting (in the genres of hip-hop and pop) does not exist. This finding, therefore, invalidates the suggestion that topic variety could differ between first and second language speakers, which was based on the studies of Akkermans et al., (2010), Caldwell-Harris and Ayçiçeği-Dinn (2009), Chen and Bond, (2007), Hsu et al. (2015), Ross et al.,

(2002). Because of this, the finding seems to be in line with some of the earlier discussed previous research about language and thought that caused the suggestion that topic variety did not exist between L1 and L2 speakers. However, this is not completely the case. Tilstra and Smakman (2017) showed in their paper that people are always influenced by ideas in other languages they speak. This would mean Dutch artists who write in English would get influenced by their mother tongue and therefore a higher difference with the native English speakers who write in English (that are not influenced by the Dutch language) would be expected.

The other finding of this study, that Dutch songs differ from English songs in terms of topic variety, regardless of the origin of the writer of the English songs, is in contrast with the study of Tilstra and Smakman (2017) as well. As said, that paper showed people's thoughts are influenced by their knowledge of other languages, which argues against a variation of topic variety between Dutch songs written by Dutch speakers and English songs written by Dutch speakers. This suggests that when Dutch artists write in English, they are influenced by their knowledge of Dutch as well so a difference in topic variety with writing in Dutch is minimal. That is, however, not the case. In addition, this finding is contrary to the study of Van Hell and Dijkstra (2002) that described a negligible difference in expressiveness between native Dutch speakers that produce Dutch speech and English speech, which incorrectly suggested that a difference between Dutch and English song written by native Dutch speakers in terms of topic variety will be absent as well.

That Dutch artists, according to this study, differ in topic variety when writing in Dutch or English is in line with the studies of Grijp (2003), Caldwell-Harris and Ayçiçeği-Dinn (2009), and Hsu et al. (2015) that showed emotions are stronger experienced in a first language and therefore suggested topic variety (remarkable is, nevertheless, that the results show that this difference is actually not caused by a large variation in words of the category 'emotion'). As described in the conclusion it could not be confirmed if this difference is actually caused by first and second language use. The fact that Dutch songs differ from English songs in general (regardless of if they are written by an L1 or L2 speaker) could mean that this variation of topic variety could also be caused by a lot of other factors. For example, Language priming could have influenced the topic variety, which would be in line with the findings of Akkermans et al. (2010). In their study, they showed how Dutch students made other decisions in a prisoner's dilemma game when playing in English instead of Dutch.

It is independent of the aim of this study, interesting to look at the frequencies in the general categorizing of the corpus groups that was based on the USAS model. It can, for

example, be seen that the categories ‘General/Abstract’, ‘Movement/Transport, ‘Social actions’, and ‘Psychologic Actions’ are highly present in every group. On the downside, not one word was labelled in the categories ‘Arts and Crafts’, ‘Food and Farming’, ‘Education’, and ‘Science/Technology’. The absence of this categories does not seem to limit the answering of the research question, unlike a couple of other factors that could possibly have influenced and limited the results that were derived. Firstly, the way the corpus was created had an impact on this. Wikipedia was used as an encyclopaedia for music artists to select artists that meet the requirements that were determined. However, Wikipedia will not always be complete in its information and therefore some errors could have occurred in the selection of artists. This would mean artists might incorrectly or might not be correctly selected, because of vague or incomplete information on their Wikipedia page. The second problem that occurred because of the use of Wikipedia as a music collection was the variation in corpora sizes. The aim was to get at least twenty artists per genre-language group selected out of a list of fitting artists by random sampling. For the group of English hip-hop songs nevertheless, only three artists were present in the lists of Wikipedia. In addition, for the English-speaking groups, the goal was to select both British and American artists. However, the Wikipedia collection contained exactly zero British hip-hop artists, so that only hip-hop musicians from the USA were included in the corpus. This limited availability of artists could have affected the validity of the total corpus and therefore the study.

After the artists were selected, a tool was used to derive lyrics of songs from the website Genius. This tool did not filter those lyrics for covers, songs that featured other artists, or songs that were not in the right language at all. Therefore, the entire selection of songs had to be manually checked. However, not all the required information to know if a song meets the requirements could be found on the internet or in other ways. Because of this, no guarantee can be given, that the corpora only contained songs that meet the requirements.

