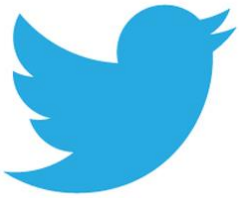
Faculteit Letteren & Wijsbegeerte

Sofie Vandenhoven



What does the bird say?

A corpus study on how Twitter language use reflects personality

Masterproef voorgedragen tot het behalen van de graad van

Master in het Tolken

2016

Promotor Dr. Orphée De Clercq

Vakgroep Vertalen Tolken Communicatie

Acknowledgements

Before elaborating upon the subject of Twitter, personality and language, I would first like to seize the opportunity to express my sincere gratitude to everyone who helped or supported me while writing my dissertation.

First of all, I would like to thank my supervisor, Dr Orphée De Clercq, for accepting me as one of her dissertation students. She supplied me with a highly interesting subject and supported me continuously throughout this whole process. Her patience, help, advice and encouragement have made my dissertation journey a pleasurable one. I could simply not have wished for a better supervisor.

I would also like to thank all twenty Twitter users who agreed to participate: without their collaboration, this dissertation would not have been possible. I would like to enumerate every single one of them, however, I promised and do respect their anonymity.

To end this word of gratitude, I would like to thank my family and friends for their encouragement and support. Moreover, Yara and Astrid, I do not think I had more pleasant moments writing my thesis than in your company.

Abstract

The aim of this dissertation is to ascertain whether the use of language in Dutch tweets can offer researchers insight into the personality of a Twitter user. A database was built, consisting in the tweets of 20 Belgian, Dutch-speaking users of this platform, with an equal representation of both genders. After a Part-of-Speech (PoS) tagging and three sentiment analyses on these tweets, of which the main message was in Dutch, three major tendencies became visible. The first tendency can be seen in the PoS tagging: female users refer more often to themselves or the groups they belong to with the pronouns *ik*, *mijn*, *wij* and *ons* than male users do. Secondly, the sentiment analyses with Duoman and Pattern did not yield particular results: there is presumably no link between personality traits and the use of words with positive or negative connotations. The Linguistic Inquiry and Word Count tool (LIWC), however, revealed ties between certain dimensions of words and characteristics of the Big Five. These results are an indication that further research on personality and the language use on Twitter is highly recommended.

Table of Contents

List of tables, figures and bar charts.

1 INTRODUCTION

Social media is an important aspect of modern-day communication, since it "enables users to create and share content or to participate in social networking" (Oxford Dictionary, s.d.). This importance of social media in the 21st century is demonstrated by the amount of accounts on social media platforms such as Facebook, where the number of users exceeds a paramount 1.65 billion monthly active users worldwide (Facebook, 2016). Facebook is thus the most popular social media platform in the world, before Qzone, which can be considered the Asian equivalent of Facebook and Myspace. Third place goes to the platform Twitter (Statista, 2015): it is a microblogging website with 320 million active users a month, who send over 500 million status-updates, so-called *tweets*[1], a day (Twitter, 2016). No official statistics are known for the Belgian users of Twitter, however, the Flemish newspaper *De Tijd* speaks of 1 million monthly active users in Belgium in 2014 (Demeester, 2014), a number which is likely to have risen over the years.

This high number of users on Twitter – and on all other social media platforms – has logically drawn the attention of researchers, since people share a lot of information about themselves online. All this data is stored in a large and easily accessible database[2]. Consequently, different sorts of research on social media have already been conducted: personality, gender and age in the language of social media (Schwartz et al., 2013), the use of social media among teens and young adults (Lenhart et al., 2010), even the motivation of older adolescents to use social network platforms (Barker, 2009) and also language (Golbeck, 2011). Especially Twitter is highly interesting for linguistic research, since its focus mainly lies on verbal messages, in contrast to other platforms such as Facebook, where there is a great focus on images as well. Twitter invites, so to speak, the user to tell (personal) stories in a personal manner and this influences consequently the language used on the Twitter accounts[3]. Those accounts offer two different types of information. On the one hand, there is the general information, including an obligatory username and optional data such as a bio(graphy), a location, a website, the year and month of registration and – recently added – a date of birth.
On the other hand, there is a second group of information which is shared through the tweets: in the form of free text, limited by 140 characters, users give insight into their relationships,

[1] The working of Twitter and the specific terms are explained in Section 3.1.
[2] It is important to know that the terms of use of both Facebook and Twitter pose restrictions to this accessibility.
[3] It should be noted that also companies or representatives for companies use Twitter as a medium to reach out to potential clients. Those accounts are not taken into account in this research: the focus lies on personal profiles of individuals tweeting in their own name.

interests, sexual preferences, employment and so on. In short, Twitter can offer a deeper insight into the online personality of a user: how they perceive the world, what they think of current events and how they react on other people are only a few examples. Even more, it might also offer a deeper insight in their personality, by revealing specific character traits. This is where the main focus of this dissertation lies.

This dissertation will not target the explicit content of the messages, however, but mainly how the message is formed by using certain language. More specifically, the main research question will be: what can personal use of language reveal about the personality of the person behind a Twitter profile? In other words, is it possible to draw an accurate personality profile by analysing someone's tweets?

In order to answer these questions, we will first discuss how the Big Five of Personality, a widely used personality model which will also be used in this research, came into existence. Subsequently, the relation between personality and social media in general, the choice of platform, language and gender will be elaborated upon. This literature study (Chapter 2) will be followed by a description of the research performed for this study (Chapter 3). It discusses the different steps that were taken to gather respondents, how the personality of these respondents was measured and how their tweets were filtered in order to construct a reliable and balanced database. Next, we explain which expert systems and techniques from the field of Natural Language Processing were used in order to conduct two linguistic analyses on the tweets. In a first analysis, a more abstract representation of the language used in tweets was created by means of Part-of-Speech tagging. For the second analysis typical sentiment and personality-charged words were derived from the tweets based on well-known lexicons. This chapter is followed by an extensive discussion of the results of all the analyses (Chapter 4). We conclude this dissertation by summarising the main findings of our research and at the same time offer prospects for future work (Chapter 5).

2 LITERATURE STUDY

2.1 THE BIG FIVE

In order to gain proper insight in personality, many researchers in psychology now rely on the so-called Five Factor Model, a term first coined by Digman (1990). Later, this model became more known under the name Big Five (Goldberg, 1990), Big Five Personality Traits (Judge et al., 1999) and Big Five Personality Domains (Gosling, 2003)[4]. This Big Five model is one of the most well-researched measures of personality structure of the last decades (Golbeck, 2011) and "provides an integrative descriptive model for personality research" (John & Srivastava, 1999, p. 122). The personality model contains five traits, marked with the anagram OCEAN or CANOE. Each trait equals a category which is labelled with one substantive. However, the category itself represents a broad range of meaning, captured within this one substantive (John & Srivastava, 1999). For example, the letter O stands for Openness, which includes among others *intellect* and *independence*. The different categories are briefly listed in Table 1 and will further be elaborated upon in the next section, where the most important research leading to the current Big Five model will be explained.

Trait	Characteristics
O for Openness	intellectual, polished, independent, open-minded
C for Conscientiousness (or Dependability)	orderly, responsible, dependable
E for Extraversion (or Surgency)	talkative, assertive, energetic
A for Agreeableness	good-natured, cooperative, trustful
N for Neuroticism (vs. Emotional Stability)	not calm, neurotic, easily upset[5]

Table 1 The Big Five as described by John & Srivastava (1999)

[4] In order to prevent any confusion, only the terms Big Five model and Big Five will be used throughout this dissertation.

[5] In John & Srivastava (1999), the trait Neuroticism was originally reversed: "Emotional Stability versus Neuroticism (calm, not neurotic, not easily upset)" (p. 105). To simplify and make use of the anagram OCEAN, we opted to name this trait 'Neuroticism' and thus to reverse the traits.

2.1.1 History of the Big Five

In essence, the Big Five is a lexical-based model which is nowadays internationally used to measure personality. The general objective which initiated the whole process resulting in this model was to create a shared, scientific taxonomy of personality. The first step to do so was to analyse natural language used for personality description (John & Srivastava, 1999). In other words, the Big Five model is entirely based on the language - more specifically on the adjectives - humans use to describe themselves and others when they are referring to personality.

2.1.1.1 First phase: Galton, Allport & Odbert

Many researchers have contributed to the current Big Five model. It all started with Francis Galton, who was the first to study personality using a lexical approach. More concretely, his research consisted in studying personality in a dictionary. In his work, Galton (1884) explains that he examined many pages of the *Roget's Thesaurus' index* and counted the words used to express various aspects of the character. He ultimately "estimated that it contained fully one thousand words expressive of character, each of which had a separate shade of meaning with some of the rest" (Galton, 1884).

It was only after half a century that Allport & Odbert (1936) continued Galton's research and conducted a psycho-lexical study on personality traits by searching the second edition of *Webster's New International Dictionary* for "all the words descriptive of personality or social behavior" (Allport & Odbert, 1936, p. 24). For all the words fitting the description, they opted for the adjectives and present or past participles. Nouns and adverbs were only used if there was no adjectival or participial alternative or when the forms bore a difference in meaning (Allport & Odbert, 1936). Consequently, it is possible to say that the trait names researched were generally trait adjectives, a term used in later research by for example Goldberg (1990). Allport & Odbert's extensive research (1936) was in fact an exact copy of Galton's study (1884). They, however, did not stop at an enumeration, and classified all traits found in four categories, called columns. The first column included the traits of personality, with *aggressive* and *introverted* as most obvious examples. The second column comprised the whole of terms of present activity, temporary states of mind and mood, for example *rejoicing* or *frantic*.

The penultimate column, which was also the longest column, consisted of evaluations of the character, with terms such as *insignificant* or *acceptable* and *worthy*. The fourth and last column was miscellaneous. This means that it included terms that are "of possible value in characterizing personality" (Allport & Odbert, 1936, p. 27), but cannot be categorised in one of the previous three columns (Allport & Odbert, 1936).

2.1.1.2 Second phase: Cattell, Tupes & Christal, Norman

Even though these first two highly important investigations marked the beginning of a lexical-based approach to personality, they were also subject to criticism. Not only Galton (1884) but also Allport & Odbert (1936) examined personality solely based on terms occurring in dictionaries. As a consequence, they did not make any link to the people whom the traits might belong to. This human aspect, however, was taken into account by Cattell (1946; Primi et al., 2014)[6], who was the first to believe that personality is the manner in which a person behaves in a specific situation. He re-organised Allport & Odbert's (1963) list of adjectives into 171 personality descriptors, which he used in his study to describe personality. He used "three kinds of basic data to capture personality dimensions" (Primi et al., 2014, p.30). The first kind of data, called Q-data, was the introspective view of oneself to one's own personality, which was obtained by a questionnaire. The second type of data, called L-data, was the view of a third party based on observations in everyday life. The last and third sort of data was the behaviour of a person studied in a lab (Primi et al., 2014). According to Cattell, primary traits would emerge in all three situations. This enabled him to identify 46 surface personality traits, based on the list of the 171 personality descriptors made before his human study, in other words: Cattell found 46 general adjectives which are mostly used to describe personality. After a thorough analysis, Cattell considered 16 personality traits as basic and constructed a personality test based on these traits called the Sixteen Personality Factor Questionnaire-16PF (Primi et al., 2014). In this questionnaire, the personality traits were made bipolar and denominated after the positive side. This implies that**,** for example, the trait Extraversion has extrovert at one side of the scale and introvert at the other (Primi et al., 2014).

[6] We were not able to retrieve the original work of Cattell, but we relied on the research of Primi et al. (2014)

The pioneering work of Cattell (1946) can be considered the start of a second shift in the research field of personality. He did not rely solely on dictionaries and lists of adjectives but he also included people in his research. Moreover, his work raised interest by other researchers to further examine personality traits (John & Srivastava, 1999) and to ultimately narrow his sixteen traits down to five.

One pair of those researchers were Tupes & Christal, who should, according to Goldberg (1993), be considered "the true fathers" (p. 27) of the Big Five for conducting their 1961 study. In that study, eight different samples of airmen, all with a different type and length of education, were analysed. Those samples consisted in peer-evaluations of personality, more precisely, ratings of the 30 bipolar personality traits found by Cattell. Each airman had to rate all his fellow airmen and was in turn also rated by them (Tupes & Christal, 1961). Despite the criticism that some airmen might have downgraded others on purpose, the researchers concluded that "five fairly strong and recurrent factors emerged from each analysis labelled as (1) Surgency, (2) Agreeableness, (3) Dependability, (4) Emotional Stability and (5) Culture" (Tupes & Christal, 1961, p. 11) and that those "five factors possess a fundamental nature and probable invariance" (Beck, 1999, p. 30).

Unfortunately, Tupes & Christal's efforts to narrow down the traits to five remained largely unnoticed for personality researchers since it was published as a technical report for the American Air Force. Therefore, other publications, among others Allport & Odbert's (1936) and Cattell's (1946), "dominated the literature on personality structure" (Digman, 1990, p. 419).

Norman (1967) consequently returned to the basis of the personality studies and used Allport & Odbert's (1936) list as a basis for his own research. His objective was to contribute to "the development of a well-structured taxonomy of personality descriptive terms" (Norman, 1967, p. 1) Therefore, his study happened on two levels: the first part consisted in making a list of trait adjectives, mirroring Allport & Odbert, and secondly running tests with university students, in order to research how known the found traits are. This will now be explained in closer detail.

Norman (1967) started with making a list of trait adjectives, which took place in four phases. In the first phase, the entire list was constructed by using Allport & Odbert's list (1936) as a basis and further completing it by adding all words relevant for behaviour from the *Webster Third New International Dictionary Unabridged.* The complete list contained roughly 10% of the entire dictionary. Secondly, certain terms were left out, such as rarely used words or vague terms. Words that were not thrown out, got categorised during the third phase into three classes, which were in their turn divided into three or four subcategories. In the fourth phase, 2,800 terms were considered suitable and "stable "biophysical" traits" (Norman, 1967, p. 6). In the second phase, Norman gave university students three different tasks. Firstly, he asked for either a synonym or short definition of the word given, or to cross the term out if it was unfamiliar to them. Secondly, the same students had to rate themselves and others with the available terms and lastly, the students "were asked to rate the degree of social desirability possessed by each trait presented" (Norman, 1967, p. 14).

Norman (1967) concluded that of his 2,800 terms' list, many were described as unfamiliar: "about 34.5% of all words were rated by all 100 respondents" (p. 22). This means that these unfamiliar terms could cause problems when used in a rating scale "for use in the general population or with less verbally sophisticated subgroups in the general population" (p. 22). However, he also requested the subjects to provide him with synonyms or definitions of the traits. This data might render it possible to refine the pool of words. Once this happens, his list can be used as "a partial basis for future investigations" (p. 19). Since this research was followed by a hiatus, the list of words was not further refined.

2.1.1.3 Third phase: Goldberg

In the years following Norman's research, publication of personality research became rather difficult. Psychologists such as Walter Mischel firmly believed that personality changed according to the situation, instead of being a static element (Mischel, 1968). Therefore, only after a hiatus of two decades, researchers among which Lewis Goldberg reviewed Tupes & Christal's (1961) recurrent five factors in an attempt to rebut arguments that the five factors had not been sufficiently generalised (Goldberg, 1990).

To prove the sufficient generalisation of personality, Goldberg (1990) conducted two separate though similar studies. The first study consisted in asking college students to describe themselves with the terms of his list, which was based on the 2,800 trait terms Norman (1967) had selected. When describing themselves, the students were instructed to use "an 8-step rating scale ranging from extremely inaccurate to extremely accurate as a self-descriptor" (Goldberg, 1990, p. 1217). The second study was a similar exercise with clustered terms, researched in both self- and peer-evaluations. Those two studies enabled Goldberg (1990) to demonstrate that "any reasonably large sample of English trait adjectives [...] will elicit a variant of the Big Five factor structure" (p. 1,223). He thus concluded that virtually all English trait adjectives can be represented within the Big Five Model. In his research, he gives these Big Five traits the following names: Surgency, Agreeableness, Conscientiousness, Emotional Stability and Intellect (Goldberg, 1990). Later, John & Srivastava (1999) opted for a double name for certain traits: Surgency was also Extraversion, Emotional Stability was connected with its opposite Neuroticism and finally, Intellect was equal to Openness. In this study, the traits forming the anagram OCEAN will be used, since they are currently the most widely used trait names. They can be found in Table 1, which was presented at the beginning of this chapter (Section 2.1).

