Topic Modelling in Spontaneous Speech Data

Bachelor of Arts, Honours in Language Studies

The Australian National University

Mr Marcel Reverter-Rambaldi

October, 2022

This thesis is submitted in partial fulfilment of the requirements for the degree of Honours in Language Studies in the College of Arts and Social Sciences.

I hereby declare that, except where it is otherwise acknowledged in the text, this thesis represents my own original work.

All versions of the submitted thesis (regardless of submission type) are identical.

Chapter 2.2 & Appendix 2 contain material submitted for the Corpus Review and Corpus Creation tasks, respectively, in LING4106. The same applies to Chapters 1.1, 1.2, 3.2, 3.3, 3.4, 3.5 & Appendix 2, but for the Research Paper in LING4009.

The compilation of the Sydney Speaks corpora was approved by the ANU Human Research Ethics Committee (protocol number: 2015-088, Travis). I obtained approval to use the Sydney Speaks corpora as a student on the project.

Table of Contents

Table of Figures

Index of Tables

Acknowledgements

I would like to thank Professor Catherine Travis, my academic supervisor, for guiding me through the complexities of academic research, and introducing me to the Sydney Speaks Project. My best wishes extend to the Sydney Speaks team, ARC Centre of Excellence for the Dynamics of Language, and Language Data Commons of Australia team. Also, to my academic mentors, family, partner and friends, thank you for supporting me in the monumental task that is producing a thesis.

I wish to dedicate this thesis to two groups of people: all those who dabble in computational linguistics, and the downtrodden searching for happiness in the world. To those that feel betrayed, disillusioned or exploited by this world, remember that your failures alone do not obstruct you. Rather, it is how you deal with them.

Agency or Nothing

Life's a game of cops and robbers, some aim to avoid drops from dollars.

Such lines are sinister, as ministers administer to us cynics where we fight to aspire.

But only strife responds as lone critics, from which we must tire.

So give life your own gimmicks, the end accepts all attire.

Abstract

The development of large-scale, language corpora has highlighted the increasing need for automated methods, to assist humans in the inefficient task of sorting and labelling language-transcripts by semantic contents (i.e. topics). One approach to semantic labelling involves using a class of unsupervised, machine-learning algorithms known as “topic modelling”. These algorithms process a document (e.g. a transcript), and identify clusters representing words that occur in proximity to each other in the document.

To date, topic modelling has been implemented widely in written language – including newspapers, academic articles, and business reports – but much less to spontaneous speech data. The linguistics literature has identified the need to apply more qualitative and analytic approaches, when judging and improving topic modelling for future use. My research applies topic-modelling algorithms to transcripts from sociolinguistic interviews, compiled for the Sydney Speaks Project. I apply certain modifications to improve topic-modelling’s performance, including the use of a custom stoplist, *a human-benchmark for measuring efficacy*, and linguistically-based, text partitioning. The findings support the idea that text partitioning and a custom stoplist, produce results that align better with the human benchmark.

Chapter 1- Introduction, Preliminaries and Literature Context

1.1 Introduction

The development of large-scale, spontaneous-speech corpora has transformed the field of linguistics. With ever-increasing numbers of speech transcripts being added to existing corpora, and new corpora coming into existence every year, the feasibility of humans to retrieve information manually from databases of transcripts is diminishing. This thesis will explore an unsupervised, machine-learning approach for extracting information – known as *topic modelling* – to consider its viability as a method for identifying topics in spontaneous speech data.

A "topic", for the purposes of *computational topic-modelling*, constitutes a "bag of words" which co-occur to encode a semantically-cohesive concept (Jelodar et al., 2018:4-5). From a language-use perspective, it is easier to think of a topic as a continuous linguistic-unit, which carries a shared concept across multiple sections of speech and between interlocutors. In this respect, the concept shares many similarities with "semantic frames", as proposed by Fillmore (1976, 1982). From a computational perspective, it is represented as a cluster of semantically-related words. Topic modelling, therefore, refers to a category of clustering algorithms, designed to condense networks of words

from arbitrarily-large documents, into interpretable and representative clusters.

The retrieval of topics from corpora of spontaneous speech is of interest to linguists, historians and sociologists, as corpora may contain people's perspectives on major time-periods, such as The Great Depression or the migration of certain groups. Researchers and policy-makers benefit from understanding how the general public respond to social, environmental, economic or political disruption, from the perspective of what people raise in discussions, and what they say about it. The value of understanding the dynamics of society through compiling collections of data in an accessible format, is visible in the existence of government agencies such as the National Archives of Australia. Another example constituting large-scale, digital-infrastructure is the Language Data Commons of Australia (2021 – 2023) project. Hence, a method which can identify the topics of discussion in millions of words, and present these in manageable and processable units (i.e. topic modelling), will greatly facilitate the classification, organisation and understanding of large-scale corpora.

While topic modelling has been largely applied to written data, very little has been done to attempt topic modelling with spontaneous speech data. This detail is crucial to the gap which this thesis is addressing, as in written texts, for example, there are explicit, *user-edited* paragraphs to divide documents by

ideas, representing a “fundamental sub-structure in which a text is organised” (de Arruda et al., 2016:3).

This thesis begins with an overview of what defines a topic, and argues in favour of topic modelling in existing corpora. The literature on the applications of topic modelling will be reviewed, and the specific, topic-modelling algorithm to be used will also be stated. The review will address the circumstances in which current, topic-modelling approaches are deficient, to motivate the modifications proposed and applied in this thesis to improve efficacy with spontaneous speech data. A summary of the remaining chapters is presented at the end of this chapter.

1.2 Applications of Topic Modelling

1.2.1 Topic Modelling in Public Perceptions and Business

Topic-modelling approaches have been implemented in many perception studies, each addressing a different research-question. Qiao & Williams (2022) applied topic-modelling algorithms to analyse Twitter conversations about Global Warming. Fino et al. (2021) and Ghasiya & Okamura (2021) investigated Twitter posts around gambling during Covid-19, and views on the Middle East in Japanese newspapers, respectively. Reports and corporate disclosures during conference calls were analysed by Huang et al. (2018).

Qiao & Williams (2022) analyse Twitter conversations about Global Warming using the Latent Dirichlet Allocation topic-modelling algorithm. They examined

all tweets retrieved containing the words “global warming” between 1st January, 2020 and 30th July, 2021 – a total of 538,478 tweets and nearly 22 million words (pp. 2 – 3). The results identified multiple word-clusters formed by the combined tweets, referencing the ideas of “factors causing global warming” (e.g. *carbon*, *pollution*), “impact of global warming” (e.g. *rise*, *weather*), “actions to stop global warming” (e.g. *fight*, *policy*), “global warming and Covid-19” (e.g. *hoax*, *Trump*), “close relations between global warming and politics” (e.g. *ideology*, *vote*), “global warming as a hoax” (e.g. *lie*, *Trump*), and “global warming as a reality” (e.g. *evidence*, *anthropogenic*) (pp. 4 – 9).

Fino et al. (2021) found that the two most-common topics in the Twitter posts (as retrieved through REST API (p. 493)), related to “gambling addiction amid the COVID-19 outbreak” (e.g. *bet, casino*) and “the user’s psychology of gambling addiction” (e.g. *impulsivity, addiction*) (p. 496). Ghasiya & Okamura (2021) explored Japan’s dependence on the Middle East for petroleum products, via media reports. Due to scarcity of these resources in Japan, it is in the country’s interests that major conflicts are avoided (p. 872). The topic-modelling findings revealed that the media often discussed the “Islamic State, the refugee crisis, the Syrian civil war, Qasem Soleimani killing, and Iran nuclear deal” (p. 871).

Huang et al. (2018) used topic modelling on analyst reports and corporate disclosures during conference calls. Their findings indicated that the topics of

"business outlook" (e.g. *new, strategy*) and "revenue and sales" (e.g. *revenue, margin*) were some of the most frequent (p. 2850). Additionally, analysts discussed topics that were not referenced in earlier conference-calls, which corresponded to "when managers have stronger incentives to withhold information" (p. 2848), and discussions on conference-call dealings were confirmatory remarks about "managers' statements", to encourage investors (p. 2848).

1.2.2 Topic Modelling in Corpus Linguistics

Applications in corpus linguistics have appeared in the last decade. Jaworska & Nanda (2016) discuss the role of topic modelling, and the need to develop an approach that balances between solely quantitative and qualitative methodologies, which "offers applied linguists new ways of exploring discourse in large collections of texts" (p. 373). They utilise a corpus of 317 reports (14,806,512 total words) in the context of corporate social responsibility in the oil sector (p. 373), to analyse changes in social policies by companies over a 12-year time-span, and demonstrate how topic modelling can be implemented in "discourse studies interested in analysing large collections of texts" (pp. 382 – 383). The most frequent topics were "future plans" (e.g. *future, increase*) and "people, community and rights" (e.g. *human, rights*) across all years. They note a decrease in the prevalence of topics about "environmental protection" (e.g. *emissions, climate*) and "research and

development" (e.g. *exploration, reserves*), but increases in "business operations" (e.g. *operations, performance*) and "corporate governance" (e.g. *management, audit*) (pp. 385 – 388) over time. With the combined topics, the authors make the inference that petroleum companies, despite mentioning human rights in public statements, "pursue, if at all, a human rights minimalism" (p. 394 – 395), due to a negligible concern for environmental impacts on communities.

Murakami et al. (2017) discuss the use of topic modelling to extract a corpus' topic contents automatically, as opposed to relying on manual annotations. The conclusion is that topic models are best suited as an exploratory tool, since they are particularly useful in the initial exploration of large-scale texts where manual reading is not feasible and the data "does not require pre-specified [semantic] categories" (p. 272). Their implementation is on 675 academic papers, or around 2.5 million words, published between 1990 and 2010 in the journal Global Environmental Change (pp. 248 – 250). The work arises from the need to determine "how to approach a specialised corpus to discover what might be said of significance about it" (p. 244), without relying on predetermined, semantic tagging, such as annotations or keywords. They extracted topics about livelihoods (e.g. *farmer, village*) and kinship (e.g. *household, family*), which directly corresponded to information from keywords (pp. 268 – 270).

1.2.3 Uses without Human Analysts and Sentiment Analysis

Beyond human language, topic modelling has proven its value and versatility in summarising computer code, such as malware detection in Stegner's (2021) thesis; the application was to identify code containing malicious commands (pp. 8 – 10). As a control, the author also applied the same algorithm to non-malicious code, with all data coming from the "Das Malwerk" and "BIG 2015" corpora of software and viruses (p. 11 – 13). The author found that, at its best performance, topic-modelling tools were able to categorise correctly a piece of software – as either malicious or non-malicious – in "97.03%" of cases (p. 21).

Additional non-human-analyst work using encyclopaedic materials is described in Skaggs' (2014) thesis, which investigates the use of topic modelling to improve the relevancy of hyperlinks within Wikipedia articles. The methodology involved obtaining written descriptions from "disambiguation" pages (i.e. a Wikipedia page that lists all possible meanings for a given word in terms of other Wikipedia pages, such as "Frank Field" referring to either the Australian or British politician (p. 41)), running the data through topic-modelling algorithms, and then determining which word-clusters were associated with each page (p. 22 – 24). By comparing the degree of overlap between clusters from disambiguation pages and their referent pages, the suitability and relevancy of hyperlinks could be determined. The

author concludes that their algorithm “avoids having humans manually assess several hundred disambiguation predictions” (p. 45).

A process called *sentiment analysis* sometimes accompanies topic-modelling during implementations (cf. Fino et al., 2021; Ghasiya & Okamura, 2021; Qiao & Williams, 2022), which involves analysing responses (e.g. tweets, reviews) for different emotions. Companies, or the public service, can then use this information to determine if there is a need to “improve their products and services” from corpora of written reviews (Ghasiya & Okamura, 2021:873). While an interesting technology, this thesis will only explore methods that deal with the literal words in a corpus (i.e. *topic modelling*). Therefore, sentiment analysis falls beyond the domain of this thesis.

1.2.4 Modifications to Topic Modelling

There is a recent body of works exploring improvements to topic-modelling algorithms. Here, I review whether some of these might be suitable to spontaneous speech data.

De Arruda et al. (2016) apply topic modelling to encyclopaedic materials – namely, a corpus of Wikipedia articles – to develop modifications for improving existing algorithms. However, some of these computational models may encounter problems with spoken data. Firstly, their “paragraph-based model” (p. 3) relies on the assumption that “paragraphs represent the fundamental sub-structure in which a text is organized, whose words are

semantically related" (p. 3). While this holds for more formal writing, the division of information in spontaneous spoken language is not as explicitly and saliently marked as paragraphs in writing. Hence, this raises the question of how – or even *if* – partitioning by paragraphs could be translated to processing informal, spoken language?

An attempt to improve the accuracy and reliability of the paragraph-based model, is presented by the "adapted paragraph-based method" (p. 3). The method works by identifying highly-frequent co-occurring words in the original document, and then comparing these to co-occurrences in a randomly-shuffled version of the same document. If the co-occurrences from the original text are more frequent than those in the shuffled version (i.e. if words are co-occurring together more than what would be expected from random chance), then the words are viewed to be related. Since the *number* of co-occurrences is a variable quantity (as opposed to a binary state), this metric compares the likelihood of terms being interrelated. A similar idea to this metric, underlies the Bayesian statistical model used in Latent Dirichlet Allocation (LDA, cf. Qiao & Williams, 2022) – a topic-modelling algorithm commonly used on written language.

While the aforementioned approach sounds brilliant for spontaneous speech, seeing as words are not produced randomly by speakers, the primary issue lies with long run-times when processing data, owing to the shuffling process to

find co-occurrences (Jelodar et al., 2018:5-6). From initial trials with transcripts, LDA's run-time grew exponentially as more were added for processing. Thus, in accordance with advice from a co-supervisor that an alternative algorithm, called *Latent Semantic Indexing (LSI)*, performs similarly to LDA in cluster contents for large-scale corpora, LSI is the chosen algorithm in this thesis. However, as later parts of this thesis investigate modifying the topic-modelling algorithm to process spontaneous speech data through partitions (which act as *separate* transcripts from the algorithm's perspective), LDA's use in large-scale, spontaneous speech corpora is an idea to explore in a future thesis or journal article.

An alternative method for determining co-occurrences is seen in Joty et al.'s (2013) article, which discusses how to automate the labelling for topics in conversations conducted via writing – these conversations being email and social-media exchanges. Though the rules of turn-taking in these environments are different to those from spontaneous discourse – as users have time to polish messages and cannot be interrupted with overlapping speech – some of the algorithms appear of use to spontaneous speech data. The first algorithm involves using a "random walk" model which "respectively captures two forms of conversation specific information" (p. 521); the first (and most relevant) piece relates to "the fact that the leading sentences in a topical cluster often carry the most informative clues" (p. 525). In an email exchange, it is understandable that an initial message would serve to

introduce any ideas to be discussed, with subsequent replies contributing to its initial claims. Additionally, when examining an emailed message in isolation, it is not unexpected that its contents would be organised by semantic value, since an author has time to edit the email's phrasing and paragraph structure. So, on the surface, it may seem that this topic-labelling method may not be compatible with spontaneous speech.

1.2.5 Critiques of Topic Modelling

Brookes & McEnery's (2019) article on topic-modelling a corpus of NHS (National Health Service, UK equivalent of Medicare) online-feedback surveys provides critique over a poor application of linguistic methodology in previous studies. They highlight "the absence of any theoretical account of what constitutes a 'topic' within topic modelling" (p. 6) in the existing literature, the majority of which comes from computer-science research and not linguistics. To address this gap, the authors describe the clusters of words which form a *topic* in topic-modelling, as being similar to "propositional topics" (p. 6) (cf. van Dijk, 1977). Propositional topics refer to mental concepts which are constructed over stretches of text (or discourse) through co-occurring words, such that "a topic of discourse cannot simply be explained in terms of semantic relations between successive sentences. Rather, each of the sentences may contribute one 'element' such that a certain structure of these elements defines the topic of that sequence in much the same way as, at the

syntactic level, words can be assigned a syntactic function only with respect to a structure 'covering' the whole clause or sentence" (van Dijk, 1977:6).

The corpus used in Brookes & McEnery (2019) contains 228,113 feedback comments from patients, which correspond to "approximately 29 million words" (p. 8). These comments represent an online, opinionated style-of-language (due to someone recounting events or experiences), like that seen in previously-mentioned articles exploring Twitter posts. In addition to the topic-modelling analysis, the authors apply a qualitative exploration of the comment data, using previous work by Gkotsis et al. (2017) on analysing online, health-related posts. This methodology involves (1) having an analyst who is familiar with a healthcare setting judge the computational topics (i.e. groups of co-occurring words) for coherency, and (2) a direct comparison of the generated topics to the original text, through manual revision. The latter, though the more robust of the qualitative approaches, is inefficient for use in all situations, and is best reserved for judging the accuracy of topic-modelling (as opposed to in-practice, analyst use).

The results from Brookes & McEnery (2019) show that the significance of topic clusters was easiest to interpret *if* an analyst had knowledge of the original texts. As an example, a blind examination of topics (i.e. when nothing is known about the original corpus) by a medical practitioner, revealed that of the 20 topics extracted from the corpus, "the majority (n = 14) [...] lacked

thematic coherence throughout and were mixed in terms of their reliability" (p. 17). Of the six which were interpreted as being coherent, only one accurately reflected a genuine grouping, namely all positive reviews. The authors conclude the underwhelming findings, by remarking that "future research employing topic modelling methods should therefore endeavour to engage more deeply with linguistic theory when inferring the presence of topics in their textual data" (p. 18), and that assessments of topic-modelling in new areas should be "followed up with qualitative, more contextualised analysis of the texts making up the topics" (p. 19). Both of these points are fundamental ideas, which will guide the methodological approaches in this thesis.

