Utrecht University

Bachelor thesis

Tigris de Leeuw, 4213866

What language can tell us about the elderly and their behaviour:
An analysis of three language features subject to age-related change

Abstract: One of the priorities in a society with an ageing population is to support senior citizens in their aging process. An analysis of their language use can shed light on their behaviour and needs. By means of a corpus of Dutch Twitter posts, that has been collected within the European project 'GRAGE', the language use of a group of elderly people over 67 will be compared with that of a group aged under 55. The analysis is carried out on the basis of three language features presumed to be subject to age-related change: pronouns, prepositions and hashtags. I will investigate the use of these features within each age group, as well as compare the uses of the groups with help of a statistical metric. Results indicate that pronoun and preposition use of the subjects is highly similar: differences in use are limited contrary to previous literature. Hashtags appear to be more informative. Contrasts in topic choice show distinct motives for using social media: the older group for leisure, the younger one to gain work-related exposure. In the discussion, I will point out the strengths and weaknesses of this exploratory study and suggest a new variable for further research: 'membership age'.

Thesis instructor: P. Monachesi

Second reader: J. Zwarts

Date: March 24, 2017

Index

1 Introduction

Aging. A natural and effortless, yet very interesting process that happens to every living individual. In this world where people have the privilege of reaching ages higher than ever before, certain stereotypes prevail about those among us who live to see these old ages. Thinking of elderly, for some the idea of loneliness, an image of someone not being able to keep up with the fast changing world or a concept of general cognitive deterioration will come up. On the other hand, a more positive picture of gaining wisdom and common sense as life progresses is also prevalent in society. To identify the effects of aging on different aspects of society is highly valued in a world where the number of elderly citizens is higher than ever and still growing. Through language, which offers us an insight into the behaviour and values of people, we can improve our understanding of the wants and needs of the elderly. In this study, I will research the effect of ageing in the context of language use.

Age as a sociolinguistic variable might be the least examined variable in research, in contrast to for example well-studied variables such as gender or social class (Llamas, Mullany, & Stockwell, 2006). There is an immense body of work on language use in the earliest stages of life. Child language acquisition and language development from birth have been studied thoroughly (for example Pinker, 1995; Tomasello, 2005). Although the earlier stages of life have seen much research, the same cannot be said about later stages. Not much research has been done on the progression of language past adolescence. This study will take the first steps in focussing on the oldest part of the population: the language use of elderly.

How a person uses language can tell us a lot about their values, attitudes and behaviour. The way of communicating reveals norms, wants and needs. An extensive use of verbs in future tense for example might indicate that the focus of this speaker is on the future, not the present or past (Pennebaker & Stone, 2003). This then could perhaps point to an active mind thinking about what lies ahead. The population of senior citizens deserves attention with respect to their wants and needs, as their number is increasing every year (Grage project, n.d.). Ageing healthily and happily is one of the primary concerns of the society of today and a study on their language use could contribute to this.

As a source of language data, social media platforms offer an opportunity for collection. Advantages of this type of data include that it exists in large amounts, is free, suitable for automatic collection, usually spontaneously produced and their content reveals traits, interests, annoyances, concerns and other personal expressions of its producers. For this paper a database containing data of Dutch users from the social media platform Twitter is employed. Twitter is used by a variety of ages on a daily basis and as a result provides us with a continuous stream of current language production. It is important to note that the role of social media in this study is solely as a resource for data and not as a variable: the differences and similarities between spoken or written forms of language and social media language will not be the focus of this paper.

Within this study, ‘elderly citizens’ are defined as people post-retirement. As the age bar for retirement in the Netherlands is set at 67, those over 67 are considered elderly. To my knowledge, this group has not received any extensive attention on the subject of language use, not to the degree of younger age groups. A reason for this is that there is often a lack of participants in this age category. Subjects over 67 will be compared to people who are still working but not yet close to retirement, ergo those under 55 years of age.

Research objectives

In the previous paragraphs I have broadly unfolded the main objective of this paper: to analyse the relationship between language and age, and more specifically between language from social media and participants aged under 55 and over 67. To specify further, this is an exploratory study that aims to identify what features of language could be revealing about the behaviour and needs of the elderly, while trying to find an appropriate methodology to this end. Literature analysing written or spoken corpora and literature on social media suggest several features, of which I have chosen three: pronouns, prepositions and hashtags.

Pronouns are included because literature on both written and spoken language as well as on social media suggests that pronouns are subject to age-related change, however there are some conflicting results and as of yet little is

known about their presence in the language of the elderly. In contrast to pronouns, which have served as a language feature in studies on age-related language change on various occasions, not many studies have incorporated prepositions as a variable. Prepositions have a strong connection to spatial memory, which declines naturally with age. In this light, perhaps preposition use of elderly participants could reveal this decline in spatial memory. For this reason, they are included as a feature in this study on elderly. Thirdly, hashtags are unique in the sense that they do not appear in spoken or written speech, and investigating how people of different ages explore this unique feature could tell us something about their perception of this function and their choices in information sharing, and therefore intrinsically their behaviour.

The question in this study is then: How are these three features of language used by participants under 55 and over 67 and what can they tell us about the behaviour of these age groups? My analysis reveals that the differences in use of pronouns and prepositions are very small and not much can be concluded in terms of behaviour. Hashtags on the contrary are quite revealing. Evident from the difference in content of the hashtags, it appears the focus of the over-67 group is on leisure time whereas that of the under-55 group is on working life. I conclude that their motives for using social media differ: for the older group uses social media is a form of amusement, while for the younger it poses as a mean for exposure of their work.

The motive for choosing the features is explained further in section 2 in a theoretical framework discussing literature on age and language use (both spoken/written language and social media language). In sections 3 the method for analysing these features is explained. The results and conclusions are presented in section 4, after which in section 5 the paper finishes with an overall conclusion and discussion.

There are several elements that make this paper innovative. As the reader shall discover later on in this essay, a handful of studies employed corpora consisting of non-social media data such as spoken interviews. In contrast to these studies, the current Twitter corpus contains information that is expressed spontaneously. This language is different from for example written or spoken language used in responses to interviews where questions might be suggestive and socially desirable answers

are not uncommon. Furthermore, the medium Twitter is used to communicate daily pastimes, concerns and personal matters and therefore its content gives us a lot of information about its users. A Twitter post (tweet), seemingly thoughtlessly written and sent out into the world, actually contains a lot of valuable information about how someone used language to do so. Additionally, the participants included in this study are much older than those of most studies in this field, which could shed light on the language use of this age group. The linguistic aspect goes a bit further than previous studies have done: different types of pronouns are considered, the prepositions are analysed because of the cognitive aspects and a hashtags analysis reveal their semantics.

2 Theoretical framework

2.1 Age

GRAGE project

In Europe, the population aged over 65 will grow considerably in the coming years, especially compared to other continents. It is likely that most of those over 65 will live in urban areas, alone. In order to help senior citizens in urban areas in ageing healthily and actively, a European Union research project has been set up called ‘GRAGE – Grey and Green in Europe: Elderly living in urban areas’. This project aims to investigate the needs of the elderly that live in urban areas. To do so, researchers within this project focus on the factors that determine the beliefs, values and behaviour of elderly citizens and on how their quality of life is affected by urban planning.

This study is written within this context. The primary focus is on elderly and how their behaviour can be understood through research. It aims to contribute to the field of language use and the effects of ageing, by taking the first steps in assessing the behaviour of elderly in terms of their language production. The age groups used in this study are based on the purposes of the project: young (under-55), pre-retirement (55-67) and post-retirement (67+). In order to observe sharper differences, the pre-retirement group is left out here and only the extremes are included.

Age

In researching age or its effects, incorporating age as a variable is of course an absolute necessity. In studies, age has often been treated as a chronological and fixed variable. One’s age is the time someone has lived from one's birthdate up until the moment of inquiry. Chronological age is simple to work with and easily measured. Additionally, it is not so prone to error as other age classification methods. Some researchers however believe a chronological approach to be too simplistic (Nguyen et al, 2014; Llamas, Mullany, & Stockwell, 2006; Eckert, 1997). Eckert (1997) distinguishes three classifications of age. The first is chronological age, which is the exact time a person has lived from his birth up until a specific moment. This chronological age gives an approximate measure of the second classification: one’s

biological age, or physical maturity. Thirdly, social age relates to life events that are not spread out evenly across one's life span, for example matrimony or the birth of a first child. Using age as a variable is difficult: should studies group their participants on chronological age or life stage to get the most meaningful result? (Llamas, Mullany, & Stockwell, 2006).

2.2 Language and age

Age as a variable in linguistic studies

Age is dynamic and ever-changing and the relationship between language and age remains complex. Perhaps because of this, this variable remains quite a mystery in the field of sociolinguistics. Child and adolescent language has been studied largely (for example Pinker, 1994; Tomasello, 2005; Eckert, 2003). Adults and elderly speakers have not received much attention from researchers. Adults are often regarded as a homogeneous group that conforms their language use to a societal norm, such as using standard language in a workplace. Therefore differences in language use tend to become smaller as age progresses (Nguyen et al, 2014; Eckert, 1997). Later in life, after retirement, seniors are less inclined to conform to these societal norms and this effect wears off (Llamas, Mullany & Stockwell, 2006). Not much else is known and their behaviour remains largely unexamined (Llamas, Mullany, & Stockwell, 2006).

The studies that consider language use often focus on chronological age (Pennebaker & Stone, 2003; Barbieri, 2008; Schler, Koppel, Argamon, & Pennebaker, 2006; Rao, Yarowsky, Shreevats, 2011; Peersman, Daelemans, & Van Vaerenbergh, 2011; Rosenthal & McKeown, 2011). Nevertheless, individuals of the same age can certainly use language in completely different ways, for example because they take different places in society. As Llamas, Mullany and Stockwell point out, an eighteen-year-old student that lives at home might use language in different ways than an eighteen-year-old parent that has a full-time job. Therefore recently, the focus of various researchers has shifted to grouping subjects on life stages rather than chronological age (Nguyen, Gravell, Trieschnigg, & Meder, 2013). Language use is based on age identity, which is not per definition the chronological age.

Therefore by treating age as a fixed variable, an analysis might lose some of the richness (Nguyen et al, 2014).

In this study, age groups are based on chronological age. However, the chronological ages of the groups are also good indicators of one's life stage. The youngest group, under-55, can be seen as 'working'. Because the official age of retirement in the Netherlands is set at 67, the group aged between 55 and 67 is 'pre-retirement' and that over 67 is 'post-retirement'. In this light, these chronologically defined age groups are basically a loose categorization into life stages as well.

Language and age studies

Of the few linguistic studies that focus on age, two articles stand out: *Words of wisdom: Language use over the life span* by Pennebaker and Stone (2003) and *Patterns of age-based linguistic variation in American English* by Barbieri (2008). These two articles will be discussed extensively in this section, as they form the main inspiration for this paper.

