To Honor our Heroes: Analysis of the Obituaries of Australians Killed in Action in WWI and WWII

Marc Cheong
Centre for AI and Digital Ethics
University of Melbourne
Parkville, Australia
marc.cheong@unimelb.edu.au

Mark Alfano
Philosophy, Macquarie University
Macquarie University
Sydney, Australia
mark.alfano@gmail.com

***Abstract*— Obituaries represent a prominent way of expressing the human universal of grief. According to philosophers, obituaries are a ritualized way of evaluating both individuals who have passed away and the communities that helped to shape them. The basic idea is that you can tell what it takes to count as a good person of a particular type in a particular community by seeing how persons of that type are described and celebrated in their obituaries. Obituaries of those killed in conflict, in particular, are rich repositories of communal values, as they reflect the values and virtues that are admired and respected in individuals who are considered to be heroes in their communities. In this paper, we use natural language processing techniques to map the patterns of values and virtues attributed to Australian military personnel who were killed in action during World War I and World War II. Doing so reveals several clusters of values and virtues that tend to be attributed together. In addition, we use named entity recognition and geotagging the track the movements of these soldiers to various theatres of the wars, including North Africa, Europe, and the Pacific.**

Keywords*— *Applications of document analysis; Historical document analysis; Human behavior analysis*; *obituary, virtue, value, natural language processing, wellbeing.

I. Introduction

Obituaries represent a prominent way of expressing the human universal of grief [1], [2]. According to philosophers, obituaries are a ritualized way of evaluating both individuals who have passed away and the communities that helped to shape them [3]. The basic idea is that you can tell what it takes to count as a good person of a particular type in a particular community by seeing how persons of that type are described in their obituaries (since 'one must not speak ill of the dead').

Obituaries of those killed in conflict, in particular, have been rich repositories of communal values, as they reflect the values and virtues that are admired and respected in individuals who are considered to be heroes in their communities.

In 19th- and 20th-century wartime, before the invention of the Internet, obituaries in newspapers have been used to motivate military perseverance, and also to provide updates to waiting families on the home front. Newspaper obituaries contain valuable metadata such as demographic traits, movement patterns, and key dates [4], as well as *values, virtues, and constituents of wellbeing* (*VVCs*, for short) [5]. Many online repositories of digitized obituaries exist, such as Trove by the National Library of Australia[1]; and Legacy.com, a large "commercial provider of online memorials"[2] in the USA.

Academic work studying obituaries is cross-disciplinary, ranging from philosophy and digital humanities studying VVCs [6] to public epidemiology [7] and various studies in the computer sciences on digitisation and natural language processing of obituaries [4], [8].

In this paper, we combine common data-mining and natural language processing (NLP) methods with experimental philosophy techniques. Experimental philosophy is an emerging field that employs empirical methods from the social and computational sciences to address philosophical questions and problems; to date, it has mostly drawn on methods from psychology, but as it matures, it has begun to employ a wider range of methods [9].

In the study of digitized obituaries in Australia during periods of war in the 20th century, our research questions are as follows:

RQ1. Using Named-Entity Recognition, what kind of information is made available in wartime newspaper obituaries that could support the identification of places and people that were important during these wars?

RQ2. What kind of VVCs correspond to the adjectives and nouns used to describe decedents' character? In other words, if one tends to get described as X, are they more likely to be also described as Y and Z?

RQ3. Using semi-structured data, how do we map spatial and temporal migration patterns of Australians involved in the war effort?

Our paper is organized as follows. In Section II, we review literature on the theory and applications of obituary analyses. Section III discusses the methodology involved, including

[1] <https://trove.nla.gov.au/>
[2] Statistics as reported in 2009 by USA Today. <https://usatoday30.usatoday.com/news/military/2009-05-21-memorialday_N.htm>

Manuscript 1933 submitted to 2020 25th International Conference on Pattern Recognition (ICPR). Received July 9, 2020.

crucial assumptions about the obituary data supplied, as well as empirical and computational techniques used and challenges faced. Section IV presents our findings with a discussion of potential future work.

II. Related Work

As mentioned in Section I, digitized obituaries can be found in databases online. However, some obituaries are provided as digital facsimiles of actual newspapers. Hence, scholarly work deals with the process of converting the facsimiles into plain text or structured records to enable processing with NLP techniques. Several studies deal with the quality of these digitisation efforts, with emphasis on optical character recognition (OCR) [10] and strategies for improving quality for large-scale digitisation [11].

Post-digitisation, an obituary contains an amalgam of information that could benefit from segmentation. Sabbatino, Bostan, and Klinger [4] have shown that research into obituaries mostly focuses on a particular sub-part of the obituary. Hence, they propose "zoning as a preprocessing step" to distinguish between "personal information… Biographical sketch ... Tribute, Family, and Funeral Information" [4], as there has been a lack of research into this area.

Given the actual text contents of the obituaries, various staple NLP techniques and tools can be applied to study any of the above facets in [4]. A thorough literature survey can be found in [4] and [12]; only a limited selection of examples most relevant to the current paper will be covered below due to space constraints.

