Getting to Grippe With Influenza: An Investigation of Why the Disease Is Called That

Maria Bekker-Nielsen Dunbar[1,2,*,†], Manex Agirrezabal[3,†] and Tønnes Bekker-Nielsen[4]

[1]Department of Infectious Disease and Tropical Medicine, Heidelberg University Hospital, Heidelberg, Germany
[2]Centre for Research on Pandemics & Society, OsloMet – Oslo Metropolitan University, Oslo, Norway
[3]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark
[4]Department of Culture and Language, University of Southern Denmark, Odense, Denmark

Abstract

We investigate influenza-related terminology to gain a deeper understanding of what may have driven the choice in disease name when competing options were available. It is unclear why influenza in English should be called "influenza" and not "grippe" as the latter is seemingly the most common term for influenza within Indo-European languages. We examined why influenza is referred to as "influenza" in English using minimum edit distance to determine the available space in a language for a new disease term. We included other European languages for comparison. Available space may part of but not the full reason for why diseases are called what they are called when competing options are available.

Keywords

influenza, grippe, digital humanities, onomasiology, nosography

1. Introduction

Influenza is a respiratory disease with pandemic potential which has existed alongside humans for a long time. There are many influenza outbreaks, but there is certain agreement that influenza pandemics occurred in 1830 (to 1833), 1889 (to 1890), 1918 (to 1920), 1957 (to 1958), 1968 (to 1969), 1977 (to 1979), and 2009 (to 2010) whereby we need language to be able to discuss, record, and report on the disease. The term "influenza" originates from Italy, where it is found as early as 1360 [19]. In English, the disease was previously known as "influenza" and "grippe" (or "grippings"), both terms first attested in 1743 [3].

CHR 2024: Computational Humanities Research Conference, December 46, 2024, Aarhus, Denmark
*Corresponding author
†These authors contributed equally
bl328@uni-heidelberg.de (M. Bekker-Nielsen Dunbar); Manex.Aguirrezabal@hum.ku.dk (M. Agirrezabal); tonnes@sdu.dk (T. Bekker-Nielsen)
https://manexagirrezabal.github.io (M. Agirrezabal)
0000-0002-7249-3524 (M. Bekker-Nielsen Dunbar); 0000-0001-5909-2745 (M. Agirrezabal); 0000-0003-4628-5411 (T. Bekker-Nielsen)

With the advent of germ theory, the agent responsible for influenza was (incorrectly) identified and named as Haemophilus influenzae in 1892 which contains "influenza" in its name and could be considered a reason why "grippe" is not used to denote influenza in contemporary English. We suggest a different reason as disease names need not match their pathological agents: there was no lexical space for "grippe" to become the dominant disease term in English while space existed for "influenza". We illustrate this with a comparison with Romance and Germanic languages to support our argument.

1.1. Related work

Menadue [11] did a descriptive analysis of outbreak-related terms in fiction. The "plague.txt" project considered synonyms for the disease name for plague (called ontology terms, an example is the term bubo as a representation of plague) and suggest descriptive analysis of text corpora may provide insights [6]. Textual investigation is a common line of inquiry to understand infectious diseases: other researchers have also examined plague alongside measles, [1] cholera [12, 1], typhus, malaria, and smallpox [12] and sexually transferred infections [10].

Custom text corpora have been constructed for the purpose of examining influenza specifically [20, 16]. We note that Solovejute and Gatherer [16] stated "we conclude that there is little or no corpus linguistic signal in the UK national press for large-scale outbreaks of unidentified respiratory disease for the period 1785 to 1890" but there seems to be an increase for the Northern Echo and Ipswitch Journal in that period of time (red and green in the upper panel of Figure 1, respectively). We suspect the increase in occurrences of "flu" around the mid 2000s is an artefact of reporting on a large outbreak of avian influenza (H5N1) in and around the time of 2004 to 2007 which is likely being referred to colloquially as "bird flu". We also consider the results of the work by Taylor and Kidgell [18] noting that it is not immediately comparable as it combines our terms of interest (grippe and influenza (plus flu)) as well as severe acute respiratory syndrome (SARS) disease. While the mentions of influenza by itself follows the timings of known pandemics well as could be expected (Figure 1, lower panel where shaded areas and large counts match from 1900 onwards with the exception of the early 2000s), there seems to be less of an obvious pattern when SARS is included. The two approaches (Solovejute and Gatherer [16] and Taylor and Kidgell [18]) seem to match well for the later part of their study periods but there are discrepancies before 1950.