Shadrova (2021) offers an additional critique of topic-modelling's applications in the realm of digital humanities. Specifically, she states that topic modelling "operates from relevantly unrealistic assumptions, is non-deterministic, cannot effectively be validated against a reasonable number of competing models, does not lock into a well-defined linguistic interface, and does not scholarly model topics in the sense of themes or content" (p. 1). Encouragingly, however, the author eventually reconciles that "a conceptual validation would require an extended triangulation with other methods and human ratings, and clarification of whether statistical distinctivity of lexical co-occurrence correlates with conceptual topics in any reliable way" (p. 1). Such

comments align with the qualitative, manual approach that I will employ as a baseline for judging efficacy.

1.2.6 Topic Modelling in This Thesis

As topic modelling has been found to yield acceptable results in *written language* – or in otherwise quite rigid linguistic-structures such as computer code – the question becomes how well do topic-modelling methods perform with spontaneous speech data, given the apparent absence of implementations with this type of data? If existing methods are ineffective, what kinds of modifications are needed to account for the specifics of spontaneous speech? In essence, this thesis will test the idea that topic modelling – following appropriate adjustments and preconditioning – should reveal what people are discussing within large-scale corpora.

Latent Semantic Indexing (LSI, cf. Landauer et al., 1998) is the topic-modelling algorithm to be applied in this thesis. Python 3 (van Rossum & Drake, 2009) is the programming language of choice. LSI functions by collecting all words from a document (for this thesis, a transcript) into a list, and then representing each word as a grid entry with two values – the word itself, and the number of occurrences in the entire transcript. Its location in the grid corresponds to the word's location in the original transcript. Then, using Singular Value Decomposition[1], the grid is deconstructed into three separate grids. One of the grids contains the "singular values" along its diagonal – a collection of

1 A matrix-decomposition technique (Singular value decomposition, 2022)

numbers encoding where clusters (i.e. topics) appear, and the words that comprise each one. The singular values encode the most prominent pieces of data from the original document; so, LSI effectively performs file-compression on a piece of text (cf. Kontostathis & Pottenger, 2006; Landauer et al., 1998). LSI's operations are entirely based on matrices, or "grids" – something which is optimised at the hardware-level and accessible through CPU and GPU intrinsics (Xia et al., 2016). This ensures a minimal run-time for large data-sets, which is beneficial when dealing with infrastructure-scale corpora.

An important feature of topic-modelling algorithms is the requirement for users to provide the number of topics and unique words per topic, as an initial step preceding any text-analysis (cf. Jelodar et al., 2018; Landauer et al., 1998). In keeping with the literature recommendations, this thesis will first investigate a qualitative approach to identifying which (and how many) topics are present and their sizes, to improve the application of subsequent, automated methods.

1.3 Summary

To summarise, this chapter has explored the relevancy of topic-modelling for spontaneous speech corpora, in addition to providing a thorough review of topic-modelling's previous applications and present gaps. As the literature supports the idea that more input from a linguistic-analyst perspective is needed to improve topic-modelling, Chapter 2 will explore the raw data from

the corpus of choice, and then analyse some transcripts manually for topics, to serve as the human benchmark for topic modelling. Chapter 3 will then develop the first automated approach to extracting topics, by creating a specialised list to remove uninformative content in spontaneous speech data. Chapter 4 will explore the optimisation of topic-modelling for spontaneous speech data, using knowledge gained from the earlier, manual coding. Chapter 5 will present the cluster results, followed by general conclusions, limitations and recommendations in Chapter 6.

Chapter 2- Data Context and Human Benchmarks

2.1 Introduction

Since topic modelling is used to analyse human language, it is important to consider the aspects of human language that provide context, such as genre and structure. This chapter first explores Sydney Speaks – the corpus to be used in this thesis – to understand better its contents, and the implications for topic modelling. An initial attempt at topic modelling will be performed, by manually reading through transcripts to extract topics. This provides a human benchmark for extracting topics, for a comparison with automated approaches.

2.2 Sydney Speaks Corpus

The data used in this thesis, originates entirely from the Sydney Speaks Project (Travis, 2014 – 2022), which comprises three sub-corpora: the NSW Bicentennial Oral History Project (1987), the Sydney Social Dialect Survey (Horvath, 1985), and Sydney Speaks 2010s (Travis, 2014-2022), abbreviated in this thesis (for convenience) as BCNT, SSDS and SydS 2010s, respectively. In total, the combined corpus contains 1.5 million words, collected via sociolinguistic interviews with 260 speakers.

A key definition is that of the *sociolinguistic interview* – the source of all spontaneous speech data in Sydney Speaks. This format was pioneered by

William (“Bob”) Labov, in a study exploring the social stratification of English in Martha’s Vineyard (Labov, 1963). In a subsequent publication, Labov (1984:33) clarifies that sociolinguistic interviews are concerned with “shifting the [speech] style towards the vernacular”. These remarks, importantly, demonstrate that interviewing a speaker *without making your variable of study explicit* converges towards what Gordon (2013:108) describes as a “dialectological approach in seeking to observe less self-conscious usage” of language, to capture speech “when the interviewee monitors his or her speech the least”. Purser et al. (2020) synonymously describe sociolinguistic interviews as exercises in “eliciting unmonitored speech”, to record “spontaneous spoken language that arises naturally in interaction” (p. 274). Rather importantly for topic modelling, Purser et al. (2020) also note that a sociolinguistic interview constitutes “an informal interview that does not strictly adhere to a pre-determined set of questions, but rather draws on a *set of topics* assumed to be of interest to the participant, with the aim of eliciting unmonitored speech” (p. 274). Therefore, seeing as (1) no apparent topic-modelling has been performed on sociolinguistic data, and (2) sociolinguistic interviews draw their elicitation from a defined “set of topics”, there is a solid viability for applying topic-modelling algorithms to sociolinguistic data.

The BCNT sub-corpus contains transcripts of 31 elderly speakers, all born in Sydney before 1910, with both of each participant’s parents also born in Sydney (Travis & Johnstone, *In Progress*:7). The recordings were collected in

1988, and take the format of monologic, oral histories. Due to the speakers' unique positions, insofar as providing a snapshot of Australia's early decades into nationhood, the interviews recount events from bygone eras in the speakers' early lives. Thus, their topics of discussion relate to health, education, employment, courtship and religion (Travis & Johnstone, *In Progress*:7). Aside from their inherent value to historians, other work using this collection includes Lonergan and Cox's (2010) study into possible rhoticity in early Australian English. An extract from a BCNT transcript is visible in Example 1.

(1, Bcnt_AEF_073, 01:16:12.581– 01:16:28.376)[2]

PNT: Marjorie	.. the only things I s- shone at in those days was ... religious knowledge @.
INT: Jackie	...(1.5) religion is very important t[o you].
PNT: Marjorie	[it w]as most important.
PNT: Marjorie	... it still is.
INT: Jackie	... hm.
PNT: Marjorie	... but um,
PNT: Marjorie	...(1.0) I got a catechism bee when I was --
PNT: Marjorie	... there at --
PNT: Marjorie	... at ~Saint ~Brigid's %.

SSDS was compiled by Barbara Horvath (1985), for the purposes of sociolinguistic and sociophonetic study. The data-collection work occurred from 1977 to 1980 with a resulting 177 interviews completed (cf. Travis & Johnstone, *In Progress*). Some noteworthy studies using the corpus, investigated the realisation of certain vowels, phrase intonation, discourse-markers and velar/alveolar variants of (ing), such as work by Horvath & Sankoff (1987), Guy et al. (1986) and Grama et al. (2020). This work could

2 This is the format for presenting Sydney Speaks transcript extracts, showing the transcript's name, followed by the relevant timestamp

compare results across adults and teenagers, as these are the two age-groups represented in the corpus (Horvath & Sankoff, 1987). Additionally, the corpus contains texts stratified by sex, socioeconomic status, and ethnicity; the three ethnic groups represented are the Anglo-Celtic, Greek and Italian communities (Travis & Johnstone, *In Progress*:4). Across all participants, SSDS interviewers asked questions about the topics of "games, layout of the school, nicknames, and language" (Travis & Johnstone, *In Progress*:4), which is a notable departure from the BCNT interviews. Example 2 presents the topic of nicknames in the SSDS data.

(2, SSDS_GTF_855, 00:28:39.611 – 00:29:00.409)

INT: Susan	...(1.0) What about nicknames.
INT: Susan	some people have been telling me about funny nicknames they had for their teachers at school.
INT: Susan	have you got any --
PNT: Isa	... Uh=,
PNT: Isa	well there's this --
PNT: Isa	um,
PNT: Isa	...(1.5) teacher at --
PNT: Isa	... I don't know.
PNT: Isa	we ha- --
PNT: Isa	% we don't --
PNT: Isa	...(1.0) at our school,
PNT: Isa	we haven't got nicknames,
PNT: Isa	we just --
PNT: Isa	... people just say things like,
PNT: Isa	.. depending on what mood they're in?

Sydney Speaks 2010s (SydS 2010s) is being compiled by Community Research Assistants through their extended networks, with a large amount gathered between 2016 and 2019 (Travis, 2021). It is designed around the same principles as SSDS, though with slightly-different age groups – twenty year-

olds instead of teenagers – and, as well as Italian and Greek Australians, includes ethnic groups that were not present in the 70s and 80s, namely Chinese Australians (Travis, 2021; Grama et al., 2020). Another deviation from SSDS is that the interviews are conducted in a more unguided style, which should elicit a more conversational dynamic; specifically, topics are more participant directed (Purser et al., 2020). The corpus contains 1 million words, but this number is expected to grow, owing to its *dynamic* nature (Travis, 2021; Barth & Schnell, 2021:36-37). An extract from SydS 2010s can be seen in Example 3.

(3, SydS_AYM_015, 00:14:57.609 – 00:15:07.607)

INT: Heather	do you watch any other zombie movies?
PNT: Zander	... Uh,
INT: Heather	.. what movies [do you like to watch].
PNT: Zander	[oh I watch --
PNT: Zander	I] watch zombie movies,
PNT: Zander	I watch zombie shows,
PNT: Zander	but I play zombie games.
PNT: Zander	.. [like],
INT: Heather	[(TSK)]
INT: Heather	.. Ri[2ght,
PNT: Zander	[2There- --
PNT: Zander	there- --
PNT: Zander	there's2] --
INT: Heather	yep2].
PNT: Zander	there's a difference.

LaBB-CAT is the primary interface for corpus queries and searches. Developed by the University of Canterbury, it allows users to search for wordforms, lexemes or collocations, with filters for parts-of-speech and syntactic

constructions (Fromont & Hay, 2008). Additionally, as LaBB-CAT stores auto-aligned recordings parallel to each transcript, a user can search by phonemes or phones, and download them in the wider context (Fromont & Hay, 2008). For additional processing, each text can be downloaded in CSV and ELAN format, allowing the researcher to process data with R or Python (Fromont & Hay, 2008).

All corpora use a standardised transcription notation, which records aspects of intonation and discourse-pragmatic features, while still adhering to English orthographic rules (Du Bois et al., 1993). Though the CSV files do not display IPA transcription by default, this information is accessible on LaBB-CAT, in addition to syntax trees and PoS labelling. In order to capture prosodic information, certain symbols are used in Sydney Speaks corpora to represent prosody, intonation and comments in transcripts (Du Bois et al., 1993). For instance, square brackets represent overlapping-speech, with numbers included inside the brackets when needed to clarify sections of overlap (e.g. "[" and "]" vs. "[2" and "2]"); "@" represents laughter; two to three full-stops indicate short pauses (a single full-stop indicates a prosodic ending, with a comma marking continuation); "%" denotes a glottal stop, and items in double round-brackets indicate a transcriber's notes. The full list is available in Appendix 1. In order to process the corpus data as text, the prosodic markers were compiled into a list, and removed in the pre-language-processing phase by *replacing* them with nothing (i.e. deleting them without a residual, white-

space character). An example of prosodic characters touching and splitting other words, when compared to an example of polished text which the computer reads, is seen in Example 4.

(4, SydS_AYM_015, 00:14:00.756 – 00:14:15.806)

Speaker	Raw Transcript	Polished Transcript
PNT: Zander	for a while I did enjoy a computer game called League of Legends,	for a while i did enjoy a computer game called league of legends
PNT: Zander	which was .. [amazing].	which was amazing
INT: Heather	[Okay].	okay
PNT: Zander	... [2I don't2] --	i don't
INT: Heather	[2What's it2] --	what's it
PNT: Zander	.. [I don't know if you've heard] of it.	i don't know if you've heard of it
INT: Heather	[What's it about].	what's it about
PNT: Zander	.. Um,	um
PNT: Zander	.. it's an online game,	it's an online game
PNT: Zander	.. purely online.	purely online
PNT: Zander	.. Uh %,	uh
PNT: Zander	.. you're --	you're
PNT: Zander	you --	you
PNT: Zander	you choose a character that you want to be,	you choose a character that you want to be
PNT: Zander	and you're paired with four other people.	and you're paired with four other people

This thesis will be analysing five transcripts from the Sydney Speaks corpus; these being two from SydS 2010s (SydS_AOM_105_Parker and SydS_AYM_015_Zander), two from SSDS (SSDS_GTF_855_Isa and SSDS_AAF_127_Janice), and a final one from the BCNT collection (Bcnt_AEF_073_Marjorie). Since the combined participants in these transcripts cover the five age-ranges found in Sydney Speaks, this should provide a holistic overview of the spontaneous speech data from a spread of participants over 100 years of apparent-time. Importantly, though, since this thesis is purely interested in evaluating the *performance* of topic-modelling algorithms with spontaneous speech data, questions that relate to sociolinguistic aspects (e.g. do males produce more topics than females?) are

beyond the scope of this thesis. Table 1 provides a comprehensive overview of these transcripts, including the number (N) of Intonation Units (IUs) and individual words ("tokens").

Table 1 Corpus Data for this Thesis

	Marjorie	**Janice**	**Isa**	**Parker**	**Zander**
N Tokens	5680	5982	5324	7271	7720
N IUs	1366	1641	1577	1678	2423
N Minutes	29.4	33.5	28.6	32.9	39.4
Subcorpus	BCNT	SSDS	SSDS	SydS 2010s	SydS 2010s
Age Group	Elderly	1970s Adult	1970s Teen	2010s Adult	2010s Young-Adult
Gender	Female	Female	Female	Male	Male
Ethnic Group	Anglo-Celtic	Anglo-Celtic	Greek	Anglo-Celtic	Anglo-Celtic

When examining Table 1, the idea that metadata is an additional tool to complement the output from topic-modelling algorithms comes to mind, as it provides context when extracting topics from corpora. Since metadata is recorded alongside sociolinguistic corpora (e.g. speaker ages, gender, ethnicity, socioeconomic status), it can offer vital clues about topic contents.

The idea of using a transcript's metadata to predict its topic contents is presented by Onyimadu et al. (2013). In their investigation, the authors apply a metadata-scraping tool into a closed search-engine; the search engine's role is to identify parliamentary transcripts from the UK, for public reading. The reason for using metadata to identify transcripts is justified by noting that "various knowledge about aspects of the parliament and MPs are assumed,

such as MPs and their constituencies, terms of office, and their changes over the years. These are often known to the users but not captured by standard keyword-based searches, which results in mismatches between user expectations and search results. The semantic search-engine aims to improve both the relevance and coverage of search" (p. 345), since the metadata includes information that may predict the topics of discussion in a respective transcript (e.g. politicians from rural seats discussing agricultural subsidies or tariffs).

While the Sydney Speaks project did not record participants' political leanings, information such as age and ethnicity were recorded, which may affect in-text topics (e.g. Greek participants discussing visiting Greece and/or Greek School). Hence, the existing metadata (related to topics of discussion) such as schooling, childhood games, and attitudes to Australian English being discussed in SSDS (Horvath, 1985), can be employed to interpret better any cluster results. Furthermore, since recording metadata about participants is important across sociolinguistic work, such a technique for predicting topics in transcripts is a general method, with applications beyond Sydney Speaks.

2.3 Manual Coding for Topics

2.3.1 Methodology

This thesis is focused on applying automated methods for topic-modelling; however, the ability to make claims of efficacy requires a comparable,

performance benchmark. Since the human brain is optimised to process spontaneous spoken language (as the direct product of human, linguistic interaction), it is therefore reasonable to use human performance as the benchmark for efficacy. The following guidelines were developed for identifying and extracting topics.

1. Read through an entire transcript, paying close attention to the content (including contributions from the interviewer and participant), recognising that, as a sociolinguistic interview, the bulk of the content will come from the participant, and the interviewer has the responsibility of eliciting information, which may include introducing topics.

2. Identify all words that refer to a related topic, and make note of each lexeme, taking into account:

    a) The function of a specific lexeme in the discourse context, which includes the semantic-frame, e.g. *buy, pay* and *sell* in Figure 1 (cf. Fillmore, 1982), but also broader contextual information, such as *magazine* and *sold* in Figure 1, which out of context would not be in the same semantic-frame.

    b) The lexeme's role as a conveyor of information and semantic meaning, as opposed to a grammatical item (e.g. prepositions *for, what* in Figure 1, and tense marking e.g. *haven't* in *haven't*

sold, in Figure 1). Nouns, and (occasionally) adjectives and verbs, are the parts-of-speech that carry most information, and only these were highlighted.

Figure 1 A collection of identified lexemes from the same topic, given the semantic context (SydS_AOM_105, 00:25:41.881 – 00:25:46.393)

PNT: Parker	every magazine that you haven't sold,	magazine	sold
PNT: Parker	... I'll buy it f- for the --	buy	
PNT: Parker	what it [what you paid for].	paid	

3. Once all relevant lexemes are identified, create a list of words pertaining to each topic, and give each list a label that names the topic to which the words correspond (e.g. “computer games”, “online”, “playing”, rendered as “Gaming”, as seen in Figure 2).

4. Review all words in the lists for each label/topic, noting that comparing their meanings to the group’s title/label provides a good basis for determining relevancy. As an example, Figure 2 presents all the “Gaming” lexemes, seen highlighted in yellow.