One example is the application of sentiment analysis and statistical methods to identify "[t]ones, emotions, terms used to describe death" in obituaries of people who have died due to overdoses [13]. The authors analyze obituaries for key themes using an *a priori* list of themes and seed words. Wildcard searches are then performed to identify the existence of seed words and their synonyms (e.g. "shame: stigma, embarrassment, disgrace" [13]), while sentiment analysis is performed using the IBM Watson Tone Analyzer package [13].

The idea of analyzing individuals' demographic properties -- such as ethnicity, age, nationality, and socioeconomic status -- predates modern data-mining packages [14]-[15]. This research angle has been ported to large-scale web crawling and automated analysis exercises, such as [7], which successfully crawled over 79,394 obituaries. The authors introduced simple yet effective text-mining rules for detecting gender, age, offspring, and types of cancer reported (for studying its epidemiology) [7].

In terms of methodology for investigating VVCs, we turn to [16]. Citing [17], it outlines the precedent of using a "lexical approach" -- in particular, statistical methods on text and NLP -- to guide philosophical investigation. By extracting "agent-level descriptions of the deceased… [and g]eneral categories of traits" [16], the authors first performed manual extraction of such terms and constructed a co-occurrence network visualization to map out most frequent descriptors and their inter-relationships[i] with other descriptors. A follow-up analysis was conducted using semi-automated identification of plausible VVC terms (with experimenter consensus) as well as part-of-speech (POS) tagging [16].

For comprehensiveness, we have also identified more recent studies that utilize modern neural-network architecture. Of interest is the extraction of kinship information using a neural-network based named-entity recognition and relational extraction model [8] that could prove beneficial for, say, genealogical analysis and social network mapping tasks. This paper ultimately provides hints for future researchers on the value of analyzing sentence-level context as well as kinship information implicitly contained within even a short textual sample.

III. Methodology

A. Data Collection

Per Section I, the National Library of Australia (NLA) has a service called Trove, which "brings together content from libraries, museums, archives and other research organisations" and provides "a repository of fulltext digital resources … [and] metadata" [18]. Trove provides digitized versions of newspaper obituaries, which consists of the image (facsimile) of the original newspaper, metadata (such as publication date and source name), and the full text extracted from the facsimile using OCR methods.[3] This is made available to researchers with an account via the Trove API.[4]

We aim to download obituaries of members of the Australian armed forces who were killed in conflict; hence we search the API for newspaper articles containing the phrase "*killed in action*" (in quotes) within the 'Family Notices' category under 'Newspapers'[5]. These newspapers range from large companies (some of which are still in circulation, at time of writing) to small regional titles that have long been defunct. The articles range from the years 1844 to 1994, with visible spikes in the frequency during the two World Wars, as expected. Articles during non-World War periods are negligible (fewer than a hundred per annum), and are thus ignored in our analysis. We also note the lack of contemporary articles on post-World War II conflict, e.g. the Vietnam War and the Gulf War; which is likely due to the unavailability of the bulk of full text articles for copyright restrictions/limits as well as winding down of operations for defunct newspapers.

As the focus of our investigation is on the two major World Wars, we narrow the range of our papers to the year

[3] <https://help.nla.gov.au/trove/for-digitisation-partners/optical-character-recognition-ocr-newspapers>
[4] <https://help.nla.gov.au/trove/building-with-trove/api-version-2-technical-guide >
[5] As of time of writing, the V2 API is the only one available in Trove. It supersedes the V1 API which returns significantly less results, and uses the 'Obituaries' category instead.

each World War began, up to and including two years following the official end of each period of war.

- World War I (WWI): 1914-1920 inclusive.
- World War II (WWII): 1939-1947 inclusive.

The articles within these time periods were then converted from Trove's structured XML format to portable flat files for analysis in Python, Tableau, and R.

B. Preprocessing and Data Quality

We intend to examine individual obituaries, in line with earlier work done in the domain [7][16]. However, the collected articles include *'rolls of honour'* (lists of casualties) that are collections of many individual entries, and *'in memoriam'* posts. These articles have two issues: (i) they are lists of names, and thus no content regarding VVCs is visible; and (ii) *in memoriams* are sometimes posted years after the actual funeral. Hence, we excluded articles containing any of the following strings:

> ['*casualty list', 'australian casualties', 'british casualties', 'roll of honour', 'roll of honor', 'honor roll', 'honour roll', 'in memoriam'*]

In the methodology of earlier papers such as [16], demographic features (gender, age, location of birth, location of death, educational attainment, veteran status, a few others) as well as all the adjectives and nouns used to describe the decedent's character (e.g., *sweet, kind, brave, larrikin,* etc.) were coded by hand. This introduces some degree of tolerance for typographical errors. By contrast, automated methods such as those in [7] will be less tolerant of any form of typographic error.