Pennebaker and Stone (2003): Language use over the life span

Pennebaker and Stone attempt to establish certain features of language that change with age. Two projects are carried out: one to identify the features, one to check them. Their search begins by focussing on certain areas of change across the life span that, according to them, lend themselves to linguistic examination. These are: emotional processes, social and identity relationships, time orientation, and cognitive complexity.

In their first project, the authors used the LIWC tool. This tool, developed by Pennebaker, Frances and Booth in 2001, uses a word-count strategy that, when encountered with any given text file, searches for over 2000 words or word stems corresponding to over 70 linguistic dimensions. These dimensions include for example standard language categories (function words), but also psychological processes (for example positive and negative emotion words) and traditional content dimensions (such as 'occupation'). Unfortunately, one of the limitations of this tool is that it does not recognize irony or consider context. Though this is an important restriction, word choice might be telling a lot about an individual's language use.

Many subsequent studies have used this tool in the study of language use in specific groups. Note that not always all dimensions need to be involved: some studies only employ those relevant to their experiments.

By use of the LIWC program, the authors examined written samples and transcripts from spoken interviews of over 3000 English speaking participants of 45 studies. These studies had in common that their method consisted of either writing or speaking about an emotional event or experience (control participants were also included). Considering that the authors wanted to examine linguistic change across the human life span, the ages of the participants ranged from 8 to 85 years old, with a mean age of about 24 years (*SD* = 12.6). Relatively few data was collected from participants older than 55 (106 participants, of which only 44 were older than 70). For this reason the conclusions drawn for ages above 55 may have a limited reliability.

By correlating the LIWC dimensions to the participants' ages, Pennebaker and Stone were able to identify certain linguistic linear and even curvilinear relationships. The features they discovered are listed per area of investigation, the first being language and emotional processes. They found that aging is associated with a decline in negative emotion words and at the same time an increase in positive ones. This increase was the strongest for the individuals older than 55, and participants of 70 years or older used positive emotion words almost twice as often as younger individuals.

For social and identity language, a decline in the use of first-person plural over the life span became apparent. However, there was a quadratic component to this relationship: from a high start in the teen years, the use of this type of pronoun declined steadily, but increased slightly from age 55 onwards, and sharply in the 70+ category. Subjects younger than 14 and older than 70 use first person plural most often, between these limits it is used less. Another significant finding was that the use of self-references (meaning first-person singular pronouns) decreased dramatically with age, with an extreme decline for subjects older than 70. Social references in general, such as words like 'friend', also declined as age went up.

Time orientation-wise, with age individuals showed a decrease in the usage of time-relevant words (like 'soon') and past-tense verbs, whilst demonstrating a

somewhat surprising increase in future-tense verbs, indicating a shift in focus over the aging process. Apparently, the older participants had their mind set on the present and future more than the younger ones.

In the area of cognitive complexity, the authors investigated the use of cognitive markers, which they identify as: cognitive mechanisms (words like ‘ought’, ‘know’), large words of more than six letters, and total cognitive (for example ‘think’, ‘recall’), causal (‘because’), insight (‘realize’), and exclusive words (‘exclude’). There was no decline in the use of total cognitive words after the age of 25, there was even an increase in insight words and a curvilinear relationship with exclusive words over time. The authors' interpretation of this finding is that as people get older, they gain wisdom and a greater understanding of their own experiences. The use of words consisting of more than six letters increased sharply with age, especially in the groups above 55.

To verify whether these features are consistent across other media and other groups, Pennebaker and Stone carried out a second project. Using the LIWC tool they examined the language of ten prominent American and British authors from different time periods through their works. Because these were written over the course of their lives and therefore written during different moments in life, this provided a within-subject analysis of age-related changes in language. The results indicated that the patterns of ageing and word use (as age increases, so does the use of positive words, declining presence of first-person singular pronouns, et cetera) are indeed consistent with the results of the first project, even across these different groups. It is important to note that this second project focused on written forms of language that are revised and thought over multiple times, in contrast to the written and spoken form in the previous project. Even across these types of media, results were consistent.

Barbieri (2008): Patterns of age-based linguistic variation in American English

Barbieri made use of a large corpus of casual conversation in American English of the mid-1990s to explore the effect of age on spontaneous speech. She formed two age groups: her younger group consisted of 85 people with ages ranging between 15

and 25, her older one comprises 54 persons between 35 and 60 years of age. Compared to the groups that will be the focus of this study, these groups are both considered young. Despite this, the manner of research and results are still relevant.

Barbieri's method differed from that of Pennebaker and Stone. She performed a key word analysis, which is defined in her work as "a word which occurs with unusual frequency in a given text" (p. 62). This analysis resulted in two major noticeable patterns that she used as a springboard for further evaluation: use of slang and use of stance and involvement markers.

On the subject of slang, it is certainly clear that younger people's vocabulary contains a wide variety of slang words, be they swear words or non-derogatory slang. With respect to stance, or expressions concerning personal feelings, attitudes, affect, or emotional involvement, there was an abundance of outcomes.

Both groups' conversation contained a wide range of inserts ('oh', 'yeah') in their conversations, however each group seems to favour different ones and the range of the youngest group is wider, as is their range of response tokens ('right', 'okay'). It would appear that the younger group makes more use of polite speech, such as 'sorry' and 'please'. The domain of personal pronouns yielded the striking result that younger people use the pronouns 'I', 'my', 'myself', 'me', and 'you' significantly more than the older group. The older group showed a higher frequency for third-person singular and plural, as well as first person plural. However, some personal pronouns collocate often with verbs like 'mean' and 'know'. This collocation might suggest that pronouns serve an intersubjective and interactive function for younger speakers. It also proves that it is important to take bigrams into account when doing this type of research.

A feature more salient for older speakers' discourse were modal verbs, which are traditionally considered as the main grammatical device for expressing stance, meaning expressing attitudes or personal feelings. These include amongst other things 'may', 'will', and 'could'. Modal verbs are not present in large numbers in the speech of the younger group, apart from the verbs 'will' and 'can'. Instead of modal verbs, younger people rely more on evaluative adjectives like 'awesome' and 'pissed' to convey stance. Another category where these groups differ is use of adverbs.

Younger people seem to favour a limited number of intensifiers (‘really’, ‘totally’, ‘so’), while older individuals possess a wider selection of stance adverbials (‘maybe’, ‘apparently’). Another notable difference is found in the use of discourse markers, where the youth uses ‘like’, ‘right’ and ‘just’ and the elder use ‘well’ and ‘okay’. Lastly, on the subject of reporting verbs, older speakers make use of the ‘traditional’ manner by employing ‘say’ and ‘tell’, where younger speakers favour ‘be like’.

Feature 1: pronouns

In short, both articles aimed to identify certain features of language that are subject to age-related change. Not all are suitable for investigation with respect to the current study. Though positive and negative emotion words, time related words and cognitive words do tell us something about the behaviour and attitudes of their users, selecting them is time-consuming and not entirely objective in the sense that choices have to be made with respect to which words belong to these categories. A word such as ‘holiday’ for example could be seen as a positive emotion word, but can also be used in a neutral way. The use of slang has been investigated thoroughly (for example Nippold, 1998; Labov, 1992). In addition, it is a clear hallmark of youth and consequently not relevant in a study on language of elderly. Most stance features Barbieri describes are quite complex to incorporate in a study, for example response tokens because context has to be taken into account.

A feature that Pennebaker and Stone as well as Barbieri mention is pronoun use. To summarize: Pennebaker and Stone observed a decline in the use of first-person singular pronouns as age increased. First-person plural pronouns showed a similar decline, but increased in the language of participants over 55 and onward. Barbieri’s research indicated that first- and second-person singulars were used more amongst people aged 15-25 than 35-60, a result similar to that of Pennebaker and Stone. First-person plural and third-person singular and plural were more frequent amongst the older group. Barbieri’s result that first-person plural pronouns are used more by people between 35 and 60 contrasts with Pennebaker and Stone who noticed a decline in the use of these pronouns until the age of 55. However, we should take into account that the authors dealt with different types of language use:

Pennebaker and Stone used interviews, while Barbieri used spontaneous speech.

Apart from these conflicting results, pronoun use is interesting because the frequencies with which certain types are used could tell us something about where the focus of the user is: substantial use of self-references through first-person singular pronouns could imply a certain way of self-representation, or many first-person plurals could point to feeling like a member of a community or simply being invested in a significant other. For these reasons, I chose pronouns as the first feature for this study.

2.3 Age and social media

Social media as data source for research

The two articles discussed previously by Pennebaker and Stone (2003) and Barbieri (2008) are based on everyday language, mostly in spoken, some in written form. Research on the manner in which different social or age groups use language has often been constrained by the time and effort it would have taken to collect sufficient data of adequate quality from the proper target audience. In recent years, social media use has grown immensely, providing freely accessible data suitable for automatic collection. Researchers take advantage of this unprecedented opportunity and the study of how language is used by particular groups has been revived (Nguyen, Doğruöz, Rosé, & De Jong, 2015).

Despite this recent growth in availability of data, in studies other variables, for example gender, bypass age in investigating their covariance with language. When age is incorporated, often the objective of the study is to predict the age of a social media user based on their language, therefore identifying certain linguistic features that are typical for certain age groups. These studies often focus on age groups younger than the ones in this study. Occasionally, due to the lack of sufficient participants in older age classes that are active on social media, results on these groups might not always be significant or representative (as in Rao, Yarowsky, Shreevats, & Gupta, 2010). In conclusion: even though social media has given new impetus to research in the area of language use and age, questions such as what features of (social media) language are markers of old ages remain unanswered.

Social media studies on language use and age

In many studies that focus on language use and age with respect to social media, the objective is often to detect users of certain age groups by means of certain features, one of them being language. An in-depth analysis of these features is often not the aim. Because social media platforms are still used most by the younger part of the population, many of the studies form age groups until ages of 40. Due to the recent popularity of social media, another aspect of these studies is that this type of research, meaning research by means of computationally collected social media data, is still at quite an early stage: methods are being refined to eliminate errors as much as possible.

Elderly and social media

Literature that incorporates social media does not usually focus on elderly. One might wonder why social media could mean something in researching this age class. Perhaps when one thinks of social media, the elderly might not spring to mind immediately. Nevertheless, increasingly more senior citizens make use of social media. The Netherlands is one of the pioneers with respect to the elderly part of the population in its internet and social media use. Tables 1 and 2 depict the behaviour of certain Dutch age groups. As can be seen in Table 1, an impressive 80% of the Dutch people in the age group 65 to 75 year olds made use of the internet in 2016, and so did nearly half of the population over 75. Table 2 clearly shows an increase in the use of social media (this includes online chatting, writing or reading weblogs, e-mailing and use of social and professional platforms) in all chosen age classes, where almost 60% of those between 65 and 75 and more than 20% of those over 75 use social media (CBS, 2017).

