# How to be Remembered: People and Concepts intertwined

Serge ter Braake and Antske Fokkens

Media Studies, University of Amsterdam / CLTL, Vrije Universiteit Amsterdam Turfdraagsterpad 9, 1012 XT Amsterdam / De Boelelaan 1105 1081 HV Amsterdam, the Netherlands sergeterbraake@gmail.com, antske.fokkens@vu.nl

#### Abstract

This paper traces the link between famous people and concepts over time. People can be remembered for being 'wise', 'kind' or 'tyrannical', or simply for one main thing or event. Sometimes these perspectives on people, or the reason they should be remembered, shifts over time, and sometimes they remain relatively stable. The link between fame and concepts can be studied systematically with off-the-shelf tools with a gentle learning curve. In this paper we will demonstrate this by tracing the link between famous people and concepts in Dutch journal *Vaderlandsche Letteroefeningen*.

Keywords: fame, concepts, word associations

# 1 Introduction

This paper traces the link between famous people and concepts over time. People can be remembered for being 'wise', 'kind' or 'tyrannical', or simply for one main thing or event. Sometimes these perspectives on people, or the reason they should be remembered, shifts over time, and sometimes they remain relatively stable. Studies on how concepts are related to 'iconic' figures over time, such as princes or other royalty, do exist (Wieldraaijer, 2014, e.g.), but to the best of our knowledge this link between fame and concepts has not yet been studied systematically with the power of digital tools. In this paper we will demonstrate that this link can be traced quite easily with existing tools, following a bottom up approach. Specifically, we will look at famous people and their related concepts in Dutch literary and review journal Vaderlandsche Letteroefeningen (1776-1876). Our methodology allows us to chart why people are remembered and how the canonization of people and concepts are intertwined. This in turn, sheds light on biographers' choices on who to write about and on their and our own preconceptions.

This paper is structured as follows: In Section 2 we provide a brief overview of related work. In Section 3, we outline our methodology. After this we discuss the suitability of our corpus for our research in Section 4. Section 5 is dedicated to the analysis of two use cases: 'Jesus Christ' and 'The House of Orange'. Finally, we discuss our findings in Section 6 and conclude our paper in Section 7. Our search queries can be found in an appendix, in Section 8.

# 2 Related Work

Why people become famous and how concepts change over time are two traditional humanities questions. Both questions are difficult to answer with traditional methods because of the sheer size of the source material that has to be consulted to obtain results that surpass the anecdotal (Wieldraaijer, 2014, for a traditional study linking the early House of Orange to concepts). Trying to combine the two questions has, to the best of our knowledge, not been attempted yet with digital methods. For more extensive literature on fame we refer to our contribution to BD2015, in which we traced the fame of famous Dutch people in several Ngram viewers and in biographical dictionaries with a bottom up approach (ter Braake and Fokkens, 2015). In the current paper, we define 'fame' as the reason why people are mentioned in texts. We base our investigation on the assumption that if a person is famous for 'something', then he/she will co-occur frequently with the word(s) referring to this 'something' within the same text. We take this idea one step further by assuming that identifying the *concepts* that co-occur frequently with a person, will give us insight into the kind of reason someone is famous for, e.g. the type of event, like a battle, invention or prize won, or a specific societal role, like a profession or political function.

The term 'concept' deserves a bit more explanation. A concept is an abstract representation of an idea. For example the word 'democracy' represents the idea of a state form in which all (adults/men/free men) have voting rights. Most notably German historians spent a lot of time and resources to see how concepts changed over time, of which the most famous product is the Geschichtliche Grundbegriffe project (Richter, 1995, for a discussion). We are not only interested in so-called 'contested' concepts, such as 'democracy' and 'freedom', but also in simpler concepts like 'man' or 'century'. Any idea represented by a word, or several different words, will be taken into account, if the sources indicate there may be an interesting link with the person in question. A common approach for investigating change in concepts is by investigating the way in which words referring to the concept are used over time. When the associations of concepts change over time, we consider this to be a case of concept drift (see Fokkens et al. (2016) for a more detailed discussion).