These plots of time series of word frequencies over time reflect trends of use as captured in the sources investigated. News media is aimed at the general public and is unlikely to contain vast amounts of medical jargon and thus is a weak proxy for general language use. We do not consider it a full proxy as journalists writing about medical subjects will want to avoid colloquialisms to bolster the authority and legitimacy of their reporting which will of course provide noise to the signal if it is supposed to be a reflection of everyday language. Medical journals have also been used to construct custom text corpora and this approach has been used to examine changes in medical language [2] though not for our terms of interest. Descriptive analysis of our terms can provide an indication of when "grippe" is used less comparative to "influenza" but not why which is what we wish

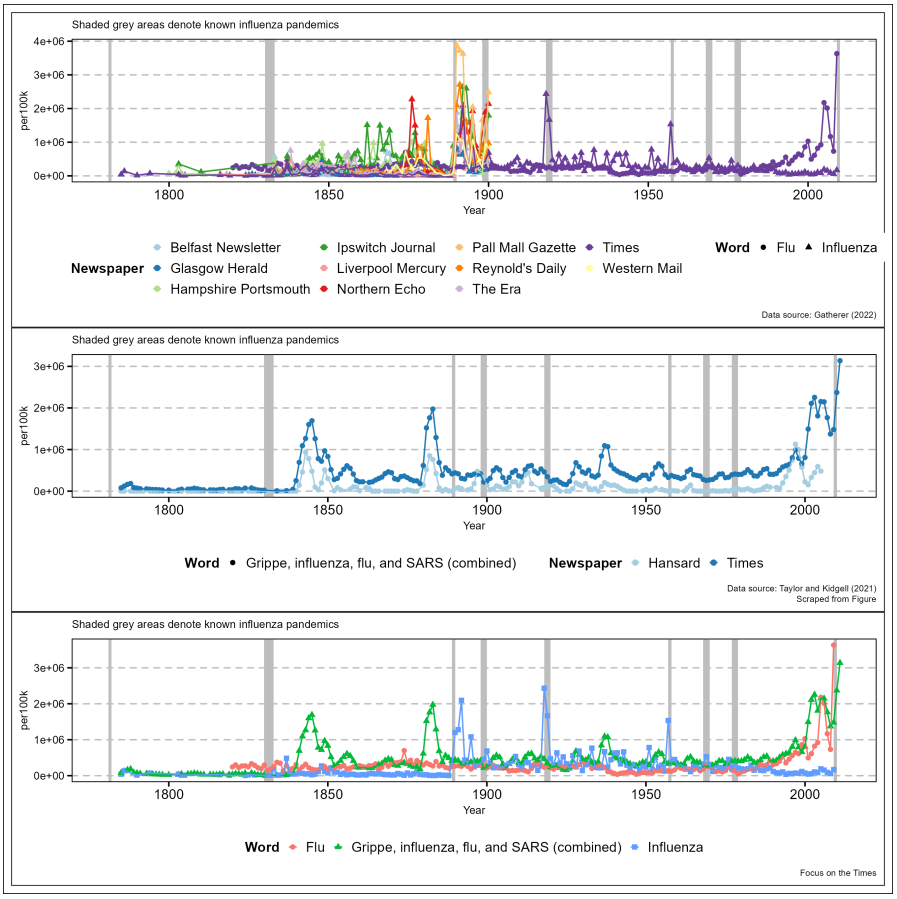


Figure 1: Data from Solovejute and Gatherer [16] on occurrence of influenza in British news media (top), scraped from Taylor and Kidgell [18] (middle), and a comparison of occurrences in the Times; a source used by both research groups (bottom)

to investigate.