Figure 2 A collection of identified lexemes from the Gaming topic, as determined by the semantic context (SydS_AYM_015, 00:13:57.107 – 00:14:11.652)

speaker	IU	Words	
INT: Heath	.. So what other --		
INT: Heath	like computer games,	computer games	
INT: Heath	do you like playing,	playing	
INT: Heath	or,		
PNT: Zand	... Uh,		
PNT: Zand	for a while I did enjoy a computer game called League of Legends,	computer game	League of Legends
PNT: Zand	which was .. [amazing].		
INT: Heath	[Okay].		
PNT: Zand	... [2I don't2] --		
INT: Heath	[2What's it2] --		
PNT: Zand	.. [I don't know if you've heard] of it.		
INT: Heath	[What's it about].		
PNT: Zand	.. Um,		
PNT: Zand	.. it's an online game,	online game	
PNT: Zand	.. purely online.	online	

As per these guidelines, I manually coded five transcripts and each transcript was spot checked (by my supervisor, Professor Catherine Travis), as a means of cross-checking the classifications I had made. This process revealed that there were multiple ambiguities in semantic distinctions between topics, and their lexical contents – in particular, in determining when to include words in one topic, or break down into smaller subtopics. An example of multiple topics, which could be viewed as belonging to a single, greater topic of "Employment", are "Magazine Distributing", "Warehouse Job", "Public Sector Work", "Driving Instructor Work" and "Retail Work" (see SydS_AOM_105_Parker as "Parker", Table 2). Thus, depending on the given interpretations, this particular transcript may contain 15 topics (assuming the "Employment" group, and others, are not collapsed, as presented in Table 2), or as low as 8, with all groups collapsed ("Employment" super-topic plus

"Funding Cuts", "Bus Life", "Bus (Mis)management" and "Interactions with Bus CEO" in a separate super-topic).

Another challenge for manual-coding is the time-consuming nature of the task. Identifying topics in these five transcripts (each containing roughly 5000 words), took between 2.5 and 3.5 hours each to complete. For large-scale corpora with hundreds of transcripts, manual approaches are unfeasible, thereby making automated solutions rather desirable. The onerous nature of the manual-coding, when combined with its novel approach as developed for this thesis, were the reasons for using just one human coder.

2.3.2 Topic and Type Counts

Returning to the variation in topic-groupings, it is clear that this demonstrates the ambiguity of topic management in spontaneous speech, and the complexity and intertwinedness of topics in natural language. It appears to be more accurate to provide a range of possible topics for a transcript (e.g. between 8 and 15 for Parker), as opposed to a single, fixed integer. This is problematic, as topic-modelling algorithms typically require a user to specify both the number of topics, and the number of *unique* words pertaining to a specific topic, or *type count*. As an example, "dog dog dog" contains three words (i.e. three tokens), but only one type. While this thesis will eventually address these issues with *coherence scores* – as a means of optimising the most appropriate number of topics, and types per topic, from a provided

transcript – the degree of variability seen suggests that future research should investigate algorithms which do not require user-input for these values.

The values for topic and type counts, as determined by the manual coding, are visible in Table 2 and Table 3. Table 2 shows the manually-extracted topics from each transcript, presented in the order in which they occurred. For example, for Marjorie (the BCNT participant), the topics of “Clothing and Sewing”, “Stores” and so forth, were identified. At a glance, the range in the number of topics per transcript is quite evident, from six with Isa, to 16 with Zander.

Table 3 summarises the variability in type-counts for the topics in each transcript. The initial statistics, the mean and median values, present a notable variability in type-counts, as the mean ranges from 18 to 36, and the median from 14.5 to 48. These values show that the length and complexity of topics, in terms of type count, do vary considerably (e.g. 6 vs. 62 types for Janice’s smallest and longest topics). Therefore, some topics are considerably more information-dense than others, which cannot be represented successfully by assuming that each topic contains an identical number of words (as existing topic-modelling algorithms currently do with their clusters). Additionally, the question of whether multiple topics are indeed distinct, or are justifiably under a broader *super-topic*, still remains (e.g. see the footnote on page 40 for a topic with few types, but an arguably distinct identity).

Table 2 Manually-coded topics for all transcripts

	Marjorie	**Janice**	**Isa**	**Parker**	**Zander**
1.	Clothing and Sewing	School Life	Games	Magazine Distributing	Gaming
2.	Stores	Games	Childhood in Greece	Warehouse Job	Movies
3.	Radios and Broadcasts	Appreciating Glass	School Life	Public Service Work	Social Media
4.	Sydney City Mission	Janice's Mother	Mum's School-Life	Retail Work	Donald Trump
5.	Miscellaneous Charity	Attitudes about Education	Greek (Afternoon) School	Family Life	Race Relations
6.	The Depression	Class Post-Interview	Language Use and Attitudes	Funding Cuts	International Views about Australians
7.	School Life	Dialectal Variation and Language Use		Bus Life	Overseas Travel
8.	Heart Problems			Driving Instructor Work	Occupations
9.	Language Attitudes and Use			Bus (Mis)Management	Socioeconomic Stance
10.	Familial Relations			Assault	Bogan Characteristics
11.	Singing			Interactions with Bus CEO	Food
12.	Alcohol Use			Train Driver Work	Attitudes on American Terms
13.				Sam's Auto Parts	Lindt Cafe Siege
14.				Foxing	School and Uni Life
15.				Race Relations	Dialectal Variation
16.					Morning Shows

Table 3 Data about the number of types per topic

	Marjorie	**Janice**	**Isa**	**Parker**	**Zander**
N Topics	12	7	6	15	16
N Types per Topic					
Mean	18	36	24	22	20
Median	15	48	14.5	20	16.5
Range	8 – 22	6 – 62	10 – 50	[3]2 – 44	8 – 43

As seen in Table 2 and Table 3, there is a degree of variability between the number of topics per transcript, and the types per individual topics. Therefore, it seems reasonable that automated, topic-modelling methods should cater to this variability, by recognising the varying lengths of topic-cohesive areas of discussion. Despite the diversity in topic lengths, the underlying idea that unifies all topics is their set-like nature, inasmuch as topic-relevant words huddle together in chunks of speech. This observation, naturally, drives the question of consistently identifying the locations of said chunks.

2.4 Summary

This chapter first explored the Sydney Speaks corpus, covering its compilation, notation and genre, in addition to the transcripts used in this thesis. Additionally, the role of demographic data in providing context for topics was also noted. Following these steps, manual coding for topics was performed on each transcript, which revealed the diversity in the number of topics per transcript, and word types per topic. While the variability in type and topic counts is noticeable, it is clear that semantically-related words tend to co-occur, and topics with more types may be easier to identify through taking *word-frequency counts*. Therefore, it is now worth exploring automated approaches, to extract topics more efficiently than humans.

3 The next largest topic, by type count, has five types. I have chosen to include the topic with two types ("Train Driver Work"), despite its exceptionally-small size, as I found it to be prominently marked from the surrounding text with (1) clear pragmatic constructions signalling a change in topic by the participant, and (2) its observable difference from the ideas described in the surrounding text.

Chapter 3- Finding Informativeness in The Most Frequent Words

3.1 Introduction

This chapter explores a procedure for removing uninformative words, for the purposes of topic-modelling. The initial sections present the importance of removing *noise* words, and the deficiencies of existing solutions. Using this basis, a list of uninformative terms – termed a ***stoplist*** – will be compiled and used to process the most-frequent words in the Sydney Speaks transcripts. The word-counts will corroborate much of the topic information seen in the manual coding.

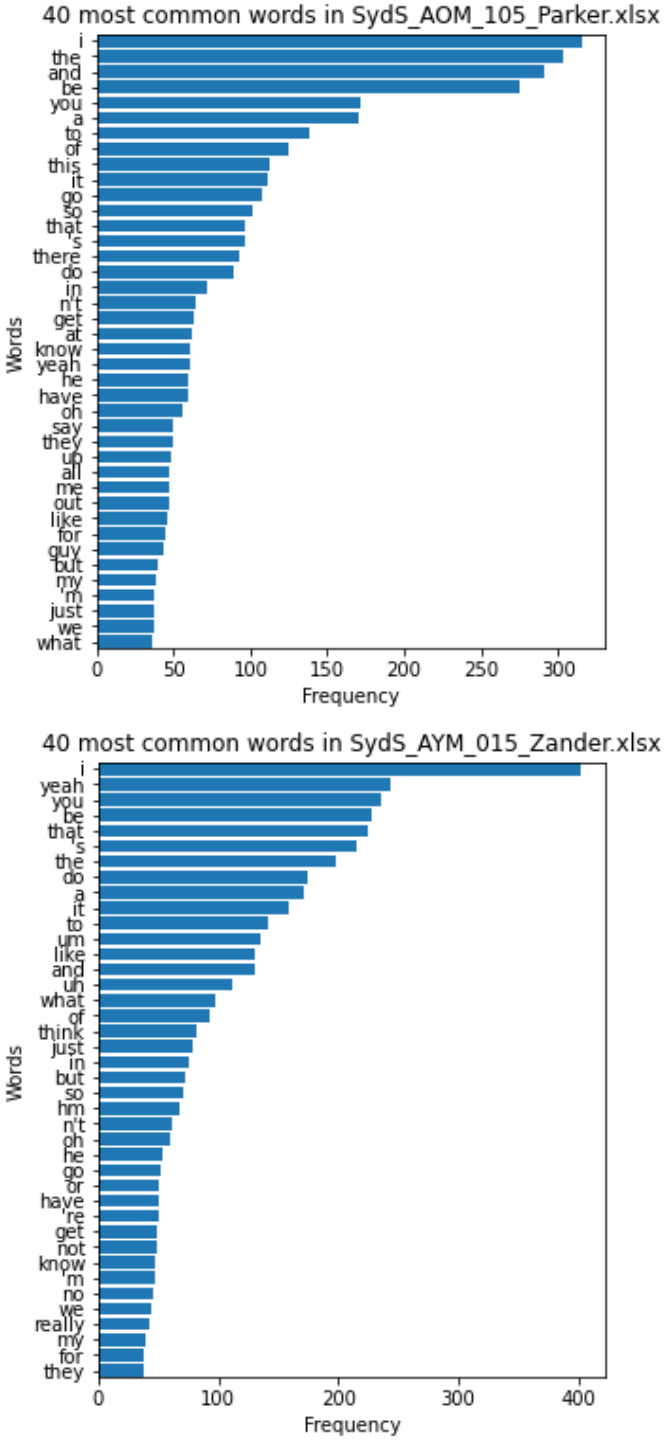
3.2 The Importance of Stoplists

To identify the most-representative words in a text, a preliminary approach is to identify the most-frequent words (cf. Xu & Zhang, 2021). However, while this may seem logically valid, it does not return very insightful information in-practice; quite simply, the high rates of grammatical words will guarantee that these items always dominate word counts, from any English text and regardless of the text's topics (cf. Hart, 1994). Clear examples of this are visible in Figure 3, which verifies that when no stoplists are used, the 40 most-frequent words provide no information as to the topics in each transcript. Also, note that the two word-counts share many of the same words (e.g. the

pronouns *I* and *you* appear in the top five, most-frequent items), despite these two transcripts covering quite different topics (cf. manual coding in Chapter 2).

An important stage, prior to applying any stoplist, involves lemmatisation. This refers to the removal of inflections from words, to collapse all inflected forms under one lexeme. Completing this is not too challenging, given Python's existing lemmatiser (NLTK, 2022a). As an example, Figure 3 displays the word *be* as the fourth most-common lexeme, across two, different transcripts. This entry, though, represents all inflected forms of the copula verb, rather than the infinitive form exclusively. The principle benefit to lemmatisation comes from providing a more representative figure of word frequency, as inflectional morphology does not change a word's class nor conceptual meaning. Also, by reducing inflected forms into one lexeme, this collapses the number of items needed on a potential stoplist, thereby reducing its length. An additional step which collapses the number of unique words (from a computer's perspective), is ensuring that all words are rendered in lowercase. This process is easily achieved using Python's "lower()" function (Bird et al., 2009).

Figure 3 *Top 40 most-frequent words in two transcripts from SydS 2010s, prior to applying any stoplists*



When faced with Figure 3, a solution is to create a list of certain words to ignore from frequency counts, such that texts which differ by content will

provide differing words. This solution, after all, relates to the fundamental idea of a *stoplist* (cf. Sinka & Corne, 2003). However, despite the stoplist's simple concept, the primary, linguistic challenge exists in deciding which words to include in one. Hence, stoplists naturally provide a unique task, insofar as determining which words contribute to semantic noise, which will vary depending on a user's research questions.

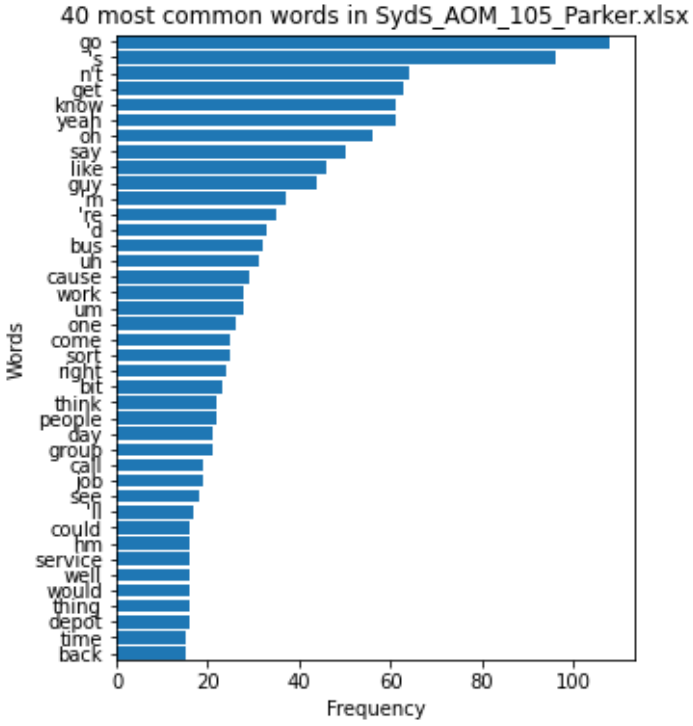
As the primary intentions for this thesis lie in topic modelling, redundancies are being defined as words which do not contribute to the ideas that a text discusses, nor its uniqueness. Uniqueness relates to the defining characteristics of a text's conveyed information. For instance, prepositions such as *to* and *of* can be used in any written or spoken piece, simply to encode grammatical relations between entities. Their use is equally valid whether a text explores children's games, or vocabulary differences between Australian and American English. Therefore, as these words are useless for distinguishing the areas of discussion within a text, they are justifiable as stopwords in topic modelling.

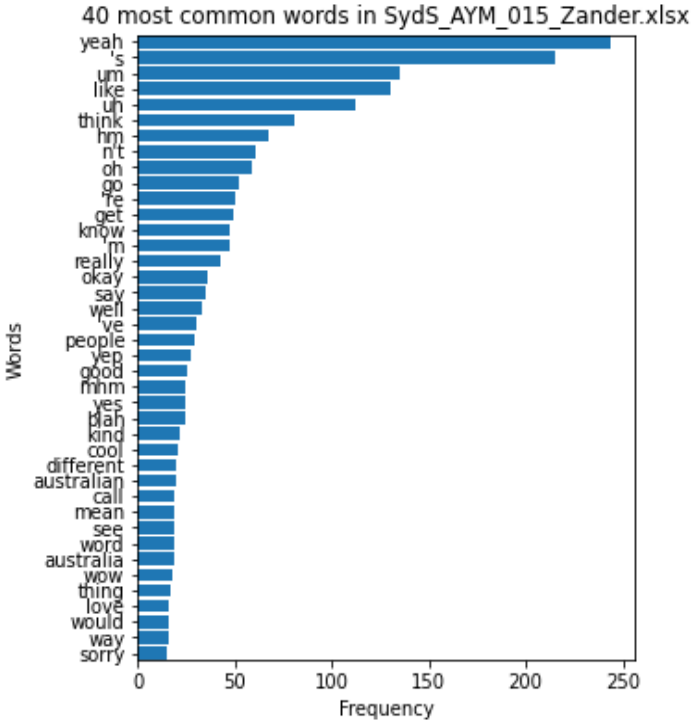
3.3 Existing Solutions

Currently, one of – if not *the* – most popular stoplists in natural-language processing for English, is the Natural Language Toolkit's (NLTK) "stopwords" list (cf. Bird et al., 2009). Traditionally, almost all existing stoplists have been designed around targeting specific parts-of-speech, including pronouns and

prepositions, due to their common and predictable occurrences across all English texts (cf. Sinka & Corne, 2003). NLTK's list of stopwords primarily consists of grammatical words, namely conjunctions, pronouns, determiners, prepositions, WH-words, modals, and copula verbs/inflections. While unproblematic for formal, written language, it fails to remove many items that frequently occur in conversation. Figure 4 presents the same word frequency-counts as Figure 3, though with the NLTK stoplist applied. The NLTK stoplist provides some improvements for topic extraction, with Parker only now containing the words *bus, job, service* and *depot*. Zander shows similar improvements, with the words *different, Australian, call, word* and *Australia* suddenly being present. However, many uninformative items are not removed, thereby hampering the ability to extract further topics. These items include clitic endings such as *'s* and *'re*, backchannels including *um* and *right*, semantically-light verbs such as *go* and discourse-markers such as *like* and *yeah*.

Figure 4 *Top 40 most-frequent words in two transcripts from SydS 2010s, once the NLTK stoplist is applied*





Therefore, given the noticeable issues with the NLTK stoplist, I will take its current entries as a basis, and then augment it considerably with a

conversational stoplist that I have developed for this thesis, based on the Sydney Speaks data.

3.4 Developing a Stoplist for Spontaneous Speech Data

3.4.1 How to Build a Better Stoplist

The following review examines literature for the design process of stoplists, including issues with applying an over-generalised stoplist without considering the corpus data or research question(s).