As the data is obtained from OCR results (and crowd-sourced corrections, if available) from the National Library of Australia's Trove project, we cannot guarantee that the digitized text is 100% correct [11]. Furthermore, extant research on OCR for Australian archivists deals with accuracy on a character-level, which "varied from 71% to 98.02%" [11]. An "average OCR accuracy" is determined to be at least 90%; however there is no "...consensus on whether these percentages referred to character or word confidences, and whether this was at page or article level" [11], which further complicates the setting of a baseline. Another report coordinated by the National Library of the Netherlands reports that the statistics from controlled experiments with two popular OCR packages are rather grim: word-level OCR accuracy has a range of 39.01% to 78.05% [19].

Hence, we introduce a simple heuristic to determine the quality of an obituary before we process it further. To that end, we use the open source *hunspell* [20] spellchecker package[6], which is currently in use in many popular proprietary and open source products. Each word in an obituary is checked by *hunspell* against the standard Australian English dictionary (*en_AU*). An *accuracy* heuristic for a single obituary is defined as:

> *AccuracyHeuristic = number of correctly-spelled words per obituary / total number of words within the obituary.*

As the accuracy range in the context of Australian digitization projects [11] is reported to be between "71% to 98.02%", we take the midpoint of the two values (84.5%) as the minimum threshold needed for our accuracy heuristic. Any obituaries lower than the accuracy heuristic are discarded; and the final corpus of filtered obituaries (*FilteredObits*) is then used to answer **RQ1-RQ3**.

C. NLP-based VVC Identification and Semantic Web Generation

To answer our first two research questions, **RQ1** and **RQ2**, we first employ common NLP techniques to automate the identification of VVC terms. In Section II, we have seen several automated and semi-automated techniques applied to obituary text, ranging from a pre-registered list of terms that are then expanded with text searches [13] to manual coding followed by POS tagging [16].

Here, we combine the best of both approaches, by first identifying adjectives and adverbs [21] in *FilteredObits*. To do this, we use both the NLTK [22] and spaCy [23] libraries, which are widely used in NLP research, to extract adjectival words from text corpora (see also [24]). The tokens in each obituary are POS-tagged, preserving their original case; as case can convey additional context.

- spaCy (version 2.2.3) was used to perform POS tagging (with the *en_core_web_sm* model[7]), to identify words tagged with '*ADJ*' and '*ADV*' (adjectives and adverbs, respectively).
- NLTK (version 3.4.5) was similarly used for POS tagging, to pick out adjectival words tagged *'JJ'*, *'JJR'*, or *'JJS'*.

The extracted list of adjectives was then read by both authors to determine which adjectives can be used as VVC descriptors for **RQ1** [16]. To streamline the process by merging overlapping terms (e.g. *brave*, *bravest*), we use the *hunspell* [20] stemming function to reduce each word to its dictionary stem; any duplicates were thus combined into a single stem. After that, the authors went through the list of 839 words and rated words as either 2 (refers to military rank, e.g., *'sergeant'*), 1 (refers to VVC) or 0 (doesn't). Inter-rater reliability was calculated using the *irr* package in R [25]. The raters agreed on 712 out of 839 words, indicating that reliability was acceptable (Cohen's kappa = .657, z = 19.5, p < .0001 -- see Landis & Koch [26] for discussion). Consensus was then reached on terms that the reviewers disagreed about, through discussion and reference to example texts, to finalize the VVC descriptors (138 total terms).

An adjacency matrix was constructed on the *FilteredObits* corpus to identify which VVC descriptors occur in each individual obituary. In this matrix, each row represents an obituary, and each column represents a VVC term. A value of 1 is placed in a cell if that VVC term in question occurs in the

[6] Using the *pyhunspell* bindings for Python <https://pypi.org/project/hunspell/>

[7] <https://spacy.io/models/en>

obituary, and a 0 is placed otherwise. We then constructed a co-occurrence matrix by multiplying the adjacency matrix by its transpose, which produces a matrix that represents all pairwise overlaps. A partial representation is included in Fig. 1 below.

	dear	kind	beloved	brave	best	true	great	loving	sweet	good	noble	loved
dear	605	124	147	130	134	122	83	88	76	67	51	35
kind	124	183	42	18	32	44	19	26	20	21	8	12
beloved	147	42	195	42	42	41	26	27	24	23	14	13
brave	130	18	42	148	59	46	31	27	17	20	25	13
best	134	32	42	59	154	51	24	34	26	28	22	18
true	122	44	41	46	51	128	21	39	30	32	16	10
great	83	19	26	31	24	21	109	15	12	15	9	5
loving	88	26	27	27	34	39	15	96	20	21	11	11
sweet	76	20	24	17	26	30	12	20	82	16	5	10
good	67	21	23	20	28	32	15	21	16	81	11	6
noble	51	8	14	25	22	16	9	11	5	11	58	7
loved	35	12	13	13	18	10	5	11	10	6	7	47

Fig.1. Extract from the co-occurrence matrix of 138 total VVC terms.

The diagonal of this table represents the frequency of each term. The remaining cells represent co-occurrence patterns (note that the top-right half of the table perfectly mirrors the bottom-left).

We then read this co-occurrence matrix into Gephi for network analysis using modularity detection, weighted degree, and visualization using the ForceAtlas layout (Fig. 2).

