Pre-processing Large Resources for Family Names Research

Adam Rambousek

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
rambousek@fi.muni.cz

Abstract. This paper describes methodology and tools used to pre-process historical archive documents in various formats and their conversion to unified format. Resources were used to investigate the origins and geographical distribution of surnames in the United Kingdom, as part of the Family Names in Britain and Ireland research project. Data extracted from the documents and their connection proved to be valuable research resource which helped to speed up the lexicographic work.

Key words: DEB platform, lexicography, big data, family names, data conversion

1 Introduction

Family Names in Britain and Ireland (FaNBI) [1] is the research project of the University of the West of England that started in 2010, the first phase finished successfully in May 2014 and the research was extended to 2016.

FaNBI aims to complete a detailed investigation of the origins, history, and geographical distribution of the 45,000 most frequent surnames in the United Kingdom. The DEB platform [2,3], developed at the NLP Centre FI MU, was selected by the University of the West of England (UWE) as the dictionary writing system for the project because of its versatility and possibility to combine various resources. This paper describes the tools and methods used to pre-process various resources needed for the research.

2 List of names and frequencies

The list of frequency for each family name is the cornerstone of the FaNBI project. It is not only the list of entries to edit, but the frequency also decides which names will be edited in each phase of the project. In the first phase, all names with more than 100 bearers were edited. The work was extended to all names with more than 20 bearers in the second phase.

At the beginning of the project, two lists were used – 1881 census report [4] and 1997 statistical data [5]. However, both lists had to be preprocessed and

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 105–109, 2016.

filtered, because they contained a lot of noise and errors (for example, spelling errors or invalid characters). Another issue with the lists provided was that all the names were written in uppercase. A straightforward solution is to leave the first letter of each word uppercase and the rest in lowercase, however, this is not true for all names. For example, Scottish and Irish names like O'Brian or McGaffin had to be considered. This type of names also produced the issue with various written spellings used. For the Mc- names, three different variant spellings were present – Mac-, Mc- and M'-. Similarly for O'- names, various apostrophe characters were used and sometimes the name was written without the apostrophe. It was decided to include only the spellings Mc- and O'- into the dictionary, and redirect readers searching for other variants to the correct dictionary entry. To make the matter even more complicated, some family names starting with the string Mac- are separate names and not the variant spellings of Mc-, for example Mach or Mackarel. To solve this issue, the list was edited in two steps. In the first step, variant spellings were detected and uncertain samples were reported. In the next step, the proposed changes were approved by the lexicographers. In case of variant spellings, the frequencies had to be summed for all the forms.

The 1881 census list was edited with the described method, and the method was updated for other lists. Lexicographers' approval was not needed anymore, because the 1881 list was included in the cleaning tool to decide the correct spellings and variant combination. Finally, only names with frequency of at least 20 were included. During the cleaning and combining of the 1881 census list, the number of records was reduced from 469,356 to 373,319 records.

With the report from recent years, the issue linked with growing immigration was discovered – both masculine and feminine forms of the names appeared for languages where these forms differ (for example, Polish or Czech). It was decided to keep only the masculine form of the family name, and the method for frequency inclusion was updated. Feminine forms are detected by the known suffix and if the masculine form is present in the database, the frequency numbers are combined.

3 Combining resources

In the aim to include as much historical evidence as possible, various existing databases are used to search for the records of the family names. A selection of records is available as a webservice from The National Archive[1], however it was needed to clean or preprocess the resources.

Very valuable resource for family names studies is the *International Genealogical Index* (IGI) [6] compiled by The Church of Jesus Christ of Latter-day Saints. The IGI contains worldwide records extracted from the parish archives and similar sources, or submitted by the members of the Church. IGI records are published on the FamilySearch website[2], however the website does not pro-

[1] `http://discovery.nationalarchives.gov.uk/`

[2] `https://familysearch.org/`

vided access to the complete collection and records may contain various errors or inconsistencies. The original database records for the Great Britain were provided to the FaNBI project. The database was transcribed from the parish archives by volunteers over the course of several decades. Because of many reasons (for example, unreadable books, different spellings by each transcriber, spelling mistakes etc.) the database had to be cleaned up before it could be included in the FaNBI research. Sometimes, several volunteers transcribed the same parish records, so the duplicate data had to be detected. The following list sums the process of the cleaning and deduplicating the IGI database.

- Original database contained 188,043,185 records. Each record contains information about the event type (birth, christening, marriage, or death), first name, surname, date, location (county, town/place name, sometimes the exact parish), and the role of the person (e.g. for marriage bride, groom, or their parents).
- Obvious mistakes were deleted, for example records claiming that the English cities are in France.
- Names of the counties were standardized from variant spellings and abbreviations.
- For each county, a list of place names was extracted. These lists were distributed amongst the volunteers from the Guild of One-Name Studies[3]. Volunteers checked if the place name on the list belongs to the given county, or provided correct spelling. As a result of this process, a standardized list of place names was created and the database records were fixed. The records that provided incorrect information about the place name were deleted.
- In the next step, duplicate records were deleted. Because the main aim for the FaNBI research was not to build complete and perfect database, but provide reliable evidence, it was possible to delete not just exact duplicates, but also suspect duplicates. The rules for duplicate detection were considering following information from the records: first name, surname, date, town, county, and event type. Records were flagged as duplicate when all information were identical, but one of the following fields was different: first name, town, county, or event type.
- At the end of the process, IGI database contained 72,187,630 records.