The first article to examine, is Hart's (1994) publication around cryptography and breaking simple ciphers; however, it also lists the most frequent words in English. Many of these words, such as *the, and* and *with* (full list is on page 103 of Hart (1994)), are semantically vacuous across all texts, due to the shared grammatical structure of the English language. Thus, it is worth ignoring these words when identifying topics in a transcript, as they do not contribute to a text's uniqueness.

The domain-specific nature of stoplist creation is highlighted in Lee et al. (2019). A stoplist, by its very construction, implies that some words provide no meaningful information, and should therefore be disregarded when determining relevant information. Meaningfulness is fundamentally dependent on a researcher's agenda, and how this relates to the source material. This view is supported by the authors' comment on the risks of over-generalised stoplists, in stating that generalised "lists are outdated and too

general for standalone or off-the-shelf use with domain-specific documents" (p. 1761). The authors employ two topic-modelling algorithms, Latent Dirichlet Allocation and Latent Semantic Indexing, and use their outputs to calibrate adequate stoplists. Regarding this, they state, "in our study, the words observed from the top-ranked topics [without any stoplist filtering applied] in the topic model are used to generate a stop words list because those words are considered more representative than the other words for the given topic" (pp. 1762 – 1763). In essence, due to the highly-frequent nature of stopwords, a topic-modelling algorithm will provide clusters of stopwords during initial runs, before the stoplist is compiled. Moreover, these remarks rely upon redundant words appearing frequently across all potential topics, such that their presence does not contribute to a topic's uniqueness nor identity.

Lastly, Sinka & Corne's (2003) article begins by noting that the creation of stoplists has largely existed in the domain of computational science and information technology. Previous research, referenced on page 396 of Sinka & Corne (2003), indicates that "over 50% of all words in a small typical English passage are contained within a list of about 135 common words" (cf. Hart, 1994). The five most common, are listed (in order) as *the*, *of*, *and*, *to* and *a*; the full, 135 word list is available on page 103 of Hart (1994). Since these grammatical words are common across many English texts, irrespective of topic, it is therefore reasonable to view them as uninformative, when identifying a text's semantic topics. Following data-extraction experiments

from web-pages, the authors conclude that “the topic of stoplists is worth revisiting” (p. 402), as “different stoplists lead to different benefits” (p. 402) and the choice of stoplist will yield varying results. Thus, the article strongly supports the idea of creating a customised, conversational stoplist, as a means of better processing spontaneous speech data.

3.4.2 Language Use in Spontaneous Speech

This section explores the role of epistemic constructions in spoken language. Specifically, comparing their valuable function as markers of speaker knowledge, with their high token-frequency and grammaticalised roles. These conditions present some unique challenges, as I will show that *epistemic phrases* are uninformative in terms of encoding topic-relevant information.

Drew’s (2018) article discusses the role of epistemics, in the establishment of information. The author approaches epistemics by noting that they are “not, in CA [conversation analysis] at any rate, about individual cognition, or about what individuals actually know or do not know. Put simply, it is about attributions of knowledge, to self and others, in the public domain that is constituted through interaction” (p. 165). This distinction between what each participant knows, and how the knowledge’s reliability is marked in speech, defines epistemic marking. Hence, epistemics are understandable as how a speaker communicates their perception and interpretation of worldly events and entities, rather than shaping or affecting them. Importantly, the property

of marking speaker perception is delegated to inflectional morphology in certain languages (cf. Quechua in Grzech (2021)), which supports the grammaticalised nature of epistemic constructions.

Since epistemic markers are, therefore, largely uninformative in providing topic-relevant information – owing to bleached semantics from grammaticalisation – the next stage is deciding which individual markers are worth excluding. Ford et al.'s (2014) paper presents the role of "remember" as an epistemic marker, as opposed to a transitive, cognitive verb (p. 122). The authors claim that "remember" is used by speakers "as an interactional marker of epistemic stance" (p. 122), but also to invite a listener's engagement/"affiliation" with the presented information (p. 122).

Similarly to Ford et al. (2014), Thompson's (2002) article notes that CTP (complement-taking predicate) phrases "provide an epistemic, evidential, or evaluative stance" (p. 155) towards their respective complements. Many of the items which constitute CTP-phrases, are provided in a list of grammaticalised markers (Thompson, 2002:138). The list details words which the author found to be uninformative, owing to their high token-frequency and grammaticalised nature. The data came from a corpus of 13 conversations (unguided exchanges, rather than sociolinguistic interviews) between adult speakers of US English (p. 126). During trial-runs of the Sydney Speaks data, many markers from Thompson (2002:138) were found to be equally

uninformative, and are therefore excluded for the purposes of topic modelling. These markers are *feel, guess, know, see, suppose, sure, say, look, like, mean* and *seem*.

Having established that a set of epistemic/grammaticalised markers should be excluded for topic modelling, it is fitting to explore whether other words deserve a similar exclusion, and why. This idea will be explored in the following section.

3.4.3 Approach

From multiple studies in discourse-pragmatics, it is clear that the use of language in informal, spoken contexts, differs quite significantly from more formal, written texts (cf. Thompson, 2002; Ford et al., 2014; Lee-Goldman, 2011). Departures in phrasing and sentence structure, are also present in word choice and usage. As an example, many verbs which act as complement-taking predicates (e.g. *think*) are used as epistemic markers in spoken contexts (cf. Drew, 2018; Ford et al., 2014). Bou Orm's (2021) work, quite notably, focused on the frequent use of *think, feel,* and *guess* as epistemic markers in the *Sydney Speaks Project*. Epistemics do not provide information about a conversation's semantic contents; instead, they indicate stance or degree of commitment, and function as a kind of discourse marker. Another discourse marker that must be removed is *like*.

Two other categories worth removing are all semantically *light* verbs, and backchannels. The former are highly frequent verbs which, as I discovered through examining numerous transcripts, commonly form compound verb-constructions, such as *go*, *make* and *come*. They are categorised as light verbs due to their semantically-bleached and grammaticalised nature, which correlates with their high frequency. Words such as *like*, however, which are employed as approximators in informal, spoken language (in the case of *like*, the additional roles of being a verb of sentiments, or quotative), belong in the discourse marker category (cf. Lee, 2020; Winter, 2002). Backchannels, including *um* and *hm*, constitute words which confirm the receipt of information by a listener (cf. Gardner, 1998). While interesting from a conversation-analyst perspective, they do not provide information about a discussion's topics. Hence, they should also be removed from word-counts. In total, I have identified a total of six, stopword Categories, which I list in Table 4.

The Categories refer to the function of a given stopword, and hence its reason for being filtered. Conversational and Interactional stopwords represent items unique to processing informal, spoken language, inasmuch as designating highly-frequent words common to all spontaneous, spoken data. Number stopwords pertain to number words, such as *one* (as used in *one of*), which do not contribute any topic-relevant information. Grammatical and Other refer to items that should always be filtered, including in formal, written language.

Thus, many of these latter items represent gaps in NLTK's existing stoplist. The remaining entries apply more specifically to informal, spoken data.

The final Category, Topic Modelling, refers to words which through trial-and-error during topic modelling, did not add any meaningful or revealing information to clusters. These can all be described as abstract nominals or adjectives, such as *people* and *good*. In addition to the six Categories, each stopword can be further classified into one of 16 S*ubcategories*, thereby giving a large degree of specificity for filtering. The entire, Category structure is presented in Table 4, with a word from each Subcategory shown as an example. In some cases, a Category only contains words from one Subcategory, even if the Subcategory in-question is present under multiple Categories (e.g. *Vacuous Adjective* under *Numbers* and *Topic Modelling*). Moreover, a Subcategory might only belong to one Category (e.g. *Backchannel* under *Interactional*). While these instances may appear to be redundancies or overlap, they are retained to highlight the importance of *word function* when assigning Subcategories, as opposed to the more generic C*ategories*. The full stoplist is included in Appendix 2.

Table 4 Categories and Subcategories in the conversational-data stoplist

Category	**Subcategories**	**Example Entries**
Conversational	Discourse Marker, Epistemic, Functional, Light Verb, Vacuous	*like, know, wanna, go, stuff*
Grammatical	Adverb, Conjunction, Contraction, Modal, Preposition, Vacuous Pronoun	*certainly, whether, 'll, might, among, everyone*

Interactional	Backchannel	*um*
Numbers	Vacuous Adjective	*one*
Other	Letter, Truncated	*k, et*
Topic Modelling	Vacuous Abstract, Vacuous Adjective	*people, good*

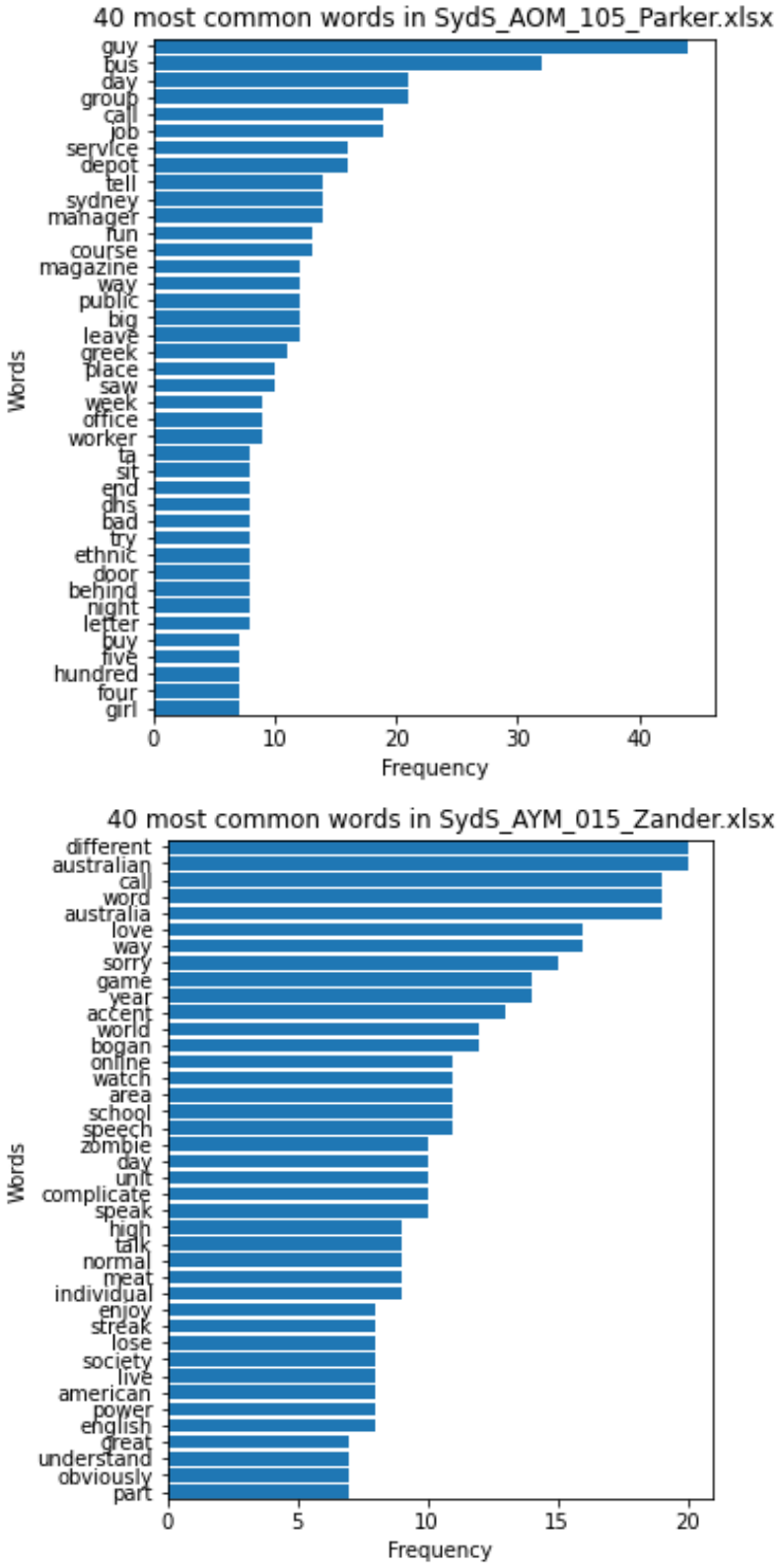
The assignment per Subcategory, provides a key benefit to users by enabling control over which parts-of-speech to filter. This ensures a great amount of flexibility for the stoplist's use, as a user can easily adjust the filtering for varying language-data or research questions.

3.5 Applying the Conversational Stoplist

Once the conversational stoplist is applied alongside the NLTK stoplist, it is possible to control word-counts for a large degree of uninformative material. To demonstrate this, Figure 5 shows how the most-frequent words from different transcripts vary noticeably in contents.

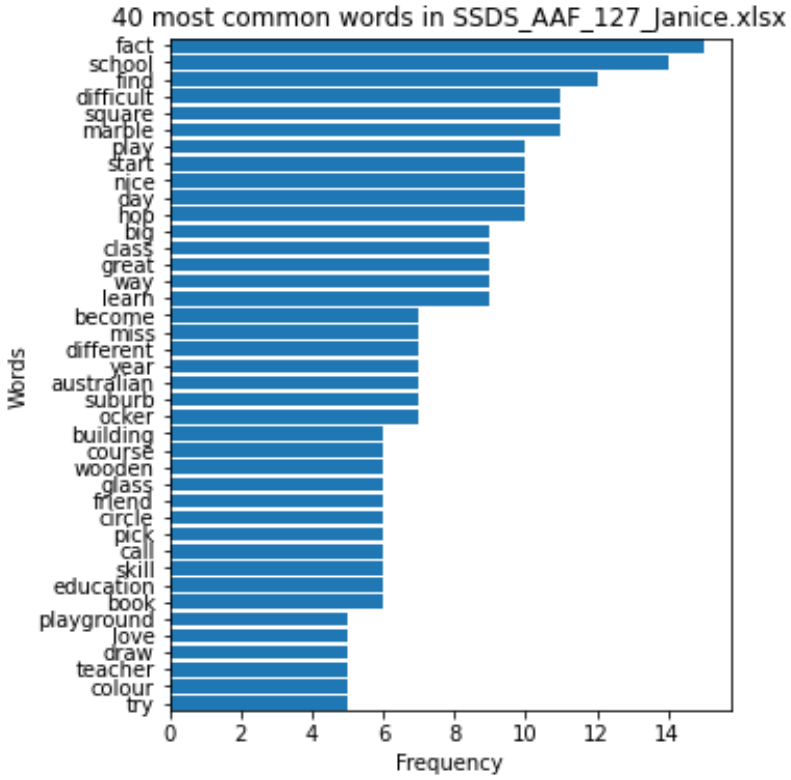
The combined conversational- and NLTK-stoplists capture topics that appear in these transcripts, with minimal noise. Notable examples include Parker's mention of bus-related work (*bus, service, depot, manager*), and Zander's discussion on Australian-American vocabulary differences (*different, Australian, call, word, Australia, accent, American*). Additional topics include Parker's public-service work (*service, public, DHS*), and Zander's interests in gaming (*game, online, zombie*).

Figure 5 *Top 40 most-frequent words in two transcripts from SydS 2010s, once the NLTK and Conversational stoplists are applied*



Further results from the conversational stoplist are visible in the most-frequent words for the three other transcripts; Figure 6 shows results from SSDS data, and Figure 7 presents an example from the BCNT transcript. Figure 6 verifies the presence of topics from Chapter 2's manual coding, including how Janice discusses games (*marble, square, play*), dialectal variation (*suburb, ocker*), school-life (*school, class, learn, teacher, playground*) and appreciating glass (*glass, colour, nice*). Additionally, Isa's manually-coded topics of school-life (*teacher, school, nickname*), Greek afternoon-school (*teacher, school, Greek, speak, lesson*) and language use (*language, English, speak, word, different*) are also conveyed. Similarly, Figure 7 supports the manual coding's assessments that Marjorie discusses familial relations (*mother, grandfather*), singing (*sing, choir*), radios and broadcasts (*radio, hear*), and charities/missions (*Sydney, mission, Saint, Vincent*).