Table 1

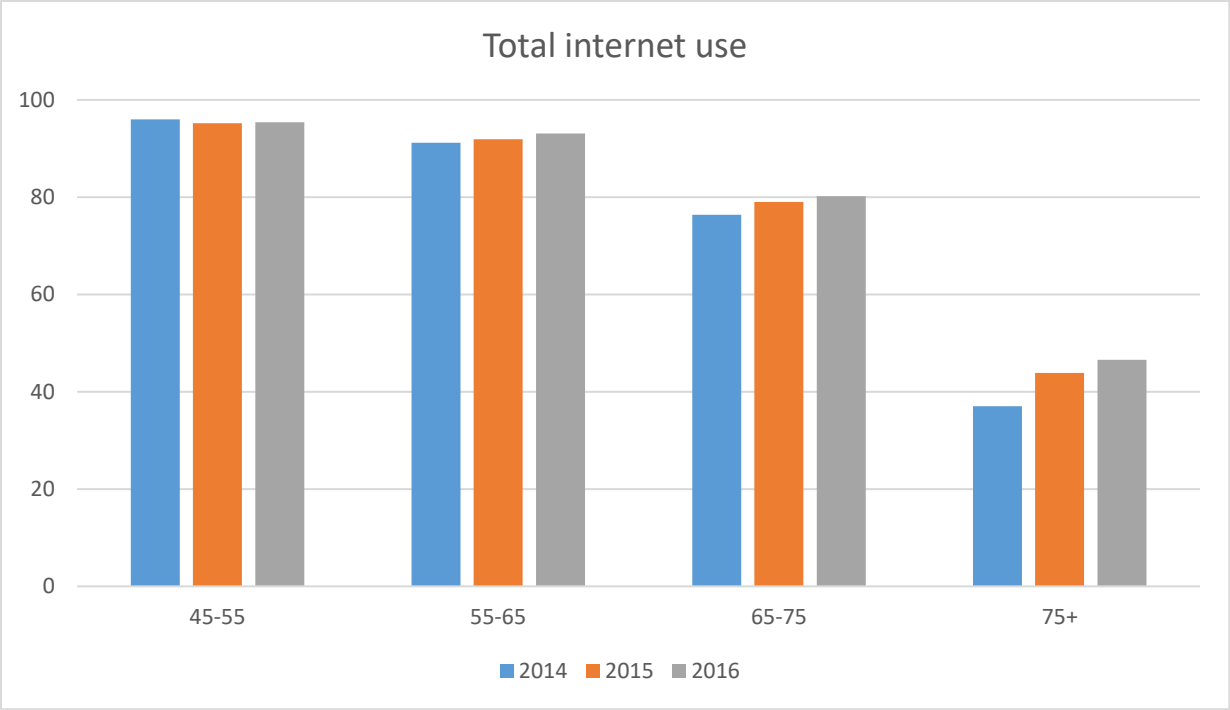
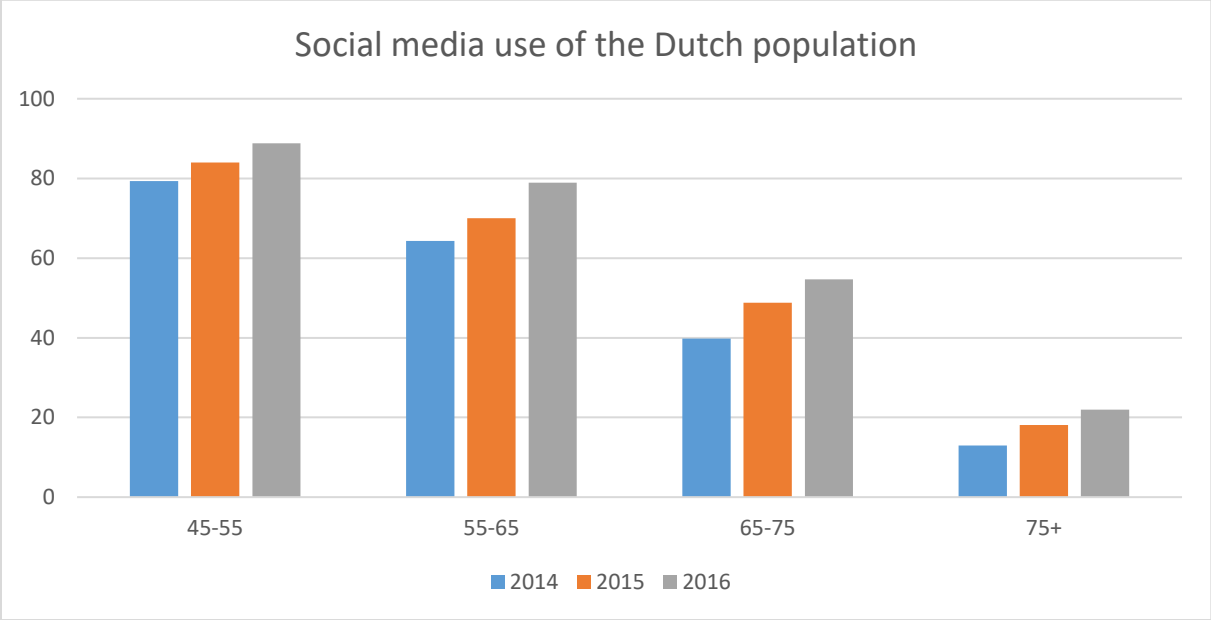


Table 2



Using social media as data source for this study on elderly language use is a meaningful move, since a lot of spontaneous data produced by elderly people can be found here. This data can be collected in large amounts and relatively quickly. Users

on a social media platform such as Twitter write spontaneously about their interests, therefore providing a lot of information about their wants and needs and behaviour. For these reasons a Dutch Twitter corpus will be employed here.

Pronouns on social media

Pronouns did not only figure as a feature in the studies of Pennebaker and Stone and Barbieri, a variety of computational social media studies incorporated pronouns as a variable as well. As one of the first to pick up on the subject of language use in social media, Pennebaker collaborated in 2006 with Schler, Koppel and Argamon and once again in 2007. In both articles, the corpus consisted of blog posts extracted from the website 'blogger.com' and considered the following features: parts-of-speech, function words and blog-specific features (for example 'blog words' and hyperlinks). The categories of participants were 10s (13-17), 20s (23-27) and 30s (33-42), which are very young compared to the ages of the participants in the current study. Still, on both occasions the authors found that as age increased, the total use of personal pronouns decreased.

Rosenthal and McKeown (2011) did research on blogging as well and investigated whether a prediction of a bloggers age could be made based on certain features. Amongst other results, they found that their younger group (18-22 years) opt for first person singular pronouns in subject position, whereas the older (38-42) seem to favour first person plural.

MySpace figured as the data source for Pfeil, Arjan and Zaphiris (2009) by which they investigated age differences and similarities in social capital of two extreme age groups (13-19 and over-60 years of age). They found that the younger group uses more self-references (in this paper these include first-person singular as well as plural pronouns) on their profiles. Along with the findings that the younger group uses more negative emotion words and cognitive words whereas the older group uses more articles and large words (words of more than six letters), the authors suggest these results demonstrate a difference in motivation for using MySpace. Younger people see themselves as protagonist and talk more about themselves, older people present themselves in a more formal way on their social

media.

Lastly, Nguyen, Gravel, Trieschnigg and Meder (2013) analysed a corpus of Dutch Twitter language, and it seems that younger people in their teens prefer first-person singular (comparable with Pfeil, Arjan and Zaphiris) and second-person singular pronouns. This usage declined after the age of 20 and stagnated at 30 years of age. Older age groups make more use of first person plural, though an exact age is not given.

Together with the articles of Pennebaker and Stone and Barbieri, this body of literature suggests a solid tendency for younger persons in their teens and early twenties to use more self-references in the form of first-person pronouns in their language (Pennebaker & Stone, 2003; Barbieri, 2008; Rosenthal & McKeown, 2011; Pfeil, Arjan & Zaphiris, 2009). At older ages, people tend to use more first person plural pronouns (only Pfeil, Arjan and Zaphiris (2009) concluded otherwise), though a concrete age limit cannot be identified as of yet (Pennebaker and Stone, 2003; Barbieri, 2008; Rosenthal & McKeown, 2011; Nguyen et al, 2013). Most of the groups in these studies are quite a bit younger than the participants that will be used in this study, so hopefully we can provide an answer to the question how elderly citizens make use of pronouns.

Feature 2: prepositions

A type of function word that does not appear in either Pennebaker and Stone's (2003) or Barbieri's (2008) study is prepositions. Prepositions are an interesting phenomenon, since they have a strong connection with temporal and spatial relationships of objects. With respect to the latter, in most languages all sorts of spatial relations between objects have to be covered by a limited number of spatial prepositions, which leads to polysemy. Spatial prepositions can be relational, meaning they specify the location of one object in relation to another, as in 'the tablecloth is over the table'. Prepositions can also be directional, which means they convey information about the direction in which an object is located or a change in position of an object. An example is 'the plane flew over the Pacific'. These two example sentences provide an additional example of just two of the many

polysemous meanings of the preposition ‘over’ (Coventry & Garrod, 2004).

With respect to spatial relationships, prepositions are particularly interesting in a study on language and age. This is because the normal aging process involves a mild or moderate decline in cognitive abilities, hence there is an age-related decline of spatial memory (Bach et al., 1999; Gallagher & Pelleymounter, 1988; Moffat, Zonderman, Resnick, 2001). The spatial memory is the memory where spatial information is stored. This cognitive decline is not clearly visible in language use until very late. In assessing the cognitive aspect of language, Pennebaker and Stone saw no decline in the use of total cognitive words (such as ‘think’, ‘recall’) after the age of 25. Pfeil, Arjan and Zaphiris (2009) also saw that overall cognitive words such as ‘ought’ and ‘cause’ were used more frequently by teenagers than their participants older than 60. Pennebaker and Stone observed that the presence of long words (>6 letters) increased with age. This last finding is replicated by Pfeil, Arjan and Zaphiris for their participants above 60 on MySpace. Nguyen and her colleagues (2013) concluded that older participants use more complex language on Twitter in the form of longer tweets (posts on Twitter) with longer words and more prepositions. There seems to be no clear substantiation of a cognitive decline in normal language production when looking at cognitive words such as ‘ought’ and ‘think’. One could ask whether this cognitive decline due to age is reflected in the use of prepositions in language of elderly speakers, since the use of certain prepositions involves an understanding of spatial relationships and the employment of the spatial memory. If so, preposition use could tell us something about the cognitive state of elderly.

Literature mentions briefly that prepositions are an area of age-related linguistic change. Both previously discussed studies by Schler, Argamon, Koppel and Pennebaker (2006; 2007) report that there is a linear relationship between age and the use of prepositions: as age increases, so does the use of prepositions in blogging. Their oldest category consisted of people in their 30s (33-42), so a significant increase could already be identified in these age ranges. This conclusion is supported by Nguyen and her colleagues (2013) for the Dutch Twitter language, their older participants (like Schler and his colleagues, also 40+) used more complex language with more prepositions. This higher use of prepositions could be the result

of the fact that more complex structures are employed, for example recursive structures (the door *of* the house *of* the neighbours). Unfortunately, neither of these studies defined or elaborated on prepositions and not much is known for people aged over 65. Therefore, prepositions will be the second feature explored in this study.