D. Migration and Journey Mapping

Another outcome we are interested in is the creation of a geographical journey map to illustrate the typical geographic migration patterns of Australian service personnel who were killed in action during World Wars I and II. These pieces of information are contained within an obituary's 'biographical sketch' and 'tribute' sections [4].

We use the geographic information -- country, town/city, state, and other place names -- within an obituary to visualize the movement patterns of an individual on a geographic map. This includes the rank of a particular service member, along with the list of countries in which he/she has travelled through in the course of duty, starting from Australia.

First, we create a frequency table of locations found within all obituaries using spaCy's [23] Named Entity Recognition (NER) annotation. Locations are nouns tagged with either *'GPE'* (for *Geo-Political Entity*) or *'LOC'* (for *'LOCation'*)[8].

As a first pass, the locations are then geocoded[9] using the GeoNames API [27]. The main objective of this first pass is to determine the country (standardized to ISO-3166 two-letter codes) in which a certain *GPE* or *LOC* entity is contained.

However, geolocation results often require manual correction (e.g. [28], [29]). We identified several issues in the context of historical WWI and WWII obituaries, due to issues such as popular place names which are similar but in different countries, and the changing names of nation-states. Hence, the extracted list of locations was manually checked for accuracy. To this end, we applied the following heuristics:

- Mistaken proper nouns: locations that are part of a noun phrase but are not locations per se. Examples: *Empire* (part of a noun phrase e.g. *British Empire*); *ANZAC* (a term for the Australia and New Zealand Army Corps).
- Mistaken common names and surnames: these include geocoding of names to obscure places. Examples: *Henry* (given name, incorrectly coded as a place in the Philippines); *McKnight* (surname incorrectly coded to Canada).
- Ambiguous and incorrectly-coded locations: cases in this category include *Richmond* (present in both the US and Australia); *Vic.* (an abbreviation of the Australian state of Victoria). In Australian obituaries, they are likely to be a part of Australia -- used in the reporting of birthplaces and hometowns of personnel -- and are treated as such. The exceptions are common locations in other parts of the world which are not likely in Australia, such as *Virginia* (USA) and *York* (UK).
- Documented sites of battle: certain documented sites and theatres of battle are manually corrected. Examples: *Gallipoli* (the Gallipoli Peninsula is actually in Turkey); *Bardia* (should actually refer to Bardia in present day Libya, based on the Battle of Tobruk); the generic *Middle East* theatre (approximated to be the center of Iraq, based on war maps). The Australian War Memorial website [30] is consulted to identify these cases.
- Minor typos: minor single-character typos due to OCR which are easily distinguishable as common place names are manually corrected. Examples: *Builecourt* (Bullecourt); *Belguim* (Belgium).

To determine the rank of an individual, we use a simple rule-based rank classifier, using the list of official military ranks provided by the Australian War Memorial [31]. The most senior rank present in an obituary (e.g. *Sergeant*, if both the words '*Sgt*' and '*Corporal*' are present within) will be used to tag the particular path on the map. For each obituary, its date, the migration path information (consisting of two or more country codes), and rank are used in the visualization to address **RQ3**.

IV. Results and Discussion

A. Data Preprocessing

As of the time of writing, we are able to obtain 96,096 total obituary records that contain digitized text, from 623 newspapers, using the methods outlined in Section III.A. Out of the total, 54,707 obituaries were from the World War periods (World War I with 35,673 + World War II with 19,034 respectively).

We then applied the following heuristics in succession to narrow down obituaries of sufficient accuracy and quality, as well as to negate the effects of false positives.

- Article year range: World War I and World War II only (35,673+19,034 respectively). Filtered total = 54,707.

[8] <https://spacy.io/api/annotation#named-entities>
[9] Using the Python *geocoder* library <https://pypi.org/project/geocoder/>