2. Methods

We examined the dominant term for influenza in European languages (Figure 2) using the term in use in Wikipedia (as a reflection of most common use) or as is translated from English in dictionaries for translation. We found that "grippe" is the term used in most European languages with the exceptions being English, Nordic languages, Uralic languages, and Italian from whence the term "influenza" originates. English and Italian are outliers on their language branches (Figure 2) and a belt of "grippe" is found between them when plotted on a map. We sought to determine why influenza is not called "grippe" in English. We postulate that is due to a minor form of synonymy avoidance by arguing that there was not space in English for "grippe".

Euphonics explains how a word fits into a language. Rather than investigating ease of pronunciation, we examine which word can be said to have the least amount of ambiguity. In this work we consider that for a word to have "space" in a language and thus fit, it needs to not be taking up space already occupied by a similar word. A contemporary example is that COVID-19 (full name: coronavirus disease 2019) was often abbreviated as "corona" in Scandinavian countries but not in Spanish as the word corona (meaning crown) already existed.

Table 1
Influenza-related terms (above) and control words (below)

English	Spanish	German	French	Italian	Danish
influenza	influenza	influenza	influenza	influenza	influenza
grippe	gripa	grippe	grippe	grippa	grippe
catarrh	catarro	Katarrh	catarrhe	catarro	katar
cough	toser	husten	tousser	tossire	hoste

In this research we consider influenza-related terms used in English, Italian, French, German, Spanish, and Danish (Table 2). We include "catarrh" and "cough" as control words, the former is in use in the 1700s [5, and Samuel Johnson's dictionary] and 1800s [17, 4] and the latter is a common respiratory disease symptom. We then investigate whether "grippe" and "influenza" could take up space in the target language, though certain words are already in use and suspected to "block" the incoming influenza-related term (some examples are given in Table 1).

We determined the space available for the words of interest (Table 2) through calculating the minimum edit distance of our influenza-related terms. We expect, under this hypothesis, that whether "grippe" or "influenza" is preferred in a language depends on whether there is space in the language for that term. We do not pass judgement on these words, i.e. we make no further assertions about their role besides potentially being present.

We examine dictionaries to determine the available space for words starting from "grippe" and "influenza". By changing a letter we investigate how many similar words we find in the dictionary. If many similar words exist, we can conclude there is little space for the word. We construct our own dictionaries based on the latest raw exports[1] from Google Books Ngram Viewer.[2] These exports allow us to construct dictionaries covering a specific time period, where we wished to focus on the nineteenth century. This is because the closer our cut-off is to the present day, the greater the risk of language bias towards English in medical communication.

To determine the available space for the disease term, we use Levenshtein distance wherein we count the minimum number of operations required to match another word contained in the dictionary. The possible operations to change a letter in the word

[1] https://storage.googleapis.com/books/ngrams/books/datasetsv3.html
[2] https://books.google.com/ngrams/

Table 2
Potential terms already taking up space for influenza-related terms. Adjectives related to influenza such as "grip(p)al" are not included in the table, nor the contemporary occupation of "influencer".

English	Spanish	German	French	Italian	Danish
influence (V) influence (N)	influenciar (V) influencia (N) influir (V) influjo (N) influyente (A) influenciable (A)	Influenz (N) influenzieren (V)	influencer (V) influence (N) influer (V) influent (N) influençable (A)	influenzare (V) influsso (N)	influere (V) influens (N) influent (N)
grip (V) gripe (V) grip (N) gripe (N)	gripar (V) gripaje (N)	Grip (N) Grips (N)	agripper (V)	grippare (V)	gribe (V) grib (N) grip (N)

are: deletion, substitution, and insertion. Using the distance metric, we determine the number of words that can be found in a language following the operations. The intuition is that given two words that have the same length, if we change the same number of letters (perform the same number of operations on the word), the word that has more alternatives in a dictionary will have less space in the target language as that space is already in use by a greater number of words.