Figure 6 Top 40 most-frequent words in two transcripts from SSDS, once the NLTK and Conversational stoplists are applied



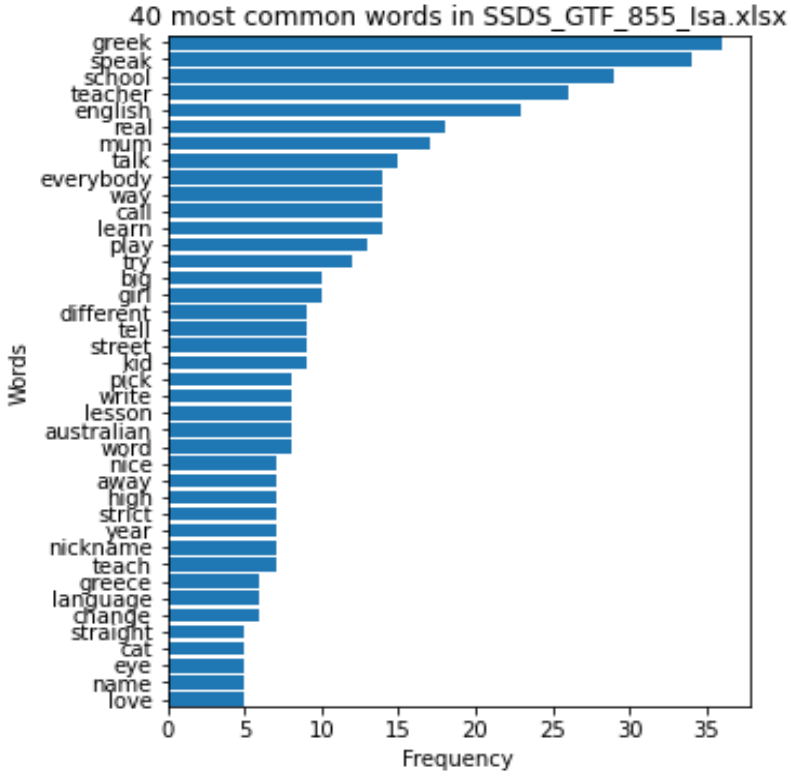
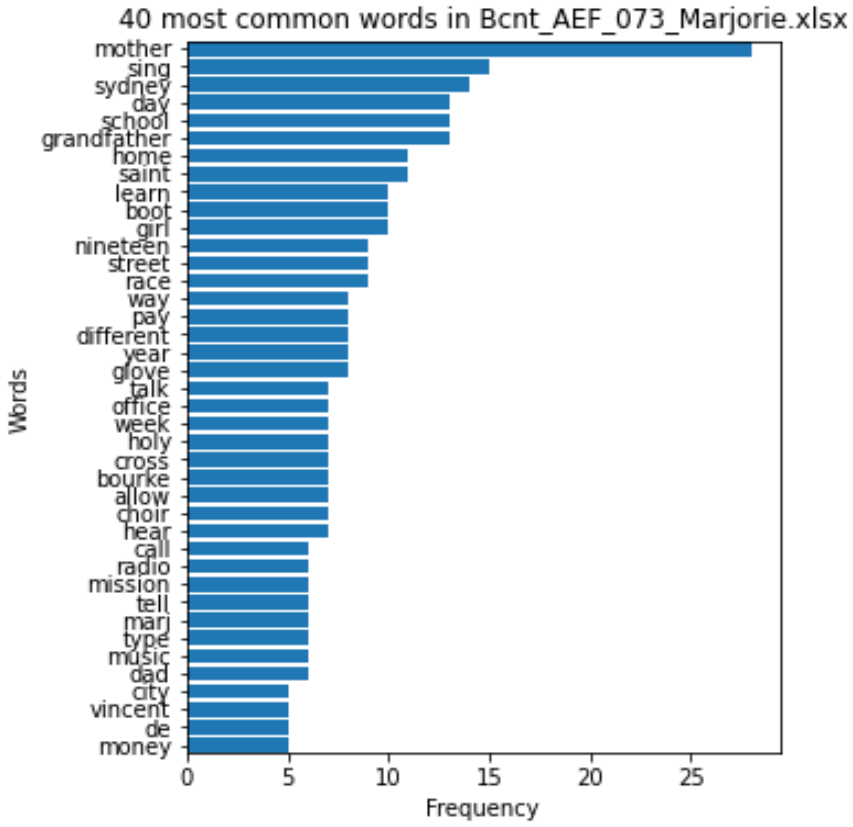


Figure 7 *Top 40 most-frequent words in a BCNT transcript, once the NLTK and Conversational stoplists are applied*



3.6 Conclusion

To conclude, the word-frequency counts demonstrate that once an appropriate stoplist has been applied, the 40 most-frequent words reveal clues about the topics in a transcript, in accordance with the information identified by manual coding. Since the removal of stopwords distils transcript data into a set of meaningful words, the following chapter will examine the use of topic-modelling, cluster algorithms, to explore how the individual words interact with each other in the transcripts.

Chapter 4- Topic-Modelling Partitioning and Optimisation

4.1 Introduction

This chapter follows the automation path established by Chapter 3. However, it explores the intricacies of topic-modelling cluster algorithms, by optimising cluster generation through applying (1) linguistic theory to identify topic-rich sections of text, and (2) automated scores for semantic coherency, termed "coherence scores".

4.2 Partitioning a Text

A topic-modelling method should identify the correspondences between words – namely, co-occurring words which correspond to a similar semantic-frame under one topic. In this thesis, I use LSI as the underlying, topic-modelling algorithm, as outlined in Chapter 1. Here, I explore LSI – both *without user-modifications* (i.e. as applied in the current literature), and with certain modifications in the form of *text partitioning*. The unmodified version identifies the specified number of clusters across the whole document; whereas, the latter treats each partition as its own, separate document, and identifies one word-cluster per partition.

The need for such partitions, comes from two observations. Firstly, running unmodified LSI on Sydney Speaks data – as seen in Appendices 4 to 6 – has yielded poor results, with many incoherent clusters. Secondly, the manual-

coding revealed the tendency for words from the same topic to co-occur in groups, and thus I hypothesise that the LSI algorithm will perform better *if* it is given chunks of texts within which to identify topics, whereby each chunk represents a single topic. The manual-coding also revealed that topics come in various lengths, as demonstrated in Chapter 2. As such, if the partitioning approaches can recognise this variability (through generating partitions of *varying sizes*), the resulting clusters should align closer to the idealised, human benchmark.

Ideally, the partitions should be drawn in linguistically-meaningful ways, to return more coherent results. To test this, I have developed two partitioning algorithms, each applying a different criteria for where to draw partition boundaries. The first method uses interviewer turns as potential boundaries, and the second creates boundaries with a list of designated, pragmatic phrases. Since partitions can vary in size (depending on where the boundaries are drawn), those which contain too few words to run LSI will be returned directly to the user; in these cases, the number of words does not exceed single digits. A third algorithm has also been created, which merely draws boundaries to create evenly-sized partitions, irrespective of the boundaries' relations to contents. It is to serve as a control, against which to compare the two other methods. The performance of these algorithms – as judged by their partition locations – will be compared to the manual coding.

Lastly, it is noteworthy that Guo et al. (2021) is the only previous research to raise the concept of partitioning in topic modelling. The authors modified Latent Dirichlet Allocation (LDA) to process long texts by individual segments. My partitioning differs from them by (1) implementing LSI, instead of LDA, and (2) their algorithm uses paragraphs as partitioning boundaries, in formal, written documents (p. 334).

4.2.1 Algorithm for Even Partitions

The first partitioning algorithm, named Even Partitions, works by dividing a transcript into evenly-sized partitions ("even", in terms of word-count per partition), and then provides each partition individually to LSI. As an example, if four partitions are specified, then three boundaries need to be placed. The algorithm will count the total number of words in the transcript, and then place boundaries to divide it into four, equally-sized partitions by word-count. In the event that the number of words is not divisible by four, the final partition will contain the remainder.

The concept behind this algorithm, is to test the effects of converting a whole transcript into smaller *mini-transcripts* of equal size, and thereby act as a control for the Turn and Pragmatic Partition algorithms, to be discussed in the following sections.

4.2.2 Algorithm for Turn Partitions

The Turn Partitions algorithm relies on the idea that longer contributions from *interviewers* are more likely to elicit greater information from a participant – as measured by a longer reply – when compared to shorter equivalents. Shorter contributions are easier to associate with backchannels and feedback, and a continuation by the participant of the same topic. Note that the number of IUs, regardless of their word contents, is the metric I am using for *length*. This decision comes largely out of practicality for large-scale applications, as IUs present data in a format which divides sections of speech by speaker.

Evidence to support the significance of long interviewer-turns, comes the Sydney Speaks data's basic statistics. Using the same five transcripts from which I have been illustrating throughout, the mean, median and maximum turn-lengths have been calculated for interviewers and participants. The results in Table 5 demonstrate that participants overwhelmingly produce longer turns, across all metrics. In particular, the median turn-length of 1 for interviewers supports the idea that most interviewer contributions are backchannels, or other confirmatory remarks. This can be seen in that, for example, out of the total of 553 single-IU interviewer contributions, the most common are *hm* (31%), *yeah* (13%), *mhm* (7%), and other more minor forms, such as *okay* and *yes*. While the statistics about the most-common contents for single-IUs are for interviewers only, the key finding is the absence of topic-

prominent material within them, *when combined with* their greater occurrence from interviewers. Therefore, multi-IU turns for interviewers are noteworthy (e.g. 28 under "maximum"), as they are likely to be more than mere expressions of emotions or engagement.

Table 5 Interviewer vs. Participant Turn-Lengths by IU Count

Speaker Role	**Mean**	**Median**	**Maximum**
Interviewer	2.04	1	28
Participant	6.81	3	120

The idea that longer turns elicit more contentful replies is supported by existing literature. Clayman (2012:160) notes that extensions to syntactically-complete turns can serve to minimise disagreement in oral exchanges (and thereby, encourage more input), with Ford & Thompson (1996:168) describing extensions to turns as being driven by a "pursuit of uptake" from a recipient.

The Turn Partitions algorithm works to identify the longest interviewer-contributions and use them as boundaries. Therefore, the algorithm's technical specifications are as follows. Using the number of specified topics (N), and the number of boundaries required to generate partitions for N topics (N-1), it locates the N-1 longest interviewer statements in the transcript, and draws boundaries at these points. As an example, if four partitions are specified, then three boundaries need to be placed. The algorithm will first identify the longest, continuous stretch of interviewer speech, and draw a boundary at the initial IU's location in that turn. Taking the interviewer's initial

IU allows the algorithm to capture the beginning of the new topic's presentation, as the interviewer contribution is likely to contain topic-relevant words. The algorithm will then proceed with the second and third longest-stretches of interviewer contributions. With the four partitions determined, the LSI algorithm is run in each partition individually.

4.2.3 Algorithm for Pragmatic Partitions

The Pragmatic Partitions algorithm relies on the hypothesis that certain discourse-pragmatic features are used by speakers to mark where topic-changes occur. Whereas written language employs paragraph breaks, these are not explicitly marked in speech, and speakers make use of other pragmatic devices to signpost changes in discourse.

There is no established list of pragmatic markers which signpost topic changes. Thus, this partitioning algorithm is necessarily exploratory, and further research into the role of words and phrases in eliciting topic-changes (including in future theses) is needed. I have identified signposting phrases through examination of the Sydney Speaks data. However, there are a few articles which offer some support to this approach, including Ford & Thompson's (1996:166-169) mention of turn-initial markers being used to display disagreement with an ongoing discussion and thereby shift its perspective, such as the question-word *what* (p. 169) and discourse-marker *so* (p. 166). Ford et al. (2014) also *indirectly* reference topic-changing

constructions, namely the question-marker *do you* (p. 126). Ford et al. (2002) detail the use of *oh* in providing "disaffiliation between the position taken by the previous speaker and the position the current speaker is about to adopt" (p. 197), and further note that "*oh*-prefaced responses to questions […] often embody a challenge to their relevance" (p. 197). In other words, a speaker may use *oh* to redirect or clarify the contents of a conversation, by signalling an upcoming question with more pragmatic-markers. Examples of *what, so, do you* and *oh* being used to mark topic shifts (or otherwise extract information from a recipient) in Sydney Speaks data, are visible in Examples 5 to 8.

(5, SSDS_GTF_855, 00:23:55.245 – 00:24:05.221, what appears to change the topic from recreation to teachers)

PNT: Isa	[yeah I l]oved the monkey [2bars2].
INT: Susan	[2mad on2] the gymnastics.
PNT: Isa	yeah.
INT: Susan	... **What** about your teachers.
INT: Susan	do you remember them?
PNT: Isa	.. Yeah,
PNT: Isa	.. u=m,
SP1: Female	XXX.
PNT: Isa	...(1.0) they were- % --
PNT: Isa	...(1.0) they were good.

(6, SydS_AOM_105, 00:53:36.483 – 00:54:02.495, so appears to change the topic from workplace assaults to languages)

PNT: Parker	...(2.0) now I'd never had anyone try to punch me out in the workplace before.
INT: Alice	No,
INT: Alice	that's @ --
INT: Alice	@unbelievable isn't it.
PNT: Parker	<@ I'm just li[ke] @>,
INT: Alice	[oh] my gosh.
PNT: Parker	... <@ sure I've been called a few names in my time,
PNT: Parker	but @>,
INT: Alice	Ye=ah.
PNT: Parker	...(1.0) The place is just ridiculous.
PNT: Parker	t- it's a shame.
PNT: Parker	it's just it just it --
PNT: Parker	it's a disgrace.
PNT: Parker	it just --
PNT: Parker	...(1.0) yeah.
INT: Alice	.. **So** I'm guessing in that --
INT: Alice	the time you were there,
INT: Alice	.. you must have heard,
INT: Alice	...(1.0) (TSK) a lot of other languages.
PNT: Parker	... Oh yeah.
PNT: Parker	.. Yeah,

(7, SSDS_AAF_127, 00:54:01.784 – 00:54:13.631, do you appears to change the topic from accent accommodation to ocker culture)

PNT: Janice	.. they --
PNT: Janice	.. I guess,
PNT: Janice	try and be more international and less,
PNT: Janice	... <VOX how're you going mate alright and meat pies VOX>.
PNT: Janice	[@@]@@@@@
INT: Angela	h[m].
PNT: Janice	[2XX2].
INT: Angela	[2**Do you** think2] ocker really exists.
PNT: Janice	... oh yes,
PNT: Janice	.. yes I do.

(8, SydS_AYM_015, 00:24:36.873 – 00:24:43.206*, oh appears to frame an upcoming question)*

INT: Heather	... that's really c- --
INT: Heather	that's really um,
INT: Heather	inspiring.
INT: Heather	.. Oh.
INT: Heather	... I want to ask you what --
INT: Heather	what was the prejudice that you felt.

Another proposed, topic-changing marker is *no*, when used "as a marker of topic shift" and a framing construction for the "projected action" of a topic shift (Lee-Goldman, 2011:2632). In essence, while *no* may not be the *sole* pragmatic-marker of a topic shift, it foreshadows the eventual pragmatic action of changing the topic. Evidence of this marker fulfilling a similar role in the Sydney Speaks data is visible in Example 9, in which the participant seemingly signals "outside of that" with it.

(9, SydS_AYM_015, 00:41:26.300 – 00:41:46.579, no appears to frame an eventual topic shift)

PNT: Zander	.. Well what's life without a bit of flair.
PNT: Zander	@@@@[@@@@]
INT: Heather	[Oh yeah well,
INT: Heather	.. I] --
INT: Heather	y- y- you- --
INT: Heather	you give me no e- --
INT: Heather	no way to .. respond to that other than,
INT: Heather	oh yeah I get [that.
PNT: Zander	[@@@@@
INT: Heather	@@@@@@
PNT: Zander	@@@@@
INT: Heather	@Yeah,
INT: Heather	I get @that.
INT: Heather	@@@@
PNT: Zander	**No** but u]m,
INT: Heather	(SNIFF)]
PNT: Zander	.. **outside of that**,
PNT: Zander	uh,
PNT: Zander	... I have a --
PNT: Zander	I'm d- also doing a= .. unit called .. Philosophy --

Through listening to the available transcripts and taking note of other discourse-pragmatic markers used in topic-shifts, it was apparent that a wide set of items was being used. Thus, I have compiled the following list of markers in Table 6, which I term Topic Transition Relevance Phrases (TTRPs). Each TTRP is assigned a numerical score, to represent its likelihood in triggering a topic change, in addition to being assigned a "Category" to justify its inclusion. Also, note that all TTRPs are recorded using fully uninflected (i.e. lemmatised) forms (as the transcript is lemmatised prior to checking for

TTRPs), and hence “do you” accounts for “do you” and “did you”, along with “I be think” realised as “I am thinking” and “I was thinking”.

Table 6 All TTRPs listed with their relevant Category and Score

Category	**TTRPs**	**Score**
Discourse Marker	*oh, now, nah yeah*	2
	so, no, yeah nah	4
Literal Reference	*intrigue*	4
	topic, subject	16
Proposition	*for instance, in all seriousness*	2
	I be think	4
	outside of that, speak of, we be talk	8
	think back	16
Question	*whereabouts*	2
	what, have you, how, can you	8
	do you, to ask you, be you	16

The scores were determined intuitively after exposure to the transcripts, having observed that for the most frequent tokens, their rate of usage to mark solely topic-shifts was exponentially higher than for other, more flexible markers (e.g. *oh* as a backchannel). This intuition is supported by statistics from the Sydney Speaks data, as, out of 1456 total instances of TTRPs, the four most-frequent are *so* (19.4%), *oh* (15.7%), *what* (15.6%) and *no* (11.2%). *So, oh* and *no* have more flexible pragmatic-roles than *what*, and are therefore assigned lower scores. *What,* despite being highly-frequent, is an unambiguous interrogative-marker and therefore gets a higher score of 8.

In accordance with Zipf’s Law (cf. Zipf, 1935) – a cross-corpora trend that as certain words become more frequent, the frequency-counts increase

exponentially (often, in powers of two) – I have used powers of two to encode exponential, yet naturalistic increases in topic-change usage. This is not to be confused with token frequency, which does not encode a TTRP's pragmatic role in context.

The Pragmatic Partitions algorithm divides a document into a specified number of partitions, and places boundaries in locations where three TTRPs co-occur. I have chosen to examine three TTRPs per group, as this provides a balance between understanding topic-changes as arising from acts of pragmatic signalling (which lends itself to examining multiple TTRPs), versus the risk of misplacing a boundary beyond where a topic-change has actually occurred, should some TTRPs occur in an IU that is beyond a speaker's first introduction of a new topic. Example 10 visualises how a single partition would be drawn when five TTRPs are present.

(10, SydS_AYM_015, 00:24:30.120 – 00:24:53.007, a transcript extract with the relevant TTRPs presented in isolation with their scores, and the boundary's location shown with blue highlighting)

PNT: Zander	...(1.5) I became a Christian,		
PNT: Zander	I just ...(1.5) gave up that prejudice.		
INT: Heather	Okay,		
INT: Heather	yeah,		
INT: Heather	yeah.		
INT: Heather	... that's really c- --		
INT: Heather	that's really um,		
INT: Heather	inspiring.		
INT: Heather	.. Oh.	*oh*	2
INT: Heather	... I want to ask you what --	*to ask you, what*	16, 8
INT: Heather	what was the prejudice that you felt.	*what*	8
PNT: Zander	...(2.0) Um,		
PNT: Zander	...(1.0) very similar to Pauline,		
PNT: Zander	uh Pauline Hanson.		
PNT: Zander	... [As in],		
INT: Heather	[Okay].		
PNT: Zander	we're being overrun,		
PNT: Zander	.. uh,		
PNT: Zander	blah blah blah.		
PNT: Zander	... [etcetera etcetera].		
INT: Heather	[Oh].	*oh*	2

The possible groups are *oh, to ask you* and *what* (score = 2+16+8 = 26), *to ask you, what* and *what* (score = 16+8+8 = 32), and lastly *what, what* and *oh* (in the excerpt's last line; score = 8+8+2 = 18). Note that as currently implemented, the algorithm is unaffected by the number of intervening IUs between TTRPs – hence the final cluster includes two items in adjacent IUs and a third 10 IUs later. Therefore, a partition will be drawn at the IU marked in blue, with a score of 32.