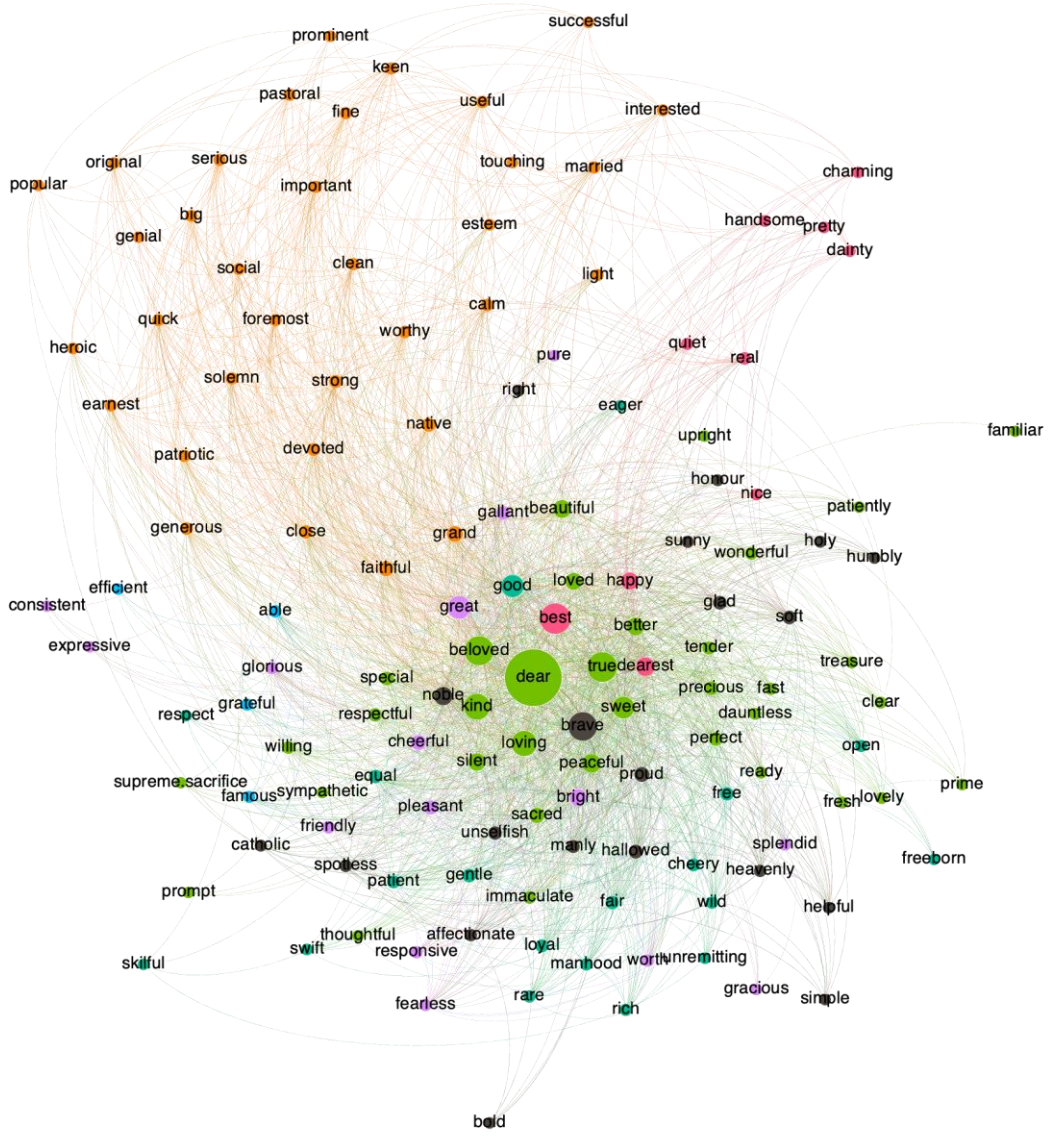

Fig.1. Semantic network of the virtues, values, and constituents of wellbeing attributed to Australian service personnel killed in action in WWI and WWII. Node size = weighted degree, node color = modularity class.

- Removing *'rolls of honor'* (lists of casualties) and *'in memoriam'* posts. Filtered total[10] = 10,914.
- Removing obituaries with *AccuracyHeuristic* of less than 84.5. Final total = 923**.**

We henceforth refer to the final corpus of 923 obituaries as *FilteredObits*.

[10] This includes a manual removal of four obituaries, as their lengths are close to 9,000 words which is atypical of an obituary. We opine that these are incorrectly segmented out of an entire page of obituaries.

B. *Semantic Web of VVC*

To answer **RQ1**, we obtained a grand total of 963 unique adjectives and adverbs (Section III.C) from *FilteredObits*. Following stemming (which yielded 839 adjective stems), the two authors independently performed manual coding to categorize each as either not a VVC descriptor for the deceased (*0*), a suitable descriptor (*1*), or a mention of military rank (*2*). Any disagreements were discussed, and the final list of adjective stems are then obtained for the next step.

The final list of words coded as *1*s are used to generate an adjacency matrix of occurrences in each of the 927 obituaries, resulting in a binary matrix of 927 rows and 138 columns (following the methodology used in [15]).

