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

## Marc Cheong, University of Melbourne Mark Alfano, Macquarie University

International Conference on Pattern Recognition (ICPR202o) Document and Media Analysis Track (PS T4.2, \#1933).

## Obituaries, Virtues and Values

## TROVE

- Obituaries - a ritualized way to express grief
- Obituaries of those killed in conflict reflect values and virtues associated with heroes.
- values, virtues, and constituents of wellbeing (VVC)
- In 19th- and 20th-century wartime newspaper obituaries as metadata.
- Interdisciplinary data-mining/NLP x experimental philosophy techniques.
- Focus: Trove by the National Library of Australia; Australian obituaries of service personnel.
- Image source: https://trove.nla.gov.au/newspaper/article/187018252


## Roll of Honour.

## Total Casualies, 938.

## 96 Queenslanders.

The 311th Australtan casuality list was released by the censor at 10 a'clock this morning. The summary reads as follows: Accidentally killed, others, 1 killed in actioni others, 97 d died of nounds, officers 1, others 6; died (other causes), others i, wounded, officers 7 , others '540; missing, others 102 : sick, officers 22 , others 90 ; injured, officers 4 , others 14 ; prisoners of war, others 48 nature of casualty not known, others 1 total, officers 34; others, 900. Xurses, ill, 4. Total, officers 38 , other: 900 Grand total, 938 . Retarned to duty, others, 14.

QUEENSLAND.
KILLED IX: ACTION.

## Background Work on Obits

1. Converting the scans/facsimiles into text/structured records [10-11]; zoning/structuring [4], fixing issues [33-35]
2. Application of sentiment analysis and statistical methods:
"[t]ones, emotions, terms used to describe death" [13];
demographic properties [14-15]|7]
3. VVCs, "lexical approach" to guide philosophical investigation [16-17]: co-occurrence network visualization to map VVC descriptors and their inter-relationships, incl. part-of-speech (POS) tagging.
4. Modern (deep) NN techniques:
e.g. kinship information with named-entity recognition [8].

## Data + RQ1: places and people during these wars?

Australian Newspaper Articles from Trove:

WWI (1914-1920 inclusive) + WWII (1939-1947 inclusive).

Obits with "killed in action" excluding Honour Rolls (list of casualties).

AccuracyHeuristic $>\mathbf{8 4 . 5 \%}$ based on the Hunspell checker.

Starting total 54,707 $>$ removed 43,793 'rolls of honor' $>$ removed 9,991 obituaries (where AccuracyHeuristic < 84.5)

* final total of 923 FilteredObits.

NLP Pipeline:
spaCy POS tagging (en_core_web_sm model):
'ADJ' and 'ADV'.
NLTK tagged 'JJ', ‘JJR', or 'JJS'.
Hunspell stemming function to reduce each word to its dictionary stem.

## Coding:

Authors manually and independently checked the list of 839 words.

712 out of 839 words agreed upon.
$($ Cohen's kappa $=.657, \mathrm{z}=19.5, \mathrm{p}<.0001)$ [26].
Differences resolved by consensus.

RQ2: what VVCs correspond to the adjectives and nouns used to describe decedents' character?

Semantic network of the virtues, values, and constituents of wellbeing attributed to Australian service personnel killed in action in WWI and WWII (using Gephi).

Node size $=$ weighted degree.
Node color $=$ modularity class.

## RQ3: spatial/temporal migration patterns of personnel in the war?

spaCy's [23] Named Entity Recognition (NER) annotation * GeoNames API [27] for initial geocode.

Manual correction heuristics:

- Mistaken proper nouns
- Mistaken names (not places)
- Ambiguous locations
- Documented sites of battle
- Minor typos


Rule-based rank classifier to determine personnel rank in obits for Tableau geo-vis dashboard.
Australian War Memorial [31] consulted as expert knowledge source for R23.

## Discussion and Conclusion

## Future work:

Data quality: optimizing OCR texts, auto segmentation and extraction
Incorporating more $\boldsymbol{n}$-grams $(n>1)$.
Data segmentation along pre-defined, theoretically-motivated lines
Diversity: supplementing our corpora with texts that talk about women and children Improving geocoding quality: e.g. crowdsourcing an authoritative place list Consulting death certificates in archival sources

## Conclusion:

Wide range of VVC descriptors to characterize and laud service members: semantic clusters serve as objects of reflection for experimental philosophers.
Much extant 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 $»$ enables ordinary people's intuitions to speak for themselves.
Useful tool for philosophers to reflect upon and widen the scope of potential ethical concerns beyond those generated a priori.