2.1.1.4 Fourth phase: establishment of the Big Five and its relation to other variables

After one hundred years of extensive research, the Big Five was established as an internationally accepted lexical-based model of personality. The model has five factors, which each represent a personality trait on a broad level. Each of these five factors equals in fact a bipolar scale: the trait Extraversion, for example, ranges from extravert to introvert. Each trait contains various aspects which in their turn enclose even more specific traits. If this is applied to Extraversion, *sociability* is an aspect of Extraversion, while *sociability* includes among others *talkative* or *outgoing* and so on (Correa, 2010).

It can be concluded that all personality research leading up to the official establishment of the Big Five shows three main phases. The first period of research, with studies by among others Galton (1884) and Allport & Odbert (1936), is characterised by research in dictionaries. Cattell (1946), however, caused a major shift in this field of research: he initiated the second period of research and was followed later by Tupes & Christal (1961) and Norman (1967).

Those researchers relied on lists of traits in combination with questioning actual people. After Norman, a hiatus took place, caused by the view that personality is not a fixed item. Research on the matter was resumed in the third period, by using lists of adjectives together with questionnaires. The main objective in this period, however, was different than in the previous two phases. The hiatus signalled and was caused by a change in the perception of personality: it was consequently necessary that new research proved that the Big Five was in fact sufficiently generalised, even if personality was regarded as non-fixed. This aim was eventually achieved by Goldberg (1990).

It can be said that a fourth phase has been initiated when the Big Five model was established as official measurement for personality. From that moment onwards, the link between the Big Five traits and other variables has been extensively researched. Examples here are job performance (Murray & Mount, 1991), career success (Judge et al., 1999), job satisfaction (Judge et al., 1999) and even romantic success (Shaver & Brennan, 1992). A brief summary of the history of the establishment of the Big Five can be found in Table 2 below.

Phase	Who
I. Research by dictionaries	• Galton (1884) • Allport & Odbert (1936)
II. Research by lists of adjectives and observation of people	• Cattell (1946) • Tupes & Christal (1961) • Norman (1967)
Hiatus: personality is not fixed (Mischel, 1968)	
III. Research by lists of adjectives and questionnaires	• Goldberg (1990)
IV. Research of personality in relation to other variables	

Table 2 Summary of the history leading up to the Big Five

This fourth phase is very important for this dissertation; however, the research questions mentioned previously are not our main interest, since in this study, the focus is on another variable, namely language. Even more, the objective is to figure out whether it is possible to create a correct personality profile of a user by analysing the language he or she uses on a social media platform. The reasons why these platforms are highly interesting for linguistic and additional psychological research will be extensively discussed in the following sections.

2.1.2 The Big Five and social media

Four main reasons make social media highly interesting for research. Firstly, the increasing popularity of social media in the last decade has created an enormous database of personal information. Moreover, and secondly, the content in this database, which is widely available through public profiles, is user-generated. This means that the information given relates "to material [...] that is voluntarily contributed by members of the public" (Oxford Dictionary, s.d.). This so-called voluntarily contributed content also implies that it is, in most cases, spontaneously generated content. Proof of this spontaneity can be seen in the language used on social media, which is completely different than what is used in, for example, a letter. This language, which fluctuates between spoken and written language but really is neither of them, forms the third reason why a database made from social media content is highly interesting: it is a new form of communication. In a 2013 TED-talk, linguistic professor John McWhorter[7] labelled this language "fingered speech". With this term, he refers to a tendency to "write as you speak". This tendency became very widespread as soon as people had the opportunity to quickly send messages either to each other or to the world. In general, McWhorter (2013) claims that the so-called fingered speech has a lack of rules and a loose structure. That is why it is more closely related to spoken language, even though the medium through which the message is spread is clearly written. The fourth and last reason why it is interesting to research language on social media is because the messages spread often contain very personal and emotional content. It is highly possible that those four elements, and in general the rise of social media, caused or at least coincided with a surge in research regarding the Big Five and social media.

However, an often heard criticism is that online profiles might also depict a false and better image of a user, making personality research on social media useless. However, in Back's (2010) Facebook study, no evidence was found to support this presumption. On the contrary, the results show that "people are not using their OSN[8] profiles to promote an idealized virtual identity" (p. 374). This would mean that the personality traits displayed online should correspond to the actual personality of the user. Many researchers have tried to predict personality and have found corresponding correlations between usage of social network sites and the characteristics of the Big Five, which will be elaborated upon in the next section.

[7] The TED-talk of professor McWorther (2013) can be found with the following link: https://www.youtube.com/watch?v=UmvOgW6iV2s.
[8] In Back's research (2010) OSN stands for Online Social Network.

2.1.2.1 Personality and choice of platform

Personality does not only influence whether one is more likely to be drawn to social media or not, it will also play a role in which social medium is used. Hughes (2012) wanted to find out what personality traits are typical for users choosing either Facebook or Twitter. He did so by recruiting 300 participants on both Twitter and Facebook, who were asked to fill in a questionnaire with "three existing personality measures, a newly developed scale measuring Twitter and Facebook usage and demographic questions concerning age, sex, employment status and continent" (p. 563). Hughes (2012) most relevant hypotheses for this research were based on previous work and are the following:

- H1: People scoring higher on Neuroticism will use Twitter more often,
- H2: People scoring lower in Extraversion might be attracted to Twitter since there is "potential for increased anonymity" (p. 562),
- H3: People scoring higher in Conscientiousness might be attracted to the "short, quick fire nature of Twitter" (p. 563).

Hughes (2012) discovered that personalities differ, depending on whether the website is used for social means or informational use. We will discuss the most important results regarding the social network site Twitter because this platform is also at the heart of this dissertation. Socially, this medium is more appealing to users scoring higher on Openness and scoring lower on Conscientiousness, the latter contradicting hypothesis 3. There was neither evidence nor contradicting data found for a correlation between Neuroticism and the social use of Twitter, stated in the first hypothesis. This might suggest that Twitter is not really used as "a tool to mitigate loneliness" (Hughes, 2012, p. 567). However, when Twitter is used for informational use, hypothesis 1 is contradicted: users score significantly low on Neuroticism. Hypothesis 2 and 3, on the other hand, are confirmed: people using Twitter for information are in fact more introverted and more conscientious.

Even though Hughes (2012) has made a distinction between using Twitter for social means or informational use, it is highly probable that many users also use the platform for both means. In that case, one could enumerate the general characteristics of people using Twitter: high on Openness and low on Neuroticism and Extraversion. Since this conclusion was not explicitly drawn by Hughes, we will also look for ways to support this generalisation in our research.

In Table 4 below, all relevant findings of Hughes' study and the generalisations (column 3) are summarised.

Twitter					
Social use		**Informational use**		**In general**	
Conscientiousness	Low	**Conscientiousness**	High	**Conscientiousness**	High/low
Openness	High	**Neuroticism**	Low	**Openness**	High
		Extraversion	Low	**Neuroticism**	Low
				Extraversion	Low

Table 3 Summary of Hughes' (2012) findings for use of Twitter

Personality plays an important role in how social media is perceived and used in general. A general study, conducted by Rosen and Kluemper (2008), revealed that people scoring high on Extraversion and Conscientiousness find social media sites not only easy to use, but also useful.

In 2000, Hamburger asked 72 Israeli students what purpose they used the internet for: social services, information services or leisure services. Consequently, the students were requested to take a personality test on Extraversion and Neuroticism[9]. Hamburger (2000) concluded that social services are generally negatively correlated with Extraversion and positively correlated with Neuroticism. In short, this means that users of social media are in general introverted and neurotic. A side note that should be made is that this was only the general conclusion of the research. Hamburger (2000) also found a significant difference between genders: where female users of social services are generally introverted and highly neurotic, men are quite the opposite. This highlights the importance of gender in personality research and social media, something that will also be taken into account for this dissertation.

Gender differences were not considered in Ross et al.'s study (2009), where almost 80% of the 97 questioned students were female. For this study, all subjects filled in a Facebook questionnaire and a NEO Personality Inventory, in order to link personality traits to Facebook behaviour. The most important conclusion that was drawn from this study is that Openness positively correlates with the general use of Facebook.

[9] The test used was the Eysenck Personality Inventory.

This study of Ross et al. (2009) formed the basis for Correa's (2010) three hypotheses, namely:

- H1: Extraverted people will use social media more frequently,
- H2: Emotionally stable people will use social media less frequently,
- H3: Open people will use social media more frequently.

For her study, 1,482 people responded to the survey by filling in a questionnaire on their frequency of social media use, a 10-Item Personality Inventory and a Satisfaction with Life Scale. After combining all these elements, Correa (2010) found that all her hypotheses were supported: social media is more easily used by people scoring higher on Openness and Extraversion. Moreover, it is less used by people who are emotionally stable.

If this last finding is rephrased, we can see a similarity with Hamburger's (2000) general findings: namely that people scoring high on Neuroticism, and thus being emotionally unstable, are more easily drawn to social media. Even though this first conclusion of Hamburger corresponds with Correa's study (2010), Hamburger's second conclusion about low scores on Extraversion is contradicted by the work of Correa. It is possible that the relatively little amount of subjects in Hamburger's study (72) caused a biased conclusion, whereas the pool of Correa, which was 20 times larger, might have generated a more general image. A summary of which Big Five personalities are quite likely to be drawn to social media can be found in Table 3 below.

Personalities drawn to social media	
High in Neuroticism	Hamburger (2000) Correa (2010)
High in Openness	Ross et al. (2009) Correa (2010)
High in Extraversion	Correa (2010)
Low in Extraversion	Hamburger (2000)

Table 4 Summary of the studies about who is generally drawn to social media

2.1.2.2 Personality and language

As discussed in the previous section, a general image of which personalities are mainly drawn to social media does exist. However, most of these studies take more than only linguistic features into account, or they study anything but language use on social media.

Golbeck (2011) claims to be the first to test whether all information displayed on a profile can predict one's personality. She first tested this on Facebook (2011a), followed by the same study on Twitter (2011b).

The information she retrieved from the Facebook profiles did not only include the status-updates, but also among others birthday, relationship status, religion, favourite TV shows and movies. Not only did Golbeck (2011a) point out some interesting connections regarding the displayed information and personality; she was also able to come forward with a few linguistic correlations between the status updates and the Big Five. Those relations, which "made up half of all features considered" (p. 257), were said to "also largely make intuitive sense" (Golbeck, 2011a, p. 257) and were retrieved by means of the Linguistic Inquiry and Word Count tool (LIWC) (Pennebaker et al., 2001).

Conscientiousness, she found, was the easiest personality trait to link with linguistic features. Her first conclusion was that more conscientious people use less words related to "perceptual processes" (p. 257), meaning words related to *seeing, hearing* and *feeling*. In other words: they write less about things they have seen or heard, but she also found that they tend to talk more about other people on their profile. She also discovered some interesting correlations when it comes to words describing feelings: using more words describing positive feelings is a sign of scoring high on Agreeableness. When she observed these correlations, it was no surprise to her that people with a high frequency of anxiety words also scored high on Neuroticism. Her findings are summarised in Table 5 below.

High on Conscientiousness	High on Agreeableness	High on Neuroticism
• less words related to 'seeing, hearing and feeling' • more words about other people	• more words that describe – positive –feelings	• more words that describe anxiety

Table 5 Linguistic correlations found in Golbeck's (2011a) Facebook study

Later, Golbeck (2011b) conducted the same study on a different social media platform, namely Twitter. Again, not only the *tweets* itself were collected, but also public account data such as among others followers and mentions were taken into account. Exactly as in the previous study, the linguistic analysis formed the major part of the study, namely 79% (Golbeck, 2011b). As with her Facebook study, she discovered some intuitively logical

correlations, which are summarised in Table 6. She found that Conscientiousness was negatively correlated with "words about death (e.g. *bury, coffin, kill*)" (p. 152), meaning that the more conscientious a user is, the less he or she will refer to death. Moreover, the same trait was also negatively correlated with negative emotions and sadness. Hence both findings suggest that highly conscientious people abandon unhappy subjects. Another finding concerning Conscientiousness reveals an equal trait on both Facebook and Twitter: the pronoun *you* is more frequently used, indicating that highly Conscientious people do indeed talk more about others. Scoring high in Agreeableness also indicated a significant use of the pronoun *you* and those users were also less likely to talk about achievements and money.

High on Conscientiousness	High on Agreeableness
• less words about death • less words about negative emotions and sadness • more use of pronoun *you*	• more use of pronoun *you* • less words about achievements and money

Table 6 A summary of the main findings in Golbeck's Twitter study (2011b)

In general, when it came to predicting personality automatically Golbeck (2011) was able "to predict all personality traits within roughly 11%" for both Facebook and Twitter. This means that the percentage estimated for a trait was never further away than 11% of the actual percentage revealed by filling in personality tests[10]. Even though the results for Openness and Agreeableness were more or less the same for both networks, Twitter produced significantly "less impressive results for Conscientiousness, Extraversion and Neuroticism" (Golbeck, 2011b, p. 154). Golbeck (2011b) also makes an important remark about using LIWC: due to misspelled words on Twitter, important language features might have been overlooked. This is certainly something that should be taken into account in our research as well.

2.1.3 The Big Five and gender

As mentioned in Section 2.1.2.1, Hamburger (2000) already made a remark on how the female and male users differ in character. Consequently, there is a difference between both genders' Big Five scores, something that has already been researched extensively. However, many of these researchers go very much into detail: they do not only research the five traits of the Big Five, they also include underlying traits. Since this dissertation focuses more on the possible relation between linguistic elements and personality traits, only three main

[10] How personality is measured will be explained in depth in Chapter 3.

researches, conducted on the general five traits of the Big Five will be discussed in order to have a general image of which gender differences have been demonstrated so far.

The oldest research, conducted by Lynn & Martin (1997), was held in order to find gender differences in 37 different countries. They did so on the traits of Neuroticism, Openness, Extraversion and Psychoticism. Only the results of Neuroticism and Extraversion are of any interest for this dissertation: they found a statistically significant result that women are more neurotic than men in all 37 countries[11]. For Extraversion, the male participants scored higher than their female counterparts in 30 of the 37 countries, however, in 5 of those remaining 7 countries, women scored significantly higher than men.

Budaev (1999) found that on average, females scored significantly higher on both Agreeableness and "low Emotional Stability", which equals scoring high on Neuroticism, already confirming what has been said by Lynn & Martin (1997).

Vianello et al. (2013), however, do say that there is a difference in personality when the character traits are explicit, which means self-reported, or implicit. When self-reported, they found that women scored higher in four of the five traits: Neuroticism, Conscientiousness, Agreeableness and Extraversion. When the implicit measures were analysed, they found little to no gender differences.

A summary of the discussed research on who scores higher or lower on specific traits is enumerated in Table 7.

Trait	Who scores higher	According to
Openness to Experience		
Conscientiousness	Female scores higher than male	Vianello et al. (2013)
Extraversion	Female scores higher than male	Vianello et al. (2013)
	Male scores higher than female	Lynn & Martin (1997)
Agreeableness	Female scores higher than male	Vianello et al. (2013) Budaev (1999)
Neuroticism	Female scores higher than male	Lynn & Martin (1997) Vianello et al. (2013) Budaev (1999)

Table 7 Summary of gender differences in the Big Five

[11] In one country, the score was not statistically significant.

