# Low Saxon dialect distances at the orthographic and syntactic level

Siewert, Janine

The Association for Computational Linguistics
2022

Siewert , J , Scherrer , Y & Wieling , M 2022 , Low Saxon dialect distances at the orthographic and syntactic level . in N Tahmasebi , S Montariol , A Kutuzov , S Hengchen , H Dubossarsky & L Borin (eds) , Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change . The Association for Computational Linguistics , Stroudsburg , pp. 119-124 , International Workshop on Computational Approaches to Historical Language Change , Dublin , Ireland , 26/05/2022 .

http://hdl.handle.net/10138/347809

# Low Saxon dialect distances at the orthographic and syntactic level

**Janine Siewert**
University of Helsinki
janine.siewert@helsinki.fi

**Yves Scherrer**
University of Helsinki
yves.scherrer@helsinki.fi

**Martijn Wieling**
University of Groningen
m.b.wieling@rug.nl

## Abstract

We compare five Low Saxon dialects from the 19th and 21st century from Germany and the Netherlands with each other as well as with modern Standard Dutch and Standard German. Our comparison is based on character n-grams on the one hand and PoS n-grams on the other and we show that these two lead to different distances. Particularly in the PoS-based distances, one can observe all of the 21st century Low Saxon dialects shifting towards the modern majority languages.

## 1 Introduction

We are investigating dialect similarity in 19th and 21st century Low Saxon based on data from Germany and the Netherlands. Traditionally, Low Saxon dialect classification has mostly been based on phonological and morphological traits, such as the ones presented by Schröder (2004). In this study, however, we focus on the orthographic and the syntactic side and compare how these relate to each other. We compare two levels as we expect the intensity and nature of the majority language influence to differ here. The choice of these two particular levels was motivated by the fact that orthography can be inspected without annotation and for syntax, we could train sufficiently reliable PoS taggers[1], which at this point is not possible for morphology and phonology. Furthermore, we investigate how the dialect closeness on both levels has changed over time.

An interesting area to pay attention to with respect to dialect distance is the Dutch-German border. Like Goossens (2019) observed, the Low Saxon dialects along the border have started to diverge under the influence of the majority languages. According to him, this divergence is most pronounced at the lexical level, but convergence towards the majority language has also been attested in phonology, morphology and syntax. While studies on the divergence of dialects along the border often focus on the occurrence and frequency of particular traits based on interviews, cf. Smits (2011), we address the overall (dis)similarity in prose texts.

Since in the 19th century school education and majority language media played a smaller role in everyday life compared with today, we assume the effect of language contact with Dutch and German to be less visible in the morphology and syntax of 19th century Low Saxon, as such changes to the language system itself take time and gradually add up. On the other hand, the border is probably already clearly discernable at the orthographic level due to reading and writing education in the majority language, which we assume to have had a more immediate influence, particularly in areas where the Low Saxon literary production had ceased (nearly) completely after Middle Low Saxon times. Therefore, from the 19th to the 21st century, we expect a greater change in distance towards the majority languages at the PoS level than at the character level. We thus hypothesize that the Low Saxon dialects will appear closer to each other on the syntactic side with distance to the majority languages decreasing over time, while 19th century dialects might already group together with the respective majority language at the orthographic level.

## 2 Background

The West Germanic language Low Saxon (also called "Low German") today is primarily spoken in Northern Germany and the North-Eastern Netherlands by around 5 million people and enjoys official recognition in both countries (Moseley, 2010). As a result of the lack of an interregional standard language, Low Saxon speakers tend to use their own dialects in all language use cases. As there is no official common orthography either, one needs to take into consideration two layers of variation: on the one hand spelling variation and on the other ac-

[1] Around 85% accuracy based on a manually annotated test set.

*Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 119 - 124
May 26-27, 2022 ©2022 Association for Computational Linguistics

tual dialect variation. People may for instance stick to their own dialect but switch writing systems depending on whom they address.[2] This multilayered variation poses challenges to the development of NLP for Low Saxon but at the same time presents an interesting case for historical dialectology of written language.