To illustrate as with Turn Partitions, let us consider a case where a user specifies four partitions. We therefore require three boundaries to be drawn. The algorithm will first identify where all TTRPs exist in the transcript, and

then calculate which group of three TTRPs has the highest *combined* score. Once this group is located, a boundary will be drawn at the final TTRP's location (as the group's end-point), since the end-point represents the theoretical completion of topic-change signalling in discourse. This process would then repeat identically for the second- and third-highest scoring groups. With the four partitions now determined, the LSI algorithm is run in each partition individually.

4.2.4 Pseudo-Code for Partitioning Algorithms

Having presented the three partitioning algorithms, Figure 8 contains their algorithms in pseudo-code (i.e. coding blueprints).

Figure 8 Combined Pseudo-Code for the Partitioning Algorithms

1. Obtain a list of stopwords. This list should only contain lemmatised forms. *Do not* filter anything yet from the transcript!

2. Lemmatise all wordforms in the transcript (i.e. remove plurals, verbal inflections, etc.)

3. Specify the number of partitions (a.k.a. topics), to be denoted as N.

4. Partition the transcript using the criteria in 4a for Even Partitions, 4b for Turn Partitions and 4c for Pragmatic Partitions

4a. For Even Partitions:

I. Create a matrix with all words in the transcript.

II. Divide the text into evenly-sized partitions ("evenly-sized" as per identical word token-counts in each partition). Should the number of words not be divisible by the number of partitions, include remainders in the final partition.

4b. For Turn Partitions:

I. Create a matrix with all words in the transcript, and note which IUs come from the interviewer and participant. This can be marked with designated boundaries at the end of each IU.

II. Find turn-clusters where interviewers speak the longest and uninterrupted, rank them by the number of interviewer turns. Using N from Part 3, draw N-1 boundaries for the N-1 longest interviewer contributions, with the boundaries at the beginning these sections (where a new topic is introduced).

4c. For Pragmatic Partitions:

I. Obtain a list of TTRPs and their respective scores (i.e. powers of 2).

II. Assign "nodes" to the TTRPs that appear in the transcript. Each node records two pieces of information – its immediate neighbours and "score" (where the scores form a discrete set, with powers of 2). The larger the score value is, the more likely that TTRP is to mark a topic shift.

III. Link each node to its immediate neighbours ("immediate" defined as the two, closest nodes within an IU, or otherwise the closest nodes beyond the current IU) and do the following steps:

a) Calculate the sum of node clusters (with 3 nodes per cluster), and store the sum values – remembering where each "sum" corresponds to in the transcript (e.g. if a transcript contains four nodes – A, B, C and D, in that order by in-text proximity – first calculate the sum for A + B + C, and then do B + C + D).

b) Using the number of partitions and cluster sums, place topic boundaries at the end of the highest scoring clusters. For N topics, place boundaries at the N-1 highest scoring clusters.

5. Feed the partitions into LSI individually, and return the clusters to the user.

4.2.5 Comparing Partitioning Algorithms and Manual Coding

Having established multiple approaches to partitioning a transcript by topic contents, we will now compare these to the human benchmark. To complete this, each partitioning algorithm was run on the five manually-coded transcripts, and programmed to list all partitions. The number of partitions was fixed across the three algorithms, as the number of manually-extracted topics per each transcript. To compare the degree to which these partitions corresponded with the manual topic-clusters, every partition was given a score (on a scale from 1 to 4), depending on its correspondence with its alignment to the manual topics. Partitions corresponding fully to the human topic-clusters (i.e. with no words from one topic on either side of the partition as in Figure 9), were given a score of 4. If 1 or 2 words overlapped on either side, a score of 3 was assigned (illustrated in Figure 10). Scores of 2 were used for 3 or 4 overlapping words (see Figure 11), with more than 4 words leading to a score of 1, seen in Figure 12, where the automatic partition falls in the middle of a manually-identified topic. Also, if a topic transition was correctly identified (without overlap), but multiple boundaries were inserted within the transition (such that the gap between boundaries contains no extracted words), all boundaries but the first were given a score of 1; Figure 13 shows such an example of this.

Figure 9 A perfect alignment between manual and automated partitions (Score = 4, SydS_AYM_015, 00:24:52.164 – 00:25:24.901)

INT: Heath	[Oh].					
PNT: Zand	... Australia's no longer a white country,	Australia	white country			
PNT: Zand	etcetera etcetera.					
INT: Heath	... What do you think about that statement.					
PNT: Zand	... Um,					
PNT: Zand	... %					
PNT: Zand	... personally I don't really care.	care				
PNT: Zand	@@[@					
INT: Heath	[Yeah you don't care,	care				
PNT: Zand	.. Ye]ah,					
INT: Heath	ye]ah?					
PNT: Zand	it's just --					
PNT: Zand	.. it's just the way it is.					
INT: Heath	.. Yeah,					
PNT: Zand	[Li-] --					
INT: Heath	[what do you-] --				Pragmatic	4
PNT: Zand	life happens.					
INT: Heath	.. Yeah.					
INT: Heath	... Yeah yeah,					
INT: Heath	yeah that --					
INT: Heath	.. that makes sense.					
INT: Heath	So what do you think Australians are	Australians				
INT: Heath	.. sh- are- --					
INT: Heath	are known for around the world.	known (for	world			
PNT: Zand	...(4.0) Hm.					
PNT: Zand	...(1.5) Around the world.	world				
PNT: Zand	...(2.5) Well,					
PNT: Zand	.. bogan's the first word that comes to mind.	bogan				
INT: Heath	.. Really.					

Figure 10 An alignment which misses the topic-change by one word, and is, therefore, assigned a score of 3 (Score = 3, SydS_AYM_015, 00:33:12.947 – 00:33:40.192)

PNT: Zand	...(1.0) Because meat.	meat				
INT: Heath	...(1.0) You've [got meat,	meat				
PNT: Zand	[@@]					
INT: Heath	why] have veggies when you can have meat.	veggies	meat			
INT: Heath	.. [2Fair enough2].					
PNT: Zand	[2That's true2].					
PNT: Zand	@@[@]				Turn	3
INT: Heath	[I understand] that.					
INT: Heath	...(1.0) I do love my meat,	meat				
INT: Heath	that's for sure.					
INT: Heath	I can't say I'd- --					
INT: Heath	.. can't deny that.					
INT: Heath	... Um,					
INT: Heath	I'm just --					
INT: Heath	... thinking back .. be- before,					
INT: Heath	something that we --					
INT: Heath	a lot of Australians do,	Australians				
INT: Heath	I don't know if you've noticed this.					
INT: Heath	... is that um when --					
INT: Heath	you said flip flops,	flip flops				
INT: Heath	and then you --					
INT: Heath	.. you corrected yourself saying no thongs,	corrected	thongs			
INT: Heath	we're not American.	American				
INT: Heath	and then I laughed because I totally agree with you.					

Figure 11 Two alignments from <u>different</u> algorithms, which miss the topic-change by three words, and therefore both are assigned scores of 2 (Score = 2, SydS_AYM_015, 00:35:08.302 – 00:35:45.204*)*

PNT: Zand	even looking at its bare essential,					
PNT: Zand	uh,					
PNT: Zand	coming down to me as an individual.	individual				
PNT: Zand	... I've always enjoyed being different from everybody else.	differently				
INT: Heatl	.. Okay ye[ah].					
PNT: Zand	[I've] always been- --					
PNT: Zand	enjoyed the spotlight.	spotlight				
PNT: Zand	... so,				Turn	2
INT: Heatl	.. Yeah.					
INT: Heatl	.. No I --					
INT: Heatl	I f- --				Even	2
INT: Heatl	.. I felt that way too about my life.					
INT: Heatl	I've ... wanted to --					
INT: Heatl	...(1.0) be a lot different to the way other people do things,	differently				
INT: Heatl	and do things because I --					
INT: Heatl	... I have a purpose or a reason for doing them.	purpose	reason			
INT: Heatl	... So I get that,					
INT: Heatl	yeah.					
INT: Heatl	.. Um,					
INT: Heatl	... (TSK)					
INT: Heatl	...(1.0) I was just thinking about things that we um,					
INT: Heatl	... have happened in the um,					
INT: Heatl	.. in ar- --					
INT: Heatl	in around us in Australia,	Australia				
INT: Heatl	that um,					
INT: Heatl	are very significant parts of um,	significant				
INT: Heatl	...(1.0) (TSK) (TSK)					

Figure 12 An alignment which entirely misses any topic-change by three words, assigning it a score of 1 (Score = 1, Bcnt_AEF_073, 01:21:39.048 – 01:22:22.954)

PNT: Marj	...(1.5) Mother said now have you got everything ~Marj,	everything				
PNT: Marj	your ke- --					
PNT: Marj	.. kes- and kes-,					
PNT: Marj	... yes Mum,					
PNT: Marj	she said,					
PNT: Marj	.. no you haven't got ~Marjorie.					
PNT: Marj	... and I said why,					
PNT: Marj	what have I forgotten Mum,					
PNT: Marj	she said your gloves.	gloves				
PNT: Marj	...(1.0) I had to sit down and put my gloves on.	gloves				
PNT: Marj	... yes <X put your X> --					
PNT: Marj	... what --					
PNT: Marj	what did you have to go and p- t- --					
PNT: Marj	.. put your gloves on for and I said,	gloves				
PNT: Marj	.. <Q I must never go out with gloves Q>,	gloves				
PNT: Marj	and do you know it's one of the things I still can't do,				Pragmatic	1
PNT: Marj	is go out without gloves.	gloves				
PNT: Marj	... but uh Mother always wore gloves,	gloves	mother			
PNT: Marj	.. I can remember --					
PNT: Marj	... going shopping with Mother and she always had gloves on.	gloves	mother			
PNT: Marj	... Mother was very tall,					
PNT: Marj	and very austere-looking p- person,					
PNT: Marj	very good-looking,					
PNT: Marj	not					
PNT: Marj	.. not like me at @all @@,					
PNT: Marj	she was quite different to me altogether.					
PNT: Marj	... and,					
PNT: Marj	.. and um,					
PNT: Marj	...(1.0) she wore those uh,					
PNT: Marj	...(1.0) high,					
PNT: Marj	... button-up boots.	Button-up boots				
PNT: Marj	... and I always wore boots,	boots				
PNT: Marj	.. which Grandfather had made for me.	Grandfather				
PNT: Marj	.. with twenty-two eyelets in the holes you know n- trrr.	Twenty-two eyelets				

Figure 13 *Scoring of partition boundaries placed adjacently, when no overlap is present (SydS_AYM_015, 00:21:03.963 – 00:21:36.371)*

INT: Heath	... um,					
INT: Heath	of trying to .. ban --	ban				
INT: Heath	... um,					
INT: Heath	Muslims entering --	Muslims				
INT: Heath	or people from ... high terrorist countries like,	terrorist	countries			
INT: Heath	.. um in the Middle East,	Middle East				
INT: Heath	from entering .. the US,	US	entering			
INT: Heath	what did you think of --				Pragmatic	4
INT: Heath	... what do you think of that one.				Pragmatic	1
PNT: Zand	...(1.0) Pf= --					
PNT: Zand	like,					
PNT: Zand	... %					
PNT: Zand	... d- I --					
PNT: Zand	@@					
PNT: Zand	.. I don't know.					
PNT: Zand	... Um,					
PNT: Zand	... I think the idea of banning a people,	banning	(a) people			
PNT: Zand	.. or- or a religion,	religion				
PNT: Zand	uh,					
PNT: Zand	.. is .. silly.					
INT: Heath	... H[m].					
PNT: Zand	[Um],					
PNT: Zand	...(1.5) I don't know what he was hoping to get out of it,					
PNT: Zand	personally.					
INT: Heath	.. Yeah.					
INT: Heath	... what do you think of like in our society though,	society				
INT: Heath	that like,					

For each algorithm, its scores were tallied per transcript, and then compiled into a table to compare efficacy quantitatively, presented in Table 7.

For each transcript, the number of partitions identified in Chapter 2's manual-coding is given in the first column. Parker was completed with 8 partitions, as opposed to 15, in accordance with Chapter 2's remarks on collapsing some topics. The second column presents the maximum, possible score for each transcript – this being the number of boundaries (one less than the number of partitions) multiplied by 4 (the maximum possible score). Using this maximum score, the following columns take the score for each algorithm and divide it by the maximum possible score, to generate the "mark". In essence, the mark is a

scaled-version of the algorithm's score, which *enables comparisons despite the different number of boundaries in each transcript*. The difference between the lowest and highest mark (i.e. the range) is displayed in the final column, for each transcript. Furthermore, the bottom row presents the average mark for each algorithm and the range values, as a means of summarising performance.

The Even Partitions method, which serves as a control for the linguistic partitioning methods, received an average mark of 41%, which leaves it in last-place of the three methods. This result, while not unexpected, is incredibly close to the mark given to the Pragmatic algorithm (43%). Additionally, the range for all the average marks is a mere 10%, which indicates that either (1) all algorithms perform similarly across all transcripts, or (2) each algorithm has its value in specific situations, which when averaged leads to a close result.

From Table 7, the evidence supports the second hypothesis, as each of the three algorithms presents higher marks across certain types of data. The Even algorithm performs well when more partitions are used. While Zander's transcript seems slightly underwhelming, because the difference between the marks is only 5%, it does not override the better performance seen with Marjorie and Janice, when compared to poor results with Isa and (especially) Parker. This result seems logical, as the more partitions are inserted, the smaller each one becomes, which minimises the amount of space for differing

content to occur within one partition. Conversely, the Pragmatic algorithm appears to favour fewer partitions, as it only performs really poorly with Marjorie and Zander – even with Zander, it only differs from the best algorithm by 5%. It is *the* best algorithm for Isa by 10%, beats Even in Parker by a whopping 32%, and only comes last in Janice by 8%.

Lastly, the Turn algorithm receives the best mark (or tied for it) in three of the five transcripts, with a stunning 93% for Parker (this being 25% higher than Pragmatic – the second best). Its downfall, however, arises with Isa's transcript – in a tie for last place with only 25%. Similarly, it comes in second place with Marjorie's transcript. These results indicate that the Turn algorithm is quite strong all-around. But, the contrast between Isa and Parker's marks, despite having a similarly-few number of partitions, suggests that Turns may be a good "first step", but another algorithm *might* outperform it, depending on the transcript.

Table 7 *Final scores for each partitioning algorithm, across the five transcripts. The highest score and mark per row are highlighted.*

Transcript	**Number of Partitions**	**Maximum Possible Score**	**Even Partitions**		**Pragmatic Partitions**		**Turn Partitions**		**Range**
			Score	**Mark**	**Score**	**Mark**	**Score**	**Mark**	
Bcnt_AEF_073_Marjorie	12	44	22	50%	15	34%	19	43%	16%
SSDS_GTF_855_Isa	6	20	5	25%	7	35%	5	25%	10%
SSDS_AAF_127_Janice	7	24	12	50%	10	42%	12	50%	8%
SydS_AOM_105_Parker	8	28	10	36%	19	68%	26	93%	57%

SydS_AYM_015_Zander	16	60	26	43%	23	38%	26	43%	5%
Average Mark				41%		43%		51%	10%

To summarise, each partitioning algorithm provides its own contribution to the "toolkit" of topic modelling. However, further research is needed to optimise the performance of the Turn and Pragmatic algorithms.

4.3 Automating Topic- and Type-Counts

The previous section has explored the alignment between computationally-drawn partitions and manually-identified topics. The theory behind this, is that a greater correspondence between the location of partitions and groups of manually-coded topics, will translate to better cluster coherency. However, by definition, this process relies on manual-coding, which, as noted, was time-consuming and imprecise. Earlier sections also revealed the need to obtain specific topic- and type-count values, in order to employ topic-modelling methods. These specific values must therefore be chosen carefully, by considering as much semantic and contextual information from a transcript as possible. Hence, I will now consider automated methods for determining the optimal topic- and type-counts, termed "coherence scores" (cf. Newman et al., 2010; Mimno et al., 2011). The first two methods to be investigated are UCI and UMass. Neither of these have any programmed information about the dictionary definitions of any words; instead, they rely on the idea that word-

proximity within the original document will translate to coherent and interpretable clusters.

When these methods are run, a numerical score is given per number of clusters, and (for some methods) the number of word types per cluster. This functionality is key to optimising the performance of topic-modelling algorithms, as all methods require a user to specify the number of topics. Therefore, as better scores translate to better coherency, they are used to advise a user on how many clusters (and types per cluster) should be requested.

4.3.1 UMass Coherence Score

The first coherence score to be applied is called UMass (Rosner et al., 2013). Note that the term “UMass” is the literal name used in publications, with its meaning not clarified. The algorithm works by checking each possible pair of words in a cluster, and calculating (1) how many times the words co-occur in the original document, and (2) the frequency of each word in the original document (Rosner et al., 2013). The list of clusters and the original document must both be supplied by the user. The algorithm then divides the number of bigram occurrences by the frequency of each individual word, resulting in two values per bigram; these scores are then averaged for a final result (Rosner et al., 2013). Due to its coding implementation, a key aspect of UMass is that

only the *number of topics* affects its score, and not the number of words per topic (cf. Rehurek, 2022; Stevens et al., 2012).

Another detail about UMass lies in the range of possible scores. Aside from being unbounded (i.e. not restricted between an upper and lower limit, such as between 0 and 1), better UMass scores are always further from 0, by *absolute value*. In other words, a UMass score of -3 implies more coherency than a 2 (Stevens et al., 2012). Variations between positive and negative values arise due to the vectorisation process when compiling the list of words and their locations; a technicality which lies beyond the scope of linguistics.