2.1.4 Conclusion on the Big Five and social media

To sum things up, there have been several studies on which personalities are drawn to social media, what aspects of language can be related to certain personality characteristics and how gender can also have an influence on personality traits. It is important to realise that these studies have been mainly conducted on English-speaking subjects. This dissertation's main objective is very similar to the one of Golbeck (2011), our research question, however, is more narrow. We want to know whether Dutch language use without any other profile information, can reveal something about someone's personality. And if this is the case, we want to find out which aspects of language are important to take into consideration.

In order to answer this main research question subjects had to be persuaded to donate their tweets and fill in a personality questionnaire. Next, a linguistic analysis on the basis of the language used in these tweets had to be conducted. The different steps that were taken to conduct this research will be explained in closer detail in the next chapter.

3 METHODOLOGY

The main research question of this dissertation is whether language use allows one to draw a realistic and correct personality profile of a Twitter user. To conduct this research, several steps had to be taken. Firstly, a social platform had to be selected (Section 3.1). Secondly, subjects were persuaded to participate in this research and their data was collected (Section 3.2). In the third phase, it was decided how the personality of the respondents would be measured and the chosen personality test was sent to the subjects (Section 3.3). Afterwards, all collected tweets were carefully and thoroughly categorised in order to end up with a balanced and reliable database (Section 3.4). After this careful selection, two linguistic analyses were conducted: for the first analysis, a more abstract representation of the language used in tweets was created by means of Part-of-Speech tagging; for the second analysis typical sentiment and personality-charged words were derived from the tweets based on well-known lexicons (Section 3.5).

3.1 STEP ONE: CHOICE OF SOCIAL MEDIA PLATFORM

If one wants to research personality from a purely linguistic point of view, without taking into account the social interaction or the available data regarding groups or amount of friends, it is useful to focus on Twitter. The reason is very straightforward: Twitter is not just another social network. It is also considered a microblogging site, which is defined by the Cambridge Online Dictionary as “a blog in the form of a short message for anyone to read, sent especially from a mobile phone” (s.d.). This blogging element automatically directs the focus of the platform on the messages shared with the world, the so-called *tweets*. Research has proven that those tweets are mostly used for “sharing of opinions and information” (Hughes, 2012, p. 561). Moreover, Twitter also embraces in general a higher level of anonymity (Hughes, 2012), making only a nickname an obligated item. All other personal information to complete the profile – bio, location, website, date of birth – can be filled in voluntarily. The linguistic focus, together with a certain level of anonymity, increases the possibility that people reveal more information through their tweets, are more honest in what they share with the world and care less about whether they phrased certain things too bluntly or not. Since this research will be focusing on what lexical items give away about personality, Twitter was considered the most appropriate choice as social medium.

But how does this social medium work? Twitter is a platform where users share status updates, the so-called *tweets*, which have a 140 character limit, as can be seen in Figure 1.

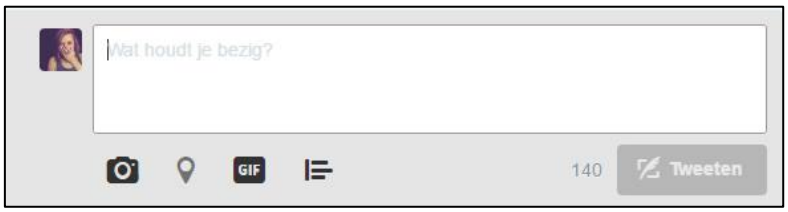


Figure 1 Screenshot of an empty status update

These tweets can include a *mention* (@) or a *hashtag* (#). A *mention* is the @-sign in combination with another username, making it a link to that user's profile. The *hashtag* (#) is used to mark keywords or topics. Tweets with the same hashtags are collected in lists, which are available by clicking on the hashtag. Both the *mention* and the *hashtag* are present in Figure 2.

Figure 2 Screenshot of a tweet with a hashtag (#tweetseats & #tvvv) and a mention (@VTMTheVoice)

When a user tweets something funny or interesting, there is an option to *retweet* that particular tweet: this means that you allow a third party's tweet to be shown on your timeline, which is illustrated by Figure 3. The retweeted tweets will be seen by your followers. It is consequently possible to compare the *retweet* function to the *sharing* option on Facebook.

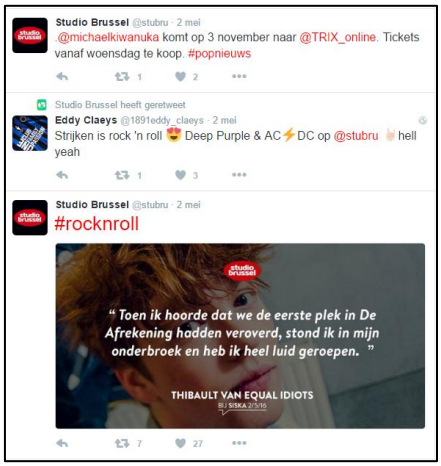


Figure 3 A retweet of the account @1891eddy_claeys on the timeline of the user @stubru

Lastly, it is also useful to know that, unlike Facebook, Twitter does not require a mutual friendship. One can choose to follow an account, in order to be able to see the updates, an action which does not require an acceptance of the user being followed[12].

3.2 STEP TWO: SUBJECT AND DATA COLLECTION

The initial goal was to persuade as many people as possible, and at least twenty subjects, to participate in this research. Subjects could participate if they were (i) willing to donate their tweets and (ii) willing to fill in a personality test.

A first request was launched via Twitter, as illustrated below.

Figure 4 Tweets used to gather respondents

As can be derived from Figure 4, we asked for respondents tweeting in Dutch, targeting both Belgian as Dutch citizens and both males and females. Realising that most previous research had been conducted on English (see Chapter 2), we envisaged to find out whether there exist differences between languages and whether the conclusions drawn from English content translate to Dutch. Our hypothesis is that the conclusions will be the same, since we believe personality to be something more or less universal.

During the subject collection we narrowed the group down to twenty final subjects based on several specific requirements: mother tongue, age and gender, language of tweets and lastly, the amount of tweets. This will now be explained in closer detail.

[12] This does not apply to the situation where a user has put his profile on 'private'. Users with a private profile get a notification with a follow-request, and can then allow or deny access to their profile.

Firstly, the mother tongue of the subjects needed to be Dutch, with the majority of tweets written in Dutch. This was already facilitated by the call for respondents written in Dutch. Since there were not enough participants of the Netherlands to make a balanced database, our focus shifted towards Dutch-speaking Belgians who tweeted mostly in Dutch.

Secondly, the participants had to be more or less representative of the typical Belgian Twitter user. At first, and as can be seen in Figure 4 above, our first call for participants targeted the age category of 20-25 years. However, the interest of elder Twitter users was raised as well. Since there was at that moment no statistic supporting the proposed age category, all interested Dutch-speaking Belgians were accepted until a fixed age category was determined. The age category was determined based on the 2015 statistics which is presented below.

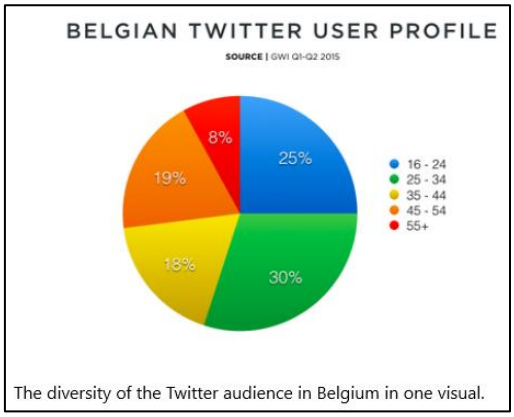


Figure 5 Twitter statistics about the Belgian twitter user profile

According to this pie chart, 25% of the Belgian Twitter users are aged between 16 and 24 and 30% between 25 and 34. This means that the age category 16 to 34 covers more than 50% of the entire Belgian Twitter population. We reflected this statistic in our database and made sure that half of the selected respondents belonged to the first category (16-24) and the other half to the second category (25-34).

We also made sure that both male and female respondents were included in our database since previous research (Hamburger, 2000) has revealed that a difference in gender might reveal a different test result regarding personality. In order to avoid a biased database, we thus selected ten women and ten men.

A penultimate requirement was the language in which the tweets were written. The very first requirement was that our subjects should have Dutch as a mother tongue. The reason why we did not include anyone whose mother tongue is not Dutch, is because of the presumption that personality will be more easily visible or traceable when one uses his or her mother tongue. However, it would be wrong to assume that the tweets collected for this study are exclusively written in Dutch. Dutch-speaking people – and certainly young adults aged between 17 and 34 – have a habit of inserting English words or phrases into both their writing and speaking: a phenomenon that is known as code-switching. In order to have a database that reflects the actual language of Dutch-speaking Twitter users, we made a careful selection of what tweets could and could not be included for our analyses. This will be explained in depth in Section 3.4.

Lastly, the minimum amount of tweets was set to 300, that is, if a subject was to be considered for this study, he or she should have posted at least 300 usable tweets. In Table 8, the data of the twenty selected respondents is presented. [13]

F	age	number of tweets	M	age	number of tweets
F1	23	3222	**M1**	27	3178
F2	30	3230	**M2**	34	4873
F3	30	3162	**M3**	17	3192
F4	20	3244	**M4**	24	2342
F5	24	1433	**M5**	33	3194
F6	22	3214	**M6**	34	3229
F7	27	3260	**M7**	32	5518
F8	22	2426	**M8**	24	3196
F9	22	2625	**M9**	24	2803
F10	27	3156	**M10**	28	2812
total		**28,972**	**total**		**34,337**

Table 8 Age and number of tweets of all the participants

All people who were selected received a Direct Message to inform them their data had been collected, whereas all people that did not make the selection were also informed. In the first column, F stands for female and M for male. In the second column, their age is presented, followed by the total amount of tweets that were automatically retrieved for each subject. This data collection took place in autumn 2015.

[13] All 63 309 tweets of the entire database can be found back in Appendix Five.

More information on the selection of the Twitter-data will be discussed in Section 3.4, but first we will discuss which model was used to measure the personality of our twenty respondents.

3.3 STEP THREE: CHOICE OF PERSONALITY MODEL

While gathering and reading previous personality research on social media, it became immediately obvious that the Big Five model would be used in this dissertation. There are two main reasons for this decision: firstly, all other researches on social media found are conducted with the Big Five as measurement for personality. By using this model, comparisons and differences will be easier recognised. Secondly, the Big Five is, as explained in Chapter 2, the most researched and valid personality model up to date.

After selecting the model for measurement, the search for a feasible test was initiated. An online test was sought, since this would facilitate the communication with the respondents. Moreover, an online test also lowers the threshold: participating in our research would only include giving us consent to download their tweets and filling in a short personality test, which could be done anytime and anywhere.

The test chosen was a general Big Five personality test with 46 questions[14]. This is a test that is thorough enough for this research and yet not too extensive to be filled in by the respondents. As all personality tests for the Big Five, the questionnaire uses a Likert-scale from 1 (strongly disagree) to 7 (strongly agree). A screenshot of what the questions look like can be found in Image 6 below.

Figure 6 Screenshot of some example questions of the Big Five test used for this research

[14] The test used can be found on the following url: http://www.outofservice.com/bigfive/.

After filling in all 46 questions, their gender and their age, the test results are shown: a percentile for every trait is given, accompanied by a verbal explanation. We are aware that the respondents could see their results before sending them back to us, which can either be considered an advantage or a disadvantage. The advantage is that by filling in the test the users also learn more about their own personality. However, it is also possible that, even though there is no 'good' or 'bad' in personality, there were users who were not satisfied with their results and might have re-taken the test. Yet we did ask the respondents explicitly to fill in the test only once and we trust their honesty.

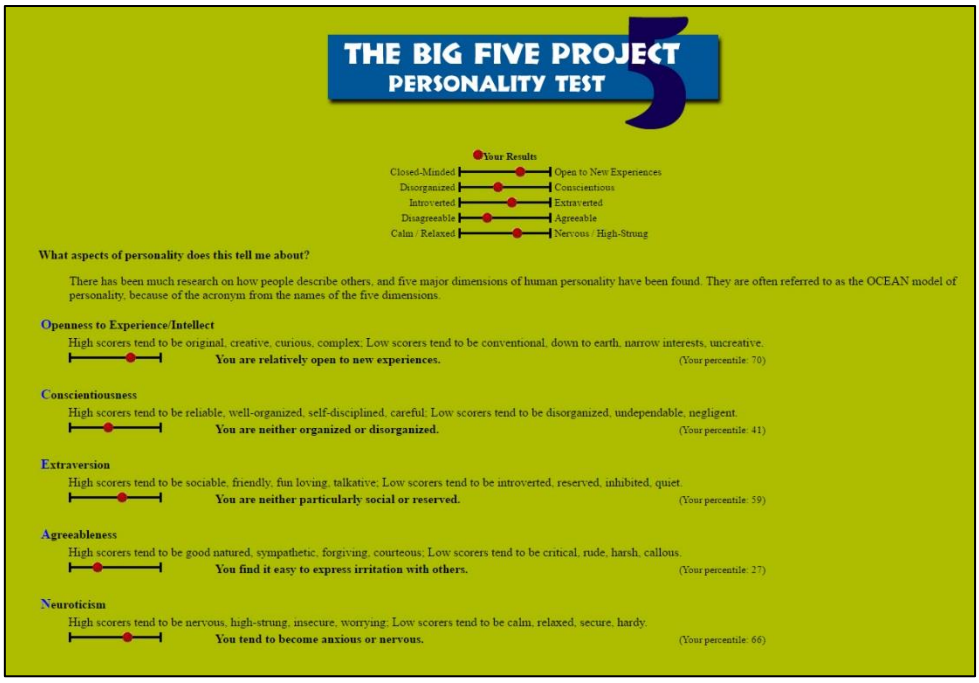


Figure 7 Fictional test results of the online Big Five test

The test results, of which an example is shown above in Figure 7, should be interpreted as follows: scoring above 50% is considered as scoring high on a particular trait and scoring lower than 50% as low. Each trait of the OCEAN anagram has been appointed two main characteristics, each at the end of the scale. In other words, there is one description for either scoring above and below 50%. Both traits are also appointed several trait adjectives. For Openness, for example, the two bipolar traits are *closed-minded* (<50%) and *open to new experiences* (>50%). If one scores high and is consequently labelled as *open to new experiences*, which is the case in the above example, the subject is also appointed several trait adjectives: *original*, *creative*, *curious* and *complex*.

In Table 9, the two main characteristics and adjectives used to describe people scoring high or low on a particular trait are summarised.

	Low score (<50%)	adjectives	High score (>50%)	adjectives
Openness	closed-minded	*conventional* *down to earth* *narrow interests* *uncreative*	open to new experiences	*original* *creative* *curious* *complex*
Conscientiousness	disorganized	*disorganised* *undependable* *negligent*	organised, conscientious (neat)	*reliable* *well-organised* *self-disciplined*
Extraversion	introverted	*introverted* *reserved* *inhibited* *quiet*	extraverted	*sociable* *friendly* *fun* *loving* *talkative*
Agreeableness	disagreeable	*critical* *rude* *harsh* *callous*	agreeable	*good natured* *sympathetic* *forgiving* *courteous*
Neuroticism	calm/relaxed	*calm* *relaxed* *secure* *hardy*	nervous/high strung	*nervous* *high-strung* *insecure* *worrying*

Table 9 Division of the Big Five on both parts of the scale

3.4 STEP FOUR: DATA SELECTION

All twenty participants gave consent to have their Twitter data collected. With a Python script, the maximum amount of tweets was downloaded for each respondent. The Twitter API allows users to download about 3,000 tweets per person. However, after the thorough selection, the set minimum amount of tweets (300) was not reached for all participants, which is why we ran our script one more time for M2 and M4. In a next step all the downloaded tweets were transformed into comma-separated files (CSV), which were then converted to Excel. In order to clean up our database for personality research, all data was manually classified into seven categories (see the summary in Table 10).

After having manually divided all tweets into categories, the database used for our personality research was created by combining the data present in category I with the data in category II.