Some of the seemingly most promising digital humanities work on tracing concepts through time has been carried out with word2vec, by among others Kenter et al. (2015) and Recchia et al. (2016). Word2vec creates vector representations of words based on their co-occurrence with other words. When words occur in similar contexts, their vector representations will be similar. The closer the (cosine of the) vectors of two words, the more related their semantics. Shifts can be traced by looking at the changing proximity of words over time. As appealing as this may sound, however, word2vec has trouble producing stable results as shown by Hellrich and Hahn (2016). In most cases, Humanities data is too small to produce vectors that are reliable and stable enough to capture the nuances researchers in Humanities are after. Wevers (2017) consciously<sup>1</sup> takes a different approach in his PhD thesis on the cultural influence of the United States on the Netherlands. He mostly traces concept change by generating word clouds and bigrams from newspaper advertisements. His computations run over relatively small, homogeneous datasets and he stays quite close to the sources for his analyses, in the sense that he follows a work flow that actively combines distant- and close reading (Wevers and Verhoef, 2017). Philosophers Betti and van den Berg also prefer to stay close to their texts (Betti and van den Berg, 2014). They stress the importance of studying concept change more structurally than has been done with traditional close reading only. They argue that concepts should not be studied in isolation, but together with related concepts within a 'system of concepts' (models). They are still working on how to model this computationally and how to support their research by computational methods.

# 3 Methodology

In this paper, we follow a strictly bottom-up and data driven approach, as we have advocated as well in ter Braake and Fokkens (2015). We start by selecting all articles that mention a person, and look at the most frequently occurring words in these articles. We take these words to trace the frequency in which they occur in the same article as a certain person over time, by using the Amsterdam Content Analysis ToolKit (AMCAT).<sup>2</sup> If word X occurs in all texts with person Y, then the 'association score' would be 1. The higher the association score the more related a term is to a person.

We selected two use cases about famous people from Dutch history and created subcorpora with the texts in which their names occur. Instead of looking for particular concepts that would be 'logical' to investigate in relation to these people (a top-down approach), we follow a bottom-up and theory neutral approach by identifying the most frequent words occurring in these texts. We then look at differences between the association scores between these concepts and people over time. Since the dataset is relatively small, we looked at periods of twenty years.

For our investigation, we are interested in discovering patterns. Since the interpretation of these patterns depends heavily on (historical) context and possible biases in the sources, we are less interested in statistical significance. A change in occurrences that is statistically insignificant may be historically very significant and the other way around. We therefore use the association scores in our tables as leads for further research, not as end results in themselves. We are especially interested in striking variations between the different selected concepts that we map to the famous people.

# 4 Our Corpus: Vaderlandsche Letteroefeningen

The journal *Vaderlandsche Letteroefeningen* (roughly translated 'Fatherland's Writing Exercises', henceforth VL) started out in 1760. Its goal was to offer anything that could serve to educate or entertain (Johannes, 1995). VL was one of the many periodicals published around this time, giving a voice to the public opinion of the literate elite (van Eijnatten, 2003, p. 10). VL did not 'belong' to any particular group within the Netherlands, but tried to stay neutral on the most delicate (religious) topics in order to appeal to a broad audience. Compared to many of its counterparts, VL was progressive and open to innovations and new ideas (van Eijnatten, 2003, p. 398-424). Logically any conclusions we draw in this paper are reflections of the contents of VL alone.

In VL, we are dealing mostly with book reviews on all kinds of topics. The topics that are dealt with seem to remain quite stable over time, dealing with state formation, relations to other countries, religion, science and medicine. The texts were brought to us in XML format from Nederlab,<sup>3</sup> from which we extracted raw text in utf-8 for our own analysis. There are a total of over 32,000 articles. There are 300 to 450 articles a year, except for the last fifteen years when VL's output radically dropped to less than 50 articles per year. Another difference between the last fifteen years of VL and previous years is that the number of words per article goes up. The association scores for our last period of 20 years will therefore be inflated, since longer articles increase the chance of two words occurring together. We did analyze these years, but omitted them from the results we are presenting in this paper.

