university of groningen / faculty of arts

Authorship Identification with shared DNA

How DNA influences writing style

R.K. Bruijn

Bachelor thesis
Information Science
Rune Bruijn
s3204758
May 30, 2019

ABSTRACT

Authorship identification (the task of identifying the author of a given text) is a hot topic amongst data scientists, computational linguistics and information retrieval. It is also applied in areas such as journalism and law where knowing the author of a document (such as a ransom note) may even be able to save lives. The author of a given text can be identified with authorship identification because of his personal writing style. The stylometry can provide unique stylistic fingerprints for any author (Baayen et al., 2002). But what about the author's DNA? Does someone's DNA influence their writing style? In this thesis I will dive into this problem. I will test this by making use of an authorship identification model.

The model I use for the authorship identification is the Groningen Lightweight Authorship Detection (GLAD) model (Hürlimann et al., 2015). This model was built for a general challenge of PAN at CLEF 2015 (PAN, 2015), concerning author identification. The main goal of this challenge was to answer the question whether two given documents have the same author or not. The performance of this model is measured on the product of AUC and c@1 score (Peñas and Rodrigo, 2011). The GLAD model had an AUC-c@1 score of 0.41 for English on the test data, which is significantly higher than the baseline of 0.25, but not ideal. I ran this model on a data set, consisting of the columns Dear Abby and Ask Ann Landers. These columns were written by two twin sisters and two other women.

The results of the research are not as significant as I expected. Even though the model classified texts written by two twin sisters as the same author more than most of the other combinations, the difference is really small compared to some other combinations. The percentage of times the model predicted two texts to be written by the same author for two twin sisters was 44 percent, which was the highest percentage. However, even though this was significantly higher than some others, this percentage was 40 for the similarity between Esther Lederer and Ruth Crowley, who do not have shared DNA.

There are several possible explanations for this, but the main problem was that the data set was probably too small and the texts used as data were too short. In future work, this should be improved. Another problem was the relatively low AUC-c@1 score of the model I used. This could be improved by testing on Dutch instead of English, as the AUC-c@1 score of the model is higher for Dutch.

CONTENTS

PREFACE

At the moment, you are reading the thesis *"Authorship Identification with shared DNA"*, written by me, Rune. Right now, I am studying Information Science at the University of Groningen and I am currently in the last year of my Bachelor and will finish this Bachelor by writing this thesis.

For the past few months I have worked on this thesis about the influence of shared DNA on writing style. I am an identical twin myself and my twin brother is named Kai, who also studies Information Science. His thesis is actually about the same subject as well. My supervisor, prof. dr. M. (Malvina) Nissim, came with the idea of doing research on authorship attribution and writing style combined with shared DNA. She has participated in a general challenge of PAN at CLEF 2015 about author identification (PAN, 2015). Together with some other people she built a model that would predict whether two or more documents were written by the same author or not. She proposed this idea to me and my twin brother because she thought it would be a really interesting idea for us, having shared DNA ourselves. Kai and I were really enthusiastic about this idea and thus chose to do our thesis about this subject. Besides, I think that this thesis might come in handy as proof that the writing style of twins could be very similar, because we have had some issues about plagiarism due to this. This is also one of the reasons why I expect that the conclusion of this thesis is that the writing style of twins is very similar due to shared DNA.

One of the difficulties I faced during my research was that the AUC-c@1 score of the model I used was the lowest for English, which made it difficult to draw conclusions. However, the model did score significantly higher than the baseline. For future research however, this is something that could be improved.

I would like to thank my supervisor, prof. dr. M. (Malvina) Nissim very much for helping me with my thesis for the past few months. Especially the idea of doing research on authorship identification with shared DNA is something I would like to thank miss Nissim for, as I had not thought of this myself, but I do find it very interesting. I would also like to thank her and her colleagues for the Groningen Lightweight Authorship Detection model, which I used for my authorship identification during my thesis.

I would also like to thank Esther Pauline "Eppie" Lederer and her twin sister Pauline Esther "Popo" Phillips for writing the columns Ask Ann Landers and Dear Abby, which I used as data during my research. Beside that, I would like to thank Ruth Crowley, who created the Ask Ann Landers originally, and Pauline Eshter Phillips' daughter Jeanne Philipps, who took over the column of her mother after she died. I also took some of their columns as data for comparison.