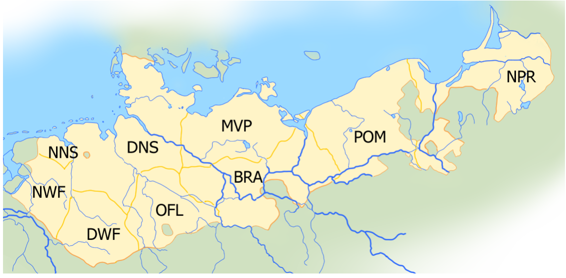

Figure 1: Major Low Saxon dialect groups: Dutch North Saxon (NNS), German North Saxon (DNS), Dutch Westphalian (NWF), German Westphalian (DWF), Eastphalian (OFL), Mecklenburgish-West-Pomeranian (MVP), Brandenburgish-South-Marchian (BRA), East Pomeranian (POM) and Low Prussian (NPR).

Figure 1 shows the major dialect groups of modern Low Saxon. The eastern dialects East Pomeranian (POM) and Lower Prussian (NPR) were spoken in these areas prior to WWII.

## 3 Data

The majority of our dataset is taken from the LSDC dataset (Siewert et al., 2020) since, as far as we are aware, this is the only dataset for modern Low Saxon annotated for dialect and century. Especially in regard to the 19th century data, we supplemented it with relevant prose texts from Leopold and Leopold (1882)[3] and the Twentse Taalbank (van der Vliet, 2021).

The overall size of the dataset is 120,720 sentences and 2,410,261 tokens and it covers eight dialect regions: Dutch North Saxon, German North Saxon, Dutch Westphalian, German Westphalian, Eastphalian, Mecklenburgish-West-Pomeranian, Brandenburgish-South-Marchian and Low Prussian. In this rough division, Dutch Westphalian includes all Dutch Low Saxon dialects except for Gronings, which consequently is identical with Dutch North Saxon here. The first five of these dialects are included in our current experiments. As we currently lack annotated data from Mecklenburgish-West-Pomeranian (MVP), Brandenburgish-South-Marchian (BRA) and Lower Prussian (NPR) for the 20th and 21st century, we cannot yet perform diachronic comparisons and thus exclude these dialects from our experiments as well. Furthermore, we do not use the 20th century data in our comparisons as it still consists mostly of data from only two dialects.

In our experiments, we thus used data from the five dialects presented in Table 1. We distinguish dialects from the 19th and 21st century and treat these as separate data points.

| | 19th | 21st |
|---|---|---|
| German North Saxon (DNS) | 3,869 | 475 |
| Dutch North Saxon (NNS) | 1,774 | 16,964 |
| German Westphalian (DWF) | 2,557 | 10,225 |
| Dutch Westphalian (NWF) | 4,925 | 9,150 |
| Eastphalian (OFL) | 278 | 7,896 |

Table 1: Sentences per dialect and century in our dataset.

For comparison, we also used UD data in Standard German (Borges Völker et al., 2019) and Standard Dutch (Bouma and van Noord, 2017) containing 153,035 and 18,078 sentences, respectively. These datasets seem to consist mostly of data from the late 20th and 21st century.

The Low Saxon data was converted to CoNLL-U format and automatically PoS tagged with the help of the Stanza tagger (Qi et al., 2020)[4] trained on UD data in Danish (Johannsen et al., 2015), Dutch (Bouma and van Noord, 2017), German (McDonald et al., 2013), and Swedish (Borin et al., 2008) in addition to manually annotated Low Saxon data.

In connection with the publication of the paper, our dataset, as well as the n-gram counts that form the basis for our experiments, will be added to LSDC-morph repository[5] on the Helsinki-NLP GitHub page.