The OCR quality of the texts is good enough to avoid problems with missing or misread terms, leaving us with the issue of spelling variation for the same term. Especially in the first thirty years of our corpus, the spelling is quite different from the other seventy years. If we for example look for a word like 'kracht' ('strength') then we see that it is spelled like 'kragt' until shortly after 1800. The same applies to 'vrijheid' ('freedom') and 'vryheid' and many other words. Fortunately, these changes in spelling are quite abrupt and without both forms co-existing for a longer period of time. Since we inspect the lists of most frequent terms manually, it is unlikely we miss potentially important observations because of this reason.

To summarize: we are dealing with over 30,000 high quality texts on a wide variety of subjects, all from one journal aimed at a general audience over a period of a hundred years. This corpus seems ideal for a first exercise in tracking the relation between people and concepts.

# 5 Selection of Use Cases

We did not have any prior demands for our use cases, other than the need for them to be related with enough data. Since

<sup>&</sup>lt;sup>1</sup>Wevers worked in the same team as Kenter at the time. <sup>2</sup>http://amcat.nl

<sup>&</sup>lt;sup>3</sup>https://www.nederlab.nl

our dataset ends in 1876 we obviously could exclude anyone who became famous after this time. In our 2015 paper, we found that the members of the House of Orange were dominant in several 'fame' lists, based on Ngram viewers from Dutch newspapers, literary sources, the Biography Portal of the Netherlands and Google Ngrams for Dutch (ter Braake and Fokkens, 2015). We also found that the most famous person without his own biographical entry in the Biography Portal of the Netherlands was Jesus Christ. A quick search in our corpus showed that both the House of Orange ('van Oranje') and Jesus Christ would generate enough hits to search for patterns in VL.

One advantage of taking Jesus as a use case is that what is known about this person is more of a mythological, or to some even fictional, nature than of a historical nature. One could even claim that Jesus Christ is not a person but a concept himself. Mapping the person/concept of Jesus to other concepts could provide us with important insights in how this person, and religion in general, was perceived over a period of a hundred years in VL. During this time there were many heated debates in the Netherlands about religion and the role of God (van Eijnatten, 2003). By mapping the concepts related to Jesus in VL over a longer period of time, we can directly contribute to the historiography on this topic.

For our second use case we decided not to single out one member of the House of Orange, but to take the House as a unit. This consideration is practical because in this House most men are called 'William', including the stadtholders and kings (from 1815 onwards) from our period of investigation. Disambiguating all these Williams would be a difficult task with our current corpus. The 'House of Orange' is not only a collection of famous individuals, but could also be considered a 'concept'. Taking the entire House of Orange as a unit for analysis obviously entails a gross generalization, since not all branches of the Orange family are valued the same. For the bigger part however, the texts deal with the contemporary Oranges, the late eighteenth century governors (stadtholders) and the nineteenth century kings. Nevertheless, we need to keep in mind that a change in concepts related to the house may have directly to do with the 'William' in charge at the time or with political circumstances.

#### 5.1 Use Case 1: Finding Jesus

Our first step was tracing all the texts in VL that have one or more mentions of the bigram 'Jesus Christ' ('Jezus Christus' or 'Jesus Christus'). By adding 'Christ' to the name of Jesus we can be relatively sure it refers to the Christian savior. We found over 1,000 articles which make one or more mentions of him. If we had expanded our search to include all references to Jesus only we would get 2637 articles and a graph that shows a similar distribution over time, which suggests we are mostly dealing with the same Jesus. We decided, however, that 1000 articles was enough for our analysis here, so we avoided the risk of 'polluting' our corpus.

We removed the stop words, looked at word frequency lists within these texts and categorized the words that make sense within this context. The first thing we noticed in

Concept	1776-1796	1796-1816	1816-1836	1836-1856
Happiness	0.268	0.274	0.236	0.202
Strength	0.448	0.433	0.511	0.595
Love	0.381	0.362	0.489	0.467
Freedom	0.316	0.271	0.456	0.435
Heart	0.437	0.466	0.300	0.429