At last, if this subject is of real interest to you, I would like to recommend the thesis of Kai Bruijn, called *"The Influence of DNA on Writing Style"*. He has done research on the same subject, the influence of DNA on your writing style. However, he used different data and also included a lot of family members in his research. For his data he actually used texts written by him and myself. So, if you are interested, please have a look at his thesis too.

But first, have a look at mine. I hope you enjoy the reading.

Rune Bruijn

1 INTRODUCTION

Authorship identification identifies the most possible author from a group of candidate authors for academic articles, news, emails and forum messages. It can be applied to find the original author of an uncited article, to detect plagiarism and to classify spam / non- spam messages (Wang, 2017). It could proof useful in many situations. For example, determining the author of a ransom note might even save lifes.

Authorship identification or authorship attribution is based on the writing style of the author. However, what if two authors have a similar writing style? What if identical twins, with shared DNA, have a very similar writing style? How much would this influence a model trained for authorship identification? In my thesis I will dive into these problems. I will do this by answering the following research question:

"How much influence does DNA have on writing style?"

For my research, I will use two online columns as data. The column Ask Ann Landers was created by Ruth Crowley in 1943. It was a column where people could ask Ann Landers (a pen name) for advice. In 1955 Ruth Crowley died and Ask Ann Landers was taken over by Esther Pauline "Eppie" Lederer. Esther's sister, Pauline Esther "Popo" Phillips was also writing a similar column, named Dear Abby in 1956. After Pauline died, her daughter Jeanne took over the column Dear Abby. I will use these columns of the twin sisters as data for my research, including some columns written by Ruth Crowley and Jeanne Phillips.

I will input this data into an existing model for authorship identification. The model I will use is the Groningen Lightweight Authorship Detection model (Hürlimann et al., 2015). This model predicts whether two or more documents were written by the same author or not.

If it turns out that the model will predict that the columns written by Esther Lederer and her twin sister Pauline Philipps are written by the same author, we can conclude that these twin sisters have a similar writing style. This way, I will use authorship identification in order to see whether the writing style of a person is influenced by DNA.

The similarity of the two written columns will be measured by the similarity score that the model gives to a text. The model compares one or more known texts to a text with an unknown author. The model will then output a similarity score between 0 and 1, 0 indicating that the texts are not written by the same author, 1 indicating that the texts are definitely written by the same author. These scores can be converted to a binary answer, with a label of 'YES' or 'NO'. Either the texts are written by the same author (a score higher than 0.5) or not (a score lower than 0.5). If the score is exactly 0.5, I will disregard it, as I only want to label a text as 'YES' or 'NO'.

This thesis is organized as followed. In Chapter 2 I describe previous research that is important for my own research. Chapter 3 describes the data that I used and the material. This is split up in collecting the data and processing the data. Chapter 4 describes which methods I have used for my research. Chapter 5 describes the results of my research. In this chapter there are two sections: Results and Discussion. The last chapter, Chapter 6 contains the final conclusion for my research.

2 BACKGROUND

There has not been done a lot of research about the writing style of twins or shared DNA. In fact, I did not manage to find any research on these subjects. This makes it a bit more difficult for me to do my research, as I have no previous researches in this field to look in to. Also, obtaining data for my research was difficult. Because no one has done research on this subject, no one has created a data set for me to use. This is why I had to scrape my own data from the internet. I had to do this manually, by hand.

The small data set that I created can also be used in the future for other people in their research. This way, I am helping colleagues in the future, so that they do not have the same problem as I had. That way they can build further on this subject and hopefully gain new insights on this field. This database can be found on GitHub.

Even though there has not been done any research about the combinations of writing style and shared DNA, there has been a lot of research about authorship identification. This can be used to gain insights on writing style. Authorship identification was first discussed in Stamatatos et al. (2000). In this research, they used Greek newspaper articles as corpus. They said that texts can be characterized by the content of the text and by its writing style. They used multiple regression to create a value as output, indicating whether a text was written by particular author or not.

As mentioned in the research of Madigan et al. (2005), authorship identification is based strongly on the authors writing style and stylography. Reading a lot of researches about authorship identification and writing style gives me a better perspective on these subjects for my thesis.