The first category is an obvious choice: it is the column with tweets completely written in Dutch. It is not important whether the hashtags are written in Dutch or any other language, since Category II, which is also used for our research, is the column where the main message is in Dutch, with addition of non-Dutch elements. The reason to include this column in our research is that it actually reflects real life speech; moreover, it is very often used online. In the 21st century, nobody has a language completely free of foreign words, especially English words. Moreover, some English (and other foreign words) are accepted in the Dutch language. To make a distinction between what is and what is not accepted would be a nearly impossible task. Therefore, all tweets with their main message written in Dutch and their non-Dutch additions were also included in the database. A fictional example is: "Ik vertelde hem vandaag wat ik van hem vond en hij liep gewoon weg. *Whatever*." Though this codeswitching as such could also reveal something about personality, we do not include this variable for the research performed in the framework of this dissertation.

Category III consists of *mentions*, these can be considered as answers to questions, an example of which is depicted in Figure 8. Although personality is definitely reflected in this type of communication, it can be presumed that the topics talked about are not entirely chosen by the user him or herself. Therefore, those tweets will not be included in the database.

Figure 8 A tweet with a mention that will not be included in the database

If a user, however, initiated the conversation and embedded his or her *mention* within the tweet, as shown in Figure 9, then the tweet will be placed in either Category I or II.

Figure 9 Tweet with mention that will be included

Leaving out Category IV (Retweets), V (Not Dutch) and VII (Miscellaneous) can be all traced back to one single reason, namely that the tweets are not written by the user him or herself. This is most obvious in the case of Category IV, the column of the retweets. These tweets might reflect how a user feels or thinks about something or someone and thus reflect personality. The content, however, is not written by them or written in their own words and since this is a linguistic research, it is very important that the tweets in the database are written by the user him or herself.

This leads us to Category V, with tweets entirely written in a different language. Leaving this column out has two main reasons: firstly, people are presumed to express themselves better in their mother tongue than in any other language. Therefore, they might be limited by their knowledge of the other language and use certain words because they do not have any alternative, something that would bias our database. Secondly, a lot of these – mostly English – tweets happened to be song lyrics. This leads us back to our first main argument to leave out this column: lyrics are not written by the user.

The miscellaneous column, Category VII, is left out for the same reason: it contained, among others, automatically generated tweets, such as check-ins on Foursquare or titles of shared posts.

Category VI, to conclude, exists of tweets with only hashtags, URLs, emoji's or a combination of two or three of those elements. Emoji's, for example, are certainly part of the online language, however, they do not bear any linguistic value, nor do hashtags or URLs. Therefore, also this column was excluded from the final database.

Category	Tweets	Explanation
I	Dutch tweets	Tweets that are completely written in Dutch. This does not take into account the used hashtags: those might be English or any other language.
II	Partly Dutch	Tweets that have a main message in Dutch, however, they are mixed with words or phrases in another language, mostly English.
III	Mentions	Tweets that are sent as a reply to somebody. These tweets always start with @username.
IV	Retweets	Tweets that are posted on one's profile, but are written by someone else. In the CSV file, they start with RT.
V	Not Dutch	Tweets that are written in any other language besides Dutch.
VI	#, url or emoji's	Tweets that have no written message, but only make use of hashtags, URLs or emoji's or a combination of those elements.

VII	Miscellaneous	Tweets that do not belong in any previously mentioned category. This column includes tweets that are check-ins, for example on Foursquare, or automatically sent Tweets after taking a certain test.

Table 10 Division of the tweets in seven columns

This division of all the tweets is presented in Table 11 for the women and in Table 12 for the men.

	I	II	III	IV	V	VI	VII
F1	**513**	**210**	2263	138	33	27	38
F2	**508**	**35**	2346	258	21	19	43
F3	**563**	**176**	2159	185	26	24	29
F4	**785**	**128**	1534	672	77	25	14
F5	**377**	**110**	615	150	169	4	8
F6	**534**	**103**	2486	74	16	1	0
F7	**1830**	**268**	927	127	72	13	23
F8	**805**	**233**	1982	99	94	14	4
F9	**306**	**52**	1304	432	519	7	5
F10	**1044**	**179**	1533	180	102	18	100
Total used	**8759**						

Table 11 Division of the female tweets in 7 categories

	I	II	III	IV	V	VI	VII
M1	**530**	**31**	2412	87	17	50	51
M2	**335**	**32**	3897	431	57	85	36
M3	**804**	**84**	1682	414	103	68	37
M4	**601**	**36**	1422	148	50	7	78
M5	**1468**	**163**	1155	254	127	0	27
M6	**1131**	**176**	1826	43	25	17	11
M7	**367**	**133**	4862	56	80	14	6
M8	**705**	**128**	1686	612	38	13	14
M9	**637**	**126**	1408	71	75	18	468
M10	**1179**	**114**	593	716	157	2	51
Total used	**8780**						

Table 12 Division of the male tweets in 7 categories

In total, our database thus comprises 17,539 tweets. It is this data that is subjected to two linguistic analyses, which will be explained in close detail in the next section.

3.5 STEP FIVE: LINGUISTIC ANALYSIS

After the careful selection of tweets, two linguistic analyses were conducted. For the first analysis, a more abstract representation of the language used in tweets was created by means of Part-of-Speech tagging (Section 3.5.1). For the second analysis typical sentiment and personality-charged words were derived from the tweets based on well-known lexicons (Section 3.5.2).

3.5.1 Part-of-Speech tagging analysis

We first analysed the data on a more abstract level, by relying on the frequencies of the different word forms used in the tweets of our test subjects, in order to derive whether personality can be connected to particular grammatical choices.

In order to tag all the tweets, an in-house expert system for Dutch linguistic pre-processing was used: the LeTs Preprocess Toolkit (Van de Kauter et al., 2013). In order for this tool to perform Part-of-Speech (POS) tagging, all text is first transformed to UTF-8 encoding, split into sentences and tokenised. Tokenisation is the process where punctuation marks are split off words.

The POS module of LeTs automatically assigns morphosyntactic labels – word forms such as nouns, adjectives, verbs…. – to each token. Since LeTs is normally used to process standard text material, the output of the tool was adapted in order to deal with Twitter-specific tokens such as hashtags, mentions, emoji's...

The output of the POS-tagging was divided into two categories: one comprising all linguistic POS-tags and another comprising Twitter-specific POS-tags. This is illustrated in Table 13.

Linguistic features		**Typical Twitter features**	
Adjectives	Adverbs	Emoji's	Hashtags
Articles	Nouns	Special[15]	URL
Interjections	Numerals	Mentions	Punctuation
Conjunctions	Pronouns		
Prepositions			

Table 13 The major division of the part-of-speech tagging

[15] This category is a miscellaneous category: if the PoS tagger cannot give the word an existing tag, than it receives a 'special' tag. These are for examples, words in a dialect, unknown English words, unknown interjections…

Our focus is on the linguistic features. After all tweets of each subject had been POS-tagged, frequencies of each of these linguistic features were automatically derived and further analysed. This analysis will be extensively discussed in the next chapter.

3.5.2 Lexicon-based analysis

The second linguistic analysis performed on the Twitter data focuses more on the occurrence of certain words, more precisely words that are known to be charged with a certain sentiment or personality on the basis of lexicons.

As sentiment lexicons, the only two existing lexicons for Dutch were used, namely the Duoman lexicon (Jijkoun & Hofmann, 2009) and the Pattern lexicon (De Smedt & Daelemans, 2012). The Duoman lexicon comprises nouns, adjectives, verbs and adverbs that have been manually labelled by two human annotators as either positive, negative or neutral. The Pattern lexicon is a list of adjectives that were manually assigned a polarity value between -1 (negative) and +1 (positive). In order to perform the analysis all tokenised tweets were processed with a Python script and all positive and negative lexicon matches were saved in a list.

From our literature overview we learned that a lexicon that has already been used for personality research is the Linguistic Inquiry and Word Count or LIWC (Pennebaker et al., 2001). Fortunately, this lexicon has also been translated to Dutch by Zijlstra et al. (2004). An analysis with the LIWC, results in a categorisation of all words used into lexical dimensions, accompanied by their relative percentages[16]. Examples of those dimensions are *negemo* for negative emotions, *future* for future tenses and *cogmech* for cognitive processes such as *cause* or *ought*. This analysis could reveal that people scoring particularly high or low on a character trait might be recognized by the use of some lexical dimensions.

For both analyses the same two steps were performed in order to relate the personality traits of our subjects to the language they use. First, a more qualitative analysis was performed where we placed the lowest or highest female and male scorer per personality trait next to each other and compared the scores. If they both score the highest or lowest percentage, those

[16] In this dissertation, we will often use the abbreviations. Therefore, the abbreviations, with their explanation and examples in both English and Dutch can be found in Appendix Four.

results will be mentioned in the result chapter, since this might be the indication of a correlation. This more intuitive analysis was followed by measuring actual Pearson correlations in a second step. It should be noted, however, that since we only have twenty participants, it is highly unlikely that we will find correlations that are statistically significant. In the next chapter we present a thorough discussion of the results.

4 RESULTS

In this result section, we first need to explain in Section 4.1 whether there actually is such a thing as a general personality image of a Twitter user, or not. Can we discover a tendency in what personalities are drawn to Twitter and is there a difference between genders? Then, in Section 4.2 and 4.3, based on the results of our two linguistic analyses we will try to answer our main research question: is it possible to build a realistic and correct personality profile of Twitter users, solely based on the language they use in their tweets?

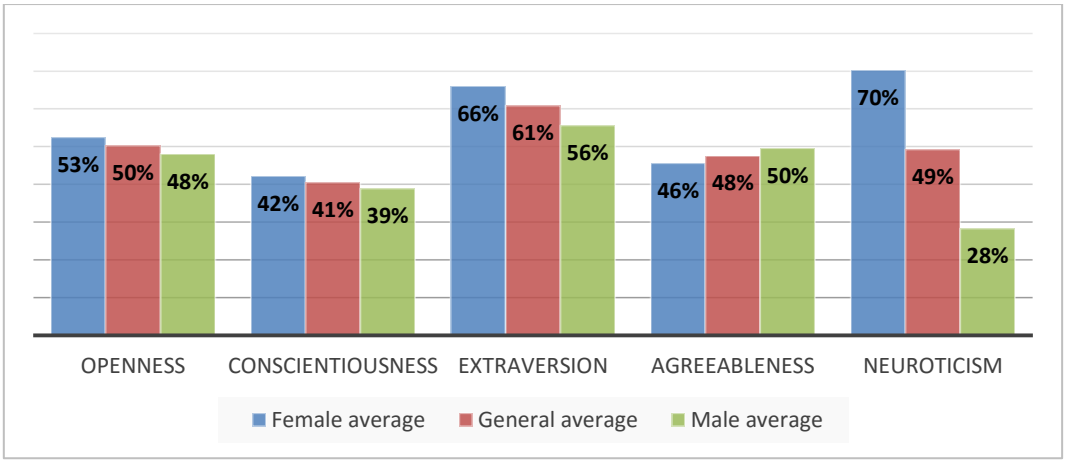
4.1 GENERAL PERSONALITY RESEARCH ON DUTCH DATA

Each respondent filled in the Big Five test, which was introduced in Section 3.3. After rating all 46 statements, such as *I see myself as someone who is talkative* or *I see myself as someone who can be tense* on a Likert-scale from 1 to 7, the respondents filled in their gender and age and passed on to their test results, which were then sent back to us. We already discussed in Section 3.3 how these results should be interpreted, but for clarity's sake a summary is added in Table 14 below.

	Low score (<50%)	High score (>50%)
Openness	closed-minded	open to new experiences
Conscientiousness	disorganized	organised, conscientious (neat)
Extraversion	introverted	extraverted
Agreeableness	disagreeable	agreeable
Neuroticism	calm/relaxed	nervous/high strung

Table 14 Division of the Big Five on both parts of the scale

We combined the personality scores of all participants into separate pie charts; these can be found in Appendix One. In this section we discuss whether it is possible to draw a general image of the average Twitter user. To this purpose, a bar chart containing the general averages assigned to each personality trait of our twenty subjects and the average male and female scores, has been constructed (Bar chart 1).



Bar chart 1 Comparison of the general, female and male average of the Big Five scores

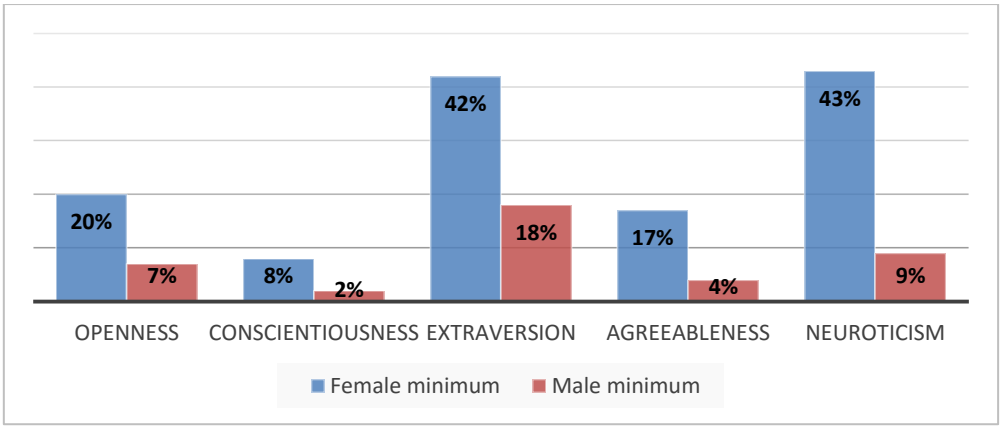
When looking at Chart 1 from a purely personality research point of view, this image corresponds almost completely with what has been said in previous research. Since there was no research found which mentioned Openness explicitly, this trait cannot be compared in this dissertation. However, for Conscientiousness and certainly for Neuroticism, all research (Lynn & Martin, 1997; Budaev, 1999; Vianello et al., 2013) concluded that women would score higher than men, which is definitely the case in our research. For Extraversion, these results confirm Vianello's study (2013): female people do score higher than their male counterparts, something which is in contradiction with the research of Lynn & Martin (1997). The only trait where females did not score higher than males, and which was predicted by both Vianello et al. (2013) and Budaev (1999) was on Agreeableness. In general, the majority of the previous findings thus correspond with our results.

When zooming in on the character traits and social media, previous research by Hamburger (2000), Ross et al. (2009) and Correa (2010) found that people scoring high on Openness, Extraversion and Neuroticism are the individuals which are more easily drawn to social media in general. It has to be said that the general averages demonstrated in our database are not convincing enough to confirm or deny these results. In general, the 20 subjects do score high on Extraversion: 61% on average. With an average score of 50%, they score nor high or low on Openness. The same is true for the trait Neuroticism: on average, the subjects score 49%. Strictly speaking, this would be called scoring low; however, scoring low by only 1% is not convincing enough.

It should be noted and kept in mind that the database used, consisting of only 20 respondents, is very limited. It is consequently necessary to make some remarks. Firstly, it is very likely that scoring high on Extraversion has something to do with social media: scoring 61% on average in a limited database is quite a convincing number and can be seen as a confirmation of previous findings. Secondly, the scores for both Openness and Neuroticism are close to 50%. It might be the case that, if this personality test was run again on a larger database, the results would be the same as in Hamburger (2000), Ross et al. (2009) and Correa's (2010) research, meaning that people would actually score high on both traits.

When examining the three traits, being Openness, Extraversion and Neuroticism, separately for each gender, it can be observed that Extraversion always scores above 50%, making both genders Extravert by definition. Both percentages of Openness fluctuate close to 50%, the women above the average with 53% and the men below average, scoring 48%. Again, if this research was performed on a larger database, both genders might score convincingly higher on both traits.