Table 1: Mentions of Jesus with feelings

Concept	1776-1796	1796-1816	1816-1836	1836-1856
Ideas/views	0.259	0.288	0.280	0.247
Gospel	0.024	0.129	0.595	0.491
Religion	0.410	0.482	0.370	0.604
Influence	0.256	0.307	0.371	0.316
Teacher	0.214	0.344	0.300	0.193
Reason	0.295	0.334	0.323	0.291
Truth	0.515	0.509	0.615	0.662
Knowledge	0.042	0.071	0.089	0.257

Table 2: Mentions of Jesus with science and religion

our texts is a frequent occurrence of the words 'kracht' (strength), 'vrijheid' (freedom), 'liefde' (love), 'geluk' (happiness) and 'hart' (heart). We therefore decided to first chart the co-occurrence of Jesus with these terms over time. Table 1 shows an increase of the co-occurrence in the same article of Jesus with 'love', 'freedom' and 'strength', while 'happiness' and 'heart' stay relatively stable, suggesting a shift to a more 'personal' Jesus in the third quarter of the nineteenth century, in the sense that he seems to become more a man you could love and be inspired by than a distant God figure. We also queried Jesus with 'people' (God, children, man, men, people), 'nature' (earth, nature, death, world), and 'science and religion'. The first two categories did not show any noteworthy shifts, but the third did.

The most striking shift in Table 2 is the link from Jesus to science/knowledge ('wetenschap'). There is also, however, a general increase in the occurrence of the word 'wetenschap' in VL, which automatically increases the chances of co-occurrence with Jesus. The associations with the following terms increase gradually, and consistently, over time: 'religion', 'gospel', and, to some extent, 'truth'. A further close reading of the texts in VL should reveal whether Jesus indeed became a more personal figure and how an increased direct link to science/knowledge, religion and the gospel could be explained.

### 5.2 Use case 2: the many colors of Orange

The House of Orange has been an integral part of Dutch history ever since prince William of Orange initiated the 'Dutch Revolt' (1568-1648) against the House of Habsburg. William's descendants continued to be leaders of what later would be the Netherlands, as governor of most of the provinces of the Dutch Republic. After the death of William III, the Oranje-Nassau branch continued this role. After the French annexed the Dutch Republic in 1795, the House of Orange was sidelined for some time, until their return in 1815 when William of Orange VI was proclaimed King William I of the Netherlands. In 1830 the Belgian Revolution caused a rift in the Kingdom of the Netherlands,

Concept	1776-1796	1796-1814	1816-1830	1832-1852
Countrymen	0.063	0.302	0.160	0.150
Netherlands	0.587	0.453	0.766	0.850
History	0.270	0.377	0.553	0.724
Nation	0.190	0.208	0.234	0.252
Father	0.238	0.245	0.309	0.417
Fatherland	0.683	0.679	0.660	0.835
People	0.429	0.377	0.213	0.433

Table 3: Associations of 'of Orange' (van Oranje) and words related to State Formation and Nationalism

Concept	1776-1796	1796-1814	1816-1830	1832-1852
Happiness	0.159	0.200	0.223	0.197
Feeling	0.048	0.188	0.394	0.276
Heart	0.238	0.329	0.468	0.354
Love	0.079	0.235	0.309	0.346
Praise	0.159	0.365	0.394	0.291
Courage	0.127	0.188	0.202	0.268

Table 4: Associations of 'of Orange' (van Oranje) and Words related to Emotions

leading to the *de facto* independence of Belgium. The image of the Oranges was crafted as brave warriors and protectors of freedom and religion from the start. The brave warrior idea was abandoned eventually, but the ascribed role as protectors of freedom and religion remained (Wieldraaijer, 2014).

A lot is written in VL about the contemporary Williams, but the Oranges of the past, both the men and women, are definitely not ignored. Since name disambiguation would be problematic we decided to study this family as a group and search for associations with the bigram 'van Oranje'. The historical events created peaks in the mentioning of the name 'van Oranje' in the years 1815 (William I crowned as king) and 1830 (rift with Belgium). We used these peaks as part of the process to see what words are most frequently related to the House of Orange, but left them out of our association analysis, to avoid atypical events distorting our results.