As different words may have different lengths, instead of considering the total number of operations we consider the proportion of a word that has been changed. If we make two operations on a 4 letter word, half a word is changed, while if we make two operations in a 10 letter word, one fifth of it will be changed. This allows us to plot the number of possible matching words in a dictionary against the amount changed.

Regarding pre-processing of the data, we removed digits, punctuation, and made all words lowercase. We considered only without any Part-Of-Speech attached (meaning no words with something after an underscore). We added the word counts for different years (previous 10 years to the moment of interest) and after summing everything, we made sure that each word appeared at least 10 times. We did this to avoid considering nonce words (these words may appear from optical character recognition errors or misspellings). The code will be made available at: xxxx.

3. Results

As the Google Books Ngram source does not contain Scandinavian languages, we were are unable to do the entire analysis (including Danish) but continued the investigation with five target languages (English, Spanish, German, French, and Italian). Figure 3 shows a plot for each language, where we can observe how many words were found in each dictionary (vertical axis) after altering a certain proportion of the words (horizontal axis). We only show the proportions where no more than the 60% of words were altered.

If a larger proportion of the words is altered, we could observe that longer words would have more matching words in the dictionary regardless of whether the word was a true word. In the future, we should further analyse the effect of word length in these plots.

This plot is constructed using the dictionaries obtained from the words that were used between 1820 and 1830. We show the plot for four words of interest, namely: "gripe", "grippe", "influenza" and "'catarrh" as a control. We include in the background of each plot 50 randomly sampled words from each target language. This is intended as an aid that provides a sense of the amount of space in the language in general and sets the infectious disease terms in a wider non-epidemiological context.

If we take a closer look at each target language, the plot with English words (Figure 3a) has the word "influenza" as the word with the fewest matching words, by a very small margin. This is identified as the curve that is closest to the lower right corner of the plot. In the case of the plot for the words in Spanish (Figure 3b) the words "grippe" and "gripe" seem to have the least number of matches in the language. The results for English and Spanish thus coincide with the actual word preference as shown in Figure 2 and support our postulation that available space can determine which term becomes used.

The plot that represents French words (Figure 3c seems rather similar to the English one, but in this case, the plot lines are closer to each other, making it harder to draw conclusions. In the case of German (Figure 3d), "influenza" seems to be the term that would be preferred, which does not coincide with the use of "grippe" according to Figure 2. The last plot (Figure 3e) shows the results for Italian, and it seems as the preferred word in this case would be the same as in Spanish, as both the words "gripe" and "grippe" seem lower than both "catarrh" and "influenza". Italian and Spanish are similar in structure so the results being similar for these languages is not surprising.

We also considered the years 2014 to 2024 to reflect the use of language further removed from the origin of the two terms (results not shown). The results are more homogenous for all languages which makes it difficult to conclude in favour of either term with recent data.

4. Discussion

An additional dimension not fully explored yet is geography. We postulate it may be the case that there are a few languages influencing neighbouring languages: Welsh vs Breton (Figure 2) suggests there is a neighbouring language trendsetter effect rather than the choice of term being driven by euphonics or space available in a language. This might explain why Estonian and Hungarian use different terms (Figure 2) though they are linguistically similar.

We were not able to investigate regional language differences, only national languages. This may have an impact as Florentine, Castilian, and High German became Italian, Spanish and German, respectively, whereby the usage is a reflection of a specific sub-national dialect rather than the entire country. This may explain why the results for Italian and German are not as obviously interpreted in favour of our argument as for

English.

We have not considered the socio-linguistic dimension of language in this work. If "grippe" was a colloquial word (which is not unthinkable since "influenza" is related to the name of the infectious agent), we would need to determine where it is used. Colloquialisms rise up through society (which is well-documented in Latin) and so popularity flows from the people to "high society". We have treated all texts considered in the data as effectively having equal weight which is to say we have not distinguished them upon inclusion.