Other important literature for my thesis was the Groningen Lightweight Authorship Detection model (Hürlimann et al., 2015). In this research, they tried to create a model that would predict whether texts were written by the same author or not. They compared one, two, three, four or five texts with a known author with one text with an unknown author. The classifier will produce a score between 0 and 1, indicating whether the texts were written by the same author or not. A score of 1 means that the texts were definitely written by the same author, while a score of 0 means that the texts were definitely not written by the same author. The research also explains how the model works and evaluates how well this model works on English texts.

The research of Hürlimann et al. was built on the PAN at CLEF 2015 challenge (PAN, 2015). Hürlimann et al. cited the overview of the authorship verification task at PAN at CLEF 2015 as follows: "In the authorship verification task as set in the PAN competition, a system is given a collection of problem sets containing one or more known documents written by author A_k and a new, unknown document written by author A_u, and is then required to determine whether $A_k = A_u$ without access to a closed set of alternatives. In this form, the task is generally interpreted as a one-class classification problem, in the sense that the negative class is not homogeneously represented and systems are based on recognition of a given class rather than discrimination among classes. This is akin to outlier or novelty detection and different from standard authorship attribution problems, where a system must choose among a set of candidate authors in a more standard, multi-class text categorisation fashion" (Hürlimann et al., 2015). This is also important literature for me, in order to understand why and how the Groningen Lightweight Authorship Detection model was built. On the website of PAN I also found freely downloadable data for training the model, which I used during my research.

3 DATA AND MATERIAL

3.1 COLLECTION

For my research, I had to obtain texts from identical twins in order to do research on the influence of shared DNA on writing style. At first, I was thinking of using texts written by myself and my twin brother. However, as my twin brother is doing research on the same topic, he had this idea as well. Since our research is already quite similar, I chose to use different data for my research than Kai. Sadly, I could not find any easily accessible data from identical twins online. Luckily, my thesis supervisor told me about two twin sisters, who wrote columns online.

The column Ask Ann Landers was an advice column. People could send in questions or problems to a fictional pen name Ann Landers. She would then respond to this person and give advice. It was created in 1943 by Ruth Crowley. After she died in 1955, Ask Ann Landers was taken over by Esther Pauline "Eppie" Lederer. Esther Lederer died in 2002, meaning the end of the column Ask Ann Landers.

In 1956 Esther's twin sister Pauline Phillips founded an advice column as well, called Dear Abby. The concept of this column was actually the same as Ask Ann Landers. In 2002, Paulines daughter Jeanne Phillips took over Dear Abby, as her mother got Alzheimer's disease. Pauline died in 2013, at the age of 94.

For my research, I decided to use the responses in the columns, written by Ruth Crowley, Esther Lederer, Pauline Phillips and Jeanne Phillips. This way, I have four people to compare writing style with. My model for authorship identification only compares texts of two people. The most important comparison will obviously be the comparison by Esther Lederer and Pauline Phillips. However, I also want to compare the texts written by Ruth Crowley and Esther Lederer, to see if Esther succeeded in taking over Ask Ann Landers without changing the writing style. I will do the same for Dear Abby by comparing the texts of Pauline Phillips and Jeanne Phillips. This is even more interesting as she is also her daughter. An overview of the relations between these four people can be found in Table 1, for clarification.

Because I could not find any database online, I had to manually find and scrape these columns online. This took a lot of time, especially for the older columns. At the end, I found some texts for all four persons, even though it is not as many as I had hoped. Eventually, I have found five texts written by Ruth Crowley, fifty-one written by Eshter Lederer, sixteen written by Pauline Phillips and fifty written by Jeanne Phillips. An overview can be found in Table 2. All texts are in English. The scraped data can be found on GitHub. I copied every single text in a .txt file. See Figure 1 for an example of a .txt file that is used as input.

Table 1: Overview of the relations between people.

	Eshter Lederer	**Pauline Phillips**	**Ruth Crowley**	**Jeanne Phillips**
Eshter Lederer	X	Twin sisters	Ask Ann Landers	
Pauline Phillips	Twin sisters	X		Mother/daughter and Dear Abby
Ruth Crowley	Ask Ann Landers		X	
Jeanne Phillips		Mother/daughter and Dear Abby		X

Table 2: Overview of the data collection.