Subsequently, the database was used to automatically add historical evidence to the FaNBI dictionary. For each family name, IGI records were extracted for each century and most prominent county, formatted according to the reference templates and saved in the entry. 40,274 family names entries were automatically enhanced with the IGI evidence. Apart from the enhancement of the dictionary, the processed IGI database is regularly consulted by the researchers as a valuable resource. For the sample of original IGI record and converted form to include as the historical evidence see Table 1.

[3] `http://one-name.org/`

Table 1: Original record from the IGI database and form included into FaNBI.

Original record (batch identification, event date, event place, event type, year, first name, surname, role, gender):
Bletsoe, Bedford, England | 05 Sep 1629 | Bletsoe, Bedford, England | Christening | 1629 | John | Darter | Principal's Father | Male
Converted record:
John <sn>Darter</sn>, 1629 in <src>IGI</src> (Bletsoe, Beds)

Another archive resource that required preprocessing were three volumes of *The Irish Fiants of the Tudor sovereigns during the reigns of Henry VIII, Edward VI, Philip & Mary, and Elizabeth I* [7]. The Fiants contain various court warrants and are available in the electronic format. Each record is clearly marked in the text and thanks to the official language, it is possible to detect persons' names, occupations, or residence. In the first step, the Word documents (results of the OCR recognition) were converted to the XML format. Each court record was converted into a separate XML entry with enhanced metadata. For example, the date of the record was converted from the regnal years system to calendar years.

Converted XML documents were later processed by the extraction tool. The tool standardized common OCR misspellings and detected frequently repeating text patterns in the warrant texts. The list of place names (created during the IGI database cleanup) was used to detect town names and match them with the correct county. Where available, also the persons' occupations were tagged in the record. Finally, all the information were formatted according to the FaNBI reference templates and are available for reference in appropriate entries. For the sample of the conversion from Fiants to FaNBI, see Table 2.

Table 2: Original Fiants record and form converted to include in FaNBI.

Original record:
1431. Pardon to Thomas Dowdall, of Dermondston, county Dublin, husbandman.—2 November, xi.
Converted record:
Thomas <sn>Dowdall</sn>, 1569 in <src>Fiants Eliz</src> $1431 (Dermondston, co. Dublin)

4 Conclusions

We have presented methodology to extract valuable information from various resources, unify the data from several documents, and combine the data for lexicographic research. Dictionary based on the results of the FaNBI projects

are scheduled for publication by the Oxford University Press on November 17, 2016 [4].

Combining various historical documents for a single family name into one dictionary entry helped to speed up the research and discover new connections in the data. Researchers and general public users also have the possibility to view much richer information in one place.

Proven methodology and tools from the FaNBI were later adapted for the creation of the Dictionary of American Family Names (2nd edition), starting in 2014 and aimed to be published by the Oxford University Press in 2017.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the national COST-CZ project LD15066.

References

1. Hanks, P., Coates, R., McClure, P.: Methods for Studying the Origins and History of Family Names in Britain. In: Facts and Findings on Personal Names: Some European Examples, Uppsala, Acta Academiae Regiae Scientiarum Upsaliensis (2011) 37–58
2. Rambousek, A., Horák, A.: DEBWrite: Free Customizable Web-based Dictionary Writing System. In Kosem, I., Jakubíček, M., Kallas, J., Krek, S., eds.: Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. (2015) 443–451
3. Horák, A., Rambousek, A.: Lexicographic tools to build new encyclopaedia of the czech. The Prague Bulletin of Mathematical Linguistics (106) (2016) 205–213
4. The National Archives: Census records, 1881 http://www.nationalarchives.gov.uk/records/census-records.htm.
5. Hanks, P., Coates, R.: Onomastic lexicography. In Fjeld, R.V., Torjusen, J.M., eds.: Proceedings of the 15th EURALEX International Congress, Oslo,Norway, Department of Linguistics and Scandinavian Studies, University of Oslo (aug 2012) 811–815
6. The Church of Jesus Christ of Latter-Day Saints: International Genealogical Index https://www.familysearch.org/search/collection/igi.
7. Nicholls, K., ed.: The Irish Fiants of the Tudor Sovereigns during the Reigns of Henry VIII, Edward VI, Philip and Mary, and Elizabeth I. Edmund Burke Publisher (1994)

[4] https://global.oup.com/academic/product/the-oxford-dictionary-of-family-names-in-britain-and-ireland-9780199677764