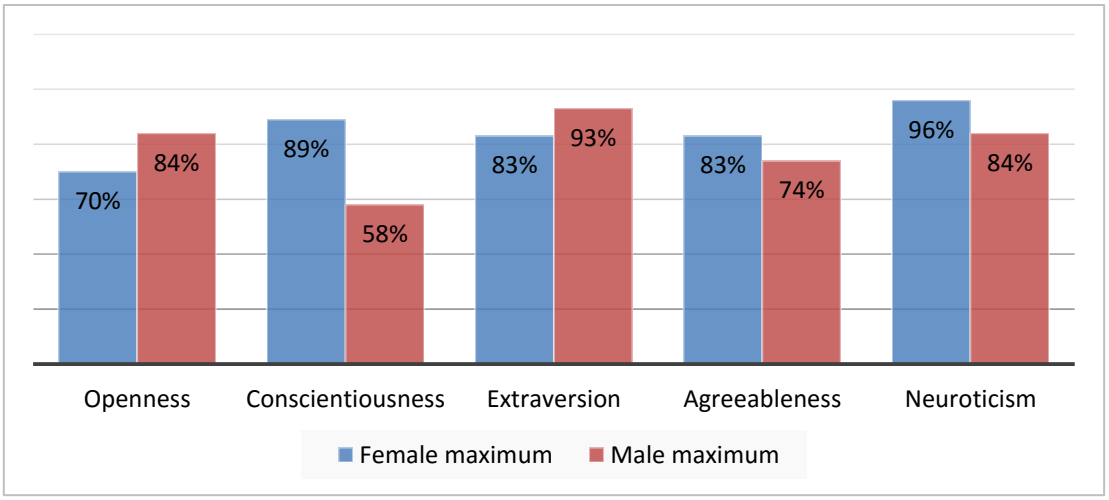
This is not the case for Neuroticism, which is quite clear from Bar chart 1. Based on the general average there might be some doubt about whether or not the typical Twitter user can be considered a neurotic person, the gender averages tell a different story. There is no doubt that female users score high on this trait: with 70% on average, they score 2.5 times as high as their male counterparts, which score only 28% on average. Scoring high on Neuroticism can thus not be seen as a general characteristic to be typical of individuals drawn to social media. This striking difference between women and men on this trait is also clearly visible when looking at the minimum scores, shown in Bar chart 2 below.



Bar chart 2 Comparison of the minimum score of Neuroticism

Whereas the male part of the database has a minimal score of 9% on Neuroticism, the female lowest score is almost five times as high, namely 43%.

What is also striking is that the differences in the minimal scores for each personality trait are often quite large between both genders. The largest difference occurs of course in the trait Neuroticism with a difference of 34%. This percentage is closely followed by Extraversion, where there is a difference of 24%. The other minimum scores have a difference, ranging from 6% (Conscientiousness) to 13% (Openness and Agreeableness).



Bar chart 3 Comparison of the maximum score of Neuroticism

This fairly big difference in minimum scores is not perceived in the maximum scores achieved for the personality traits in both sexes. These range from 9% (Agreeableness), 10% (Extraversion), 12% (Neuroticism), 14% (Openness) and 31% (Conscientiousness).

Hughes (2012) tried drawing up a more specific image of the Twitter user, yet he divided his characteristics into users of Twitter for social or informative purposes. Since we have not asked our respondents how they use Twitter, it is difficult to compare our results with Hughes' study (2012). Moreover, the traits found by Hughes (2012), such as low in Neuroticism and Extraversion, are also the opposite of the general personality drawn to social media as explained in Hamburger (2000), Ross et al. (2009) and Correa (2010). One trait worth discussing, however, is Openness. Both the general image of social media users and Hughes' Twitter users mention scoring high in Openness as a typical characteristic. In our database, however, we do not observe this. The general average score for this trait in our database is 50%, making it difficult to say whether the average tends more to being open to new experiences or being closed-minded.

In general, the personality profile of the average Twitter user that is part of our database is the following: he or she is nor closed-minded or open to new experiences (Openness - 50%), disorganised (Conscientiousness – 41%) and definitely extraverted (Extraversion – 60%). He or she is also slightly disagreeable (Agreeableness – 48%) and relaxed (Neuroticism – 49%). A summary for this general image can be found in Table 15, where we compare the general Twitter user in our database also the general male and female.

	Female		Male		General	
Openness	open	53%	closed-minded	48%	closed-minded / open	50%
Conscientiousness	disorganized	42%	disorganized	39%	disorganized	41%
Extraversion	extraverted	66%	extraverted	56%	extraverted	61%
Agreeableness	disagreeable	46%	(dis)agreeable	50%	disagreeable	48%
Neuroticism	nervous / high strung	70%	calm / relaxed	28%	calm / relaxed	49%

Table 15 Average profile based on the database of this dissertation

To conclude, it is difficult to say whether there is a general combination of the five character traits that are typically drawn to social media and more in particular to Twitter. It is remarkable that the subjects do not score extremely high on Openness, since this was expected and predicted by previous research on both the general use of social media and Twitter. A general remark which can be made is that the subjects do lean close to the profile of people that are in general drawn to social media, which was presented by Hamburger (2000), Ross et al. (2009) and Correa (2010). A larger database is required in order to corroborate these findings.

4.2 LINGUISTIC ANALYSIS

In order to answer our main research question we performed two linguistic analyses, the results of which will now be discussed in closer detail. For the first analysis, a more abstract representation of the language used in tweets was created by means of Part-of-Speech tagging (Section 4.2.1). For the second analysis typical sentiment and personality-charged words were derived from the tweets based on well-known lexicons (Section 4.2.2).

4.2.1 Results of the Part-of-Speech tagging analysis

As explained in Chapter 3, our focus was on the frequencies of the linguistic parts-of-speech. After some preliminary analyses our focus was shifted towards one of the largest linguistic categories present in our database: the pronouns. This class was with 11.07% for the women and 10.32% for the men the fourth largest category, only topped by nouns (F: 14.06% - M: 15.05%), verbs (F: 13.27% - M: 13.25%) and punctuation (F: 11,63% - M:13.73%). Moreover, the use of pronouns has already proven to reveal some interesting findings (Pennebaker et al., 2003) we thought it interesting to investigate this category in closer detail. Golbeck's (2011) research on Twitter already revealed that the pronoun *you* is more often used by people scoring higher on Agreeableness and Openness. Therefore, it might be interesting to focus on which groups people talk about: if someone uses more personal and possessive pronouns referring to him or herself alone or within a group, does this say something about personality?

In order to do so, the personal and possessive pronouns were divided into three categories, namely a category of Self, Others and Thirds. Category I, Self, is the category in which the Dutch pronouns *ik, wij, mijn* and *ons* were included, together with all derived forms. A remark for this column has to be that the pronoun *ik* is often written as *k,* in most cases even attached to the verb itself. It is consequently not excluded that not all first person singular pronouns have been calculated. Category II, Others, consists of the second person singular and plural pronouns, being *je, jullie, jou, jouw* and all derived forms such as *gij* or *ge*. This category represents the 'other' person and a group where the user him or herself is not included. It represents a contrast and a feeling of exclusion. Finally, Category III, Thirds, includes all references to other people which are more or less unknown to the user: *hij, haar, zij, hun* and all derived forms are part of that category.

In the following table, the three categories for the personal pronouns are presented, together with the corresponding percentages.

Personal pronouns											
Category I (Self: *ik - wij*)											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**average**
F	64,72	64,53	59,03	48,86	75,47	65,36	60,38	69,89	51,21	67,47	**62,69**
M	59,04	50,67	42,94	66,01	44,15	63,66	59,77	58,59	49,76	48,99	**54,36**
Category II (Other: *jij - jullie*)											
F	7,31	12,08	13,37	22,03	8,86	13,04	17,34	9,20	25,76	13,78	**14,28**
M	18,67	20,00	22,70	12,64	23,59	15,64	16,91	12,74	14,35	20,18	**17,74**
Category III (Third: *zij/hij - zij*)											
F	27,97	23,40	27,60	29,11	15,67	21,60	22,28	20,92	23,03	18,74	**23,03**
M	22,29	29,33	34,36	21,35	23,26	20,70	23,32	28,67	35,89	30,38	**26,96**

Table 16 Use of the personal pronouns

As can be derived from Table 16, both genders score over 50% when it comes to Category I, referring to themselves or a group of people in which they are included. Both categories which talk about other people, either in the second or the third person, do not even reach 30% as a maximum score. Both genders thus talk more about themselves than about others. What is striking is that female users talk even more about themselves and the groups they are included in, with a difference of 8%. This difference is striking since it is the only category where the female part of the database scored higher than the male part, and the gap is larger as well. For the second and third category, the men do score higher on average, but the difference fluctuates only around 3%.

Possessive pronouns											
Category I (Self: *mijn - ons*)											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**average**
F	69,35	72,50	63,01	63,33	70,94	78,48	67,84	78,69	57,65	73,26	**69,51**
M	79,45	60,00	53,16	71,29	38,44	62,28	55,00	61,95	68,87	54,05	**60,45**
Category II (Other: *jouw*)											
F	16,13	11,25	19,92	20,67	11,33	11,92	11,96	10,65	21,18	19,44	**13,45**
M	8,22	15,56	15,19	5,94	20,00	16,37	23,00	11,71	15,89	25,00	**15,69**
Category III (Third: *hun*)											
F	14,52	16,25	17,07	16,00	17,73	9,60	20,20	10,65	21,18	7,29	**15,05**
M	12,33	24,44	31,65	22,77	41,56	21,35	22,00	26,34	15,23	20,95	**23,86**

Table 17 Use of the possessive pronouns

For the possessive pronouns, shown in Table 17, the same tendency is visible. Women tend to talk more about their own possessions or the possessions of the group they belong to, whereas men talk more about the possession of others. It is striking that the use of the first person singular and plural in general exceeds 60%, whereas the second and third category score in between 10% and 25%. In the first category, the women score, exactly as with the personal

pronouns, almost 10% higher than their male counterparts. For the other two categories, the male percentage is again higher; however, the difference between the use of *hun* is greater (8%) than the difference in use of the second category (2%) with *jouw*.

After this first qualitative comparison, we checked whether there are any correlations that indicate the relation between the personality traits and the use of certain pronouns. Previous research by Golbeck (2011) mentions a higher use of the pronoun *you* by people scoring high in Agreeableness and Openness. However, in our correlations, which can be found in Appendix Two, these findings were not corroborated. Almost no correlations higher than 0.5 or -0.5 were demonstrated and none of the correlations was significant. Agreeableness and the pronoun *you* is in our database negatively correlated with -0.33, whereas the trait Openness is also negatively correlated by -0.15. This is the complete opposite of what has been found in previous research and is most probably due to the small size of our database.

The highest correlations found were with the trait Neuroticism: the use of the possessive pronoun *hun* is negatively correlated (-0.54) with the trait, meaning that highly neurotic people will use this pronoun less. This result is also the closest it gets to a statistically significant result (p-value of 0.09). Other correlations are the relation between *zij, haar, hij,* which is positively correlated with 0.38 and the use of first person possessives with 0.36. Again, these results are not statistically significant.

To sum up, we did observe a difference in our database when it comes to the use of particular pronouns. However, this difference cannot be attributed to differences in personality but only to differences in gender. Since Twitter is a microblogging site, it is quite logical that own opinions and own points of view are more present, thus that the use of first person pronouns is more used by both genders. However, what is striking is that there is a female tendency to use this self-centred point of view, either in the singular or plural form, visibly more than men.

4.2.2 Results of the lexicon-based analysis

As mentioned in Chapter 3, three different analyses based on lexicon look-ups were performed. We relied on two Dutch sentiment lexicons: Duoman and Pattern and one well-known lexicon for personality research, the Dutch version of the LIWC.

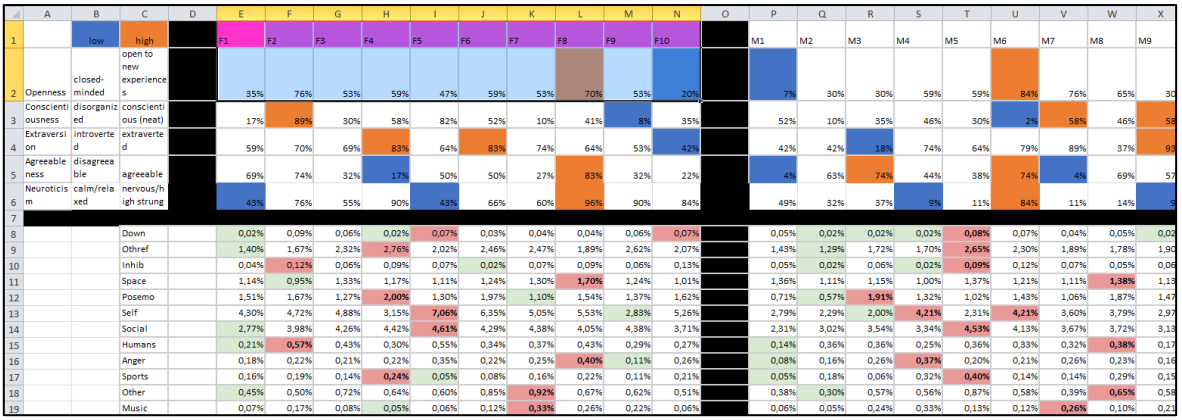
Both the Pattern and Duoman lexicons reveal the amount of words used in the tweets by our subjects having a negative or a positive connotation percentagewise. The results of our twenty subjects are presented in Table 18, where a negative score reveals that more negative words were used by a person and a positive score the opposite. Quite to our surprise, every respondent uses more positive than negative words, with one exception: M2 got a ratio of 0.25% more negative words when analysed with the Pattern lexicon. If we have a closer look at the personality of this person (pie chart M2 in Appendix One), we see that this person did not score exceptionally high or low on any trait. This is why we decided to shift our attention more towards the analysis of our data with the LIWC lexicon.

Pattern: positive words minus negative words										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
F	0,59%	0,60%	0,14%	1,33%	0,06%	0,35%	0,23%	0,43%	0,02%	0,53%
M	0,21%	**- 0,25%**	0,37%	0,51%	0,22%	0,26%	0,21%	0,97%	0,60%	0,60%
Duoman: positive words minus negative										
F	1,53%	1,10%	1,03%	2,57%	1,87%	1,86%	0,52%	1,54%	0,92%	1,66%
M	0,82%	0,91%	0,99%	0,63%	1,39%	0,57%	0,71%	1,13%	1,88%	1,70%

Table 18 Comparison of the results of the Pattern & Duoman analysis

When processing the data with the LIWC lexicon, the outcome is a table listing all LIWC categories that were found in the data, accompanied by a relative percentage.