The next step was to (manually) create wordlists of the most common words for each group that consisted of nouns, verbs, and adjectives (content words). Although highly regarded dictionaries were used, this still is a slightly subjective activity, because words had to be judged one by one. However, because this was done for every group in the corpus, this task did not cause differences in the wordlists.

The last factor that limited the findings of this study is the (non)correctness of the judgment of the used USAS tool, which tags words with certain themes. Although this tool provides a scientifically based topic categorization, its assessment of themes of words is not flawless. This is especially caused by the fact that words could have more than one meaning.

However, to draw a clear line, only the first (superior) topic that was tagged by the tool was taken into consideration. For example, a word like ‘am’ was labelled in the category time (after midnight), while in the songs it was mostly used as a verb as in ‘I am’. In addition, the tool turned to out to have problems with tagging more informal or urban words.

This high susceptibility for interpretation within a language also caused some problems regarding the tagging of the Dutch corpora. The tool that was used tagged only English words and therefore the lists of Dutch words had to get tagged semi-manually. This includes the labelling of words by means of translations that were put in the tool or compared to the already available tagged English words. After this, the executor of this study (as a second coder) manually checked the reliability of the tool, which was done by classifying ten percent of the words in one of the 21 categories of the model, purely based on intuition and without the help of the tool. This provided satisfactory intercoder reliability. When an external third coder tagged the words based on the USAS model, the intercoder reliability nevertheless was only moderate. This could among others be a result of the subjectivity of language. It can, however, also be a consequence of the fact this third coder was, unlike the second coder, unfamiliar with the restrictions and the way of categorizing that was used by the tool. Therefore, the instructions that were given to this coder should have been more concrete with less space for interpretation.

Other researchers that are willing to recreate a similar kind of study, would be well advised to keep the limitations that appeared in the current research in mind. In that way, an improved version of the method described in this paper could be used to find new or additional results that provide the possibility to make even more extensive statements about the influence of first and second language on topic variety in music. Further research could for example dive deeper into the question if the difference in topic variety between Dutch songs written by Dutch artists and English songs written by both Dutch and English artists (as found in this paper) is really influenced by first and second language use.

To conclude, this study delivered some new insights on the general topic of language and the influence of first and second language use on topic variety. It discussed previous studies that handled the influence of L1 and L2 and talked about how research regarding music lyrics can be computed. In this way, it adds to other previous corpus studies like that of Tilstra and Smakman (2018), Verspoor et al. (2012), Kreyer and Mukherjee (2007), Waszink et al. (2018), and Hanser et al. (2022). In addition, a blueprint was sketched of how topic variety in songs could be further assessed when some changes to the method are made. Lastly, this paper showed the usefulness of the UCREL semantic analysis system (Wilson & Rayson,

1993) and the corresponding tool. The framework had proven to be an interesting and suitable model for this kind of research, but it has been found in this paper that there are a couple of snags present as well. This study, therefore, adds new things to previous research that was done in this scientific field.

LITERATURE

Akkermans, D., Harzing, A. W., & Van Witteloostuijn, A. (2010). Cultural Accommodation and Language Priming. *Management International Review*, *50*, 559–583. https://doi.org/10.1007/s11575-010-0053-0.

Archer, D., McEnery, T., Rayson, P., & Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. *Proceedings of the Corpus Linguistics 2003 conference*, *16*, 22-31.

Archer, D., Wilson, A., & Rayson, P. (2002). Introduction to the USAS category system. *Benedict project report*.

Caldwell-Harris, C. L., & Ayçiçeği-Dinn, A. (2009). Emotion and lying in a non-native Language. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *71*(3), 193–204. https://doi.org/10.1016/j.ijpsycho.2008.09.006.

Chen, S. X., & Bond, M. H. (2007). Explaining language priming effects further evidence for ethnic affirmation among Chinese-English bilinguals. *Journal of Language and Social Psychology*, *26*(4), 398–406. https://doi.org/10.1177/0261927X07306984.

Daniels, M. (2017). *The Language of Hip Hop*. The Pudding. https://pudding.cool/2017/09/hip-hop-words/

Debatrix. (2022, 26 Augustus). *Woordanalyse: welke woorden gebruik jij veel?*. https://www.debatrix.com/nl/tools/woordanalyse-welke-woorden-gebruik-jij-het-meest/

Grijp, L.P. (2003). Eigenheimers en meezingers. De muzikale taalkeuze van Nederland. In J. Stroop (Ed.), *Waar gaat het Nederlands naar toe? Panorama van een taal* (pp. 45-52). Amsterdam: Bert Bakker.