Feature 3: hashtags

Language on social media often reveals a lot about the interests, values and behaviour of people, because users talk about this on these platforms. What could be revealing with respect to this study on the use of language on social media is the use of the features specific to social media. These are not present in speech or written texts and therefore add something new to the investigation. Assessing how different age groups pick up on these features could be telling about the way these groups communicate. It is important to note that features specific to social media language should not be confused with stylistic writing characteristics that are used often in social media, for example capitalization of words (‘WOW’) or alphabetic lengthening (‘niiiiice’). Even though these features are mainly seen in social media language, they are not social media specific.

Sharing links and/or images, chat words and tagging are examples of such features. The first one is studied on various occasions. Burger and Henderson (2006) were among the first in the research of social media and age on a large scale. Their data consisted of blog posts, in which it is possible to add an image or URL link to the text. The results showed there was no trend for image sharing, but did show a gentle increase in usage of URLs in posts with respect to age: apart from an inexplicable peak at the age of 24, link sharing increased with age with users older than 35 posting the most. This result on URLs is supported by previously mentioned Schler, Koppel, Argamon and Pennebaker (2006; 2007), who continued research on blogging and found that the sharing of links increases with age. To remind the reader, the age categories they formed were 13-17, 23-27 and 33-42, so data on elderly people was absent. Nguyen and her colleagues (2013) on the other hand saw a sharp rise in the use of links for Dutch Twitter users in their 20s, that stagnates in their 30s. They associated this finding with information sharing and impression

management. Rosenthal and McKeown (2011) cannot be in complete agreement to the previous authors either: the use of links and images in their blog data varied across all ages. It seems conclusions on this subject are not unequivocal and there are none for ages over 67.

Another aspect of social media use is chat language. This is not completely the same as basic slang, which appears in spoken or written language and is a clear hallmark of the language of young people (Barbieri, 2008). Research on social media language investigated this topic as well and in agreement with Barbieri concluded that slang is used the most by teen users (Argamon et al, 2007; Goswami, Sarkar, & Rustagi, 2009; Rosenthal & McKeown, 2011). Chat words however are neologisms that emerged with the rise of social media. They are often abbreviations (‘grts’ for ‘greetings’), acronyms (‘lol’ for ‘laughing out loud’) or a different way for writing something (‘haha’ to indicate laughter, ‘ur’ for ‘you are’). Literature suggests chat language follows the same trend as slang and is used most by younger users, with teens using it the most (Schler et al, 2006; Peersman, Daelemans & Van Vaerenbergh, 2011; Nguyen et al, 2013).

One feature included in certain platforms that is semantically interesting is tagging. Tagging is described by Golder and Huberman (2006) as “the process by which many users add metadata in the form of keywords to shared content” (p. 198). On certain social media networks such as Twitter or Instagram, tagging takes the form of a hashtag (‘#’) followed by the keyword. The choice of keywords could tell us about the core topics of interest and also about the motive for using social media. Hashtags are personal: from them we can derive what is important to the people using them, which reveals a lot about their wants and needs. Therefore the form of tagging on Twitter, hashtags, is included as the third and last feature in this study. Because hashtags are a broad concept, they are treated more extensively in the next section of this paper.

2.4 Hashtags

Background

As stated before, the use of social media specific features is very interesting since these do not appear in speech. An appealing opportunity presents itself to investigate how people make use of this feature and whether different ages do so in different ways, with different motives. In order to analyse the use of tweeters of the social media specific feature of Twitter, hashtags, some further information on this feature is required and provided in the following sections.

In 2007 Chris Messina suggested a new feature for Twitter as to "improve contextualization, content filtering and exploratory serendipity within Twitter" (Bruns & Burgess, 2011). By means of a pound or hashtag sign ('#'), people could add their tweets to certain topic conversation. They could expose their opinions or follow the conversation by searching for the hashtag heading it. Twitter themselves defines hashtags as something that can be used "to index keywords or topics on Twitter. […] [P]eople can follow topics that interest them in an easy way" (www.support.twitter.com). The feature was added to Twitter and has been a great success.

Function and purpose

Although the addition of the hashtag-feature was an innovation for Twitter, the basic concept of the hashtag, namely tagging, had existed for a while. A website that incorporated tagging early on was Del.icio.us, on which users could tag shared website bookmarks to their own preferences. Golder and Huberman (2006) in their research article on tagging in this website describe collaborative tagging as "the process by which many users add metadata in the form of keywords to shared content" (p. 198). In discussing the function of tags for Del.icio.us they explain that users can use tags to identify what a certain bookmarked item is or what or who it is about, who owns it, to refine categories, identify qualities or characteristics, to reference to one's self, or as task organizing. Some of these functions are personal and only relevant to the user, such as the last four. The first three are not necessarily highly personal and therefore more likely to be used by a wide variety of people. The most used tags are meaningful to the general public, those with only meaning to an

individual are used less frequently.

Bruns and Burgess (2011) also considered the purpose of tagging, tagging by means of hashtags in Twitter. In the first place, hashtags are used to “mark tweets that are relevant to specific known themes and topics”. They argue that when polysemy problems (the same hashtag for different topics) or synonym problems (different hashtags for the same topic) arise, the users themselves resolve these. They believe a user will be persistent in exposing ‘their’ hashtag to the preferred group of people and not let it be spoilt by ending up in the wrong group. Even though the initial function of the hashtag was that people could follow conversations to their interest, not all people using a specific hashtag will also follow the conversation that entails this hashtag. This is mainly the case for hashtags with a very high frequency. Hashtags can also be used to simply emphasize something, resulting from a lack of methods (such as different font styles) to express stress in Twitter.

Creation of hashtags

Hashtags are flexible and created when the need for them emerges among users. This can be *ad hoc* (in the moment) or *praeter hoc* (in anticipation of a foreseeable event, such as presidential elections) (Golder & Huberman, 2006). Cunha and colleagues (2011) draw a similarity between the creation of hashtags and linguistic innovation: just like linguistic innovation, a hashtag might be created when an individual has the need for it. And parallel to linguistic innovation, the hashtag may be accepted amongst the members of a community, or might be used only by its creator and maybe even only once. These authors also identify a rich-gets-richer pattern in which the popularity of already popular items increases faster than less well-known ones. They found this pattern in their study: 90 percent of the hashtags are used fewer than ten times, 60 percent only once, while the top-scoring hashtags are used many times. This indicates that many hashtags are restricted to one user or a small group of users, but the most used ones have a very high frequency amongst a larger community.

Popularity

Why is it that certain hashtags are more popular than others? Cunha and colleagues (2011) found a relationship between the length of a hashtag and the number of times they are used. Apparently, popular hashtags are simple, direct and concise: the longer they are, the less likely they are to become popular. A hashtag which is made up of a complete sentence is unpopular, due to the possibility of more variation in the sentence which makes the frequency of a particular sentence-hashtag less high (compare #thankyoumichael and #thankyoumj). In addition, a sentence is more difficult to remember than just a single word, and the chances of spelling errors are higher which also makes the frequency and thus popularity lower. Hashtags which include an underscore ('_') in them, which is the only punctual sign Twitter hashtags can incorporate, are a lot less popular than their signless equivalents (#michael_jackson occurred a lot less frequently than #michaeljackson).

Problems with tags

Golder and Huberman (2006) explored the difficulties of tagging. Some problems of tagging are semantic in nature: synonyms can cause problems. Content under the tag of a certain word will not be shown when someone searches for content under a tag that is a synonym of this word. Tagging is similar to filtering: only results with the exact same spelling are returned, other results that can be relevant to the inquiry made might not be shown due to synonymy. Polysemous words that have more than one related meaning are also problematic, since (contrary to synonyms) results that are irrelevant might also come up. Let us for example someone is looking for information on insects and searches for content tagged under the word 'bug', articles on computer problems might also show up..There is also the complication describing an item more specifically or generally than its 'basic level': a tag such as "javascript" might be too specific for some, while one like "programming" might be too general.

Usage of hashtags

Conover and colleagues (2011) investigated the network of political communication of retweets and that of mentions (references to other Twitter users through the use of

an ‘@’ sign) in Twitter and found that the first network is highly segregated into two groups that are politically either left or right, while the other does not show such a polarization. The reason they provide for this is the role of hashtags. In their tweets with political statements, users often include hashtags whose primary audience is a politically opposed group. This group is exposed to something that they did not necessarily want to be exposed to and to which they might want to react (using mentions), but not rebroadcast (retweet). Therefore the retweet network is highly polarized, but the mention network not.

Age and hashtags

When looking at different age groups and their use of hashtags, Nguyen and her colleagues (2013) found that hashtags are used more often by older Twitter users: low usage in teens, a steep climb in the 20s, the highest and continuous use through the years up until the oldest participants category (over 60 years of age). According to these authors, they are, similar to links, connected to the sharing of information and older tweeters apparently are more concerned with information sharing than younger users. This could be connected to the fact that older social media users appear to declare more interests than younger users (Burger & Henderson, 2006), so both groups might be using their social media network with different motives. Younger people seem to display a certain kind of online image, something older people are less concerned with (Pfeil, Arjan, & Zaphiris, 2009). Jang, Han, Shih and Lee (2015) researched the behaviour of two groups on Instagram, a social medium consisting of photo posts in which users can use hashtags to add keywords their photos, similar to adding keywords to tweets in Twitter. The group of teens (13-19 years) posted fewer photos, but added more hashtags to them. The writers concluded they expose themselves more than adults (25-39 years) do.

On the content of the hashtags in relation to age Jang and his colleagues found a difference between their subject groups: the adult group (25-39) displays a wider range of interests in topics and are very diverse: arts/photos/design, locations, mood/emotion, nature, social/people. The majority of the teens’ (13-19) hashtags concern mood/emotion and follow/like. Jang and her colleagues contemplate that this

can be attributed to the difference in financial situations between the groups, teens are often on a limited budget and cannot always afford to follow their interests as much as adults can (for example expensive purchases or holidays).

2.5 Hypotheses

This study aims to take first steps into examining the use of specific language features in younger and older participants in order to get a better understanding of their behaviour. From literature on written and spoken language or social media language in relation to age, three of these features were identified. The first one was pronouns, as they have already figured as a variable in previous studies, but not in studies on elderly language use. They might reveal how people represent themselves on their social media and on who (themselves, someone else) they focus on. In studying old ages and behaviour, identifying an indicator of a cognitive decline through language could be valuable. This study poses as a pilot to see whether prepositions could indicate this cognitive decline. Lastly, hashtags could reveal user information, because the choice of hashtags users make says something about what they deem important. The main question in this paper is how pronouns, prepositions and hashtags are used by participants in two different stages of their life and what this tells us about the behaviour of these participants.