We first performed a more qualitative analysis for which a general overview of the retrieved percentages was created. The percentages of the personality traits were put into a column, together with the percentages of the LIWC analysis, which is illustrated in the image below. Per gender, the highest and lowest score was highlighted per LIWC dimension found in the analysis. In the image, the green highlighted numbers represent the lowest scores and the red highlights cells the highest scores.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		low	high		F1	F2	F3	F4	F5	F6	F7	F8	F9	F10		M1	M2	M3	M4	M5	M6	M7	M8	M9
2	Openness	closed-minded	open to new experiences		35%	76%	53%	59%	47%	59%	53%	70%	53%	20%		7%	30%	30%	59%	59%	84%	76%	65%	30
3	Conscientiousness	disorganized	conscientious (neat)		17%	89%	30%	58%	82%	52%	10%	41%	8%	35%		52%	10%	35%	46%	30%	2%	58%	46%	58
4	Extraversion	introverted	extraverted		59%	70%	69%	83%	64%	83%	74%	64%	53%	42%		42%	42%	18%	74%	64%	79%	89%	37%	93
5	Agreeableness	disagreeable	agreeable		69%	74%	32%	17%	50%	50%	27%	83%	32%	22%		4%	63%	74%	44%	38%	74%	4%	69%	57
6	Neuroticism	calm/relaxed	nervous/high strung		43%	76%	55%	90%	43%	66%	60%	96%	90%	84%		49%	32%	37%	9%	11%	84%	11%	14%	9
7																								
8			Down		0,02%	0,09%	0,06%	0,02%	0,07%	0,03%	0,04%	0,04%	0,06%	0,07%		0,05%	0,02%	0,02%	0,02%	0,08%	0,07%	0,04%	0,05%	0,02
9			Othref		1,40%	1,67%	2,32%	2,76%	2,02%	2,46%	2,47%	1,89%	2,62%	2,07%		1,43%	1,29%	1,72%	1,70%	2,65%	2,30%	1,89%	1,78%	1,90
10			Inhib		0,04%	0,12%	0,06%	0,09%	0,07%	0,02%	0,07%	0,09%	0,06%	0,13%		0,05%	0,02%	0,06%	0,02%	0,09%	0,12%	0,07%	0,05%	0,06
11			Space		1,14%	0,95%	1,33%	1,17%	1,11%	1,24%	1,30%	1,70%	1,24%	1,01%		1,36%	1,11%	1,15%	1,00%	1,37%	1,21%	1,11%	1,38%	1,13
12			Posemo		1,51%	1,67%	1,27%	2,00%	1,30%	1,97%	1,10%	1,54%	1,37%	1,62%		0,71%	0,57%	1,91%	1,32%	1,02%	1,43%	1,06%	1,87%	1,47
13			Self		4,30%	4,72%	4,88%	3,15%	7,06%	6,35%	5,05%	5,53%	2,83%	5,26%		2,79%	2,29%	2,00%	4,21%	2,31%	4,21%	3,60%	3,79%	2,97
14			Social		2,77%	3,98%	4,26%	4,42%	4,61%	4,29%	4,38%	4,05%	4,38%	3,71%		2,31%	3,02%	3,54%	3,34%	4,53%	4,13%	3,67%	3,72%	3,13
15			Humans		0,21%	0,57%	0,43%	0,30%	0,55%	0,34%	0,37%	0,43%	0,29%	0,27%		0,14%	0,36%	0,36%	0,25%	0,36%	0,33%	0,32%	0,38%	0,17
16			Anger		0,18%	0,22%	0,21%	0,22%	0,35%	0,22%	0,25%	0,40%	0,11%	0,26%		0,08%	0,16%	0,26%	0,37%	0,20%	0,21%	0,26%	0,23%	0,16
17			Sports		0,16%	0,19%	0,14%	0,24%	0,05%	0,08%	0,16%	0,22%	0,11%	0,21%		0,05%	0,18%	0,06%	0,32%	0,40%	0,14%	0,14%	0,29%	0,15
18			Other		0,45%	0,50%	0,72%	0,64%	0,60%	0,85%	0,92%	0,67%	0,62%	0,51%		0,38%	0,30%	0,57%	0,56%	0,87%	0,58%	0,39%	0,65%	0,58
19			Music		0,07%	0,17%	0,08%	0,05%	0,06%	0,12%	0,33%	0,26%	0,22%	0,06%		0,06%	0,05%	0,24%	0,33%	0,13%	0,12%	0,26%	0,10%	0,21

Figure 10 Screenshot of the percentages of the personality traits and LIWC analysis

Next, the highest scorers – one male and one female – per trait were compared to each other and the same was done for the lowest scorers. Whenever both subjects had two green or red highlighted numbers, meaning they both scored the lowest or highest on a particular LIWC dimension and on a particular trait, they will be further discussed because this might hint at a connection between LIWC and personality.

Let us demonstrate this with an example for the trait Openness. The very first step is to compare all female scores for each LIWC dimension and highlight the highest and lowest percentages. The same is done for the male scores. Secondly, the scores of the character traits are considered, in this example Openness. To see what could be typical for closed-minded people, we have to compare both the female and male lowest scorer on that particular trait, which in our database are V10 who scores 20% on Openness and M1 who scores 8%. In order to perform this analysis a second table is made, as illustrated below.

	A	B	C	D
1		M1	F10	
2	Down	0,05%	0,07%	
3	Othref	1,43%	2,07%	
4	Inhib	0,05%	0,13%	
5	Space	1,36%	1,01%	
6	Posemo	0,71%	1,62%	
7	Self	2,79%	5,26%	
8	Social	2,31%	3,71%	
9	Humans	0,14%	0,27%	
10	Anger	0,08%	0,26%	
11	Sports	0,05%	0,21%	
12	Other	0,38%	0,51%	
13	Music	0,06%	0,06%	
14	Discrep	1,25%	1,88%	
15	Nonfl	0,36%	0,32%	
16	You	0,59%	1,09%	
17	Article	2,68%	5,65%	
18	Incl	3,92%	5,05%	
19	Sexual	0,03%	0,09%	
20	Achieve	0,05%	0,14%	
21	Relig	0,02%	0,11%	
22	Cogmech	2,44%	4,17%	
23	We	0,30%	0,08%	
24	Senses	1,03%	1,34%	
25	Eating	0,22%	0,35%	

Figure 11 Percentages of the lowest scorers on Openness

When comparing both columns, it is necessary to look for similar scores in the LIWC dimensions, meaning either the highest or lowest scores. This step of the analysis is not so difficult because these scores were already highlighted in the first step. If both columns are highlighted in the same colour, such as row 24 in Figure 11, this might be an interesting finding and the Pearson correlations between this LIWC dimension and personality trait should be checked carefully. In this example, the dimension s*enses* had two times the lowest score, which might be an indication of a link between being closed-minded and not talking about sensory and perceptual processes such as *see*, *touch* or *listen.*

After this first more intuitive phase, these findings were compared with Pearson correlations.

First phase: intuitive, qualitative results

Table 19 summarises the main findings of our qualitative study. For the exact analyses we refer to Appendix Three. The results presented in Table 19 are thus the result of both genders scoring particularly high or low on a specific dimension while scoring high or low on a specific trait. Even though saying that people scoring generally high or low on a trait will score in a certain way on a specific dimension might lead to a more general conclusion, it should be repeated that this is a very limited database and that some of these findings might only occur in this database by coincidence.

Trait	High / low score	Dimension found by LIWC
Openness	Low	• talk less about sensory and perpetual processes (*see*, *touch*, *listen*)
	High	• use less future tenses
Conscientiousness	Low	• use more articles
	High	• talk more about eating, food and dieting • talk more about physical states and functions • talk more about grooming • talk less about insight (*think*, *know*) • talk less about anxiety
Extraversion	Low	• talk more about death • talk more about sadness • talk more about television, movies and games • talk more about negative emotions in general
	High	• talk less about being down (space) • talk less about discrepancies, fights, arguments

		• talk with less nonfluencies • talk more about inclusiveness (space)
Agreeableness	Low	• talk les about cognitive processes (*cause*, *ought*, *know*) • talk less in present tenses • talk less with negations • talk less about anxiety
	High	• talk more about sensory and perceptual processes (see, touch, listen) • talk more about communication (*talk*, *share*, *converse*) • talk less in future tenses
Neuroticism	Low	• talk less about being down (space) • talk more about total first person • talk more about friends • talk more about time • talk more with numbers • talk more about certainties • talk more with negations
	High	• talk less in future tenses

Table 19 Summary of the findings of the LIWC analysis

It should be repeated that we relied on a limited database and that some of these findings only occurred by coincidence. Nevertheless, some do make logical sense. People who are more introvert scored the highest in very logical LIWC dimensions: they talk more about death (e.g. *bury*, *coffin*, *kill*)[17], sadness (*crying*, *grief*), television, movies and games and more about negative emotions in general (*hurt*, *ugly*, *nasty)*. People scoring low on Neuroticism, and therefore calmer people, talk more about friends (*buddy, friend, neighbour*), time (*end*, *until, season*) and certainties (*always*, *never*), whereas they surprisingly also talk more about themselves, something which is in contrast with our highly neurotic female half of the database, who score high on Neuroticism and also on the use of the first person pronouns. What was striking as well, is that highly organised people, or highly conscientious people, talk more about their physical appearance in general: they talk a lot about eating, food and dieting, about physical states and functions and grooming[18].

[17] The examples of the dimensions are collected from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.600.7227&rep=rep1&type=pdf

[18] Grooming means "the things you do to make your appearance clean and neat" (Cambridge Dictionary, s.d.).

Second phase: Pearson correlations

As mentioned previously, the found data from the first phase would be controlled against Pearson correlations. When interpreting a Pearson correlation, any p-value lower than 0.05 can be considered statistically significant. This means that any correlation with a p-value under 0.05 can and will almost certainly be correct on other databases as well. We will start by comparing the results of our qualitative analysis with an analysis of the correlations. A summary can be found below:

Trait	Dimension	Correlation	P-value	Corresponds?
Openness	senses	0.6242	* 0.0473	no
	future	- 0,0593	0.8698	yes
Conscientiousness	article	- 0,3742	0,2802	yes
	eating	0,1600	0,6562	yes
	physical	0,5088	0,1254	yes
	groom	0,2216	0,5347	yes
	insight	- 0,2042	0,5681	yes
	anx	- 0,0612	0,8657	yes
Extraversion	death	- 0,1042	0,7730	yes
	sad	- 0,1483	0,6804	yes
	tv	0,0013	0,9970	no
	neemo	- 0,1436	0,6901	yes
	down	-0,0526	0,8845	yes
	discrep	0,1833	0,6093	no
	nonfl	- 0,4868	0,1460	yes
	incl	0,3793	0,2730	yes
Agreeableness	cogmech	0,1540	0,6685	yes
	present	0,1879	0,6002	yes
	negate	0,2283	0,5219	yes
	anx	0,2025	0,5715	yes
	senses	0,2752	0,4369	yes
	comm	0,3487	0,3173	yes
	future	- 0,3214	0,3595	yes
Neuroticism	down	0,2607	0,4624	yes
	self	0,3721	0,2831	no
	friends	- 0,0526	0,8844	yes
	time	0,1551	0,6662	no
	numbers	0,1487	0,6795	no
	certain	0,0052	0,9883	no
	negate	- 0,0795	0,8258	yes
	future	- 0,1647	0,6468	yes

Table 20 Found correlations of the first phase with the Pearson correlation

We will start by discussing the general image: only 7 out of the 31 traits found and highlighted in bold have a correlation which does not correspond to the relation found during the first phase, one even with statistical significance (indicated with a star). People scoring high on Openness do talk more about sensory and perpetual processes (*senses*, e.g. *see, touch, listen*), which is illustrated by the rather high correlation of 0.62, which is also statistically significant. This is the opposite of what was seen in the first phase, where scoring low on Openness was thought to coincide with talking less about this dimension. All what was found in the first phase about Extraversion and television (*tv*) or discrepancies (*discrep*, e.g. *should, would, could*), or Neuroticism with talking about one's own (*self*, e.g. *we, I, me*), time (*time*, e.g. *end, until, season*), numbers (*numbers*, e.g. *second, thousand)* or certainty (*certain*, e.g. *always, never*) is shown to be incorrect with non-statistically significant and overall very low correlations. What should be noted here is that the positive correlation of Neuroticism and talking with first person personal pronouns is a tendency that we have seen before in the PoS tagging: the female half of our database, who were highly neurotic, do use these pronouns more.

All other traits do show the same tendency as was found on the basis of the qualitative analysis. However, none of these correlations is statistically relevant or particularly high, with the correlation between Extraversion and the physical states (0.50) being fairly high and relatively close to being statistically relevant (0.12).

In short, we can say that the comparisons made between male and female scores on traits can give an image of the general tendencies; however, they are not 100% accurate and sometimes even wrong. It is thus important, in further research, to rely on Pearson's correlations.

But maybe there were also other high correlations between a certain personality trait and an LIWC dimension that could found in our database. This is why in a final analysis we list all correlations scoring above 0.4 in absolute numbers (Table 21).

Trait	Dimension	Correlation	P-value
> 0.5			
Openness	social	0.6193	* 0.0497
	humans	0.5254	0.1112
	senses	0.6242	* 0.0473
	hear	0.5042	0.1296
	present	0.5227	0.1134
	communication	0.5284	0.1087
Conscientiousness	physical	0.5088	0.1254
Neuroticism	inhib	0.5237	0.1126
0.4 – 0.5			
Openness	cogmech	0.4226	0.2165
	eating	0.4074	0.2355
	tentat	0.4914	0.1415
	swear	0.4139	0.2273
	pronoun	0.4244	0.2143
	assent	- 0.4008	0.2442
	negate	0.4127	0.2288
Conscientiousness	sexual	0.4106	0.2315
	home	0.4072	0.2358
	friends	0.4904	0.1425
	body	0.4936	0.1394
	negate	- 0.4163	0.2243
	sleep	0.4321	0.2050
Extraversion	pronoun	0.4548	0.1791
	nonfl	- 0.4868	0.1460

Table 21 All dimensions scoring higher than 0.4 with Pearson correlations

All correlations which score over 0.5 are considered high. In our database, we found 8 out of 31 dimensions to be highly correlated, mostly with the trait Openness: social processes (*social*, e.g. *mate, talk, they, child*), humans (*humans*, e.g. *baby, adult, boy*), sensory and perpetual processes (*senses*, e.g. *see, touch, hear*), hearing (*hear*, e.g. *listening*, *hearing*), present tenses (*present*) and communication (*comm*). For Conscientiousness, talking about physical states (*physical*) was found to correlate positively and people scoring higher on Neuroticism tend to talk more about inhibitions (*inhib*, e.g. *block, constrain, stop*). Much to our surprise, two of these results were actually statistically relevant (these are indicated with a star in Table 21), namely the positive correlation between Openness and the mentioning of social processes (*social*) such as *mate, talk, they, child*…; and the positive correlation with the description of sensory and perceptual processes (*senses*) such as *see, touch* or *listen.* This is surprising because our database is only built on the data of twenty people. It is thus definitely

worthwhile to conduct a more elaborate study and see whether the highly correlated items mentioned in our result will also return in a research with a larger database.

In conclusion, the similarities found between low and high scorers on a particular trait on the basis of our qualitative and that did hint at a general tendency is, after measuring and comparing the Pearson correlations, not always accurate and sometimes even plain wrong. In other words, the intuitive research conducted in the first phase might already depict a somewhat general image; however, the method is far from flawless and only feasible in a small database. Relying solely on the Pearson correlations; two dimensions do show statistically significant results. Those dimensions are sensory and perceptual processes (*senses*) and social references (*social*). What has been said in Golbeck's 2011-study was not supported by our database: people high on Conscientiousness did not necessarily have a high negative correlation with words about death (*death*; - 0.0341), a high positive correlation with negative emotions (*negemo*; 0.1080) or words about sadness (*sad*; 0.0438). We do not have a dimension concerning achievements and money, so the finding about a negative correlation with scoring high on Agreeableness cannot be confirmed nor contradicted on the basis of this study.

5 CONCLUSION

The goal of this dissertation was to find out whether language alone can draw a realistic and correct personality profile of a Twitter user. In order to answer this question we first discussed how the Big Five of Personality, the personality model used for this research, came into existence and how it has been used to measure the relation between personality and social media, the choice of platform, language and gender.

Next, we explained how twenty respondents, 10 male and 10 female persons, were persuaded to participate in our research. These subjects filled in a personality test and gave their consent to have their Dutch tweets downloaded and analysed. On these tweets two linguistic analyses were then performed: a more abstract analysis by means of Part-of-Speech tagging and an analysis where sentiment and personality-charged words were derived from the tweets based on well-known lexicons. A close analysis of all available data led to some interesting results.

Firstly, since the results of the Big Five personality test of all 20 subjects were available, our findings were compared with previous research on the link between personality, gender and social media. We tried to answer the question whether it is possible to draw a general image of a social media or Twitter user.

The differences found in gender were confirmed, namely that men score higher on Agreeableness and women higher on Conscientiousness, Extraversion and Neuroticism. However, confirmation for the trait Extraversion could only be found when compared to Vianello's research (2013), which is in contradiction with Lynn & Martin's older 1997-study.

When it comes to the Big Five and social media in general, which was researched by Hamburger (2000), Ross et al. (2009) and Correa (2010), one trait corresponds completely, namely scoring high on Extraversion. For the traits Openness and Neuroticism, however, our database might have been too small: the numbers fluctuate around 50%, which makes it impossible to say whether scoring high on both traits is something frequent on social media. On the other hand, this result does not contradict the previous results either.

What is remarkable is that both the social media and Twitter user are said to score high in Openness, which is not supported by our database: our subjects score on average 50% on this trait.