Hanser, W. E., Mark, R., E. & Vingerhoets, A. J. J. M. (2022). Music and lyric characteristics of popular dutch funeral Songs. *OMEGA - Journal of Death and Dying*. https://doi.org/10.1177/00302228221075471.

Hsu, C. T., Jacobs, A. M., & Conrad, M. (2015). Can Harry Potter still put a spell on us in a second language? An FMRI study on reading emotion-laden literature in late bilinguals. *Cortex; a*

Journal Devoted to the Study of the Nervous System and Behavior, *63*, 282–295. https://doi.org/10.1016/j.cortex.2014.09.002.

Kreyer, R., & Mukherjee, J. (2007). The style of pop song lyrics: A corpus-linguistic pilot Study. *Anglia - Zeitschrift für englische Philologie*, *125*(1), 31-57. https://doi.org/10.1515/ANGL.2007.31.

Löfberg, L., Archer, D., Piao, S., Rayson, P., McEnery, T., Varantola, K., & Jantunen, J. P. (2003). Porting an English semantic tagger to the Finnish language. *Proceedings of the Corpus Linguistics 2003 conference*, *16*, 457-464.

McArthur, T. (1981). Longman Lexicon of Contemporary English. http://ci.nii.ac.jp/ncid/BA00415859

Mudraya, O., Babych, B., Piao, S., Rayson P., & Wilson, A. (2006). Developing a Russian semantic tagger for automatic semantic annotation. *In proceedings of Corpus Linguistics 2006.*

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R. M., Knight, D., Kren, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P. L., & Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. *In proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016),* 2614-2619.

Ross, M., Xun, W. Q. E., & Wilson, A. E. (2002). Language and the Bicultural Self. *Personality and Social Psychology Bulletin*, *28*(8), 1040–1050. https://doi.org/10.1177/01461672022811003.

Schipper, N. (2022, 13 December). Boos-uitzending over The Voice dit jaar het meest bekeken op YouTube. *Trouw*. https://www.trouw.nl/cultuur-media/boos-uitzending-over-the-voice-dit-jaar-het-meest-bekeken-op-youtube~bd2c36bf9/?referrer=https%3A%2F%2Fwww.google.com%2F

Spotify. (2022). *Top Tracks 2022 NL*. https://open.spotify.com/playlist/37i9dQZF1DWZ58KqXkQf1S

Tilstra, K., & Smakman, D. (2018). The spoken academic English of Dutch university lecturers. *English Studies*, *99*(5), 566–79. https://doi.org/10.1080/0013838X.2018.1483620.

Van der Hoeven, A., Janssen, S., & Driessen, S. (2016). Articulations of identity and distinction: The meanings of language in dutch popular Music. *Popular Music and Society*, *39*(1), 43–58. https://doi.org/10.1080/03007766.2015.1061344.

Van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review*, *9*(4), 780–789. https://doi.org/10.3758/BF03196335.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263. https://doi.org/10.1016/j.jslw.2012.03.007.

Waszink, V., Reuneker, A., & Van der Wouden, T. (2018). *Als ik praat, dan praat ik money: De hiphopste woorden*. Neerlandistiek. https://neerlandistiek.nl/2018/07/als-ik-praat-dan-praat-ik-money/.

Wilson, A., & Rayson, P. (1993). Automatic Content Analysis of Spoken Discourse. *Corpus Based Computational Linguistics*, 215-226.

APPENDIX A – UCREL Semantic Analysis System (USAS)

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and public	**H** Architecture, housing and the home	**I** Money and commerce in industry
K Entertainment, sport and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Language and communication	**S** Social actions, states and processes
T Time	**W** World and environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammar			

APPENDIX B – The lists of most used words + frequencies for each corpus group

NL_NL_HIPHOP

Word	Frequency
Ben	881
Heb	501
Wil	435
Kan	398
Al	360
Doe	335
Weet	295
Alleen	265
Bent	261
Moet	244
Gaat	225
Ga	205
Gaan	196
Leven	194