4.3.2 UCI Coherence Score

The second coherence-score to be tested is UCI (cf. Lin et al., 2020; Fontana, 2020; Rosner et al., 2013). Note that as with UMass, UCI is the literal name used in publications, with its meaning not clarified. This algorithm takes each cluster, and then calculates the probability of finding each constituent word in the original text, assuming random sampling (Fontana, 2020). As with UMass, the original document and clusters are user provided. It then takes this probability, and determines the likelihood of finding bigrams within the original text – bigram combinations are limited to the words per cluster – referred to as the Pointwise Mutual Information (cf. Lin et al., 2020; Fontana, 2020). Through dividing the probability of each bigram, by the probabilities of finding the two, individual words by random sampling, and taking the log of

the resulting fractions, the combined probabilities can be averaged per cluster. This average represents the coherence score, for that individual cluster (Rosner et al., 2013). Once all individual scores are determined, the average of those scores gives the final UCI score. Thus, the primary difference from UMass, is UCI's comparison of bigram occurrences against *a randomly-shuffled version of the original text;* this makes UCI an extrinsic algorithm, when compared to UMass' intrinsic status, since UMass only references the original, unshuffled document (Pleple, 2013).

The logic behind this algorithm is to determine the *coherency* of a cluster using the existing word-pairings in the original text. As a property of the probability calculations, the resulting score is always between 0 and 1 (Fontana, 2020). Moreover, since UCI averages bigram pairings across clusters, the number of words per cluster will affect the final score, unlike UMass.

4.3.3 WordNet: A Coherence Score based on Semantic Assignment

"Wordnet" is an unsupervised, machine-learning tool for the automated assignment of semantic information to words and their contextual use (Princeton University, 2022); in particular, its relationship to the human-based assignment of semantic frames, makes it suitable for generating accurate coherence-scores. Since WordNet draws on an existing, large-scale dictionary and thesaurus, it can determine the coherency of clusters through the codified meanings of their words, based on their dictionary and thesaurus data, and

usage across the searchable internet (Princeton University, 2022). This functionality will be utilised to develop a third coherence-score, which employs the pre-existing neural-network to evaluate semantic coherency. The philosophy behind this algorithm is to ensure that the coherence score rewards good, internal coherency within clusters, while punishing any semantic overlap across different clusters.

The combination of the words *teacher*, *building, classroom* and *playground*, for example, would likely create an association with the topic of "school". This example highlights the existence of semantic networks, tied to appropriate, semantic frames (cf. Fillmore, 1976, 1982). Fillmore (1976) presents *frames* between individual words in a lexicon, and their tacit function in formulating our understanding of larger concepts by stating that "the language-user interprets his environment, formulates his own messages, understands the messages of others, and accumulates or creates an internal model of his world" (p. 23). The idea of presenting a set of words to invoke a larger topic is also explicitly mentioned by Fillmore (1976:25), in which he states that the frame of "commercial event" is "activated in the mind of anybody who comes across and understands any of the words *buy, sell, pay, cost, spend, charge,* etc., even though each of these highlights or foregrounds only one small section of the frame". It is clear how this quote elegantly relates the idea of semantic frames to clusters from topic-modelling.

In this thesis, we have relied on the manual identification of semantic frames (that is, as a human-produced task) to judge accuracy. However, Princeton University's (2022) neural-network dictionary, "WordNet", provides an automated solution to assigning networks of semantically-similar words, which ideally should produce similar results to human-constructed semantic-frames. The tool is "widely used by the NLP [Natural Language Processing] community for applications including Information Retrieval and Machine Translation" (Miller & Fellbaum, 2007:211), making it a prime candidate that has not been explored to date for topic-modelling research.

WordNet contains a dictionary of labelled vocabulary, such that each entry is assigned information on its semantic meaning, and how that meaning relates to other entries (Tengi, 1998). Lexicographers are responsible for the addition of new words into the database, which are restricted to nouns, adjectives, verbs and adverbs (Tengi, 1998:105-106). The words are connected using a unified, network structure, with links constructed if words are synonymous (from a dictionary perspective), used with other words in frequent contexts, or if the word refers to a subordinate concept which belongs to a greater entity (such as *tyre* for *car*) (Tengi, 1998:107-109; Miller & Fellbaum, 2007:210). For subordinates, the mapping is generated by linking the word to *classes* (categories that hold respective subordinates), such that, for example, "*nations* are a class and Spain is an instance of that class" (Miller & Fellbaum, 2007:210).

Cases of polysemy are also addressed, since, as an example, the word *paper* can be used in the context of (1) a flat, processed material from plants used as a base for holding imprinted text or images, (2) a scholarly article, or (3) a shortened-form of *newspaper* (Tengi, 1998:107). In these instances, the database entry for *paper* is connected to every other entry that relates to *paper's* uses (e.g. *articles, news*), such that the presence of *paper* in a text, would support the presence of "synonymous" words related to scholarly works, writing, drawing or the news (Tengi, 1998:109).

I will now test the ability of WordNet to form networks across word-clusters, when compared to links as judged by human intuition. WordNet's Python package contains a function called "wup_similarity", which returns a score of similarity between two words (NLTK, 2022b). The score is always between 0 and 1, with a larger number indicating a stronger link between the words in WordNet's database; that is, being part of the same semantic-frame. As an initial trial, I have tested the example semantic-frame proposed by Fillmore (1976:25), by running the words *buy*, *sell, pay, cost, spend* and *charge* through WordNet's "wup_similarity" function, and thereby generating a network of scores for their similarities. I have also included the word *tree* as an outlier item, as a further means of testing WordNet's deductive powers. Since *tree* deviates from the words denoting financial transactions, a successful result from WordNet should give this word the lowest score in the network.

The scores are shown in Table 8, with the score for each pair of words shown in their intersecting grid-square. Formally, Table 8 represents a *Cayley Table*, with "wup_similarity" being the operation performed between two elements. As can be seen, a score of 1 only occurs when a word is compared to itself, indicating that – as would be expected – these "two" words are maximally similar. The next highest score is 0.4, seen with *spend* and *buy*, with third place going to a tie between *spend* and *sell*, *spend* and *pay*, and *spend* and *charge*, as all pairings have a score of 0.33. However, note that *tree* consistently scores the lowest for every single pairing. This implies that using WordNet's similarity score, a computer would (correctly) determine that *tree* is the outlier word in this cluster.

Table 8 Cayley Table of the WordNet Similarity Scores between Word Pairs

wup_similarity	Buy	Sell	Pay	Cost	Spend	Charge	Tree
Buy	1	0.29	0.29	0.18	0.4	0.29	0.14
Sell	0.29	1	0.25	0.17	0.33	0.25	0.13
Pay	0.29	0.25	1	0.17	0.33	0.25	0.13
Cost	0.18	0.17	0.17	1	0.2	0.17	0.12
Spend	0.4	0.33	0.33	0.2	1	0.33	0.15
Charge	0.29	0.25	0.25	0.17	0.33	1	0.13
Tree	0.14	0.13	0.13	0.12	0.15	0.13	1

As a second example, the word cluster *job, interview, distribute* and *magazine* comes from the following transcript, seen in Example 11.

(11, SydS_AOM_105, 00:26:44.930 – 00:26:55.607)

PNT: Parker	I had a job interview,
PNT: Parker	... they were looking for people,
PNT: Parker	... to go around to,
PNT: Parker	distribute these .. the magazines,
PNT: Parker	it was just magazines back then,
PNT: Parker	now they do everything else.
PNT: Parker	But,
PNT: Parker	.. it was just magazines,
PNT: Parker	but I never got the job,

When these four words are compared against each other through WordNet's *wup_similarity* function, with *people* serving as an additional, control entry taken from the same transcript, the following network appears.

Table 9 Cayley Table using a Sydney Speaks Example

wup_similarity	Job	Interview	Distribute	Magazine	People
Job	1	0.53	0.14	0.11	0.31
Interview	0.53	1	0.13	0.1	0.29
Distribute	0.14	0.13	1	0.13	0.2
Magazine	0.11	0.1	0.13	1	0.14
People	0.31	0.29	0.2	0.14	1

Table 9 reveals that *job* and *interview* are deemed to be the most closely-related words, which is understandable as the compound term "job-interview" is a prominent bigram; however, the word *people* also generates high-scores with *job* and *interview*. From what I suspect, the high token-frequency of *people* (noting its inclusion in Chapter 3's conversational stoplist) in WordNet's training data has made it co-occur with *job* and *interview* a sufficient number of times, such that WordNet views its presence as allowing

for *job* and *interview* to co-occur with it (even if the dictionary entry for *people* would not automatically be linked to these two words, unlike *job* and *interview* being grouped together for job-interview). Therefore, while a human reading the transcript would likely not include *people* in the semantic-frame encapsulated by the four other words, the high frequency of *people* does mislead WordNet. In essence, WordNet would need to be aided by further human-input (i.e. a stoplist), to determine which words are relevant for generating groups. But, in doing so, it is hoped to outperform UCI and UMass in generating coherent and representative clusters.

Another word of interest is *magazine*, as it received the lowest scores when compared to *job, interview* and *distribute*. Hence, it appears that while a human reader understands that, given the transcript's contents, *magazine* relates to the context of a delivery job, WordNet did not recognise this link too strongly. A means of addressing this issue, and similar misidentification problems, would be to train WordNet with Sydney Speaks data; but, owing to the closed access of WordNet's lexical database, this process cannot be completed in this thesis.

The specifics of the WordNet coherence-score are as follows. Firstly, it calculates the proportion of identical (overlapping) words across all clusters, as a fraction of all words. Secondly, within each cluster, the words are compared against each other to determine the most centrally-located word in the cluster

(i.e. find the semantic “midpoint” of the cluster), as per the semantic data from WordNet. A synonym for this midpoint is the cluster’s “medoid”. Then, the third step involves calculating the distance between each pair of medoids, and obtaining from this the average distance between medoids. The greater this average value, the better the final coherence-score will be. The final step involves averaging the proportion of overlapping words (see the first step) with the average distance between medoids, to give the final score. The resulting score is also bounded between 0 and 1 (unlike UMass), with a larger number indicating better coherency.

4.3.4 Coherence Score Results

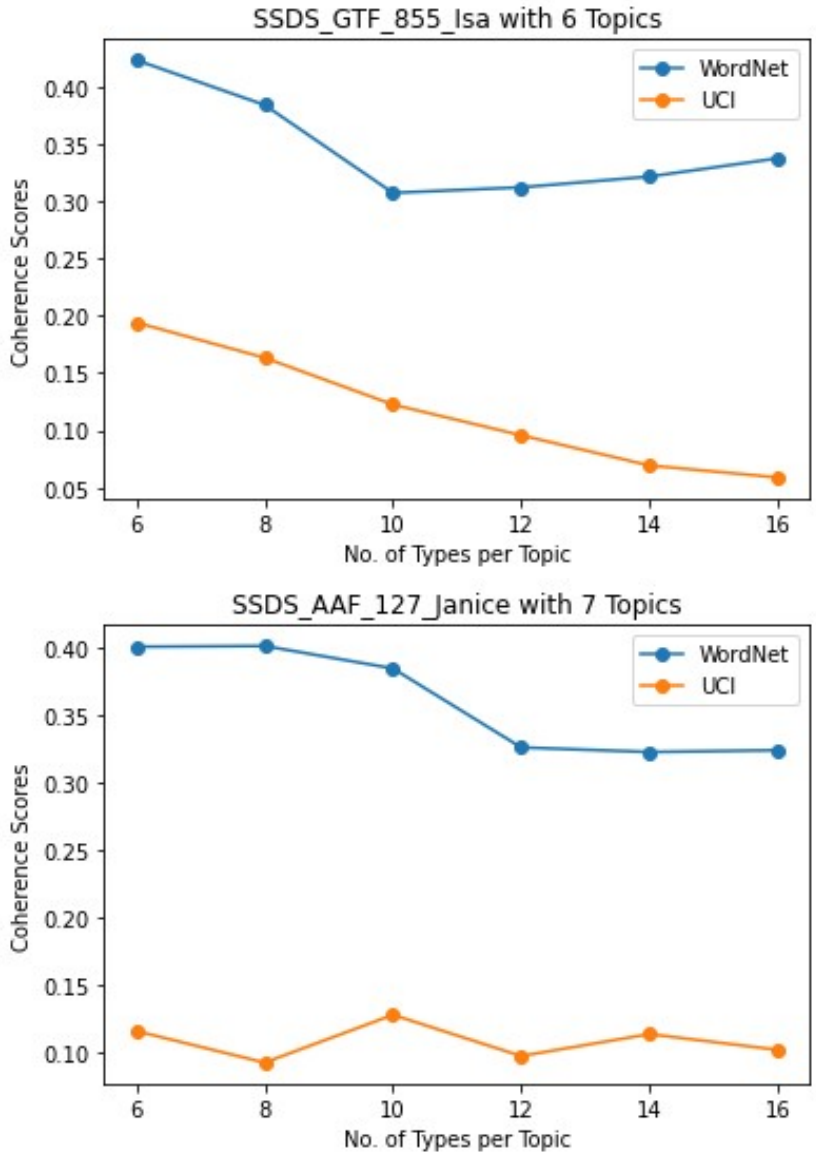
4.3.4.1 Types per Cluster

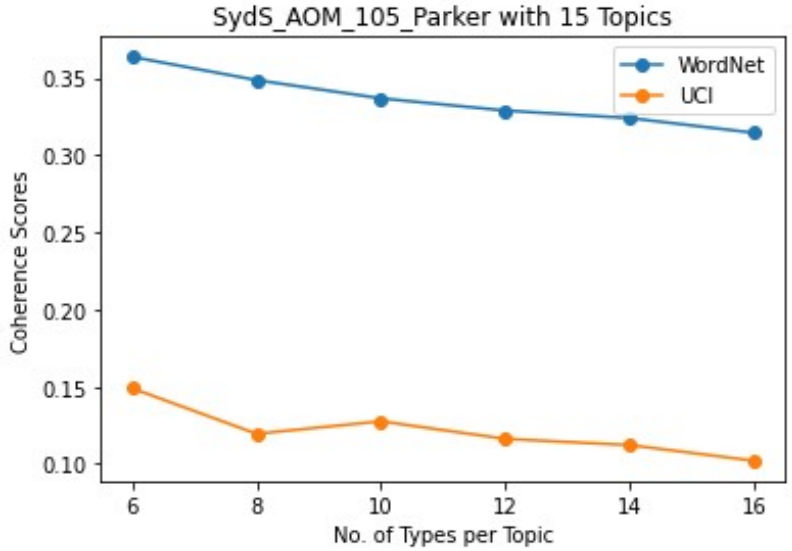
To determine how many types to use per cluster, the following graphs present the coherence-scores of clusters generated with the stock (i.e. no partitions) version of LSI. For this exercise, the number of clusters has been controlled using manual-coding numbers shown in Chapter 2; these values being 6 for Isa, 7 for Janice, 12 for Marjorie, 15 for Parker, and 16 for Zander. Having fixed the number of clusters as a control variable, the optimal number of types can be determined more objectively. Subsequently, in a later section, the number of types (as optimised here) will be fixed, to test for the ideal number of clusters.

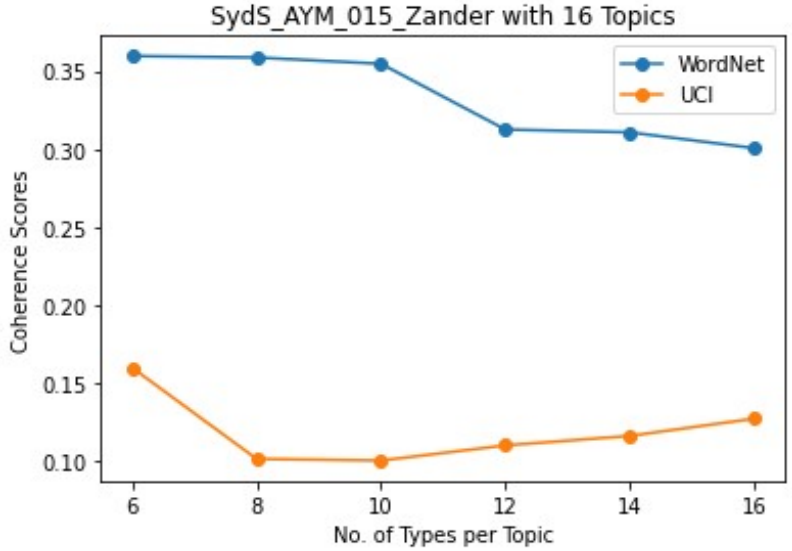
Figure 14 presents the coherence scores for each transcript, with the number of types on the x-axis. When reading the graphs, note that a higher score indicates better coherency. Additionally, the consistent gap between UCI and WordNet scores is due to the differences in these algorithms, and not better performance from WordNet in all situations. Therefore, results should be read as the variation in scores *within each algorithm*. Lastly, do note that because UMass is unaffected by the number of types per topic (as mentioned in its overview section), it has been excluded from these analyses.