	Number of texts	Average number of tokens per text
Esther Lederer	51	537
Pauline Phillips	16	484
Ruth Crowley	5	312
Jeanne Phillips	50	416

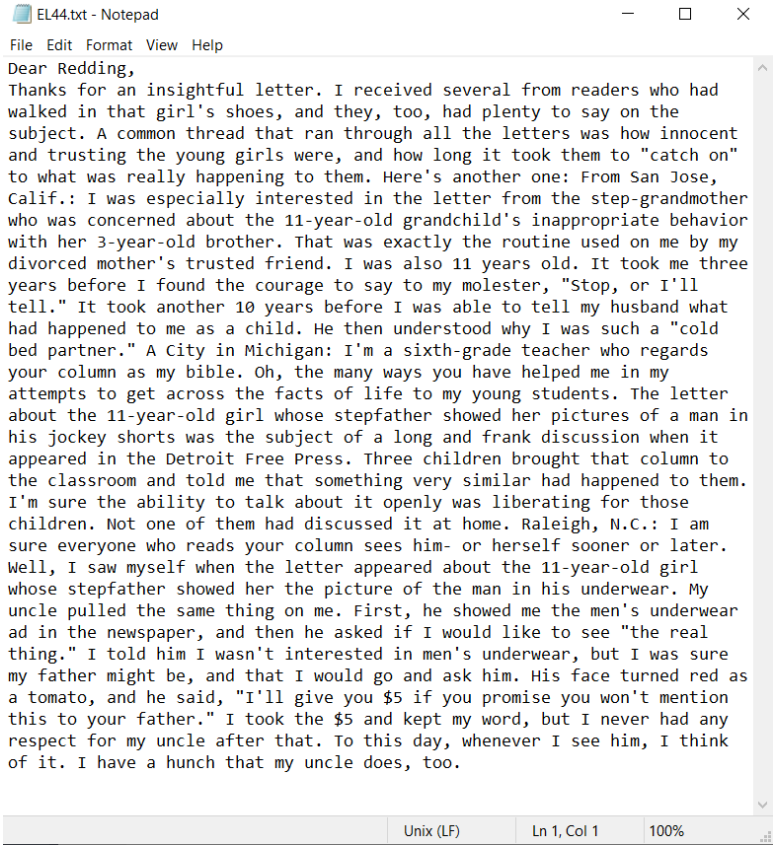

Figure 1: Example of a column from Ask Ann Landers in .txt format.

Additionaly to the previously described test data I also need training data for my model. Luckily, this is freely available and downloadable from the PAN at CLEF 2015 challenge website (PAN, 2015).

3.2 PROCESSING

Because I scraped my data from the internet, there are some problems with the encoding. I have to make sure that the encoding is in UTF-8 because otherwise my model cannot use this text as input. I have to write a script that does this automatically for all texts. I will do this in the same programming language as the Groningen Lightweight Authorship Detection model; Python. This script is called `processtexts.py`.

In this script I also have to make sure that there are no duplicate texts in my database. The texts of Ask Ann Landers are sorted into different genres like 'Money' and 'Work'. However, some texts have multiple genres. Because I scraped all texts from all genres, there are some duplicates in my database. There could be a text that is in both the 'Money' and the 'Work' genre. I have to make sure that these duplicates will be removed before testing the model on this data.

The Groningen Lightweight Authorship Detection model also has some pre-processing on its own, for example punctuation. This made it easier for me, as I did not have to do this myself.

4 METHOD

4.1 CLASSIFICATION

The research question for my thesis is: *"How much influence does DNA have on writing style?"*. In order to answer this question, I have to make a step aside from writing style to authorship identification. Because authorship identification is based very strongly on the writing style of a text, I can use this as an indication of writing style. For example, if reliable a authorship identification classifier classifies a relatively large amount of texts from the twin sisters as the same author, I can conclude that the twin sisters have a very similar writing style. If the classifier does not do the same for two "random" people, say for example Ruth Crowley and Jeanne Philipps, I can conclude that they do not have the same writing style and thus get insights on the influence of DNA on writing style.

The classifier for this research was already built, as I use the Groningen Lightweight Authorship Detection model of Hürlimann et al. (2015). This is how they described the classifier works. The system is implemented using Python's scikit-learn machine learning library as well as the Natural Language Toolkit (NLTK). They used an SVM with default parameter settings in all final models, with an implementation based on libsvm. The model predicts the output based on several visual features.

The first visual feature is that the model looks at the punctuation of the text. It registers the frequency of question marks, exclamation marks, comma's etc.

Secondly, the model looks at line endings of a text. It checks the frequency of full stops, comma's question marks etc.