Based on the literature I hypothesize that as users get older, they might have a reduced social network and therefore focus less on people (Pennebaker & Stone, 2003), resulting in an overall decline of pronoun use with age. The most used category amongst those under 55 might be first-person singular, whereas those over 67 use first-person plural more often. My guess that the differences that will be observed are very small is similar to that of Nguyen and her colleagues (2014) and Eckert (1997).

Research suggests that the higher the age, the higher the number of prepositions in their language. However, there is no information about the type of prepositions used. As cognitive abilities, including spatial memory, decline with age, this might be reflected in preposition use. It is difficult to form an hypothesis, but based on literature I do expect older people to use more prepositions, but possibly different types than their younger counterparts.

With respect to hashtag use I expect elderly participants to use more hashtags, and that the content of the hashtags is less related to working life, which might reflect a different motive, different way of communicating and different behaviour of elderly. These hypotheses will be tested in an analysis of the collected data, as explained in the following sections.

3 The analysis: method

This section will discuss the analysis carried out in order to examine the use of the features of language that change with age of the two groups. The creation of the Twitter database and method of research will be explained. Pronouns, prepositions and hashtags within the corpus are examined. Results and conclusions are presented together. The paper concludes with an overall conclusion and discussion.

Corpus

The corpus used consisted of posts (tweets) of the social media platform Twitter. The corpus was built in the form of a database as part of the Grage project and this data was used in this study. On Twitter, the name, Twitter name (screen name) and a short biography of Twitter users (tweeters) are visible to everyone. The age of users is not directly visible, unless users choose to reveal it in their profiles. This complicates finding users of the desired ages, therefore this corpus was constructed by focus crawling. Twitter accounts of Dutch companies that target an elderly audience were located. Their followers, meaning audiences, were checked out and a handful was collected in May 2016.

Followers had to meet several requirements to be included in the corpus. Their tweets should be publicly accessible and written in Dutch. If their number of followers was rather high, the user is likely to be a celebrity and not included in this research. In order to create a balanced corpus the Twitter ID, gender, and age, be it exact or approximate, was annotated manually by me and a third party. The Twitter ID was found through a website tweeterid.com, which converts Twitter names into Twitter IDs. The gender of a user was evident from their name or profile picture. The annotation of age was possible using various tactics: at times users display their date of birth in their Twitter name (for example ‘waterman1947’, who is at least 68 in 2016). Some people post their birthdate on Facebook, Twitter, their blog or another website, or appeared in a newspaper article that stated their age. One can also get a good approximation of someone’s age when exploring their LinkedIn account: if someone attended high school in 1982, being probably about 12 years old at the time, one could make an educated guess that there age is about 45 or 46 in 2016.

These user characteristics were checked by me and a third party.

Three groups were created: under 55 years of age (7 female, 11 male), between 55 and 67 (7 female, 2 male), and over 67 (2 female, 8 male). These users along with their tweets were put in a database, which functions as the corpus of this study. The total number of tweets is 17738. The under-55 users tweeted a total of 103.097 words, those over-67 tweeted 52.150 words (URLs excluded).

Procedure

The class of people with an age between 55 and 67 was eliminated in order to increase the chances of identifying different behaviour when comparing different ages. From the created database, the plain text of all tweets of the two remaining age classes was extracted. The two texts left hashtags and URLs. As a start, to gain a general idea of what words were used most, these texts were put in Wordle via wordle.net. This is a tool that shows the words of a text by frequency in the form of a picture: the most frequent words have the largest font, and the lower the frequency, the smaller the font. Figure 1 shows the results from Wordle for the under-55 age group, figure 2 for the over-67 group. Note: in both groups, the word '*via*' ('via') was extremely large: the high number of times it is used turns out to be the result of citing for example a newspaper article or internet website through Twitter, which then quotes the title of the item followed by '*via @user*'. After checking, it shows that the word '*via*' is only very rarely used as a self-generated preposition and virtually solely by this mechanism for citing articles. It can therefore be excluded from the prepositions analysis.

Figure 1.

A Wordle representation of the most frequent words used by Twitter users under 55

The Dutch articles ‘*de*’ (‘the’), ‘*het*’ (‘the’) and ‘*een*’ (‘a’) are very large and thus very common in the text. What is striking is the preposition use: ‘*in*’ (‘in’), ‘*van*’ (‘of’, ‘from’), ‘*voor*’ (‘before’, ‘in front of’), ‘*op*’ (‘on’) and to a lesser degree ‘*over*’ (‘over’) and ‘*bij*’ (‘by’) are all used frequently. Another noticeable element is that the pronoun ‘*je*’ (‘you’ sing.) is larger in font than ‘*ik*’ (‘I’).

Figure 2.

A Wordle representation of the most frequent words used by Twitter users over 67

The articles is the word class that occurs most often in the text produced by the over-67 group. The demonstrative/relative/personal pronouns '*die*' ('this') and '*dat*' ('that') are also used quite often. With respect to personal pronouns, the first-person singular occurs two times in a rather large font as the tool distinguishes capitalized letters: '*ik*' and '*Ik*' (both 'I') appear most often. This group also makes much use of first-person plural '*wij*' ('we').

Wordle has shown interesting use of pronouns and prepositions. The two texts were then put through an online word count tool: Woordentellen, via woordentellen.be. This tool computes and returns the frequency and length of all words in a text. A useful characteristic of this tool is that it keeps punctuation intact. This means that it counts '*top*' and '*#top*' as two different individual items, so it counts the number of times a hashtag is used in the given text. To test the reliability of this tool, this data was checked by a third party with a Python program and proved to be valid.

This study will focus on the relative frequencies of words within each age group as well as comparing the two groups. To assess whether one group uses a word more frequently than the other, a statistical metric will be used invented by Church, Gale, Hanks and Hindle (1991). Tjong Kim Sang (2011) used this metric in his article to compare word frequencies between men and women from two Twitter corpora. This analysis is similar to the one in this paper, where word frequencies between two age groups are compared. The first step in this method, based on one from Church, is to calculate the relative frequency of one word that occurs in two given text samples by dividing the number of times N this word occurs in a text sample by the total number of words in that same sample. Next, the variance is estimated by dividing each of the two relative frequencies by the total number of words N in the corresponding text, then adding up these results and subsequently taking the square root. Lastly, the score t is determined by computing the difference between the two relative frequencies and dividing this by the estimated variance.

To illustrate:

$P_{ov67}(w)$ = relative frequency of word w in tweets of subjects over 67

$P_{un55}(w)$ = relative frequency of word w in tweets of subjects under 55

variance = $\sqrt{(Pov67(w)/Nov67 + Pun55(w)/Nun55)}$

t = ($P_{ov67}(w) - P_{un55}(w)$)/variance

If this score t is positive, the word is used more often in the tweets written by people older than 67, if negative, the word is more frequent amongst the under 55 group. This study will use this metric for pronouns and prepositions. This method has its limits: it can only be used if a word occurs in both text samples and this brings along some problem with respect to prepositions.

Pronouns & prepositions

The focus will be on first- and second-person singular and plural pronouns. These are not ambiguous, so a PoS-tagger was not employed here as it was not necessary for these pronouns and to avoid further complicating this pilot study. Therefore, first- and second-person singular and plural pronouns will be the focus, since they appear to be the most interesting ones and additionally are not ambiguous in the Dutch

language. They were manually extracted from the texts. I followed Nguyen et al. (2013) in including Dutch and English pronouns in the analysis since Dutch people often seem to tweet in English. Prepositions were also extracted manually from both texts, with help of a list of all Dutch prepositions from taaladvies.net (Taaladvies, n.d.). Due to limitations on time and effort and in order to avoid cross-linguistic complications, English prepositions are not included.

For both pronouns and prepositions, the normalized frequency was calculated by dividing the number of times a word was used by the total amount of words in the text of the age-class. By means of this normalized frequency these topics were analysed within the age groups. Next, the metric explained above was applied to check whether the differences in usage between the two age groups were statistically significant.

Hashtags

The extraction of all hashtags from the database was done by a Structured Query Language (SQL) query, which returned the hashtag along with the name and age-class of the individual that used it, as well as the tweet in which it appears and the date this tweet was written. Once again, all data from the age class between 55 and 67 was excluded. 3 of the 100 most frequent hashtags of the under-55 list were excluded: these were '*hardlopenmetevy*' ('runningwithevy'), '*groenerstad*' ('"greener"city') and '*l1*', because these tag appeared only in tweets that are shared via Twitter from other media, such as apps or websites. These hashtags were not written voluntarily.

The hashtags were considered and compared on straightforward numeral data, such as amount of use amongst groups. Next, the top 100 most frequent hashtags in each group were divided into categories, loosely based on the categories illustrated by Jang and colleagues (2015) and compared once again. As an additional analysis of behaviour of elderly, I checked whether the assumption of Conover and colleagues (2011) is right: whether hashtags could be playing a role in the political polarization of a mention network. To do so, the tweets in which the names of Dutch political parties appear as a hashtag were examined on positive or negative attitudes.

If the tweet is negative, there is a good chance that Conover and colleagues (2011) are right.

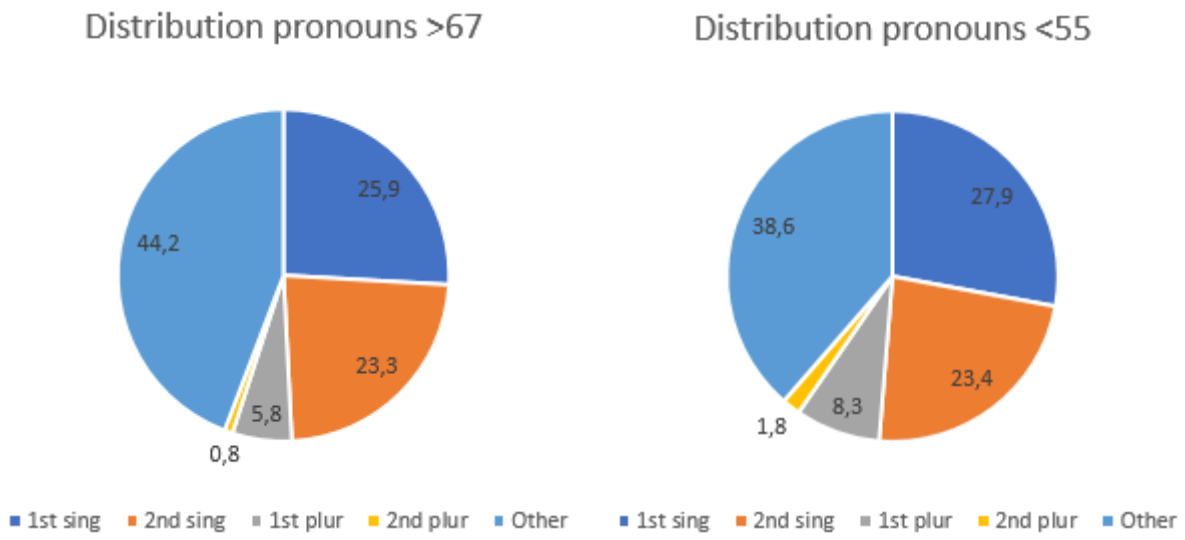
4 Results & conclusions

4.1 Pronouns

Results

The total number of pronouns used by the over 67 group is 3999, which makes up 7,6% of the total text. The most used of the 4 classes of pronouns that figure as research variable in this study was first-person singular with 2% of the total word use, followed by second-person singular (1,8%), then first person plural (0,4%) and second-person plural (0,06%). If these percentages of the total word use are translated to percentages of the total pronoun use for this group, 25,9% was first-person singular, 23,3% second-person singular, 5,8% first-person plural and 0,8% second-person plural. The remaining 44,2% are of course third-person singular and plural pronouns. Figure 3 displays the distribution of pronouns across each age class.