A strength of our work is that we considered multiple languages. This comparison allows us to examine the words not just in their English context but also the surrounding contexts, which strengthens our argument. The inclusion of more languages and more data could strengthen the results further. Showing that there was space for "influenza" but not "grippe" in English and "grippe" but not "influenza" in Romance languages (bar Italian where it originates) suggests our instincts have merit or at least that this could be part of the explanation for why influenza is called "influenza" in English. There are differences in how the word is used as an adjective related to disease, as regards the virus we speak of "virus grippeaux" in French, "Grippevirus" in German and "influenza-like virus" in English.

Currently diseases are named by the World Health Organization who follow specific guidelines to avoid the unnecessary stigmatisation of peoples [21] and viruses and bacteria are named by International Committee on Taxonomy of Viruses and International Committee on Systematics of Prokaryotes, respectively. The naming is sometimes revised [22, 9] but we believe historical investigations will provide greater insight than ones for present-time.

5. Conclusion

We introduced names used for influenza as well as its historical context. We placed our work in relation to work by others to provide an overview of investigations done by others using similar data sources. We investigated which term is used in Europe by country and suggested the use of minimum edit distance to determine whether this could explain why one term is used over the other in specific languages, focusing on the words "grippe" and "influenza" as terms to denote influenza. We examined English, Spanish, German, French, and Italian.

Based on character-level Levenshtein distance, we calculated and visualised the number of matching words in a dictionary against the proportion changed. Building on this, we found that a possible explanation for which of the terms is preferred in a language could be affected by the number of words with similar spellings. More words with similar spellings would lead to increased ambiguity whereby a less ambiguous term would be favoured. This simple method can be used for other comparisons of words and languages that have an associated dictionary. We outlined limitations of our work–that this is but one dimension of how language is determined–as well as strengths such as using control vocabulary and multiple languages.

5.1. Future directions for work

In this project we considered the birth and berth of influenza-related terms. We did not consider the death of a word, i.e. when a word ceases to be used and so ceases to be included in a dictionary, which would be the other end of the chronology. The lexicology of epidemiologic terms has to our knowledge not been examined and could be a line of future inquiry. In future investigation to strengthen the work, we will investigate other terms to evaluate whether the method works for other diseases. Infectious diseases of a certain age (meaning having been with humans a long time) are the ones likely to experience changes, so terminology for smallpox and typhus may also be worth investigating.

The shortened form "flu" is not solely used to refer to influenza and may also be used for gastro-intestinal infections "stomach flu and intestinal flu" [8]. This is a similar problem to that encountered by McEnery and Baker [10] in their investigation of "pox" as it pertains to sexually-transferred infections. Further confusion arises from hand, foot, and mouth disease being referred to as "tomato flu" [15, 14]. Researchers will need to determine whether "flu" is related to influenza or other terms.

Similarly to the chai/tea divide, with one term arriving via land routes (chai) and one via sea routes (tea), one possible line of investigation for further research are the records of physicians on ships. Epidemic diseases were a major concern on ships and in 1782, a squadron of the British Royal Navy was forced to abandon its mission and return to base after an outbreak of influenza on board [13]. The vessels transporting enslaved African people to the Americas were accompanied by ships physicians whose observations made important contributions to epidemiology [7].

Additionally, the biology of the disease has largely been neglected in this work. The north/south divide in terms with "influenza" mostly in the north and "grippe" in the south (Figure 2) could be due to where the disease is historically endemic which may be different between northern Europe and central and southern Europe and an investigation of historical terms taking into account disease endemicity could be warranted. This may become evident after examining the medical records.

References

[1] S. Abbasi Sardari, M. S. Khaghaninejad, and A. Moloodi. "Investigating the Social Representations of Contagious Diseases in the 19th and 20th Centuries: A Corpus-based Analysis of Cholera, Measles, and Plague in American Historical Texts". In: Journal of Research on History of Medicine 9.2 (2020), pp. 123–154.