Thirdly, the model checks the letter case. It calculates the ratio of uppercase characters to lowercase characters and the proportion of uppercase characters.

Also, the model takes line length into account. Here, it looks at the following features: sentences per line, words per line and proportion of blank lines.

Finally, the model looks at the block size of a text. This is split up in two parts. The first part is the number of lines per text block and the second part is the number of characters per text block.

Apart from all the visual features, the model also works with N-grams features, token features, sentence features, entropy features, compression features and (morpho)syntactic features.

Before the running the model on the test dataset, I have to train the Groningen Lightweight Authorship Detection model first. I will not train the model on any data of Dear Abby or Ask Ann Landers for two reasons. The first reason is that I simply do not have enough data to do this. The second reason is that I want to prevent overfitting the model by training too long on the test dataset. This is why I wil train the model on the freely available and downloadable training dataset from PAN (2015). I chose this training dataset because this is the same training dataset that Hürlimann et al. used. This way I can guarantee that the accuracy of the model is the same on the test data as for them, and thus acceptable.

4.2 EVALUATION

I will compare all possible pairs of combinations of Ruth Crowley, Esther Lederer, Pauline Phillips and Jeanne Phillips. I will compare multiple known texts written by one person to an unknown text written by someone else. The classifier will predict

this text as written by the same author or not. It will give a score between 0 and 1, 0 indicating that the unknown text was definitely written by someone else and 1 indicating that the unknown text was definitely written by the same person. For each comparison between two people, I will do this multiple times, depending on how much data is available. For example, I will compare five known texts of Esther Lederer with one unknown text of Pauline Philipps, ten times. I will then take the average score of these ten scores as final score between the two persons.

If a similarity score of the compared texts is higher than 0.5, I will label this as 'YES', meaning that the model predicted that the compared texts were written by the same author. If a similarity score is lower than 0.5, I will label this as 'NO', meaning that the model predicted that the compared texts were not written by the same author. If a similarity score is exactly 0.5, I will disregard this score because it is not classified as either 'YES' or 'NO'.

Finally, I will calculate how much percent of the time the model predicted the texts to be written by the same author. I do this by counting the number of times the model predicted 'YES', divided by the sum of 'YES' and 'NO'.

The Groningen Lightweight Authorship Detection model has a clear way of what the input should look like. In order to make the model work, this needs to be done in the right format. The model can be run using the following command: `python3 glad-main.py -training [TRAIN] -testing [TEST]`, replacing [TRAIN] with the training data set and replacing [TEST] with the test data set. As training data set you can simply use the training data set from PAN (2015) website. The test data set should be in the same format as the training data set. This means that it should be a folder containing multiple folders called `EN001`, `EN002`, `EN003`, etc. Inside the `EN001` folder there should be .txt files. There should be one, two, three, four or five known texts called `known01.txt`, `known02.txt` and so on. There also has to be a file called `unknown.txt`, which is the text that the model will classify. Inside the folder containing the multiple folders, there should be a file called `truth.txt`. This file specifies for all folders whether the texts inside that folder were written by the same person or not. An example of what this file should look like is shown in Figure 2. This is really important for the final score and thus prediction that the model gives. These final scores are put in a newly created text file called `answers.txt`. An example of what the file `answers.txt` looks like is shown in Figure 3.

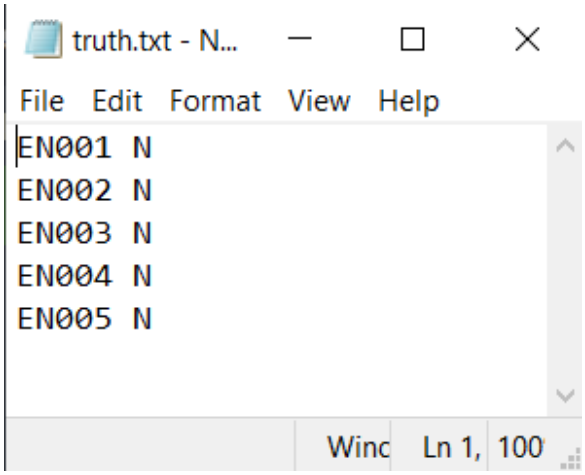

Figure 2: Example of `truth.txt`.

All the data and Python codes that I use for this thesis can be found on GitHub.