Figure 3



For the under 55 group, that used 6641 pronouns in total, pronouns form 6,4% of the total word production. Of this percentage, 1,8% featured as first-person singular pronouns, followed by second-person singular, first-person plural and second-person plural with 1,5%, 0,5%, and 0,1% respectively. In relative percentages of pronoun use, maintaining the order: 27,9%, 23,4%, 8,3% and 1,8%, and 39,5% all

third-person pronouns (Figure 3).

Table 3 gives the results of the normalized frequency and statistical metric carried out on pronouns.

Table 3

Pronoun usage in tweets of people over 67 and under 55

		> 67		< 55		
Pronoun in Dutch/English	Translation in English	Number of times used	P(w)	Number of times used	P(w)	T-score
Ik	I	663	0,01271	1204	0,01168	0,00958
I'm	-	0	0	30	0,00029	-0,01706
Me	Me	103	0,00196	128	0,00124	0,02082
Mezelf	Myself	2	3,84E-05	4	3,87E-05	-7,18E-05
Mij	Me	64	0,00123	121	0,00117	0,00156
Mijn	My	197	0,00378	298	0,00289	0,01650
Mn	My	0	0	11	0,00011	-0,01033
M'n	My	8	0,00015	47	0,00046	-0,01417
My	-	0	0	14	0,00014	-0,01165
Je	You, your (sing.)	791	0,01517	1199	0,01163	0,03281
Jezelf	Yourself (sing.)	20	0,00038	0	0	4,47214
Jij	You (sing.)	68	0,00130	141	0,00137	-0,00172
Jou	You (sing.)	24	0,00046	57	0,00055	-0,00394
Jouw	Your (sing.)	14	0,00027	47	0,00046	-0,00878

You	-	15	0,00029	73	0,00071	-0,01580
Your	-	0	0	22	0,00021	-0,01461
You’re	-	0	0	13	0,00013	-0,01123
We	We	110	0,00211	327	0,00317	-0,01887
Wij	We	214	0,00410	64	0,00062	0,13978
Ons	Us, our	45	0,00086	88	0,00085	0,00032
Onszelf	Ourselves	1	1,95E-05	0	0	1
Onze	Ours	38	0,00073	71	0,00069	0,00152
Jullie	You, your (plur.)	31	0,00059	118	0,00114	-0,01626

Conclusions

Contrary to the expectations, the group with the highest percentage of language production consisting of pronouns is the over-67 group (7,6% for the over-67 versus 6,4% for the under-55). The hypothesized decline in overall pronoun use with age is not supported. Of the pronouns used, the under-55 group holds a higher percentage in the use of first-person singular pronouns than its slightly older counterpart (27,9% versus 25,9%). Though this might appear as if the younger group is the highest user of first-person singular pronouns in their tweets, t-scores indicate otherwise. For the first-person singular subjective ‘*ik*’ (‘I’), the t-score is positive (t = 0,00958), meaning that the word is used more by the over-67 group. By a small margin (t = -0,002), ‘*jij*’ (‘you’ sing.) is more frequent amongst the under-55 group. This means that, in comparison to each other, the first-person singular subjective pronoun is used, contrary to the expectations, more often by people above 67, whereas the subjective second-person singular has the highest frequency amongst the users under 55.

Another surprising finding is that within both groups, the use of first- and then second-person singular pronouns surpasses that of first-person plural (see Table 3). The hypothesis that within the over-67 group first-person plural pronouns are used

more often than the singular is not supported. Even when all possessives, reflexives and objective pronouns are excluded and only subjective pronouns are considered, the result is more or less the same. When comparing these two groups, they appear very similar. The ranking of most used categories is the same (first-person singular, second-person singular, first-person plural, second-person plural; see Table 3) and the percentages are more or less similar as well. This supports the expectation that the differences between groups become smaller and less visible as age progresses.

The number of times the second-person plural '*jullie*' is used is higher for the under-55 group than the above-67 (t = -0,016). When examining the data, it becomes clear that this pronoun typically occurs in combination with mentioning another user/other users. In light of this, it might be possible that the under-55 group replies to other users with personal questions or mentions other users more often.

It is evident that the over-67 group does not use any English pronouns besides 'you', whereas English pronouns in the regarded categories make up 2,3% of the total pronoun use of the under-55 language. This does not come as a surprise, since the use of languages other than the mother-tongue is usually not very high amongst elderly.

The group of possessives is an eye-catcher within the first-person singular pronouns. Both groups use '*mijn*' ('my') a lot, the positive t-score reveals that the tweeters over 67 use it more than the younger group. The 'short version' or abbreviation of this word is spelled officially as '*m'n*', or more informally '*mn*'. These two are both more frequent amongst the under-55 group as indicated by negative t-scores. The subjects above 67 do not use the rather colloquial '*mn*' at all, the few times they opt for the abbreviation they use punctuation accordingly. Possibly, this group is more likely to be more accurate or formal in their writing.

What is striking is that there is a big difference between the use of the two subjective options for first-person plural ('*we*' and '*wij*'). Within the under-55 group, '*we*' is used a great deal more (327 times) than '*wij*' (64 times), while in the other group it is exactly the other way around ('*we*' 110 times and '*wij*' 214 times). If the data is looked at more closely, the latter group contains one user that has tweeted the exact same tweet containing the word '*wij*' (no other pronouns) 176 times. This

means that the number of times '*wij*' is used in original tweets is far less than it appears at first sight, yet quite a lot in comparison to the number of times '*we*' is used. Within groups, people under 55 apparently tend to make use of '*we*', whereas those over 67 opt for '*wij*'. According to the statistical metric, '*we*' is used more often by users under 55 than those over 67, and '*wij*' more by users over 67, even with the number adjusted (t-score = 0,005). When checking the data manually on who the users mean by '*wij*', no clear differences between groups can be observed. Both groups use '*wij*' to put emphasis on the fact that the writer and another person/other persons are the subject of the sentence, or to underline a certain group, for example '*wij mannen*' ('us men'). Perhaps the over-67 group intends to emphasize more often in their tweeting.

When focussing on the pronouns used only when reflexivity is optional (reflexive pronouns containing '*-zelf*' ('-self')), it can be concluded that the over-67 tweeters display a more diverse range of these pronouns. Despite the fact that the under-55 group uses '*mezelf*' ('myself') more frequently, this group does not demonstrate any other usage of pronouns of this category, when in fact the over-67 group does not only use '*mezelf*', but also '*jezelf*' ('yourself' (sing./plur.)) and '*onszelf*' ('ourselves'). '*mijzelf*' ('myself') does not appear at all in either group. Use of these reflexive pronouns is often meant to put emphasis on the pronoun, so this conclusion is in line with the one from the previous paragraph: perhaps the pronoun use of people over 67 reflects an intention to emphasize subjects.

4.2 Prepositions

Results

The language of the group with participants under 55 consisted for 12,5% of prepositions (12.868 out of 103.097 words). With 6321 prepositions for the older group that was 12,1%. The younger group used more different kinds of prepositions. 55 original prepositions could be distinguished, of which the younger group uses 53 whereas the older group used only 45.

The results of the prepositional use and the statistical metric of both groups are presented in Table 4.

Table 4

Preposition usage in tweets of people over 67 and under 55

		>67		<55		
Preposition in Dutch	Translation in English	Number of times used	P(w)	Number of times used	P(w)	T-score
À	At	0	0	1	9,7E-06	-1
Aan	At, by, on, to	307	0,00589	508	0,00493	2.39361
Achter	After, behind	12	0,00023	39	0,00038	-1.64831
Af	Off, down	33	0,00063	72	0,0007	-0,47693
Behalve	Except (for), beside	2	3,80E-05	1	9,7E-06	0,99481
Beneden	Below, beneath	0	0	4	3,9E-05	-2
Betreffende	Concerning	0	0	1	9,7E-06	-1
Bij	At, by	307	0,00589	675	0,00655	-1,57235
Binnen	Within	15	0,00029	49	0,00048	-1,86482
Boven	Above	19	0,00036	10	9,7E-05	3,00264
Bovenop	On top of	0	0	3	2,9E-05	-1,73205
Buiten	Out of, outside, except	13	0,00025	26	0,00025	-0,03422
Conform	In accordance with	0	0	2	1,9E-05	-1,41421
Cum	Cum	0	0	1	9,7E-06	-1
Dankzij	Thanks to	5	9,60E-05	23	0,00022	-2,01082

Door	By, for	135	0,00259	250	0,00242	0,60554
In	In	908	0,01741	2143	0,02079	-4,61201
Jegens	Against, opposite	0	0	1	9,7E-06	-1
Langs	Along	14	0,00027	17	0,00016	1,26079
Linksboven	Top left	0	0	1	9,7E-06	-1
Met	With	471	0,00903	1.204	0,01168	-4,94503
Na	After, behind	46	0,00088	154	0,00149	-3,45169
Naar	At, to	155	0,00297	427	0,00414	-3,75193
Naast	Next (to)	8	0,00015	15	0,00015	0,11989
Namens	On behalf of	2	3,80E-05	17	0,00016	-2,61886
Om	On, by, at	209	0,00401	439	0,00426	-0,72864
Omtrent	Around, concerning	1	1,90E-05	0	0	1
Ondanks	Despite	4	7,70E-05	12	0,00012	-0,77848
Onder	Under, amongst	21	0,0004	58	0,00056	-1,39282
Onderaan	At the bottom	0	0	4	3,9E-05	-2
Ongeacht	Regardless	3	5,80E-05	0	0	1,73205
Op	On, upon	603	0,01156	1.285	0,01246	-1,53961
Over	Concerning, beyond	186	0,00357	514	0,00499	-4,15281
Per	With, per	11	0,00021	36	0,00035	-1,60376
Richting	In the direction of	5	9,60E-05	7	6,8E-05	0,55993
Rond	About	12	0,00023	13	0,00013	1,38552

Rondom	Around	2	3,80E-05	6	5,8E-05	-0,55047
Sinds	Since	7	0,00013	11	0,00011	0,45832
Te	in, to, at	390	0,00748	731	0,00709	0,84237
Tegen	Against	53	0,00102	59	0,00057	2,80607
Tegenover	Opposite	0	0	2	1,9E-05	-1,41421
Ten	At, by	6	0,00012	19	0,00018	-1,09563
Tijdens	During	23	0,00044	68	0,00066	-1,79306
Tot	Until	33	0,00063	143	0,00139	-4,71521
Tussen	Between	21	0,0004	38	0,00037	0,32083
Uit	From, off, out of	327	0,00627	263	0,00255	9,76821
Van	Of, from	1.352	0,02593	1958	0,01899	8,3997
Vanaf	Since, from	10	0,00019	47	0,00046	-2,93495
Vanuit	Out of	8	0,00015	23	0,00022	-0,97529
Vanwege	Because of	8	0,00015	16	0,00016	-0,02684
Volgens	According to, by	35	0,00067	45	0,00044	1,79432
Voor	To, for	506	0,0097	1.361	0,0132	-6,24212
Voorbij	Beyond, past	8	0,00015	20	0,00019	-0,58443
Wegens	Because of	2	3,80E-05	3	2,9E-05	0,29003
Zonder	Without	17	0,00033	37	0,00036	-0,33353

Conclusions

Unfortunately, not many conclusions can be drawn from the results on preposition use. The percentage of prepositions in terms of total word use, the difference

between groups is nearly negligible: 12,5% for those under 55, 12,1% for those over 67. The clear presence of either an increase or decrease with age is not visible in this result.