In general, it can be said that it is not quite possible to confirm previous findings since the database is limited, however, it should be said that the findings are, certainly for the Big Five and gender and social media, very much in line with previous studies.

Secondly, based on the Part-of-Speech analysis of the tweets, we found that in general the use of pronouns in general did not seem to reveal any correlation with a particular trait; therefore, a deeper research was conducted on the use of personal and possessive pronouns. This more thorough analysis did not reveal any particular link with personality; however, there is a clear difference in use between both genders. Both male and female users talk more about themselves and groups they belong to, in other words, they use more first person pronouns, both singular and plural. This can easily be explained by Twitter being a microblogging website: it is very logical to talk more about one's own opinion and comments. What was striking, however, is that women use more first person pronouns, with a difference of 10%. This is the case for both the personal and the possessive pronouns. An addition that should be made here is that, even though the PoS tagging did not reveal any link with a personality trait, the LIWC analysis did: the first person personal pronouns, equal to the dimension *self*, shows a not statistically significant positive correlation with Neuroticism.

Previous findings in research (Golbeck, 2011) claiming that the use of the pronoun *you* did not have a positive correlation with Openness and Agreeableness were not corroborated. Actually, quite the opposite was true in our database. It should be noted, however, that he correlations found in our database did not exceed 0.35 and were not statistically significant.

Thirdly, based on the lexicon analyses no clear results were conveyed with two Dutch sentiment lexicons, i.e. Pattern and Duoman. No trait had a specifically high or low use of positive and negative words. Moreover, all but one respondent used more positive words than negative words. Since that one respondent did not score particularly high or low on a trait, we can only guess what the origins are of this difference.

The analysis with the Dutch Linguistic Inquiry Word Count (LIWC) tool, however, did provide us with some interesting findings on how often certain dimensions of words are used with a particular personality trait. These were achieved after first performing an intuitive qualitative research, after Pearson correlations were measured. Of the 31 links between linguistic LIWC dimension and personality traits, found intuitively during the first phase, 24 dimensions were also found to correlate when looking at the Pearson correlations. However, almost none of those were particularly high correlations. Even more, a comparison between a 21-item list of the highest Pearson correlations (scoring over 0.4) and the 31 intuitive correlations resulted in only three remaining LIWC dimensions. These were the occurrence of words relating to sensory and perpetual processes (*senses*) with the personality trait Openness (0.62), Conscientiousness with physical states (*physical*; 0.50) and Extraversion with non-fluencies (*nonflu*; 0.48). In other words, our first intuitive research did already provide a generally correct image regarding personality, however, the correlations found were not outspoken.

As explained above, a great challenge lied in working with such a limited database. However, much to our surprise, we did discover two statistically significant correlations. The trait Openness, which also conveyed most of the highest correlations, is positively correlated with social terms, such as *family* and *friends* and also with sensory and perceptual processes such as *see*, *touch* or *listen.* This finding is a great stimulus to continue this research on a greater database: the high correlations could even be more outspoken if only they were researched on more data.

To conclude, if we ask ourselves whether Twitter updates can tell us something about the personality of a user, the answer is precariously yes. Having statistically significant results in a small database raises hopes that more of these can be found in a larger database. In future research, it is definitely recommended to collect more data: this will help in defining more concretely the general image of a social media user and, of course, in discovering which language items are typical for specific Big Five traits. It is certain that this field should be explored even further, since those significant results might be a good first indicator of what is important for online personality prediction. Who knows if that research will enable us to go one step further and predict personality automatically? Our database definitely forms a valuable gold standard to conduct such research in the near future.

6 BIBLIOGRAPHY

Allport, G.W. & Odbert, H.S. (1936). Trait-names, a psycho-lexical study. *Psychological review publications*, *47*(1), i-171. doi:10.1037/h0093360

Back, M.D., Stopfer, J.M., Vazire, S., Gaddis, S., Schmuckle, S.C., Egloff, B. et al. (2010). Facebook Profiles Reflect Actul Personality, Not Self-Idealization. *Psychological Science, 21*(3), 372-374. doi:10.1177/0956797609360756

Barker, V. (2009). Older Adolescents' Motivations for Social Network Site Use: The Influence of Gender, Group Identity, and Collective Self-Esteem. *Cyberpsychology & Behaviour, 12*(2), 209-213. Doi:10.1089/cpb.2008.0228

Beck, J. (1999). *Jesus & Personality Theory: Exploring the Five-Factor Model.* Illinois: InterVarsity Press

Budaev, S.V. (1999). Sex differences in the Big Five personality factors: testing an evolutionary hypothesis. *Personality and Individual differences, volume 26,* 801-813.

Cattell, R.B. (1946). *Personality structure and measurement of personality.* Oxford, England: World Book Company.

Correa, T., Hinsley, A. & Gil de Zúñiga, H. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2), 247-253. Doi:10.1016/j.chb.2009.09.003

De Smedt, T. & Daelemans, W. (2012). Vreselijk mooi! Terribly beautiful: a subjectivity lexicon for Dutch adjectives. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, 3568–3572.

Demeester, S. (2014). Twitter opent kantoor in Brussel. *De Tijd*. Retrieved from http://www.tijd.be/ondernemen/media_marketing/Twitter_opent_kantoor_in_Brussel.9569376-3133.art

Digman, J.M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology, 41*, 417-440. doi: 10.1146/annurev.ps.41.020190.002221

Facebook. (2016). Company info. Retrieved on May 12th 2016 from http://newsroom.fb.com/company-info/

Galton, F. (1884). Measurement of Character. *Fortnightly Review, 36.* 179-185. doi: 10.1037/11352-058

Golbeck, J., Robles, C. & Turner, K. (2011a). Predicting personality with social media. *CHI '11 Extended Abstracts on Human Factors in Computing Systems,* 253-262. doi:10.1145/1979742.1979614

Golbeck, J. Robles, C., Edmondson, M. & Turner, K. (2011b). Predicting personality with Twitter. *Privacy, Security, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, and IEEE International Conference on Social Computing (SocialCom).* 149-156. doi:10.1109/PASSAT/SocialCom.2011.33

Goldberg, L. (1990). An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216-1229. doi:10.1037/0022-3514.59.6.1216

Goldberg, L. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26-34. doi:10.1037/0003-066X.48.1.26

Gosling, S.D. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528. doi:10.1016/S0092-6566(03)00046-1

Grooming. (s.d.). In *Cambridge Dictionaries online*. Retrieved from http://dictionary.cambridge.org/dictionary/english/grooming

Hamburger, Y.A. & Ben-Artzi, E. (2000). The relationship between extraversion and neuroticism and the different uses of the Internet. *Computers in Human Behavior, 16*(4), 441-449. doi: 10.1016/S0747-5632(00)00017-0

Hughes, D.J., Rowe, M., Batey, M. & Lee, A. (2012). A tale of two sites, Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior, 28*(2), 561-569. doi:10.1016/j.chb.2011.11.001

Jijkoun, V. & Hofmann, K.: 2009, Generating a non-English subjectivity lexicon: Relations that matter. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, 398-405.

John, O.P. & Srivastava, S. (1999). The Big-Five Trait Taxonomy: History, Measurement and Theoretical Perspectives. In Pervin, L. & John, O.P. (Eds.), *Handbook of personality* (pp. 102-138). New York: Guilford.

Judge, T.A., Higgins, C.A., Thoresen, C.J. & Barrick, B.R. (1999). The Big Five Personality Traits, General Mental Ability, And Career Success Across The Life Span. *Personnel Psychology, 52*(3), 621-652. doi:10.1111/j.1744-6570.1999.tb00174.x

Lenhart, A., Purcell, K., Smith, A. & Zickuhr, K. (2010). Social Media & Mobile Internet Use among Teens and Young Adults. Millennials. *Pew Internet & American Life Project.*

Lynn, R. & Martin, T. (1997). Gender differences in Extraversion, neuroticism and psychotism in 37 nations. *The Journal of Social Psychology, volume 137*(3), 369-373. doi: 10.1080/00224549709595447

McWorther, J. (2013) *Txtng is killing language. JK!!!*. TED-talk. Retrieved from https://www.youtube.com/watch?v=UmvOgW6iV2s

Microblog. (s.d.). In *Cambridge Dictionaries Online*. Retrieved from http://dictionary.cambridge.org/dictionary/english/microblog

Mischel, W. (2013). *Personality and Assessment*. Psychology Press.

Murray, B.R. & Mount, M.K. (1991). The Big Five Personality Dimensions And Job Performance: A Meta-Analysis. *Personnel Psychology, 44*(1), 1-26. doi: 10.1111/j.1944-6570.1991.tb00688.x

Norman, W.T. (1967). *2,800 personality trait descriptors: normative operating characteristics for a university population.* S.l.: University of Michigan, Dept. of Psychology.

Pennebaker, J. W., Francis M.E. & Booth R.J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001.* Mahwah: Lawrence Erlbaum Associates.

Pennebaker, J.W., Mehl, M.R. & Niederhoffer, K.G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual review of psychology, volume 54*(1), 547-577. doi: 10.1146/annurev.psych.54.101601.145041

Primi, R., Ferreira-Rodrigues, C.F. & De Francisco Carvalho, L. (2014). Cattell's Personality Factor Questionnaire (CPFQ): Development and Preliminary Study. *Paidéia, 24*(57), 29-37. doi:10.1590/1982-43272457201405

Rosen, P.A. & Kluemper, D.H. (2008). The Impact of the Big Five Personality Traits on the Acceptance of Social Networking Website. *AMCIS 2008 Proceedings*. Paper 274. Retrieved from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1276&context=amcis2008

Ross, C., Orr, E.S., Sisic, M., Arseneault J.M., Simmering, M.G. & Orr, R.R. (2009). Personality and motivations associated with Facebook use. *Computers in Human Behavior, 25*(2), 578-586. doi:10.1016/j.chb.2008.12.024

Schwartz, A.H., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M. et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PloS one, 8*(9), e73791. Doi: 10.1371/journal.pone.0073791

Shaver, P.R. & Brennan, K.A. (1992). Attachment Styles and the "Big Five" Personality Traits: Their Connections with Each Other and with Romantic Relationship Outcomes. *Personality and Social Psychology Bulletin, 18*(5), 536-545. doi:10.1177/0146167292185003

Social media. (s.d.). In *Oxford Dictionaries.* Retrieved from http://www.oxforddictionaries.com/definition/english/social-media

Statista. (2016). Leading social networks worldwide as of April 2016, ranked by number of active users (in millions). Retrieved on May 12th 2016 from http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Tupes, E. C. & Christal, R. E. (1961). Recurrent personality factors based on trait ratings (USAF ASD Tech. Rep. No. 61-97). Lackland Air Force Base, TX: U.S. Air Force.

Twitter. (2016). Bedrijf About. Retrieved on May 12th 2016 from https://about.twitter.com/nl/company

User-generated content. (s.d.) In *Oxford Dictionaries.* Retrieved from http://www.oxforddictionaries.com/definition/english/user-generated?q=user-generated+content

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal, volume 3,* 103-120.

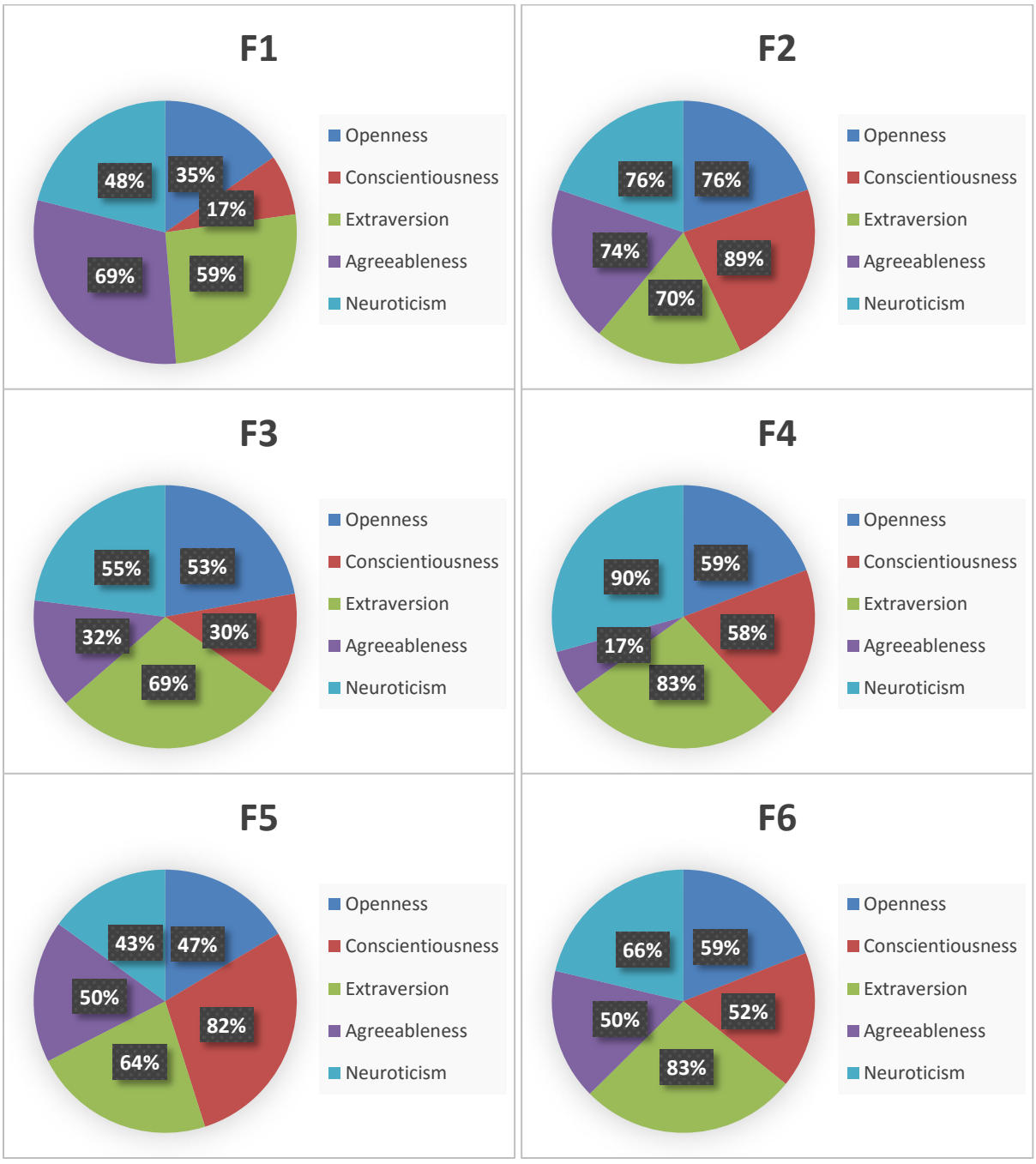
Vianello, M., Schnabel, K. Sriram, N. & Nosek, B. (2013). Gender differences in implicit and explicit personality traits. *Personality and Individual differences, volume 26,* 994-999.