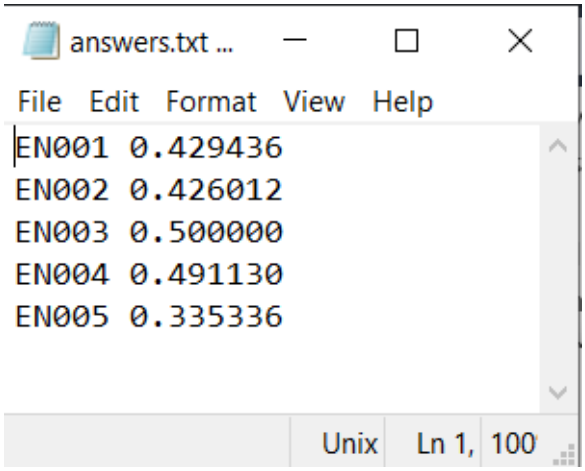

Figure 3: Example of `answers.txt`.

5 RESULTS AND DISCUSSION

5.1 RESULTS

The results of my research are the numbers that the model outputs for the test data. I ran the model on all possible combinations of pairs between Ruth Crowley, Esther Lederer, Pauline Phillips and Jeanne Phillips. The most interesting combination for my research is the comparison between Esther Lederer and her twin sister Pauline Phillips. Here I compared five texts of Esther with one text of Pauline, ten times. The model gave the following ten similarity scores: 0.455021, 0.577263, 0.424858, 0.500000, 0.282622, 0.526945, 0.208182, 0.651995, 0.180404, 0.562132. This means that the average score of my model for the comparison EL-PP (Esther Lederer - Pauline Phillips) is 0.436942. Each time the score is above 0.5, the model predicted that the compared texts were written by the same author. Each time the the score is lower than 0.5, the model predicted that they were not written by the same author. Disregarding the score of 0.5, the model classified the texts written the two twin sisters as same author 4 out of 9 times. This means that the model classified the texts from two authors with shared DNA as written by the same author 44 percent of the time. One might conclude that this is quite a high percentage for two different authors, but to make sure that this is significantly higher than the scores of different people, I also ran the model on different pairs.

The results of all scores for all combinations can be found in Table 3. In the first row, the combinations are written as abbreviations. For example, EL-PP means the combination of Esther Lederer and Pauline Phillips. RC stands for Ruth Crowley and JP stands for Jeanne Phillips. The computed results of all the combinations can be found in Table 4.

Table 3: All scores of the model.

	EL-PP	EL-JP	EL-RC	JP-PP	JP-RC	PP-RC
Scores	0.455021	0.411763	0.440201	0.412092	0.429436	0.403420
	0.577263	0.418131	0.514532	0.532823	0.426012	0.516249
	0.424858	0.454584	0.767869	0.366931	0.500000	0.337535
	0.500000	0.418888	0.349883	0.442628	0.491130	0.508039
	0.282622	0.449405	0.279930	0.199622	0.335336	0.385013
	0.526945	0.466128		0.611656		
	0.208182	0.422181		0.283872		
	0.651995	0.564720		0.491378		
	0.180404	0.425226		0.067593		
	0.562132	0.458901		0.711234		
				0.418562		
				0.446873		
				0.437005		
				0.482861		
				0.419424		
				0.603983		

Table 4: All computed scores of the model.

	EL-PP	EL-JP	EL-RC	JP-PP	JP-RC	PP-RC
Average Score	0.44	0.45	0.47	0.43	0.44	0.43
Number of 'YES'	4	1	2	4	0	2
Percentage of 'YES'	44%	10%	40%	25%	0%	40%

5.2 DISCUSSION

Looking at Table 4, there are some remarkable results. First of all, the average score that the model gave for the combination of Esther Lederer and Pauline Phillips is not higher than the rest. This can be explained by looking at Table 3. Here you can see that in the column EL-PP there are two outliers, with a very low score of 0.28 and 0.18. When looking at the files relating to this score, I found out that the texts used as input here are very short. It could be that this influenced the model and thus might be the reason of the very low scores. If I had more data to use as input, it would be easier to figure out whether these low scores are outliers or not.