The House of Orange has played a large role in Dutch state formation and as 'fathers and mothers' of the country. It is therefore not very surprising that we see many words related to state and national identity occurring frequently with the House of Orange. Our first query set out to chart how the house is related to these concepts, following the same methodology as before. Looking at Table 3, we see that 'of Orange' is increasingly related to the concepts of 'history', 'nation' and 'father'. These findings suggest that the house of Orange became more anchored as a way to unify the people of the Netherlands under the symbol of the House of Orange, with its members as fathers of the country with a long, deeply connected, history. In this respect we can also note that the association scores of 'of Orange' with 'century' ('eeuw') and 'battle'/'struggle' ('strijd') are going up significantly as well.

Interesting to note about Table 4 is that all emotions, all positive, get higher association scores until 1830. Then in the years 1832-1852, during which Belgium separated,

there is a drop for all of them, except for 'courage' and 'love'. These higher association scores also can be seen the other way around (from 'love' to 'of Orange' and from 'courage' to 'of Orange').

Finally, we looked at the category of religion and the House of Orange. Here we see a steady increase of the association scores with 'God' and 'Church' (this also is the case the other way around), while the scores with 'belief' and 'religion' remain quite stable. The association with 'truth' sees a rise between 1776 and 1830 and then remains quite stable. To summarize, the House of Orange gets stronger associations with national identity, as fathers of the country, with love and courage, with God and Church. This fits the role which, according to historiography, was attributed to them more strongly as a binding factor of the newly formed Kingdom of the Netherlands.

# 6 Discussion

One of the aims of the field of Digital Humanities is finding patterns in big data. In our view one of the main problems with digital humanities is that the data is often too small for the most powerful computations, whilst bigger datasets easily run the risk of becoming too heterogeneous for conscientious research. In this paper we therefore analyze a relatively small, homogeneous dataset with a straightforward methodology that stays close to the original texts. While this has clear benefits, it also limits the scope of the conclusions we can draw. An obvious next step in this line of research is to take the probability of terms co-occurring by chance into account while investigating the increase or decrease of co-occurrences. Methods that provide such statistics, such as Pointwise-Mutual Information scores, tend to introduce biases towards low-frequency words. We therefore see such an approach as complementary to investigating direct co-occurrences, which we plan to explore in the future.

Most importantly our conclusions are based on only one source. We cannot say the concepts of 'Jesus' and the 'House of Orange' changed in a certain way in the Netherlands, but only how they changed in VL. For future research we could expand our dataset to include many other sources as well, and we could carry out larger computations with more powerful tools, but we do not think this approach would make much sense. It would be like throwing five different pieces of fruit into one blender and then investigate how the concentration of vitamin C has changed compared to five years earlier. With such an approach it would be impossible to distinguish what fruit, or what factor, caused any changes and how we should interpret these changes. In order to be able to interpret our findings, it makes more sense to do the exact same exercise for other sources, such as more religiously orientated journals from the same period, and look at where we can see the differences with VL, if any. The results in this paper do tell us about what happened in one of the leading Dutch journals over a period of one hundred years and provides us with a starting point for further research and hypothesizing.

# 7 Conclusions

In this paper we provided an initial demonstration of how word associations can show certain concepts shifting away or moving closer to famous people over time. Our method is simple, but theory neutral and effective. We do not let preconceptions or historiography determine what we are looking for. Instead we use a completely bottom up approach to start our investigations. Furthermore, this method can be applied without too many reservations to smaller datasets as well.

We chose to look at concepts related to Jesus Christ and the House of Orange, not because we are necessarily interested in them, but because previous bottom up research on fame showed us that these people are among the most famous in Dutch history. We then generated word frequencies for the texts in which their names occur and, going from these frequencies, we selected the concepts to investigate further. Finally, we generated association scores with AMCAT. We primarily looked at the chances of the concepts occurring in the same texts as 'Jesus Christ' and 'of Orange', but also looked at it the other way around: the chance of Jesus Christ and 'of Orange' appearing in the same texts as the related words, to confirm if we would see the same pattern.