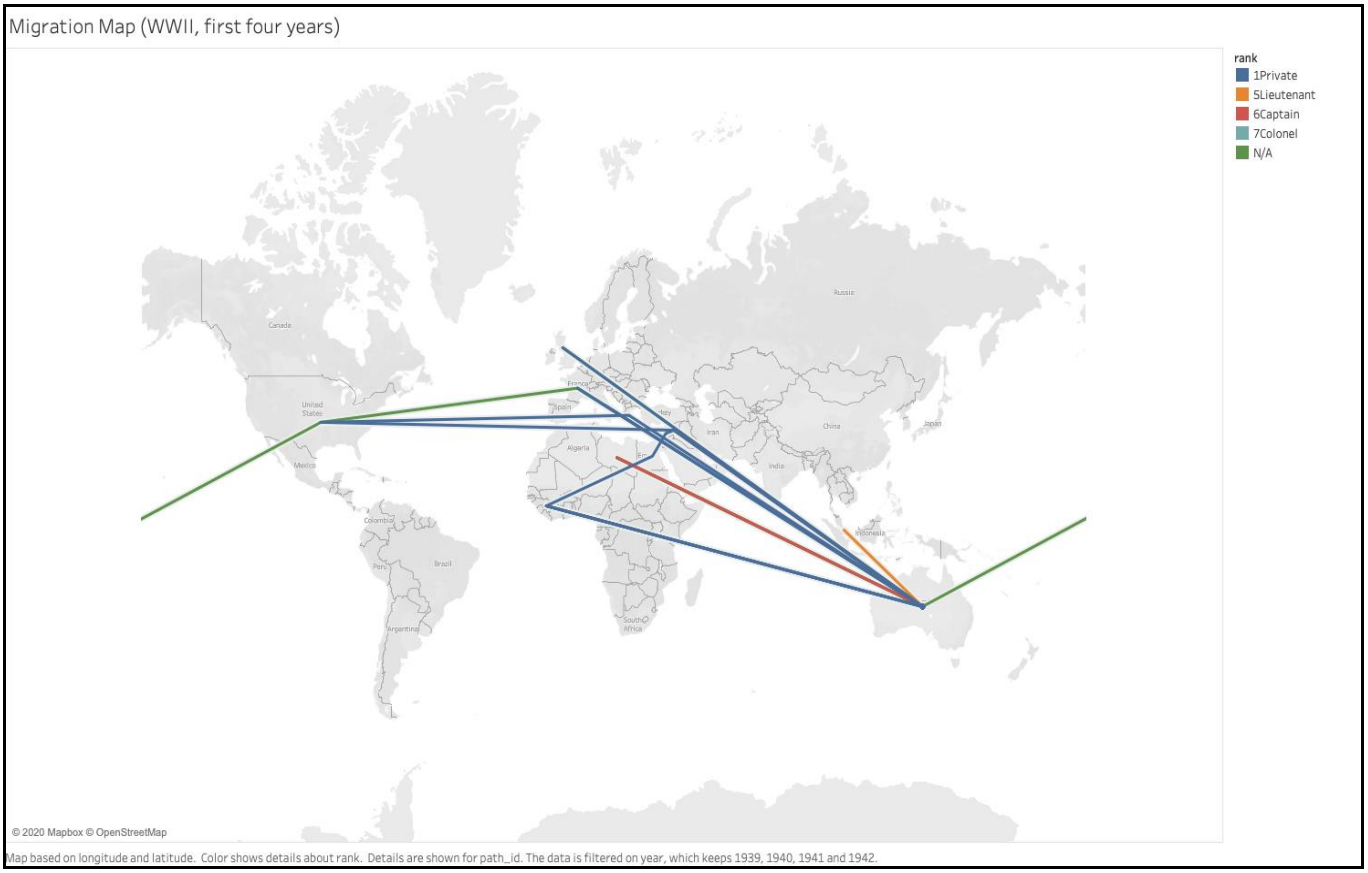

Fig. 3. Visualization of Australian Service Personnel Migration from the FilteredObits corpus (WWII, 1939-1942 inclusive).

Fig. 2 illustrates the generated semantic web of VVCs, with different modularity classes shown in separate colors. The central cluster, in green, primarily includes terms referring to intimate relationships with, e.g., family members. Next, in light blue on the left-hand side of the image, we see a range of terms referring to competences. In pink at the center and bottom of the image is a cluster that emphasizes positive attitudes. There are two clusters that refer to manly or masculine virtues: one in gray in the center and middle of the figure, the other in green at the bottom of the figure. Finally, there is a cluster that focuses on aesthetics in magenta at the top right and a cluster that focuses on loyalty in orange at the top of the figure.

C. *Migration Map*

In answering RQ3, we generate a journey map to indicate the movement patterns of service members killed in action in the *FilteredObits corpus*. By applying the semi-automated NER and geotagging process in Section III.D, we obtain an overall list of 952 unique place names found. By applying the list of heuristics in the same section, all 952 places were checked for accuracy.

In our proof of concept, we use Google's public dataset of country coordinates [32] to provide the latitude and longitude for plotting the country movement/migration paths in Tableau.

Fig. 3 depicts the dashboard generated, illustrating the example of the first four years of World War II (1939-1942 inclusive). The example illustrates that a majority of service personnel moved from Australia to European and North African theatres of war, based on the occurrences of place names and location identifiers in the obituary. The majority of obituaries found during this time period are of soldiers with the rank of Private. Also note that there are paths from Australia (or an intermediate country) to the United Kingdom.

This is due to the mention of the UK (e.g. *Britain*) within the obituary, symbolizing a patriotic connection to the Queen (who is the head of state of Australia).

V. Future Work and Conclusion

This paper illustrates how automated NLP methods, in conjunction with semi-automated heuristics and coding to bootstrap and/or fix data quality issues, can be used in digital humanities investigations, with attention to experimental philosophy.

Through the course of this paper, the authors have identified several areas of improvement that may lead to future work.