## 4 Methods

Dialect similarity at the orthographic level based on character n-grams[6] will be compared to dialect

[2] Personal observation from conversations on social media.

[3] Digitised by dbnl: https://dbnl.nl/tekst/leop008sche00_01/

[4] We use the stand-alone version of the tagger available at https://github.com/yvesscherrer/stanzatagger.

[5] https://github.com/Helsinki-NLP/LSDC-morph

[6] Character n-grams, of course, do not purely represent the orthography as they will also capture actual dialect characteristics such as inflectional suffixes, but this is the closest one can get without adding a phonological or phonetic layer.

similarity based on PoS tag sequences to investigate if these lead to different dialect groupings.

Malmasi and Zampieri (2017) observed in their experiments for identifying Swiss German dialects that approaches based on character n-grams outperform word-based ones and, in their study on British dialects, Wolk and Szmrecsanyi (2016) have employed part-of-speech n-grams for corpus-based dialectometry, concluding that this approach can achieve results comparable to manually selected features.

### 4.1 N-grams

We extract character bigrams and trigrams from tokenised and lower-cased text. Trigrams consisting of the last letter of the previous word, a space sign and the first letter of the following word are included. As for PoS bigrams and trigrams, we exclude n-grams containing the tags 'SYM', 'X' and '_'. We remove PoS and character n-grams with an overall frequency of 5 or below and the counts of the remaining n-grams are normalised with tf-idf.

### 4.2 Distance measures

For dialect distance measuring, we make use of scikit-learn (Pedregosa et al., 2011) PCA with k-means clustering with cluster sizes ranging from 2 to 5.[7] The input for our experiments are matrices with raw n-gram counts which we first normalise using tf-idf and subsequently reduce to two dimensions with PCA for visualisation purposes. The results to be seen in Figure 2 and 3 are based on this PCA-reduced data. We ran the models several times and observed marginal changes only for a larger number of clusters, when cluster borders divided very close dialects. Consequently, the random initialisation did not have a substantial effect on the results. Additionally, we compared these results to k-means clustering without PCA reduction and to hierarchical clustering and obtained similar results, cf. appendix A.

## 5 Results

As expected, the PCA-based closeness and the clustering at the character-based level differ clearly from the PoS-based results, but not all of the divergences correspond to our expectations.

### 5.1 Character n-grams

As can be seen from Figure 2, in a two-cluster case based on character n-grams, the varieties group according to country borders, with German Low Saxon clustering in the lower left corner and Dutch Low Saxon and Dutch (NDL) in the lower right corner. German (DEU) at the top is grouped into the same cluster with German Low Saxon, but at a substantial distance from the dialects. When using three clusters, German is the first to be separated into its own cluster (cf. appendix A). In case of Dutch Low Saxon, the greater closeness to standard Dutch in 21st century Low Saxon compared with 19th century Low Saxon suggests that the Low Saxon dialects in the Netherlands increasingly conform to the principles of the Dutch orthography. Such a general tendency, however, cannot be observed for German Low Saxon.

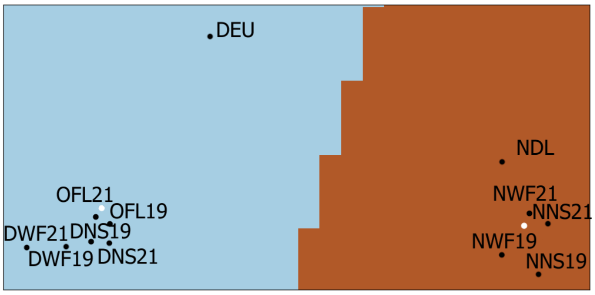

Figure 2: Dialect distances based on character n-grams

### 5.2 PoS n-grams

Compared to character n-grams, the PoS n-grams as presented in Figure 3 show a greater closeness of the Low Saxon dialects from both sides of the border. Specifically, when clustering into three groups, 19th century Low Saxon forms the left cluster, 21st century Low Saxon the middle one, and standard Dutch and German cluster on the right hand side.