Echt	184
Kom	184
Laat	180
Money	170
Tijd	161
Kijk	137
Had	133
Zeg	130
Lekker	127
Zie	124
Doen	122
Wordt	122
Goed	120
Weg	117
Man	115
Dag	113
Zit	113
Bitch	110

Maken	110
mensen	109
Geef	108
Net	108
Hoor	103
Hele	100
Zien	97
Alle	96
Blijven	96
Hou	95
Komt	95
Up	93
Maak	92
Zeggen	91
Maakt	90
Zou	88
Hoofd	86
Blijf	84

Bom	84
Jongen	83
Weten	82
Eens	80
Heeft	80
Wereld	79
Kleine	78
Lang	78
Denk	76
Zet	76
Fuck	75
Ogen	75
Shit	74
Doet	73
Willen	72
Liever	70
mannen	70
Even	69

Heel	67
Voel	67
Hart	66
Zegt	66
Papa	65
Vind	65
Dingen	64
Neem	64
Blijft	63
Komen	63
Dom	62
Pillen	62
Kunnen	61
Licht	61
Hebt	59
Geld	57
gemaakt	57
Sta	57

Één	57
Anders	56
Rennen	56
Ziet	56
dromen	55
Dagen	54
Gek	54
Werk	54
Beter	53
Schat	53
Wacht	53
Bang	52
Ging	52
Liefde	52

NL_NL_POP

Word	**Frequency**
Ben	740
weet	424

Heb	423
Laat	421
Wil	407
Kan	390
Al	372
gaan	318
alleen	287
bent	267
even	231
Kom	209
leven	208
Weg	208
Tijd	207
komt	205
moet	196
Ga	187
Zie	183
Gaat	176
goed	170
denk	169
zou	164
doe	157
hart	156

hou	150
had	143
hoofd	143
liefde	142
meiden	141
dag	140
doen	134
lekker	129
ogen	126
kijk	122
wordt	120
anders	118
zeg	118
zit	115
staan	114
wereld	109
eens	108
echt	102
liever	102
los	102
doet	101
zeggen	100
voel	99

zal	99
nacht	98
zien	98
heeft	97
lang	96
blijf	95
zon	92
mis	84
zegt	83
heel	81
mag	79
beter	78
samen	78
mooi	76
morgen	76
neem	75
staat	75
hoop	74
licht	73
hand	72
hebt	72
laten	72
man	72

voelt	72
vergeten	70
kon	68
Baby	67
Lijkt	67
maakt	66
kunnen	64
houdt	63
mensen	63
Lief	62
Zei	62
Ligt	61
Plek	61
Dans	60
Stil	60
ander	59
Loop	58
Geeft	57
geweest	57
willen	57
best	56
blijft	56
ging	56

lopen	56
wacht	56
geef	55
kwijt	55
weten	55
zag	55

NL_EN_HIPHOP

Word	**Frequency**
am	135
Do	97
Is	92
got	80
Are	77
Know	70
All	68
See	66
Be	59
Go	57
Let	57
On	55
Can	53
Like	51

Up	45
Gotta	38
Time	35
Day	32
Gonna	32
Have	31
Will	30
favorite	29
Just	28
Might	28
Need	28
Way	26
Gone	24
Love	24
Get	23
choice	21
Down	20
Fuck	20
Song	20
wiggle	20
Right	19
Say	19
burning	18

things	18
breathe	17
light	16
could	15
make	15
please	15
really	15
same	15
shit	15
baby	14
choose	14
hear	14
keep	14
place	14
wanna	14
been	13
back	12
feels	12
forgive	12
hope	12
Left	12
Put	12
space	12

days	11
Give	11
gold	11
loved	11
crazy	10
dance	10
little	10
made	10
soul	10
Try	10
would	10
beat	9
feel	9
guess	9
hands	9
life	9
play	9
alone	8
eyes	8
going	8
Heart	8
Leave	8
Mind	8

only	8
step	8
take	8
tell	8
trying	8
wave	8
change	7
matter	7
new	7
said	7
stop	7
thing	7
want	7
care	6
city	6
double	6
end	6