[2] M. Abdalla, M. Abdalla, S. Abdalla, M. Saad, D. S. Jones, and S. H. Podolsky. "Insights from full-text analyses of the Journal of the American Medical Association and the New England Journal of Medicine". In: eLife 11 (2022), e72602. DOI: 10.7554/eLife.72602.

[3] C. de Bell. "Of the influenza and luxury". In: Gentleman's Magazine 13 (1743), pp. 206–207. URL: hdl.handle.net/2027/mdp.39015013465870.

[4] G. Birt. “Influenza: Wanted a Definition”. In: British Medical Journal 2 (1895), p. 1391. DOI: 10.1136/bmj.2.1822.1391.

[5] M. Bresalier. “Naming Flu: Classification and Its Conflicts”. In: Modern Flu. Medicine and Biomedical Sciences in Modern History. Springer, 2023. DOI: 10.1057/978-1-137-33954-6_2.

[6] A. Casey, M. Bennett, R. Tobin, C. Grover, I. Walker, L. Engelmann, and B. Alex. “Plague Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894-1952)”. In: Journal of Data Mining & Digital Humanities jdmdh:6071 (2021). DOI: 10.46298/jdmdh.6071.

[7] J. Downs. Maladies of Empire: How Colonialism, Slavery, and War Transformed Medicine. Harvard University Press, 2021.

[8] M. Kay. “Lexemic change and semantic shift in disease names”. In: Culture, Medicine and Psychiatry 3.1 (1979), pp. 73–94. DOI: 10.1007/bf00114693.

[9] K. Kupferschmidt. “Alaskapox virus renamed to avoid stigmatizing state and hurting tourism”. In: Science (2024). DOI: 10.1126/science.zluvavg.

[10] T. McEnery and H. Baker. “A geography of names: A genre analysis of nationality-driven names for venereal disease in seventeenth-century England”. In: Corpus Pragmatic Studies on the History of Medical Discourse. John Benjamins, 2022, pp. 23–48. DOI: 10.1075/pbns.330.02mce.

[11] C. B. Menadue. “Pandemics, epidemics, viruses, plagues, and disease: Comparative frequency analysis of a cultural pathology reflected in science fiction magazines from 1926 to 2015”. In: Social Sciences & Humanities Open 2.1 (2020), p. 100048. DOI: 10.1016/j.ssaho.2020.100048.

[12] A. Moloodi, Z. Montaseri, and M. S. Khaghaninejad. “Public Attitudes on Deadliest Diseases from 15th to 17th Centuries: A Corpus-driven Analysis of Historical Texts”. In: Journal of Research on History of Medicine 8.2 (2019).

[13] J. W. Moore. “Remarks on the Compulsory Notification and Registration of Infectious Diseases”. In: British Medical Journal 1.158 (1880). DOI: 10.1136/bmj.1.996.158.

[14] S. S. Paul, M. Soneja, and N. Wig. “Tomato flu: misnomer for a common disease”. In: Lancet Respiratory Medicine 11.1 (2023), e3. DOI: 10.1016/s2213-2600(22)00371-x.

[15] P. Satapathy and R. Sah. “How a mutation earned a common viral infection a new name”. In: Nature India (2022). DOI: 10.1038/d44151-022-00098-2.

[16] R. Solovejute and D. Gatherer. “Hunting for Previous Coronavirus Pandemics Using Corpus Linguistic Analysis of 19th Century British Newspapers”. In: Preprints 2020120817 (2020). DOI: 10.20944/preprints202012.0817.v2.

[17] F. C. Steward. “Dr. Stewart’s Account of the Epidemic Influenza Prevailing at Paris; Known There under the Name of “La Grippe””. In: Medical Examiner (Philadelphia) 5.23 (1842), pp. 353–354.