Preposition use is very similar. When comparing the lists of prepositions ranked on frequency of use from each age-class, the top 25 of either list contains almost the same prepositions with nearly similar frequencies. After that, contrasts emerge and the differences are slightly bigger, but none stand out.

As the results stated, in the two texts 55 different prepositions were identified. The under-55 group uses all but two: '*omtrent*' ('around') and '*ongeacht*' ('regardless'). A lot of prepositions however are used only once, but still that is revealing in a text sample that is not that large. The ones that are unused by the over-67 group are '*à*', '*beneden*', '*betreffende*', '*bovenop*', '*conform*', '*cum*', '*jegens*', '*linksboven*', '*onderaan*', and '*tegenover*' (see Table 4 for translations). These prepositions are all rather unusual or quite formal, but fact remains that the under-55 group have used more types of prepositions in their tweets. It is difficult to say whether this reflects some cognitive decline because the amount of prepositions in languages is limited and the difference between the two groups is rather small.

What is interesting though is that in comparing the two groups per preposition by means of the metric, just 18 of the 55 prepositions gain a positive t-score. This means that for 37 prepositions the t-score is negative and therefore statistically used more often by the under-55 group. This is an intriguing result, because it seems that in comparison to people under the age of 55, those above 67 make less use of prepositions in their language. Once again, whether this is a concrete piece of evidence for an age-related deficit of the brain is hard to say, since the differences are so small.

Because a cognitive decline might be more visible in prepositions connected to spatial relationships, an additional analysis is carried out to check whether there is a difference in use of these prepositions. The categorization is based on that of Broekhuis in *Syntax of Dutch: Adpositions and adpositional phrases*. The prepositions '*in*', '*met*', '*om*', '*op*', '*rond*', '*tegen*', '*tot*', '*tussen*', '*van*', '*vanaf*', and '*voorbij*' (see Table 5 for translations) are excluded, since they can have a spatial as

well as temporal meaning depending on context. Table 5 displays which age group uses the spatial preposition the most (for translations, check Table 4).

Table 5

Preposition	Group that used it most	Preposition	Group that used it most	Preposition	Group that used it most
Aan	>67	Door	>67	Over	<55
Achter	<55	Langs	>67	Rondom	<55
Bij	<55	Linksboven	<55	Tegenover	<55
Binnen	<55	Naar	<55	Uit	>67
Boven	>67	Naast	>67	Vanuit	<55
Bovenop	<55	Onder	<55	Voorbij	<55
Buiten	<55	Onderaan	<55		

Of these 20 spatial prepositions, 6 are used most by the oldest age group. This is comparable to the results of all prepositions together, and does not offer any further evidence for a detection of an age-related cognitive deficit through prepositions.

4.3 Hashtags

Results and conclusions

Out of the 103.097 words for the group under-55 years of age, 5318 were hashtags. This comes down to approximately 5,2%. For the over-67 year old users, this was about 1% (532 hashtags out of a total of 52.150 words), which is quite a bit less than its younger equivalent. It seems likely that the younger group uses more hashtags than the older group. One possible explanation for this higher usage of hashtags could be that Dutch citizens over 67 are usually retired, whereas those under 67 years of age (ergo, including those under 55) still have a working life. Many of the hashtags of those under-55 seem to have some relation to professions or the labour market. People over-67 have stopped being part of the working world and do not appear to target this audience in their tweets by using hashtags. For those who have a job, exposure is very useful and social media is a valuable tool in this.

It appears as if some hashtags are used very often, however if inspected more closely, almost all of them (with the exception of ‘NDnl’, ‘top2000’ in the under-55 group, and ‘top’ in the over-67 group) are used by only one person in each age group. This suggests that the overlap of interests within each corpus is rather small to virtually non-existent.

The content is very telling on the focus of these age groups. Table 6 depicts the topics of the hashtags per age-class and their ratio.

Table 6

Distribution of the 100 most frequent hashtags amongst Twitter users under 55 and over 67 years of age into content categories

Category	Examples in English	% over-67	% under-55
Arts/photos/design	Photography	4,7	1,2
Companies	[names of companies]	0	11,6
Economics	Financing	0,9	3,1
Entertainment	top2000	18,2	12,3
Events/conferences	congresIC	0,6	4,1
Sustainability	40dayssustainable	0	1,3
Locations	Iceland	20,2	6,6
Mood/emotions	Proud	3,2	1
Nature	Bees	0,6	4,3
News	Brexit, [names of Dutch newspapers]	9,4	29,8
Occupational terms	website, qualitofhealthcare	0	14
Politics	[Names of political parties and laws]	30,8	0,8
Research/university	RadboudUMC	0,3	0,8

Social/people	AbeltAsman	2,6	0
Sports	NedMex, OS2012	2,9	2,3
Twitter-tags	dtv (*durftevragen,* daretoask)	0,3	4
Other	Wastepaper	5,3	1,2

As can be seen from Table 7, contrasts in topics are apparent. The most popular topics amongst elderly users are politics, locations, entertainment and news. Together, these categories hold over 75% of the hashtags. The under-55 group focuses more on work- and world-related tweets, with news, occupational terms, entertainment and companies as the contenders for the top spot.

The result for most used topics in each age-class is a little distorted: in the 67+ group, the hashtag '*50plus*' is used 84 times, by far the most-used hashtag in the over-67 group (with '*bedum*' in second place with 28 times). The person that incorporates this hashtag in his tweets turns out to be a member of the political party '50plus', promoting his own party by means of Twitter. That is why the score for the politics category for the over-67 group is so high. In the other age group, the top hashtag '*pedoverhoor*' ('paedo[phile]hearing') is used a striking 326 times, by someone who was following the hearing closely and tweeting excessively about this news item. This explains the high percentage of news topics by people under 55. If we exclude these extremes, the news category is still the leading category for the younger tweeters (17,4%). The politics category moves into third place for the older group with 8,2%.

The most noticeable difference between the two groups in terms of topics is that in the over-67 tweets, none of the topics of the hashtags were occupational terms or companies, and the references to congresses or fairs were much lower than for the under-55 group. Users under 55 write more professionally-themed tweets, whereas those for whom jobs ~~are a thing of~~ belong to the past restrain from this. The difference in values, focus and behaviour is clear in the hashtag use of these two groups. This difference becomes even more obvious when other topics are included: hashtags on the topic of sustainability are only present amongst users under 55.

When checked, these users tweet about this subject in the context of the company they work for. The over-67 group depicts an extensive use of location-tags, more than three times as much as the younger group, and entertainment. One might conclude that for those over 67 Twitter is a means of communicating leisurely to the world where the interests of the users are, instead of exposure of opinions or work-related tweets.

There are 30 hashtags that are used by both groups. It is no surprise that 3 of these are political parties, 5 are television programmes, 6 are news items and a few newspapers make up another 3, for these are broad topics about which any age group could have an opinion.

Generally it is observed that the main purpose of most of the hashtags is to add the tweet in which it appears to the conversation the hashtag entails, as is the case in for example news items and television shows. This purpose often coincides with the desire to let the followers of the tweeter know what occupies their life at the moment. If one tweets "*ze kunnen zomaar stunten*" ('they might just win'), a reader does not know what or who this tweet is about. However, if the hashtag '*porIJS'* is included, one understands that the tweeter is watching a sports game between Portugal and Iceland. The user probably follows the Portugal-Iceland Twitter discussion, but also wants to let his audience know what he is doing. Bruns and Burgess (2011) mention that some tagging is the result of a lack of means to express stress, such as different font styles. To convey stress, a hashtag is used instead. This is also visible is both groups. Consider for example the following tweet from a younger participant:

Ja, die hadden we nog niet bedacht inderdaad #kuch #opendeur[1]

Yes, we had not thought of that yet #ahem #statingtheobvious

[1] This tweet included a hyperlink at the end, that is left out here.

It seems unlikely that the writer intentionally wants to join a discussion about the topic of onomatopoeic coughing, but rather wishes to express sarcasm and uses a hashtag to do so. Some hashtags arise as the writer has the need for it and might not be used more than once, such as the very innovative '*GezijtgeenIJSLANDers*' ('youare(informal)noIcelanders'). The likelihood of the user feeling the need to employ this hashtag again is not very high and that of other tweeters accepting it and using it as well might be even lower. When searching Twitter for this hashtag, this assumption is verified, since no other (public) tweeter except the one mentioned above has used this hashtag.

On the popularity of hashtags, for the over-67 group, 99,5% of the hashtags is used fewer than 10 times, and 75% even only once. For the other group in this study, the amount of hashtags used more than once is 11%, and 1,3% more than 10 times. The most popular hashtags often consist of one word that is not a creation, but a regular word. Hashtags that consist of sentences are rare in the top regions of the frequency lists. The underscore sign ('_') is used in just 3 hashtags, and all of these appear only once in the combined texts. The assumption that the most popular hashtags are short and functional is reflected in these results.

Semantic complications are pervasive. A hashtag that shows polysemy is '*50plus*'. This hashtag is used by a politician older than 67 as the name for a political party, whereas a subject of the under-55 group uses it to label all those over the age of 50. These two obviously do not mean the same thing, yet the same hashtag is used. Another example of polysemy is '*SGP*', the name of another political party, which is coincidentally also the abbreviation of the name of a British music festival, the movement 'Smart Girl Politics' and an acronym for 'single girl problems'. For this reason, as Bruns and Burgess (2011) suspected, the user attempts to sort out this problem by switching to the hashtag '*SGPnl*':

> *SGP zou ook moeten twitteren met nl achter #. Net als #NDnl. #SGPnl dus. In VS twitteren ze er ook lustig op los mbt SGP #verwarrend*
>
> SGP should also tweet with nl after #. Just like #NDnl. #SGPnl, in that way. In [the] US they tweet extensively with respect to SGP #confusing

Multiple ways of writing words for the same hashtag are also problematic. In these corpora, both '*durftevragen*' ('daretoask') and its abbreviation '*dtv*' are used. '*dtv*' is used more often and is shorter, and will therefore probably survive. The counterpart of polysemy, synonymy, is not directly detectable in this data.