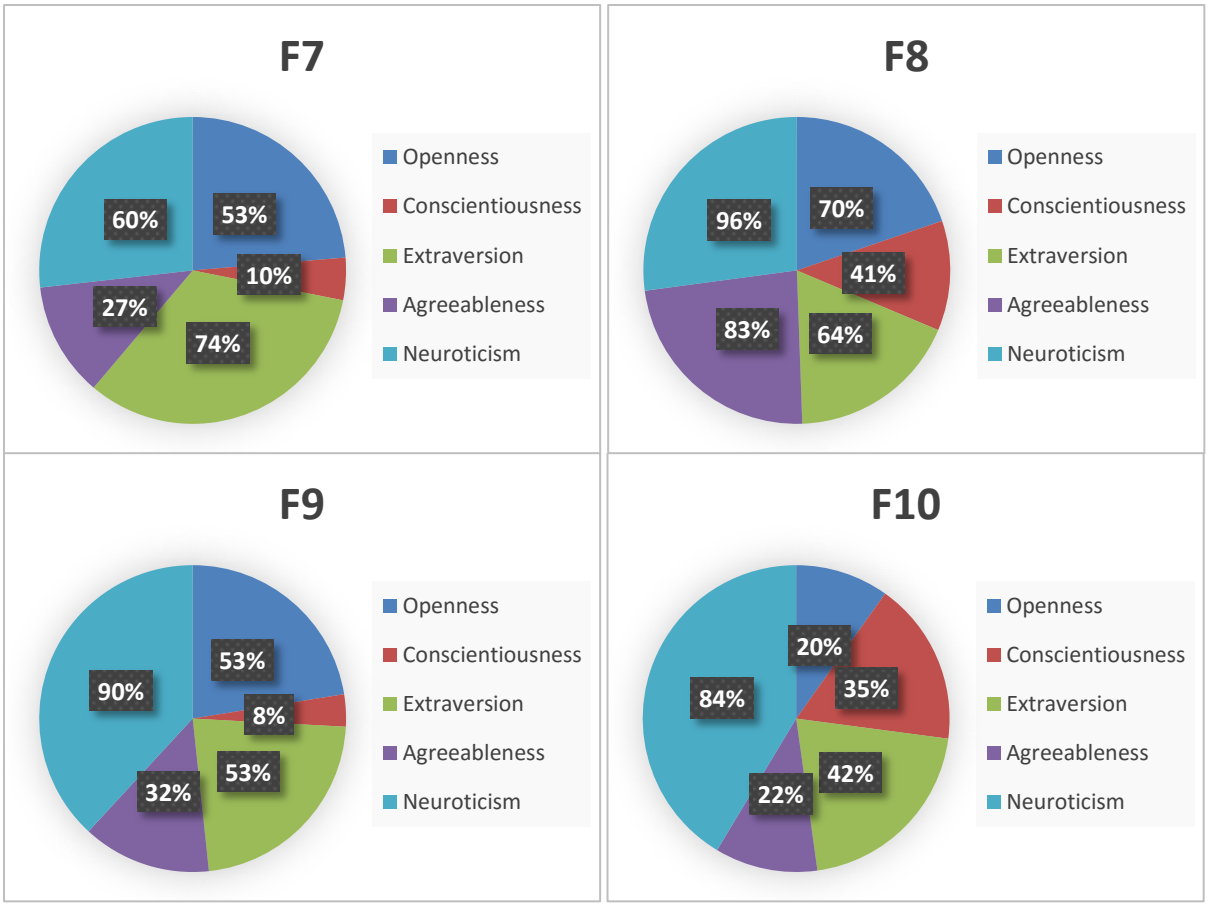
Zijlstra, H., van Meerveld, T., van Middendorp, H., Pennebaker, J.W. & Geenen, R. (2004). De Nederlandse versie van de 'Linguistic Inquiry and Word Count' (LIWC). *Gedrag & Gezondheid, volume 32*(4), 271-281.

7 APPENDICES

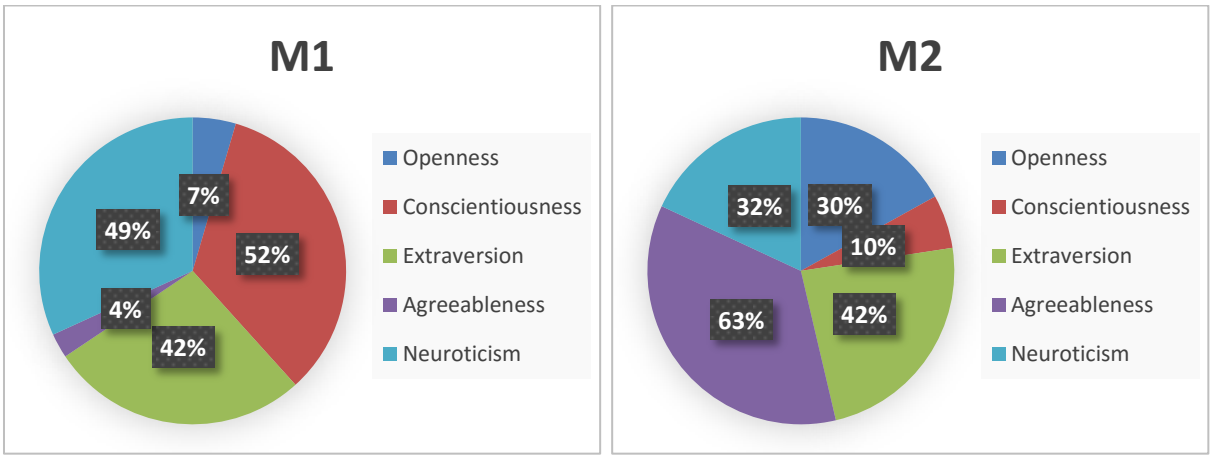
I. APPENDIX ONE: PIE CHARTS OF THE PERSONALITY TESTS

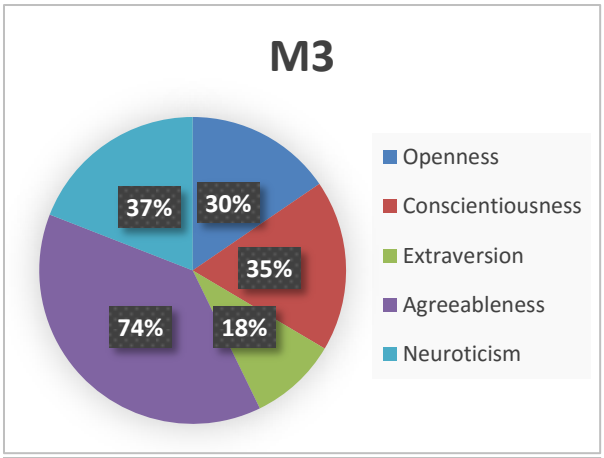
- Pie-charts of the personality of the female respondents

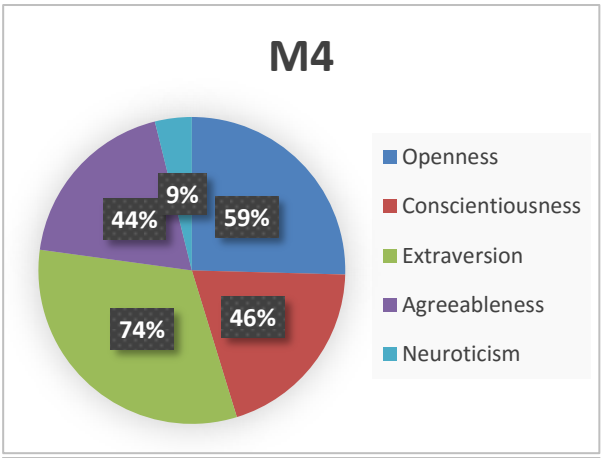


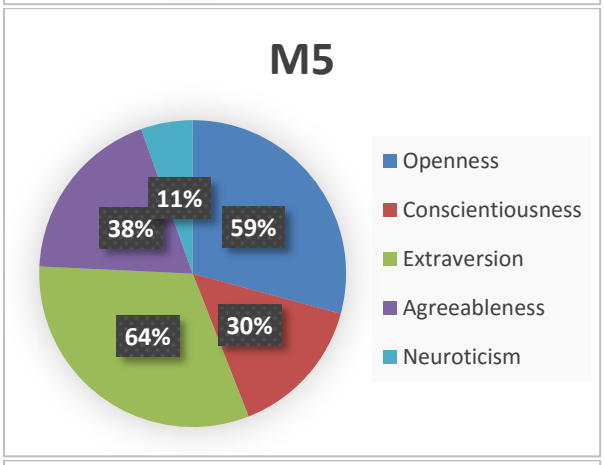


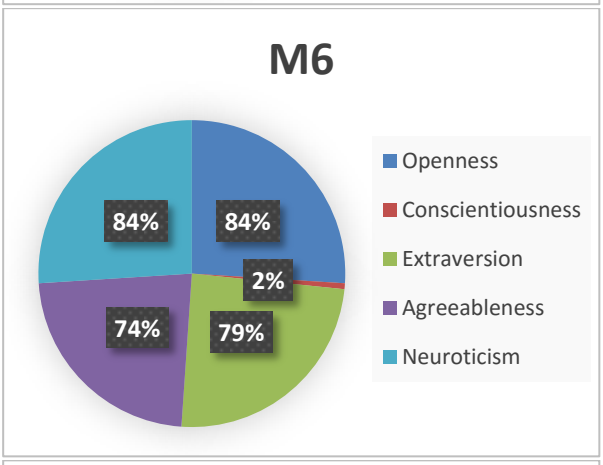
- Pie-charts of the personality of the male respondents

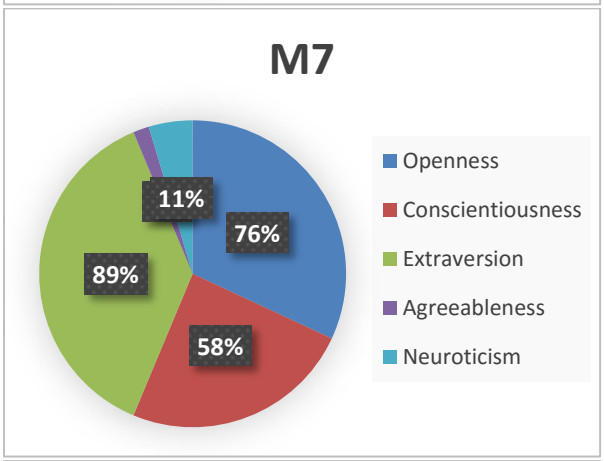


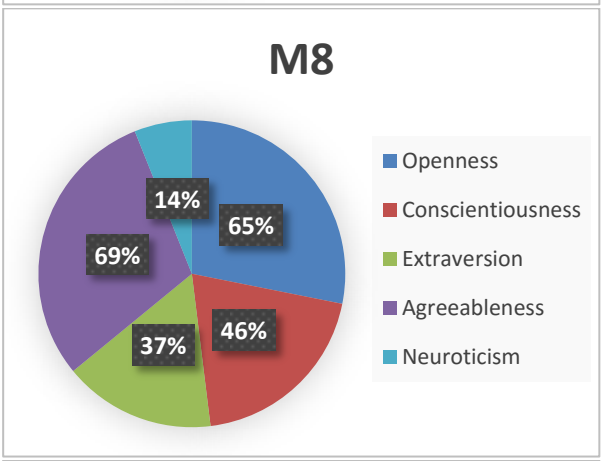

M3
30%
35%
18%
74%
37%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

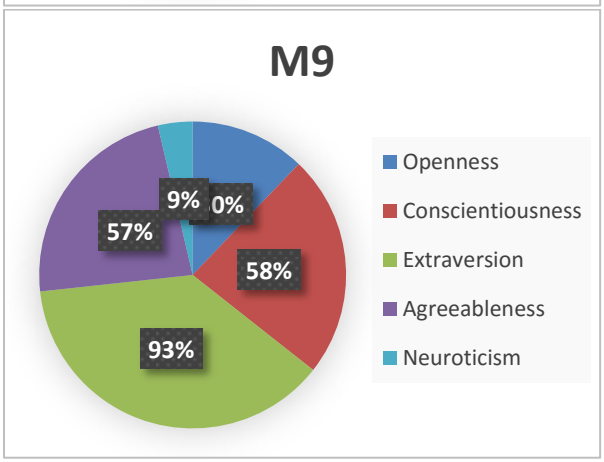

M4
59%
46%
74%
44%
9%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

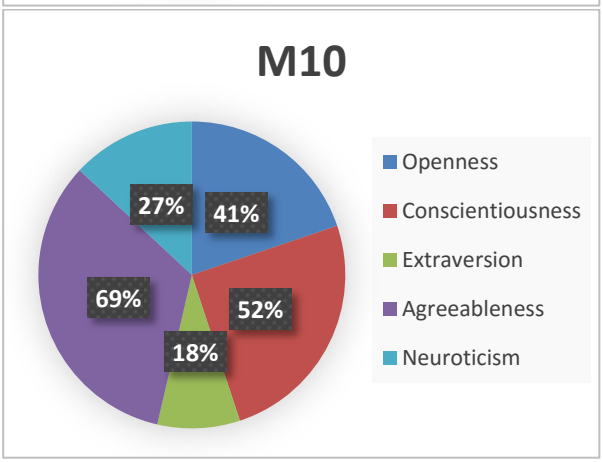

M5
59%
30%
64%
38%
11%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism


M6
84%
84%
2%
79%
74%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism


M7
76%
58%
89%
11%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism


M8
65%
46%
37%
69%
14%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism


M9
0%
58%
93%
57%
9%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism


M10
41%
52%
18%
69%
27%
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

II. APPENDIX TWO: STATISTICS OF THE PoS TAGGING

This appendix can be found in the electronic version of this dissertation.

III. APPENDIX THREE: STATISTICS OF THE LIWC ANALYSIS

This appendix can be found in the electronic version of this dissertation.

IV. APPENDIX FOUR: EXPLANATION OF THE LIWC DIMENSIONS

Abbreviation	Full name	Examples in English	Examples in Dutch
anx	anxiety	*worried – fearful - nervous*	*zenuwachtig - bang - gespannen*
article	articles	*a – an - the*	*de – het - een*
assent	assent	*ok – agree - yes*	*ja – oké - goed*
body	body states	*ache - heart - cough*	-[19]
certain	certainties	*always - never*	*volstrek – absoluut - vastbesloten*
cogmech	cognitive processes	*cause – know - ought*	*oorzaak – weten - denken*
comm	communication	*talk - share - converse*	*interviewen – gesprek - gerucht*
death	death	*bury – coffin - kill*	*dood – treuren - sterfbed*
discrep	discrepancies	*should – would - could*	-
down	down	*down*	*dieper – beneden - laag*
eating	eating, food & dieting	*eat - swallow - taste*	*drinken – honger - voeding*
friends	friends	*buddy – friend - neighbour*	*vriend – vriendschap - vriendin*
future	future tenses	*will - gonna*	*zal – wens -wil*
groom	grooming	*wash – bath - clean*	*douche – wassen - make-up*
hear	hearing	*listening - hearing*	*geluid – luisteren - klank*
home	home	*family*	-
humans	humans	*baby – adult - boy*	*meisje – mens - volwassen*
incl	inclusive	*and – with - include*	*en – inbegrepen - ook*
inhib	inhibitions	*block – constrain - stop*	*blokkeren – hinderen - inhouden*
insight	insight	*think – know - consider*	-
negate	negations	*no – no -, never*	*nee – nooit - niet*
negemo	negative emotions	*hurt – ugly - nasty*	*bedroefd – vijandig - wanhoop*
nonfl	nonfluencies	*er – hm - umm*	-
numbers	numbers	*second - thousand*	*één – dertig - miljoen*
physical	physical states	*sleep - breast - ache*	-

[19] Some dimensions were not elaborated upon in the Dutch LIWC, therefore not providing us with Dutch examples.

present	present tenses	*is - does - hear*	*afleren - bewaren - gebeuren*
pronoun	pronouns	*I - them - itself*	*ik - jij - onze*
sad	sadness	*crying - grief - sad*	*huilen - somber - teleurstelling*
self	total first person	*I - we - me*	*ik - mij - mijn*
senses	sensory and perpetual processes	*see - touch - hear*	*zien - voelen - horen*
sexual	sexual references	*horny - love - incest*	*flirten - zoen - beminnen*
sleep	sleep	*asleep - bed - dreams*	*dromen - slaperig - wekken*
swear	swear words	*damn - piss - fuck*	*verdorie - goddomme - shit*
tentat	tentative	*maybe - perhaps*	*misschien - waarschijnlijk - voorlopig*
time	time	*end - until - season*	*zomer - vroeger - zodra*
tv	tv, games and movies	*tv - sitcom - cinema*	*film - video - televisie*
social	social processes	*mate - talk - they - child*	*communiceren - delen - helpen*

V. APPENDIX FIVE: THE DATABASE

This appendix can be found in the electronic version of this dissertation.