Secondly, when looking at how much percent of the time the model predicted the texts written by two different authors as the same author, the percentage of the two twin sisters is the highest, as I expected. However, the percentage of EL-RC and PP-RC are also quite high. This could be explained by the fact that there are only five texts in my database written by Ruth Crowley. Sadly, this makes it difficult to make solid conclusions about the results, as there is not enough data. When looking at the other combinations of people with more data (EL-JP and JP-PP), the percentage of 'YES' is significantly lower compared to the percentage of 'YES' for the twin sisters (EL-PP), namely 10 percent and 25 percent versus 44 percent. Another reason for the high score of EL-RC might be that Esther Lederer took over the column of Ruth Crowley. When taking over the column, she tried not to change the column too much. This would mean that the writing style of Ruth Crowley and Eshter Lederer should be quite similar.

Finally, it is also interesting to see that the percentage of 'YES' lables of JP-PP is only twentyfive percent. I would have expected this score to be a little higher, as Jeanne Phillips took over the column Dear Abby from Pauline Phillips and thus probably try to remain roughly the same writing style. Above that, she is also Paulines daughter, so I expected their writing style to be more similar.

6 CONCLUSION

6.1 CONCLUSION

In this thesis I have done research to see how much influence shared DNA has on writing style. I have used a model of authorship identification in order to research this. While I expected that the model would classify the texts written by two twin sisters as the same author significantly more often than the texts written by other people, the results were a bit disappointing.

Even though the percentage of texts classified as the same author is the highest for the two twin sisters, the difference between other combinations is not always as much as I had predicted. As mentioned in Section 5.2, there are several possible explanations for this. The two main explanations are the fact that I did not have much data and the fact that the columns I used as input are sometimes very short.

The answer to the research question of this thesis *"How much influence does DNA have on writing style?"* is as followed. According to research based on the columns of two twin sisters, the influence of DNA on writing style is not so high. I had expected that the percentage of texts classified as the same author would be significantly higher for twin sister than for other people. The difference in percentage is however not as high as I expected.

6.2 LIMITATIONS

Even though I tried to make my research as firm and solid as possible, there are some limitations in my work and thus some improvements to be made in future work. Most of these limitations are due to the fact that I did not have enough time or resources during my research.

First of all, the results of this thesis are not very reliable, due to the small dataset. Even though I spend hours searching online for columns written by Ruth Crowley, Esther Lederer, Pauline Phillips and Jeanne Phillips, I did not manage to find enough columns for all of them. I only found five columns written by Ruth Crowley, which is really too few for a reliable result. Apart from that, there is also the problem that a lot of columns are shorter than fifty words.

Secondly, even though the Groningen Lightweight Authorship Detection model of Hürlimann et al. outperforms the baseline of 0.25, the score for English is actually the lowest, with a score of 0.41. This means that the model is not as reliable as I had hoped, especially not in combination with the small data set.

6.3 FUTURE WORK

For future work, I would recommend to use a much larger data set als test set. This would make the results of the research a lot more reliable.

Also, I would recommend use a model with a higher AUC-c@1 score. For example, there could be a similar research done with the same Groningen Lightweight Authorship Detection model, but on a Dutch data set. The AUC-c@1 score of Dutch is 0.61 in the model of Hürlimann et al. (2015), meaning the results would be more reliable due to the higher AUC-c@1 score.

BIBLIOGRAPHY

Baayen, H., H. van Halteren, A. Neijt, and F. Tweedie (2002). An experiment in authorship attribution. In *6th JADT*, pp. 29–37.

Hürlimann, M., B. Weck, E. van den Berg, S. Suster, and M. Nissim (2015). Glad: Groningen lightweight authorship detection. In *CLEF (Working Notes)*.

Madigan, D., A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye (2005). Author identification on the large scale. In *Proc. of the Meeting of the Classification Society of North America*, Volume 13.

Nederhoed, P. (2010). *Helder rapporteren: Een handleiding voor het opzetten en schrijven van rapporten, scripties, nota's en artikelen*. Bohn Stafleu van Loghum.

PAN (2015). Pan at clef 2015. https://pan.webis.de/clef15/pan15-web/author-identification.html.

Peñas, A. and A. Rodrigo (2011). A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1415–1424. Association for Computational Linguistics.

Stamatatos, E., W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M. A. Sanchez-Perez, and A. Barrón-Cedeño (2014). Overview of the author identification task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pp. 1–21.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2000). Automatic text categorization in terms of genre and author. *Computational linguistics 26*(4), 471–495.

Wang, L. (2017). News authorship identification with deep learning.