When restricting the number of clusters to two, Dutch and German form one cluster and the Low Saxon dialects from both centuries form another.

For the PoS n-gram case, the century seems to play a greater role than the state border, since the clustering suggests that Low Saxon has become closer to the majority languages in terms of syntax.

It is remarkable that the overall distance between Dutch Low Saxon and German Low Saxon does not seem to have changed drastically over time. Dutch North Saxon and Dutch Westphalian seem to have approached each other and the same appears to be true for German North Saxon and Eastphalian.

[7] Inspired by this example: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html

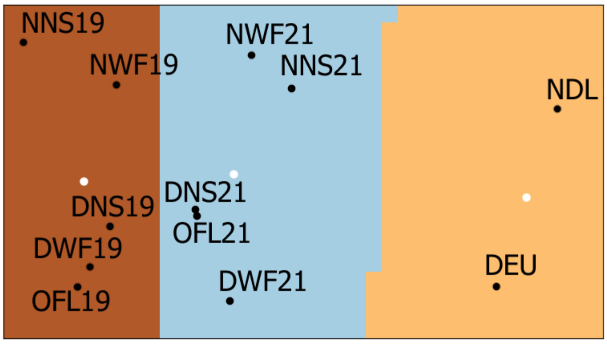

Figure 3: Dialect distances based on PoS n-grams

## 6 Discussion

Based on our knowledge of and about Low Saxon dialects, the overall results appear meaningful despite the comparatively low tagging accuracy of 85%.

In the PoS-based experiments, the fact that a noticeable distance between neighbouring dialect regions divided by a country border can already be observed in the 19th century data raises the question of how representative the written dialect is of the actual Low Saxon spoken by the average population. Given that written Low Saxon is commonly produced by people who have received their education in the majority language, this may have an influence on the kind of written language produced. On the other hand, one needs to keep in mind the size of the dialect regions. Both the German Westphalian group (DWF) and particularly the German North Saxon (DNS) group stretching from the Dutch border to Schleswig-Holstein are on their own larger than the whole Dutch Low Saxon area and not all of the texts included are written in varieties particularly close to the border. A more fine-grained dialect subdivision, where e.g., the Groningen dialect could be compared with East Frisian, would therefore be desirable for the future as well. However, this does not seem feasible in our research project at this point due to the lack of sufficient data sources for many of these dialects.

The noticeable distance between German Low Saxon and German in the character-based experiments compared with the closeness of Dutch and Dutch Low Saxon might partly be explained by the greater phonological differences between German and Low Saxon, but in addition to that, one might also consider that local writing systems for German Low Saxon tend to adhere to certain orthographic principles not found in the German orthography. One of these is that even the umlauted vowels *ä*, *ö* and *ü* may occur as digraphs, especially in closed syllables, e.g., in the words *däänsch/däänsk* 'Danish', *sööt* 'sweet' and *düüster* 'dark', according to both the Sass[8] spelling (Kahl and Thies, 2009) and the Münsterland spelling (Kahl, 2009).

The overall PoS-based distance of Dutch Low Saxon and Standard Dutch appears to be comparable to the overall distance between German Low Saxon and Standard German. This is interesting as, due to the greater phonological similarity (e.g. no High German consonant shift) on the one hand and the character n-gram results on the other, one might expect the distance between Dutch Low Saxon and Dutch to be relatively smaller on the syntactic level as well.

The relatively greater distance of 21st century German Westphalian to the other two German Low Saxon dialects deserves some attention, too. One possible explanation could be the Westphalian dialects' more conservative morphology. Whereas several dialects of German Westphalian still inflect nouns in three cases and have preserved subjunctive forms of verbs (Lindow et al., 1998)[9], it might be the case that Dutch Low Saxon, German North Saxon and Eastphalian more commonly resort to prepositions and auxiliary verbs.