On the political aspect that Conover and colleagues (2011) mention, the results are compelling. The use of political parties as hashtags are displayed in Table 7.

Table 7

Political parties used as hashtags by people over 67 and under 55

Political party	over-67	under-55
50PLUS	84	0
CDA	0	1
ChristenUnie	1	2
GroenLinks	0	3
PvdA/PvdASchiedam	3/1	0/0
PVV	2	1
SGP/SGPnl	0/0	2/2
VVD	1	0

It appears that one of the under-55 users is a journalist for a Dutch newspaper and shares articles he has written on Twitter, therefore some of the hashtags are neutral ideologically. However, it can also be observed that users use hashtags to target either a politically opposed audience or one that agrees with the user. An example of the latter is from the politician that is active on Twitter and uses this medium to inform tweeters and gain sympathy for his party *50PLUS*, a political party that promotes the interests of the older part of the population. He does not use the hashtag '*50plus*' as an attempt to target users that oppose his political views. There are users that do so. For example, this over-67 user:

> *#PVV heeft links, D66, PvdA, buitenlanders, islamieten gedemoniseerd. Alleen dieren zijn veilig. En de eigen partijleden (huh?) natuurlijk*

> #PVV has demonized left, D66, PvdA, foreigners, Muslims. Only animals remain safe. And its own partymembers (huh?) of course

Evident from this tweet, this person has a negative opinion about the party PVV, but still chooses to use the name as a hashtag. He deliberately exposes his tweet to the conversation about this political party and targets users that search for this hashtag. He wants them to know his negative views about this party. There are other examples of intentionally tagging a political party in a tweet that has negative views on the party's statements. If people with contrary opinions react to these statements, the mention network will probably indeed consist of mixed political beliefs, replicating the findings of Conover and colleagues (2011). The online behaviour of this group of users is then in line with Conover and colleagues (2011).

5 Overall conclusion and discussion

This paper figured as an exploratory study in which the language use of the elderly part of the Dutch population was analysed on specific features identified in literature to see to what extent they reveal information about their values and behaviour could be detected. Using readily available Twitter language as data source, language of people over 67 and under 55 was examined and compared.

Contrary to the hypotheses, overall pronoun use went up with age and the pronoun category with the highest frequency in both groups was first-person singular. This could mean that the motive of people for using social media is to convey one's own interests and experiences, even as age goes up. The fact that elderly use the stress-reinforcing pronouns '*wij*' and reflexives more could maybe indicate that their focus is more on people than their slightly younger counterparts.

Sadly, the results for the experiment on prepositions were not so informative. Preposition use was highly similar. The younger group used more different kinds of prepositions and it also uses prepositions statistically more often when comparing the two groups. The results remain the same when only spatial prepositions are considered. However, this is not hard evidence that an age-related cognitive decline is visible through preposition use.

It seems that those under 55 use more hashtags on Twitter, though this could not be statistically tested. The most important result of the hashtag analysis is the difference in topics of hashtags. This reflects the focus on working life of the under-55 group and on leisure time of the over-67 group.

Discussion

This study had an exploratory character. It took the first steps in assessing what features of language can reveal elderly user information, therefore it was quite simple and had some shortcomings. Overall, differences observed are very small. Although use of certain words might be statistically different, this might not mean it is that significant as well. The reason for this could be that this study did not incorporate that many elderly participants. This was simply a first step into the direction of language use and old age, and the sample size remains quite small. Even though an

increasing amount of people over 67 use social media, locating these users in order to incorporate them in research is still difficult. It would be better to include more participants, then clearer differences in behaviour could perhaps be observed.

Another reason for the small differences is based on the idea of Danescu-Niculescu-Mizil et al. (2013). These authors propose that members of online communities follow a life cycle: after entering the community with completely different language use, users will gradually adapt their language to that of the community they have become a member of. After a while, their language moves away from the community and in the next phase, the user abandons it. In the stage where members simulate the language of the community, differences between individuals are not that large anymore because members approach the same language. The language of this study dates from 2016: perhaps most users of the Twitter community have adopted the same language patterns and that is why the differences are so small.

To verify this hypothesis, I propose a new variable should be taken into account in further research on this subject: membership age, meaning the amount of time someone has been a member of a community. Perhaps this accounts for the small differences in language use and can provide for a more meaningful comparison between groups.

Strong points of this study are that older age groups are at the centre of attention, which is not the case in other research on this subject. This paper regarded more carefully the different categories of pronouns instead of generalizing them. An addition to the field is the examination of prepositions with respect to the language use of elderly participants.

Since this was a pilot study, it is important to identify weak points and propose several improvements for further research. An unfortunate limitation of this paper is that due to time restrictions and complexity elements, gender was not included as a variable in this study, though research does suggest this is important and significant (Nguyen et al., 2013; Peersman, Daelemans, & Van Vaerenbergh, 2011). A PoS-tagger could be employed to expand research on pronouns. Another limitation is that bigrams are not considered. Barbieri (2008) reports that the percentage of pronouns in her paper is rather high because of the combination of some of them with verbs

like ‘mean’ and ‘know’, more close to phatic utterances. With respect to the previous point of not considering bigrams, the method employed in this study does not regard complex prepositions consisting of more than one word, such as ‘*door middel van*’ (‘by means of’) or whether prepositions are in fact prepositions and not postpositions or particles. In short: when considering individual words, collocations are not taken into account even though this might be meaningful. This study tried to identify a method to statistically analyse word use of two groups. The metric used in this study resulted in very small numbers, so there is a possibility that drawing conclusions from these might not be very meaningful.

Because this was just a first step into considering prepositions and age, the analysis with respect to this subject was a little simplistic. There is a lot of research on prepositions and here not all different categorizations are considered. The concept of the cognitive decline was a little shallow: spatial memory was regarded as a whole, though a division into spatial working memory and other categories could be relevant. Then again, this was just a pilot, further research could expand on this subject.

One of the problems when working with social media as data is that some of the data can be ‘polluted’ due to automated bots. These produce non-human data that can cause distortions in the results. According to Tweakers, a website on technology and electronics, approximately 9 to 15 percent of the monthly active accounts on Twitter are bots (March 2017), which is a relevant amount. This is certainly an aspect to take into account when incorporating social media in research.

6 Bibliography

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, *12*(9).

Bach, M. E., Barad, M., Son, H., Zhuo, M., Lu, Y. F., Shih, R., Mansuy, I., Hawkins, R., & Kandel, E. R. (1999). Age-related defects in spatial memory are correlated with defects in the late phase of hippocampal long-term potentiation in vitro and are attenuated by drugs that enhance the cAMP signaling pathway. *Proceedings of the national academy of sciences*, *96*(9), 5280-5285.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English1. *Journal of sociolinguistics*, *12*(1), 58-88.

Broekhuis, H. (2013). *Syntax of Dutch: Adpositions and adpositional phrases*. Amsterdam, the Netherlands: Amsterdam University Press.

Bruns, A., & Burgess, J. E. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, 1-9.

Burger, J. D., & Henderson, J. C. (2006). An Exploration of Observable Features Related to Blogger Age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 15-20.

Centraal Bureau voor de Statistiek (CBS) (2016, October). Internet; toegang, gebruik en faciliteiten [data]. Retrieved 2017, March 22 from: http://statline.cbs.nl /Statweb/publication/?DM=SLNL&PA=83429NED&D1=0,12-14,25-45&D2 =0,11-14&D3=0&D4=2-4&HDR=T&STB=G1,G2,G3&VW=T. Den Haag, the Netherlands: CBS.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *ICWSM*, 133, 89-96.

Coventry, K. R., & Garrod, S. C. (2004). *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Hove, United Kingdom: Psychology Press.

Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2011). Analyzing the dynamic evolution of hashtags on twitter: a language-

based approach. In *Proceedings of the Workshop on Languages in Social Media*, 58-65.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013, May). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 307-318). ACM.

Eckert, P. (1997). Age as a sociolinguistic variable. In *The handbook of sociolinguistics*, 151-167. Oxford, United Kingdom: Blackwell.

Eckert, P. (2003). Language and adolescent peer groups. *Journal of language and social psychology*, *22*(1), 112-118.

Gallagher, M., & Pelleymounter, M. A. (1988). Spatial learning deficits in old rats: a model for memory decline in the aged. *Neurobiology of aging*, *9*, 549-556.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of information science*, *32*(2), 198-208.

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In *Proceedings of the 3rd International ICWSM Conference*, 214-217.

GRAGE project (n.d.) About GRAGE [information]. Retrieved 2017, March 22 from http://www.grageproject.eu/

Hendrikman, M. (2017). *Onderzoek: 9 tot 15 procent van maandelijkse actieve Twitteraccounts zijn bots*. Retrieved from: https://tweakers.net/nieuws /122241/onderzoek-9-tot-15-procent-van-maandelijks-actieve-twitteraccounts-zijn-bots.html

Jang, J. Y., Han, K., Shih, P. C., & Lee, D. (2015). Generation like: Comparative characteristics in Instagram. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4039-4042.

Labov, T. (1992). Social and language boundaries among adolescents. *American Speech*, *67*(4), 339-366.

Llamas, C., Mullany, L., & Stockwell, P. (2006). *The Routledge companion to sociolinguistics*. Oxford, United Kingdom: Routledge.

Moffat, S. D., Zonderman, A. B., & Resnick, S. M. (2001). Age differences in spatial

memory in a virtual environment navigation task. *Neurobiology of aging*, *22*(5), 787-796.

Nippold, M. A. (1998). *Later language development: The school-age and adolescent years*. Austin, United States.

Nguyen, D. P., Gravel, R., Trieschnigg, R. B., & Meder, T. (2013). " How old do you think I am?" A study of language and age in Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 439-448.

Nguyen, D. P., Trieschnigg, R. B., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., & de Jong, F. M. G. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,* 1950-1961.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2015). Computational sociolinguistics: A survey. *arXiv preprint arXiv:1508.07544*.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 37-44.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, *85*(2), 291-301.

Pfeil, U., Arjan, R., & Zaphiris, P. (2009). Age differences in online social networking–A study of user profiles and the social capital divide among teenagers and older users in MySpace. *Computers in Human Behavior*, *25*(3), 643-654.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37-44.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 6*, 199-205.

Taaladvies (n.d.). Retrieved 2016, November 11 from: http://taaladvies.net/taal

/advies/tekst/32/voorzetsels_algemeen/

Tjong Kim Sang, E. (2011). Het gebruik van Twitter voor taalkundig onderzoek. *TABU: Bulletin voor Taalwetenschap*, *39*(1/2), 62-72.

Twitter (ca. 2016). *What zijn hashtags (#)?* [Support]. Retrieved from: https://support.twitter.com/articles/20169394