Firstly, data quality remains a main issue, which causes problems in many NLP pipelines, requiring "tools for NE or POS [to be]... either retrained or adapted" [33]–[35]. As seen in Section II, modern OCR techniques still generate non-negligible amounts of typographical errors (spelling mistakes) and misinterpretation of characters (e.g. an uppercase '*I*' can be interpreted as a lowercase *L*, the number one, or a vertical bar character). As of the time of writing, Trove provides a crowdsourced facility[11] for human users to correct digitized text in a manner. Therefore, there is a tradeoff between accuracy (i.e. higher *AccuracyHeuristic*) and sample size: adding more obituaries to our corpus will result in more terms captured (both adjectives and locations) at the expense of accuracy. Future work in this area involves improvements and automatic correction of OCR texts, as well as automated segmentation techniques to make use of honor rolls which have been excluded from our sample.

With the above in mind, our analysis focused on one-gram (single word) VVC descriptors, and does not take into consideration bigrams (or higher n-grams), such as '*many friends*', '*no fear*', or '*home associations*'. In our preliminary investigations into Trove data, we find that these are found in obituaries with *AccuracyHeuristics* higher than the one we have used (some with 90% or higher). Future work can be done on bigrams or more.

Other directions for future work include the segmentation of the obituary corpora along pre-defined, theoretically-motivated lines. For instance, it would be interesting to see whether the VVCs celebrated in obituaries from the WWI period differ in important ways from those in the WWII period, whether the obituaries of service members at different military ranks differed in important ways, and whether obituaries of service members who took different paths through the map differ in important ways. We also note that these corpora completely ignore the roles played by women and children during the World Wars. Supplementing our corpora with texts that talk about women and children would be valuable both in its own right and in terms of inclusivity.

In terms of visualization of migration patterns, the quality of the location-coding depends on the geocoding service provider [28], [29]. Certain heuristics have to be applied (Section III.D) to automatically correct for incorrect geocoding, or results which do not reflect that of the sociohistorical context (e.g. place names that can be coded as both in the UK and Australia are likely to be Australian, given the sources of the obituaries and the hometowns of soldiers). Nonetheless, a degree of manual checking and cross-referencing with authoritative sources [30] need to be performed. Future work in this area includes the potential for curation of an authoritative list of areas mentioned in wartime reports, possibly with crowdsourcing platforms such as Mechanical Turk and Prolific.

In conclusion, we have presented a novel approach to obituary analysis in terms of VVC descriptors that allows all terms in the corpus to count as potential VVC descriptors rather than working from a pre-loaded dictionary of VVC terms. In so doing, we find that a wide range of descriptors tends to be used to characterize and laud service members who were killed in action. These semantic clusters in turn serve as objects of reflection for researchers in the field of experimental philosophy. Much philosophical work employs *a priori* methods to sketch the virtues and values that make for a good life. By contrast, this project uses a more empirical, bottom-up approach that enables ordinary people's intuitions to speak for themselves. While it would take further philosophical work to reach any normative conclusions from this methodology, it remains a useful tool for philosophers to reflect upon and to widen the scope of potential ethical concerns beyond what can be generated *a priori.*

[11] <https://help.nla.gov.au/trove/digitised-newspapers/text-correction-guidelines>

REFERENCES

[1] M. C. Nussbaum, Poetic Justice: The Literary Imagination and Public Life. Beacon Press (MA), 1995.

[2] J. Butler, Precarious Life: The Powers of Mourning and Violence. Verso, 2006.

[3] L. T. Zagzebski, "Virtues of the Mind." 1996, doi: 10.1017/cbo9781139174763.

[4] V. Sabbatino, L. Bostan, and R. Klinger, "Automatic Section Recognition in Obituaries," Valentino Sabbatino and Laura Bostan and Roman Klinger. 2020, Accessed: Jun. 10, 2020. [Online]. Available: https://arxiv.org/abs/2002.12699.

[5] M. Alfano, Moral Psychology: An Introduction. John Wiley & Sons, 2016.

[6] A. Bardi, R. M. Calogero, and B. Mullen, "A new archival approach to the study of values and value--Behavior relations: Validation of the value lexicon," Journal of Applied Psychology, vol. 93, no. 3. pp. 483–497, 2008, doi: 10.1037/0021-9010.93.3.483.

[7] G. Tourassi, H.-J. Yoon, S. Xu, and X. Han, "The utility of web mining for epidemiological research: studying the association between parity and cancer risk," J. Am. Med. Inform. Assoc., vol. 23, no. 3, pp. 588–595, May 2016.

[8] K. He et al., "Extracting Kinship from Obituary to Enhance Electronic Health Records for Genetic Research," Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. 2019, doi: 10.18653/v1/w19-3201.

[9] M. Alfano, D. Loeb, and A. Plakias, "Experimental Moral Philosophy," The Stanford Encyclopedia of Philosophy (Winter 2018 Edition), 2018, Accessed: Jul. 03, 2020. [Online]. Available: https://plato.stanford.edu/entries/experimental-moral/.

[10] T. P. Kimmo Kettunen, "Measuring Lexical Quality of a Historical Finnish Newspaper Collection–Analysis of Garbled OCR Data with Basic Language Technology Tools and Means," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 956–961, Accessed: Jun. 10, 2020. [Online].