[18] C. Taylor and J. Kidgell. “Flu-like pandemics and metaphor pre-covid: A corpus investigation”. In: Discourse, Context & Media 41.100503 (2021). DOI: 10.1016/j.dcm.2021.100503.

[19] M. Villani. Cronica. 1825. URL: gutenberg.org/cache/epub/69901/pg69901-images.html.

[20] J. Y. Wang. “An Empirical Analysis of Extended Meanings of Lexical Items in a H1N1 Corpus”. In: International Journal of Computer-Assisted Language Learning and Teaching 4.3 (2014), pp. 52–67. DOI: 10.4018/ijcallt.2014070104.

[21] World Health Organization. WHO issues best practices for naming new human infectious diseases. 2015. URL: who.int/news/item/08-05-2015-who-issues-best-practices-for-naming-new-human-infectious-diseases.

[22] World Health Organization. WHO recommends new name for monkeypox disease. 2022. URL: https://www.who.int/news/item/28-11-2022-who-recommends-new-name-for-monkeypox-disease.

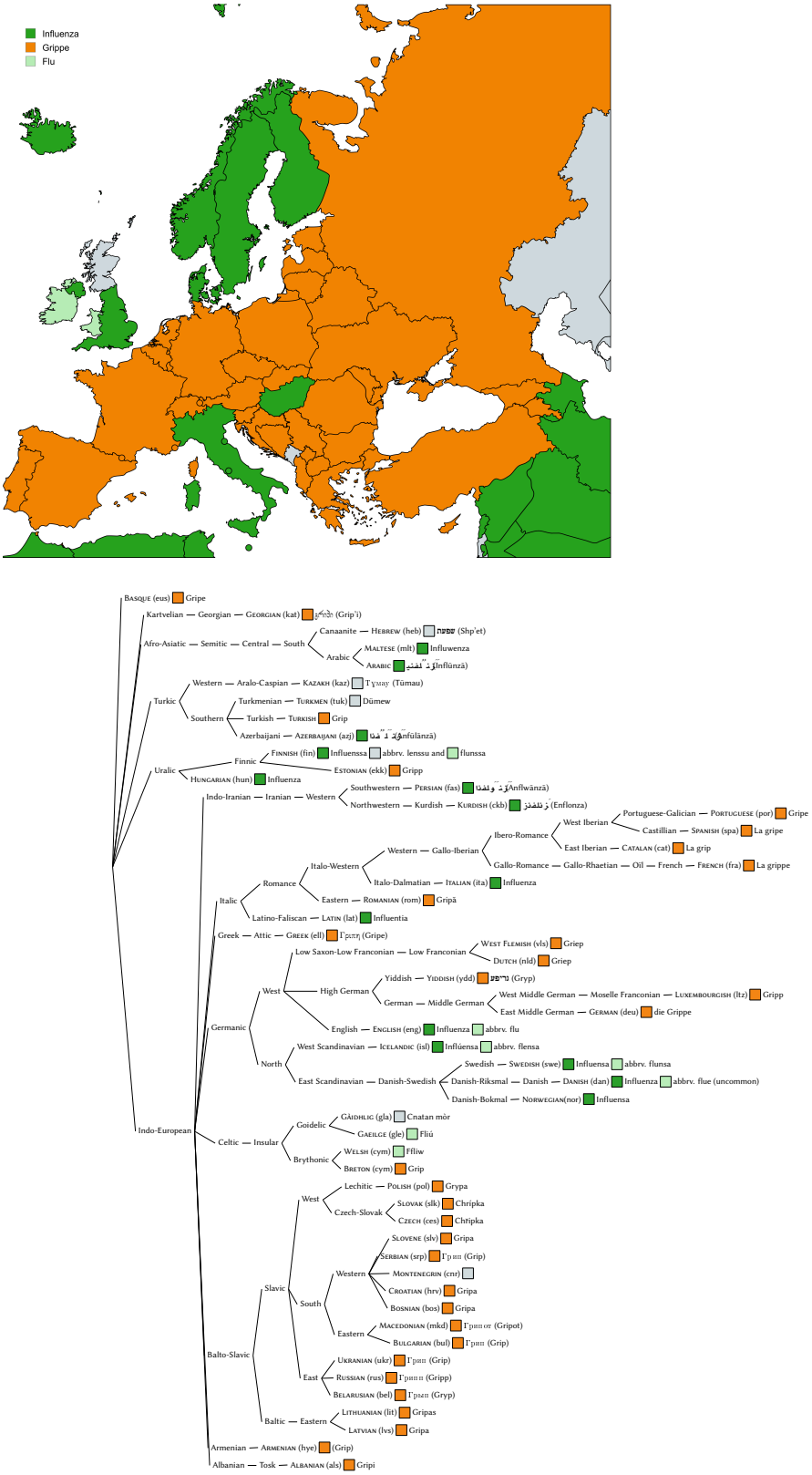
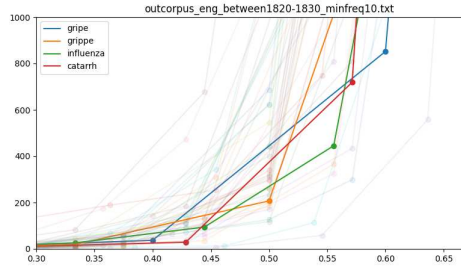
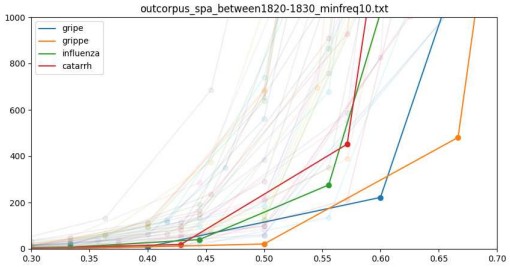