The relative closeness of German and Dutch in the PoS-based results came as a surprise as well, but the genre might play a role here: Whereas Dutch and German data largely represents more formal language from non-fiction texts such as news texts, much of the Low Saxon data sources belong to various forms of literature. While the possibility of an influence of genre differences on the distance between 19th and 21st century Low Saxon dialects cannot be completely ruled out either, it seems less likely as the majority of the data from both centuries consists of fiction texts and stories.

Due to the relatively modern data in Dutch and German, the conclusions to be drawn from our comparison are restricted. For a more meaningful comparison, one should include 19th century Dutch and German as – even though gradual assimilation to the majority language is what one would expect – it might still be the case that the distance between 19th century Low Saxon and the Dutch and German

[8] Named after the creator Johannes Saß.

[9] While, according to Lindow et al. (1998, 152), the treefold case distinction is still in use in parts of Southern Eastphalian as well, our dataset does not include texts from this region as far as we are aware.

of that time was not as significant as the distance presented here would suggest.

## 7 Future research

In our future research, we will include more Low Saxon dialects, especially Mecklenburgish-West-Pomeranian, and add the 20th century as well as Dutch and German data from relevant time periods. The eastern dialects like Mecklenburgish-West-Pomeranian would constitute a meaningful addition since we could then examine the extent to which the common division into West Low Saxon and East Low Saxon / East Low German is apparent at the levels of language under scrutiny.

Morphological tagging would be a valuable addition as well, which we plan to include in the future. At this point, the accuracy is still too low, at around 60-70%, which is why more annotation work is required. In the future, we will create more training data for both PoS and morphological tagging through manual correction of the automaticcally tagged data.

Regarding dimensionality reduction, we intend to more closely inspect which features are considered most central by the model to investigate whether the dialect distances are based on actual dialect characteristics or if the results have been influenced by artifacts of the dataset.

We hope that the datasets gathered and annotated by us will facilitate the development of NLP tools for and research into Low Saxon.

## Acknowledgements

This work has been supported by the Academy of Finland through project No. 342859 “CorCoDial – Corpus-based computational dialectology”.

## References

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. Saldo 1.0 (svenskt associationslexikon version 2). *Språkbanken, University of Gothenburg*.

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden. Association for Computational Linguistics.

Jan Goossens. 2019. „Dialektverfall" und „Mundartrenaissance" in Westniederdeutschland und im Osten der Niederlande. In Gerhard Stickel, editor, *Varietäten des Deutschen: Regional- und Umgangssprachen*, pages 399–404. De Gruyter.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 157–167.

Heinrich Kahl and Heinrich Thies. 2009. *der neue Sass - Plattdeutsches Wörterbuch*. Wachholtz Verlag, Neumünster.

Klaus-Werner Kahl. 2009. *Wörterbuch des Münsterländer Platt*. Aschendorff Verlag, Münster.

Joh. A. Leopold and L. Leopold. 1882. *Van de Schelde tot de Weichsel*. J.B. Wolters, Groningen.

Wolfgang Lindow, Dieter Möhn, D Stellmacher, H Taubken, and J Wirrer. 1998. Niederdeutsche grammatik.

Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3 edition. UNESCO Publishing, Paris. Online version: http://www.unesco.org/culture/en/endangeredlanguages/atlas.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ingrid Schröder. 2004. Niederdeutsch in der Gegenwart: Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher, editor, *Niederdeutsche Sprache und Literatur der Gegenwart*, pages 35–97. Georg Olms Verlag, Hildesheim, Zürich and New York.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. 2020. LSDC - a comprehensive dataset for low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, page 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Tom Smits. 2011. Dialectverlies en dialectnivellering in nederlands-duitse grensdialecten. *Taal en Tongval*, 63(1):175–196.

Goaitsen van der Vliet. 2021. Twentse taalbank. `http://www.twentsetaalbank.nl/`. Accessed: 2021-12-15.