[11] R. Holley, "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs," D-Lib Magazine, vol. 15, no. 3/4, Mar. 2009, Accessed: Jun. 10, 2020. [Online]. Available: http://eprints.rclis.org/12908/1/ANDP__How_Good_Can_it_Get.pdf.

[12] V. Sabbatino, "Automatic recognition of structures in obituaries," Bachelorarbeit, University of Stuttgart, 2019.

[13] K. Rajesh, T. J. Crijns, and D. Ring, "Themes in published obituaries of people who have died of opioid overdose," Journal of Addictive Diseases, vol. 37, no. 3–4. pp. 151–156, 2018, doi: 10.1080/10550887.2019.1639485.

[14] D. J. Wright and A. P. Roberts, "Which doctors die first? Analysis of BMJ obituary columns," BMJ, vol. 313, no. 7072, pp. 1581–1582, 1996.

[15] A. Marks and T. Piggee, “Obituary Analysis and Describing a Life Lived: The Impact of Race, Gender, Age, and Economic Status,” Omega , vol. 38, no. 1, pp. 37–57, Feb. 1999.

[16] M. Alfano, A. Higgins, and J. Levernier, “Identifying Virtues and Values Through Obituary Data-Mining,” The Journal of Value Inquiry, vol. 52, no. 1. pp. 59–79, 2018, doi: 10.1007/s10790-017-9602-0.

[17] J. Graham, J. Haidt, and B. A. Nosek, “Liberals and conservatives rely on different sets of moral foundations,” J. Pers. Soc. Psychol., vol. 96, no. 5, pp. 1029–1046, May 2009.

[18] National Library of Australia, “About Trove | Help centre,” Trove. https://help.nla.gov.au/trove/using-trove/getting-to-know-us (accessed Jun. 10, 2020).

[19] M. Heliński, M. Kmieciak, and T. Parkoła, “Report on the comparison of Tesseract and ABBYY FineReader OCR engines,” Poznań Supercomputing and Networking Center, Poland, 2012. Accessed: Jul. 01, 2020. [Online]. Available: https://www.digitisation.eu/fileadmin/Tool_Training_Materials/Abbyy/PSNC_Tesseract-FineReader-report.pdf.

[20] L. Németh, “Hunspell: About,” Hunspell. http://hunspell.github.io/ (accessed Jun. 25, 2020).

[21] P. Bouillon and E. Viegas, “The description of adjectives for natural language processing: Theoretical and applied perspectives,” in Proc. of the TALN’99 workshop on Description des adjectifs pour les traitements informatiques, Cargèse, France, 1999.

[22] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. “O’Reilly Media, Inc.,” 2009.

[23] M. Honnibal and I. Montani, “spaCy · Industrial-strength Natural Language Processing in Python,” spaCy. https://spacy.io/ (accessed Jun. 30, 2020).

[24] A. Okulicz-Kozaryn, “Cluttered writing: adjectives and adverbs in academia,” Scientometrics, vol. 96, no. 3, pp. 679–681, Sep. 2013.

[25] M. Gamer, “irr package | R Documentation,” R Documentation. https://www.rdocumentation.org/packages/irr/versions/0.84.1 (accessed Jul. 09, 2020).

[26] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” Biometrics, vol. 33, no. 1, p. 159, Mar. 1977.

[27] M. Wick and C. Boutreux, “GeoNames.” https://www.geonames.org/ (accessed Jun. 30, 2020).

[28] S. K. Singh and D. Rafiei, “Strategies for Geographical Scoping and Improving a Gazetteer,” in Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18, Lyon, France, 2018, pp. 1663–1672.

[29] D. Ahlers, “Assessment of the accuracy of GeoNames gazetteer data,” in Proceedings of the 7th Workshop on Geographic Information Retrieval - GIR ’13, Orlando, Florida, 2013, pp. 74–81.

[30] The Australian War Memorial, “Home | The Australian War Memorial,” The Australian War Memorial. https://www.awm.gov.au/ (accessed Jun. 30, 2020).

[31] “Rank | The Australian War Memorial,” The Australian War Memorial. https://www.awm.gov.au/learn/understanding-military-structure/rank (accessed Jun. 30, 2020).

[32] Google, “countries.csv | Dataset Publishing Language | Google Developers,” Google Developers, 2012. https://developers.google.com/public-data/docs/canonical/countries_csv (accessed Jul. 01, 2020).

[33] D. Lopresti, “Optical character recognition errors and their effects on natural language processing,” International Journal on Document Analysis and Recognition (IJDAR), vol. 12, no. 3. pp. 141–151, 2009, doi: 10.1007/s10032-009-0094-8.

[34] R. Wudtke, C. Ringlstetter, and K. U. Schulz, “Recognizing garbage in OCR output on historical documents,” in Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data - MOCR_AND ’11, Beijing, China, 2011, p. 1.

[35] M. Généreux and D. Spano, “NLP challenges in dealing with OCR-ed documents of derogated quality,” in Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI 2015, Jul. 2015, Accessed: Jul. 01, 2020. [Online]. Available: https://www.researchgate.net/publication/281112670_NLP_challenges_in_dealing_with_OCR-ed_documents_of_derogated_quality.