NL_EN_POP

Word	**Frequency**
Is	825
Are	790
Do	789

on	756
Am	671
all	666
Will	636
be	589
love	553
like	547
know	541
Can	488
up	409
go	358
Have	350
down	258
let	256
feel	250
way	249
got	240
might	229
time	227
Would	211
come	207
see	201
been	188

need	184
keep	181
get	180
baby	177
gonna	176
home	164
take	163
make	158
could	157
heart	154
day	152
mind	144
wanna	144
fall	140
light	137
hold	133
give	128
say	126
find	116
life	115
want	115
right	109
cold	107

coming	107
tell	106
world	103
real	100
some	100
think	99
around	97
better	97
running	97
night	96
little	95
another	93
long	92
head	88
bad	85
call	84
eyes	82
alone	80
really	79
leave	78
looking	78
wind	78
look	77

falling	76
good	76
lost	76
feels	75
rain	74
fire	73
high	72
gotta	71
Said	71
Run	70
Hear	69
She	69
Sky	68
were	68
place	66
trouble	66
Did	65
other	65
things	65
change	64
End	64
work	64
Gold	63

Had	62
start	62
Try	62
fine	61
man	61

EN_EN_HIPHOP

Word	**Frequency**
Is	825
Are	790
Do	789
on	756
Am	671
all	666
Will	636
be	589
love	553
like	547
know	541
Can	488
up	409
go	358
Have	350

down	258
let	256
feel	250
way	249
got	240
might	229
time	227
Would	211
come	207
see	201
been	188
need	184
keep	181
get	180
baby	177
gonna	176
home	164
take	163
make	158
could	157
heart	154
day	152
mind	144

wanna	144
fall	140
light	137
hold	133
give	128
say	126
find	116
life	115
want	115
right	109
cold	107
coming	107
tell	106
world	103
real	100
some	100
think	99
around	97
better	97
running	97
night	96
little	95
another	93

long	92
head	88
bad	85
call	84
eyes	82
alone	80
really	79
leave	78
looking	78
wind	78
look	77
falling	76
good	76
lost	76
feels	75
rain	74
fire	73
high	72
gotta	71
Said	71
Run	70
Hear	69
She	69

Sky	68
were	68
place	66
trouble	66
Did	65
other	65
things	65
change	64
End	64
work	64
Gold	63
Had	62
start	62
Try	62
fine	61
man	61

EN_EN_POP

Word	**Frequency**
Do	2740
Is	2332
be	1878

All	1645
Are	1595
On	1566
Love	1542
know	1517
Like	1497
Can	1468
Will	1437
Have	1211
Up	1127
Would	892
Got	795
Let	747
baby	746
Get	745
Go	706
want	665
Say	659
wanna	632
See	628

down	608
make	592
could	585
Time	560
Feel	523
back	511
need	509
gonna	503
Take	501
Way	499
Will	494
Tell	466
right	457
been	444
think	400
Life	371
come	359
heart	355
keep	349
were	346

Look	322
night	320
Girl	314
Did	312
good	310
Give	286
better	273
Said	273
Had	256
Hold	256
eyes	245
Man	239
Call	236
Find	236
home	236
Am	235
things	234
Stop	220
mind	219
world	219

Well	217
Made	215
Bad	213
Day	213
Try	208
really	204
Fall	203
Little	202
Leave	200
Thing	189
Same	177
Alone	176
tonight	176
Stay	175
New	170
Face	168
Should	168
Long	167
Fire	165
Put	164

Gotta	160
Wish	155
Going	154
Head	153
Hand	151
Hands	151
Maybe	145
Show	145
Light	142
Told	142
Kiss	141
Boy	140
Talk	140
another	138
goes	135
sun	134
Change	133

APPENDIX C – STATEMENT OF OWN WORK

Statement of own work

Sign this *Statement of own work* form and add it as the last appendix in the final version of the Bachelor's thesis that is submitted as to the first supervisor.

Student name: Stan Vermeeren

Student number: s1058940

PLAGIARISM is the presentation by a student of an assignment or piece of work which has in fact been copied in whole or in part from another student's work, or from any other source (e.g. published books or periodicals or material from Internet sites), without due acknowledgement in the text.

DECLARATION:

a. I hereby declare that I am familiar with the faculty manual (https://www.ru.nl/facultyofarts/stip/rules-guidelines/rules/fraud-plagiarism/) and with Article 16 "Fraud and plagiarism" in the Education and Examination Regulations for the Bachelor's programme of Communication and Information Studies.

b. I also declare that I have only submitted text written in my own words

c. I certify that this thesis is my own work and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication.

Signature: Stan Vermeeren

Place and date: Wijchen, 8-6-2023