Jesus Christ appears to become a more personal and loving person in the course of the nineteenth century, while the House of Orange and its members become more related to national identity and are attributed with the role of fathers of the Netherlands. All of this seems to be in agreement with historical events and developments, which gives us confidence that our findings make sense and that our method can be used for other, less famous people as well. Even though we did not detect any 'events' related to our use cases, events will undoubtedly show when we use the same methodology to investigate other people. It goes without saying, however, that the patterns we are seeing are limited to only one corpus and could be influenced by many factors that cannot be immediately detected by computational methods. The final steps for an exercise like this should therefore always be going back to the texts and see what could explain our findings to be able to interpret them conscientiously.

### 8 Appendix: Key words in each table

Table 1: Terms that were searched for: 'Jezus Christus or 'Jesus Christus', 'Geluk', 'Kracht' or 'Kragt', 'Liefde', 'Vrijheid' or 'Vryheid'

Table 2 'Jezus Christus' or 'Jesus Christus', 'denkbeelden', 'evangelie', 'geloof', 'invloed', 'leeraar' or 'leraar', 'verstand', 'waarheid' or 'waerheid', 'wetenschap' or 'weetenschap'.

Table 3: 'van Oranje', 'vader', 'vaderland\*', 'Nederland\*', 'volk', 'geschiedenis', 'landgenooten' or 'landgenoten', 'natie'.

Table 4: 'van Oranje', 'geluk', 'gevoel', 'hart', 'liefde', 'lof', 'moed'.

#### 9 Acknowledgements

This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities and NWO VENI grant 275-89-029. We are grateful for the comments and suggestions of three anonymous reviewers.

# 10 References

- Arianna Betti and Hein van den Berg. 2014. Modelling the history of ideas. *British Journal for the History of Philosophy*, 4(22):812–835.
- Antske Fokkens, Serge ter Braake, Isa Maks, and Davide Ceolin. 2016. On the semantics of concept drift: Towards formal definitions of concept drift and semantic change. In Laura Hollink, Sndor Darnyi, Albert Meroo Peuela, and Efstratios Kontopoulos, editors, Drift-a-LOD 2016 Detection, Representation and Management of Concept Drift in Linked Open Data. Proceedings of the 1st Workshop on Detection, Representation and Management of Concept Drift in Linked Open Data co-located with the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016) Bologna, Italy, November, 2016., pages 10–17.
- Johannes Hellrich and Udo Hahn. 2016. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In *LaTeCH 2016Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*@ ACL. 2016.
- G.J. Johannes. 1995. De Barometer van de smaak. Tijdschriften in Nederland 1770-1830. Den Haag.
- Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1191–2000.
- Gabriel Recchia, Ewan Jones, Paul Nulty, John Regan, and Peter de Bolla. 2016. Tracing shifting conceptual vocabularies through time. In Laura Hollink, Sndor Darnyi, Albert Meroo Peuela, and Efstratios Kontopoulos, editors, Drift-a-LOD 2016 Detection, Representation and Management of Concept Drift in Linked Open Data. Proceedings of the 1st Workshop on Detection, Representation and Management of Concept Drift in Linked Open Data co-located with the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016) Bologna, Italy, November, 2016., pages 2– 9.
- Melvin Richter. 1995. *The History of Political and Social Concepts. A critical introduction*. Oxford University Press, New York and Oxford.
- S. ter Braake and A. Fokkens. 2015. How to make it in history. working towards a methodology of canon research with digital methods. In S. ter Braake, A.S. Fokkens, R. Sluijter, T. Declerck, and E. Wandl-Vogt, editors, *Proceedings of the first conference on Biographical Data in a Digital World (BD2015)*, pages 85–93.
- J. van Eijnatten. 2003. *Liberty and concord in the United Provinces : religious toleration and the public in the eighteenth-century Netherlands*. Brill, Leiden.
- M. Wevers and J. Verhoef. 2017. Coca-cola: An icon of the american way of life. an iterative text mining workflow

for analyzing advertisements in dutch twentieth-century newspapers. *Digital Humanities Quarterly*, 11(4).

- Melvin Wevers. 2017. Consuming America. A Data-Driven Analysis of the United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890-1990. PhD Thesis, University of Utrecht.
- M. Wieldraaijer. 2014. Oranje op de Kansel. De beeldvorming van ORanjestadhouders en hun vrouwen in preken, 1584-179. VU University Amsterdam, Amsterdam, phd thesis edition.