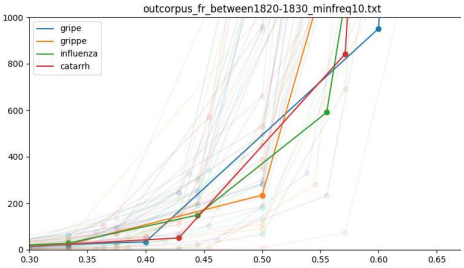
Figure 2: European languages distinguished by the term used for influenza by location (above) and language tree branch (below). Green denotes "influenza" and orange denotes "grippe". Light green denotes "flu". Grey is neither term or undetermined. Figure source: 10.6084/m9.figshare.27210597



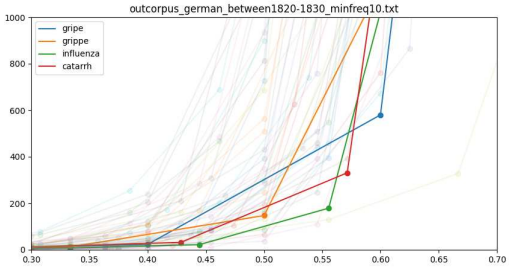
(a) English words



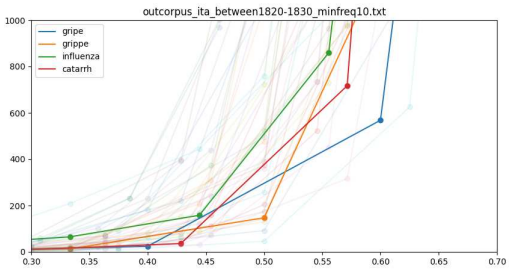
(b) Spanish words



(c) French words



(d) German words



(e) Italian words

Figure 3: The number of words found in each dictionary when a given percentage of the word was altered through insertion, deletion or substitution. Green denotes "influenza", orange denotes "grippe", blue denotes "gripe", and red denotes "catarrh" (the control) using the words outlined in Table 2. Grey curves denote randomly selected words. In these plots we use the words that appeared between 1820 and 1830.