Christoph Wolk and Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. *The future of dialects: Selected papers from Methods in Dialectology XV*, 1:225.

# A Results of other clustering approaches

In this appendix, we list the outcomes of other clustering approaches.

## A.1 K-means clustering

| Dialect | Clusters | | | | |
|---|---|---|---|---|---|
| | 2P | 2C | 3P 1st | 3P 2nd | 3C |
| 19th DNS | 1 | 0 | 0 | 2 | 1 |
| 19th DWF | 1 | 0 | 0 | 2 | 1 |
| 19th OFL | 1 | 0 | 0 | 2 | 1 |
| 19th NNS | 1 | 1 | 0 | 0 | 0 |
| 19th NWF | 1 | 1 | 0 | 0 | 0 |
| 21st DNS | 1 | 0 | 2 | 2 | 1 |
| 21st DWF | 1 | 0 | 2 | 2 | 1 |
| 21st OFL | 1 | 0 | 2 | 2 | 1 |
| 21st NNS | 1 | 1 | 2 | 0 | 0 |
| 21st NWF | 1 | 1 | 2 | 0 | 0 |
| Dutch | 0 | 1 | 1 | 1 | 0 |
| German | 0 | 0 | 1 | 1 | 2 |

Figure 4: Results of k-means clustering based on data without PCA-based dimensionality reduction. The overall results are similar, only in the case of three PoS-based clusters, there was variation between runs as to whether the Low Saxon dialects cluster according to century or according to state. P = PoS, C = character.

## A.2 Hierarchical clustering

The hierarchical clustering[10] uses the following dialect numbering: 0 = 19th DWF, 1 = 19th DNS, 2 = 19th OFL, 3 = 19th NWF, 4 = 19th NNS, 5 = 21st NWF, 6 = 21st DWF, 7 = 21st NNS, 8 = 21st OFL, 9 = 21st DNS, 10 = DEU, 11 = NDL.

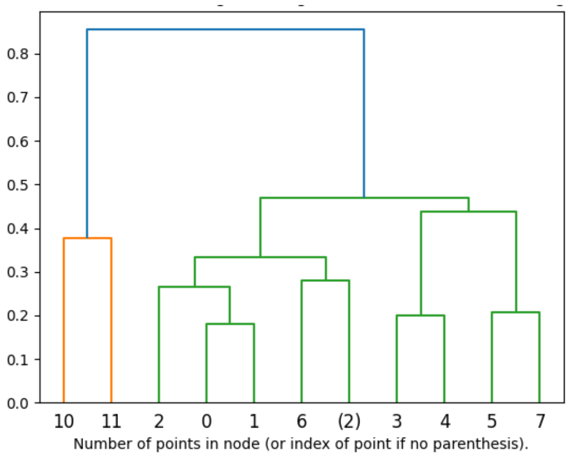

Figure 5: PoS-based hierarchical clustering using Euclidean metric and ward linkage.

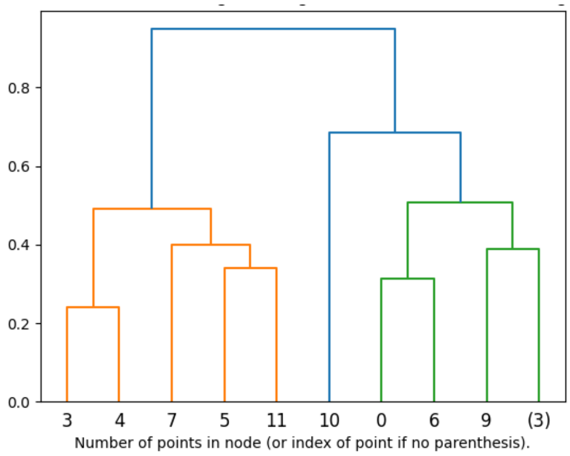

Figure 6: Character-based hierarchical clustering using Euclidean metric and ward linkage.

[10] Partly based on this example: `https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html`