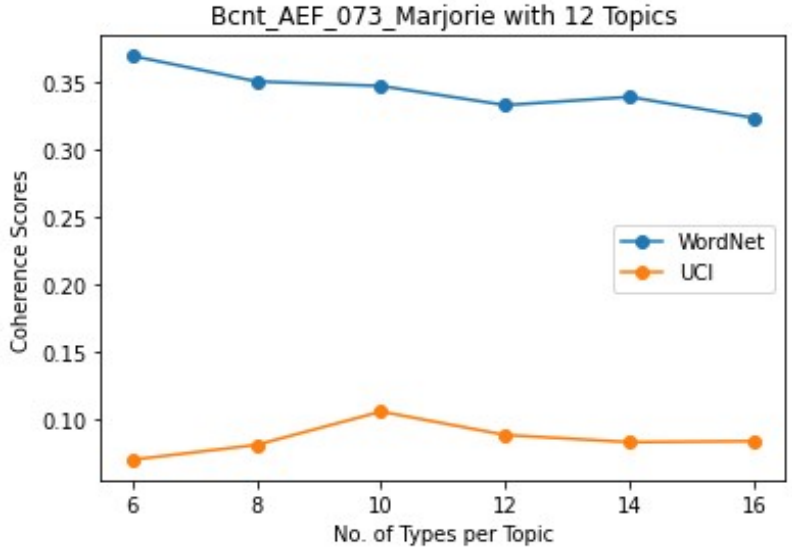
Isa, Parker and Zander's scores, across both algorithms, indicate that six types should be used per topic. In these transcripts, using more types leads to a decrease in coherency. Janice's transcript shows a similar pattern, with WordNet supporting six or eight, and UCI supporting six or ten. As six is the common value, I will use this number of types with Janice's transcript. Lastly, Marjorie's transcript yields the highest score with six types, as per WordNet, or 10 types, as per UCI. While this transcript appears to produce the only draw, I will go with six types for consistency with the other transcripts, in addition to easier reading by a human analyst.

Figure 14 *Coherence scores to identify number of word types to include per cluster, determined by WordNet and UCI for the <u>five</u> transcripts (highest score represents greater coherence)*




SydS_AOM_105_Parker with 15 Topics
WordNet
UCI
Coherence Scores
0.35
0.30
0.25
0.20
0.15
0.10
6
8
10
12
14
16
No. of Types per Topic


SydS_AYM_015_Zander with 16 Topics
WordNet
UCI
Coherence Scores
0.35
0.30
0.25
0.20
0.15
0.10
6
8
10
12
14
16
No. of Types per Topic


Bcnt_AEF_073_Marjorie with 12 Topics
WordNet
UCI
Coherence Scores
0.35
0.30
0.25
0.20
0.15
0.10
6
8
10
12
14
16
No. of Types per Topic

4.3.4.2 Number of Clusters to Extract

With the number of types set to six as a control variable, the optimal number of clusters will now be determined, using a combination of the three coherence-scores – UMass, UCI and WordNet. The topic clusters, as with the previous graphs, come from the unmodified, no partitions topic-modelling method. I will use the coherence scores to reduce the number of viable clusters down to two options. These groups of clusters are shown in Appendices 4 to 6, for transparency. This decision is in support of easing qualitative and linguistic-based approaches to topic modelling, as emphasised in Chapter 1's literature review.

The coherence scores for each transcript are presented in Figure 15 to Figure 19, with two graphs per transcript. An important note is that *lower* dots indicate a better score with UMass, unlike UCI and WordNet. This is because better UMass scores correspond to being further from zero, in the positive and negative axes. Thus, in the graph on the top – showing WordNet and UCI – the highest point represents the best coherence score, while on the bottom, for UMass, the lowest point represents the best score.

The results show a high rate-of-agreement across all transcripts, expect Janice, that four to eight (i.e. fewer) clusters is the ideal amount. WordNet supports this view across all transcripts, with Figure 15, Figure 16 and Figure 17 all having the highest scores with four clusters. Figure 18 and Figure 19 present a

tie between four and eight clusters, and favouring just six clusters, respectively. UCI seems to favour slightly more clusters, with Figure 15 being a close tie between eight and ten, Figure 16 favouring six and fourteen, Figure 18 supporting six and Figure 19 supporting eight. Figure 17 presents a big outlier, with twenty clusters being the favoured amount. Lastly, UMass presents a compromise, with eight clusters favoured in Figure 15, four in Figure 16 and Figure 18, six in Figure 19, and curiously, twenty (as with UCI) in Figure 17.

By taking a compromise between the three algorithms' scores, two cluster amounts have been selected. Since at least two algorithms agree with each other in every transcript, it is easy to decide two amounts which are supported by the coherence scores. Janice's transcript (see Figure 17) has the notable agreement between two starkly-different cluster amounts – namely, four (with WordNet) and twenty (with UCI and UMass). Given that Chapter 2's manual coding revealed that Janice had the highest mean and median types per topic (36 and 48, respectively), this might explain the unusual, cluster-amount result.

Figure 15 Coherence scores to identify number of clusters to use for Marjorie's transcript, as per WordNet, UCI and UMass: 4 and 8 topics look the best

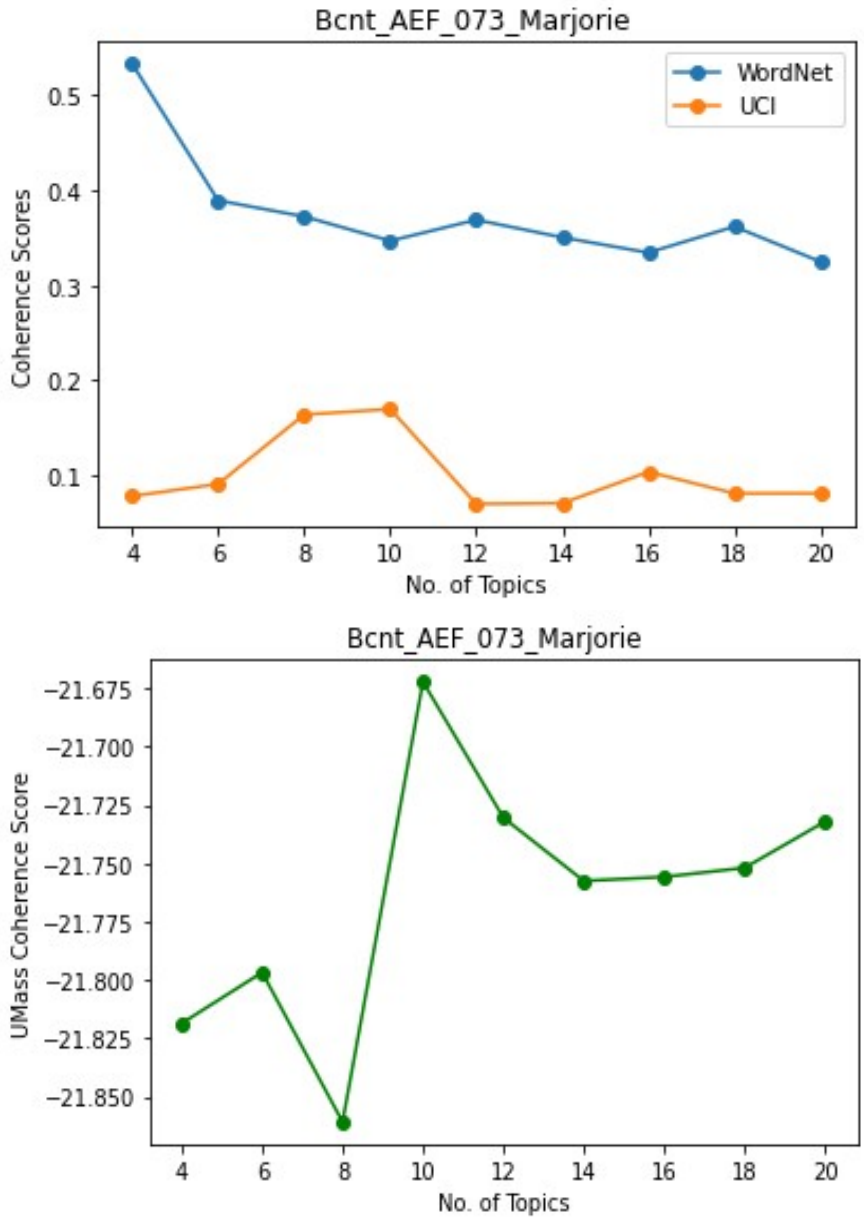


Figure 16 Coherence scores to identify number of clusters to use for Isa's transcript, as per WordNet, UCI and UMass: 4 and 6 topics look the best

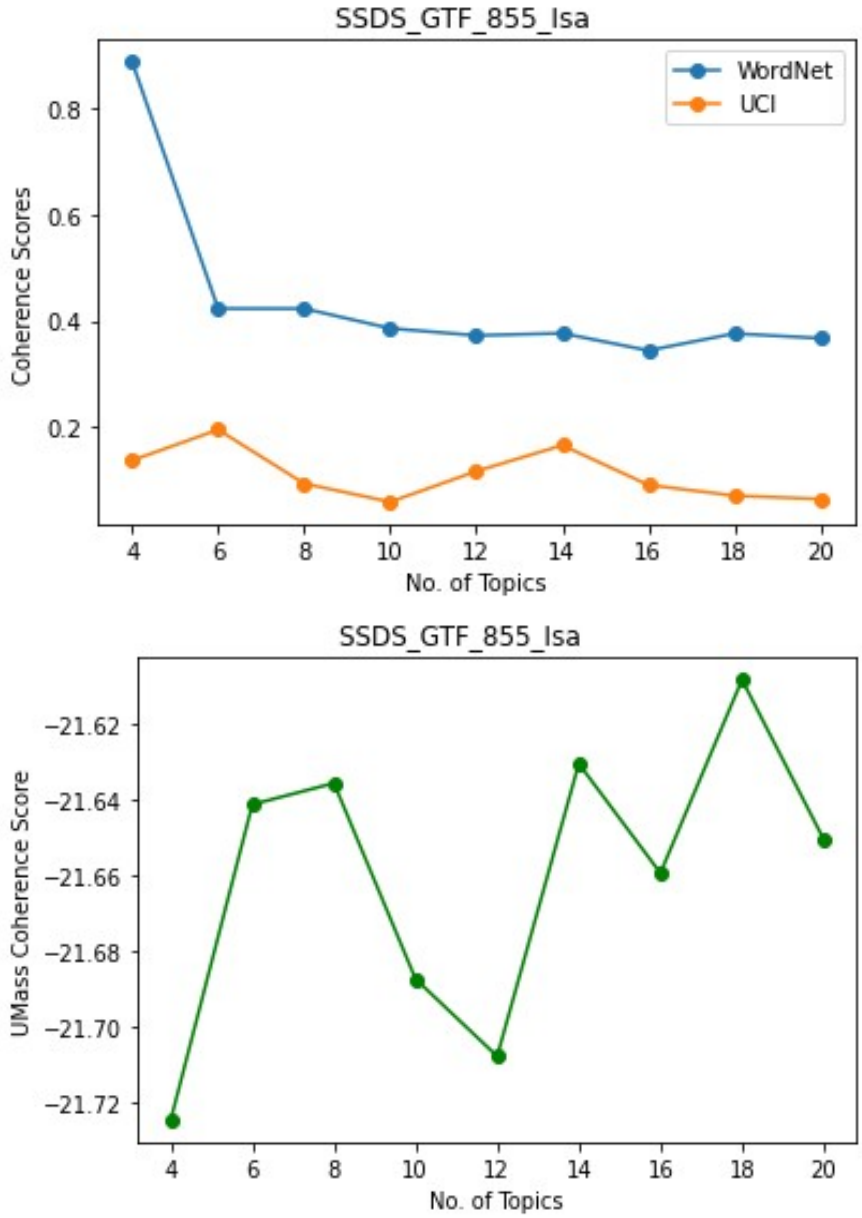


Figure 17 Coherence scores to identify number of clusters to use for Janice's transcript, as per WordNet, UCI and UMass: 4 and 20 topics look the best

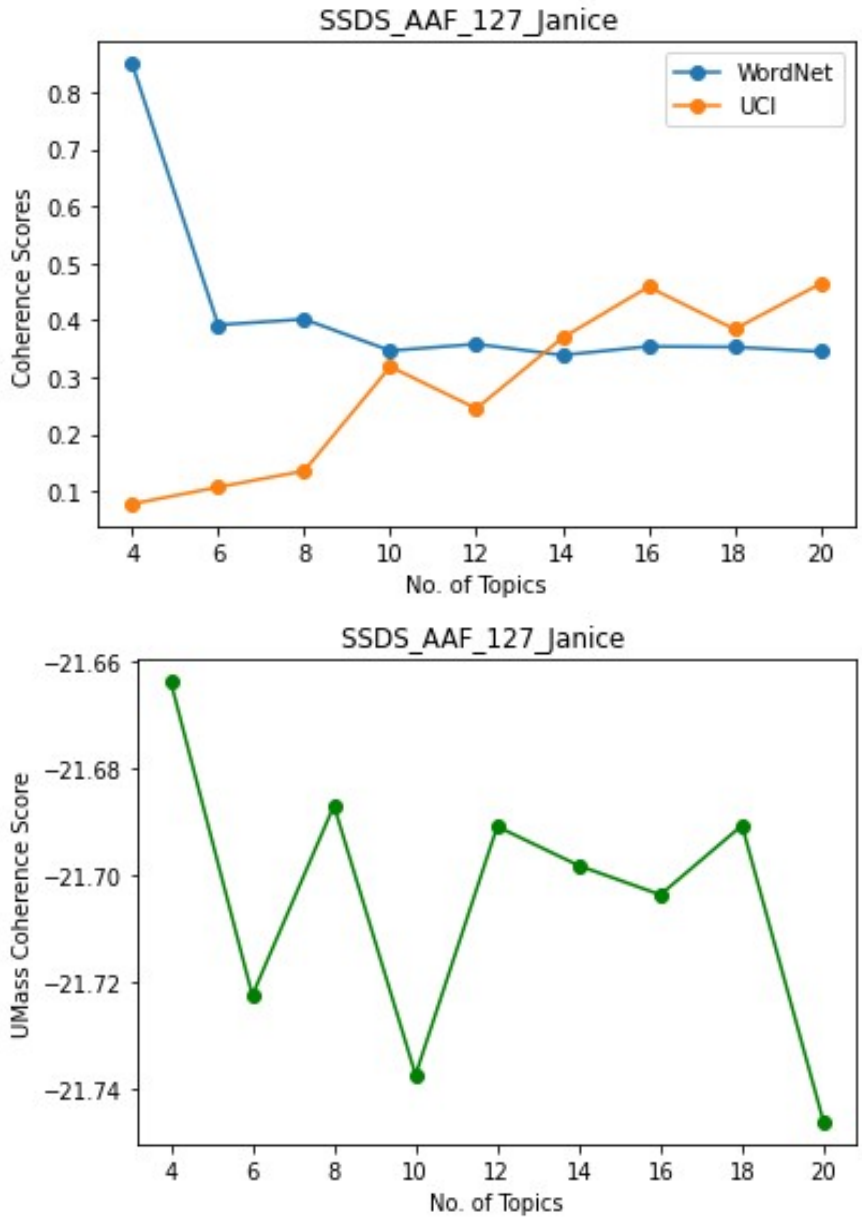


Figure 18 *Coherence scores to identify number of clusters to use for Parker's transcript, as per WordNet, UCI and UMass: 4 and 6 topics look the best*

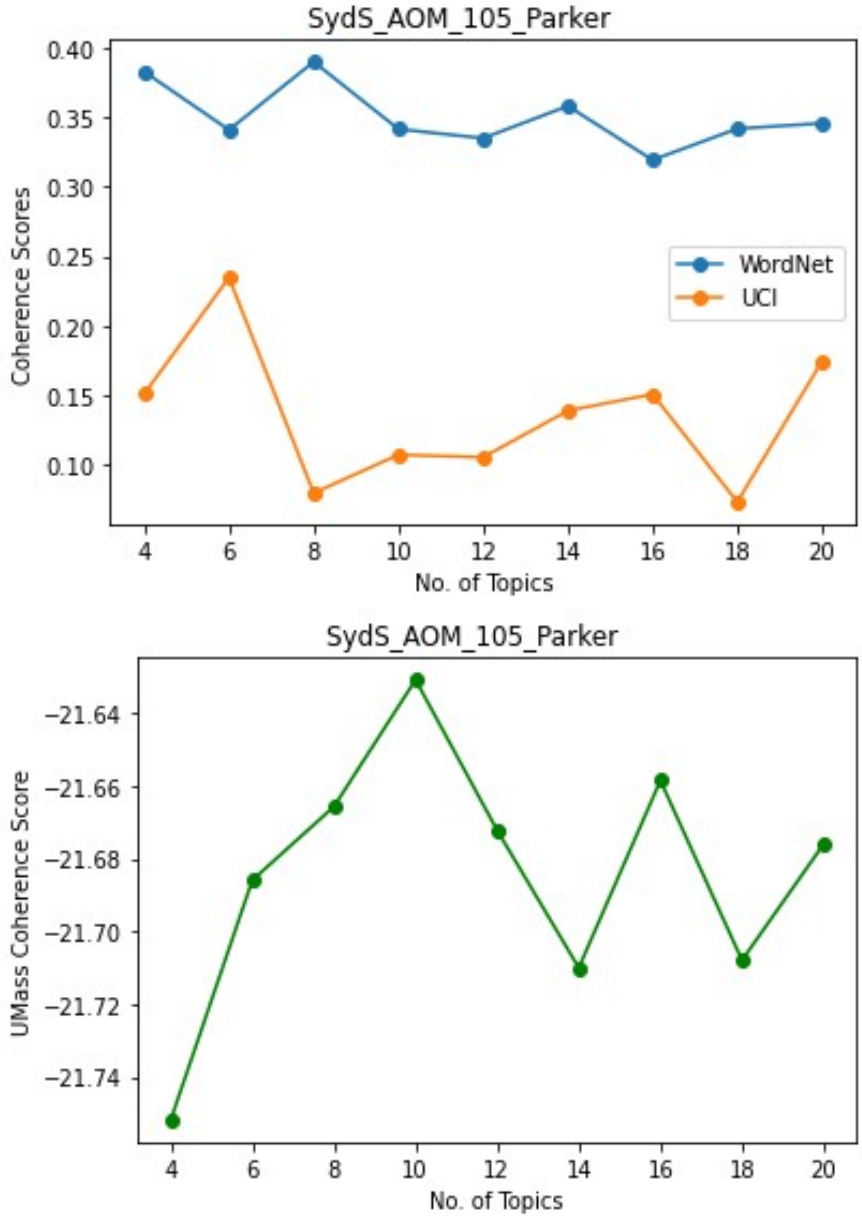
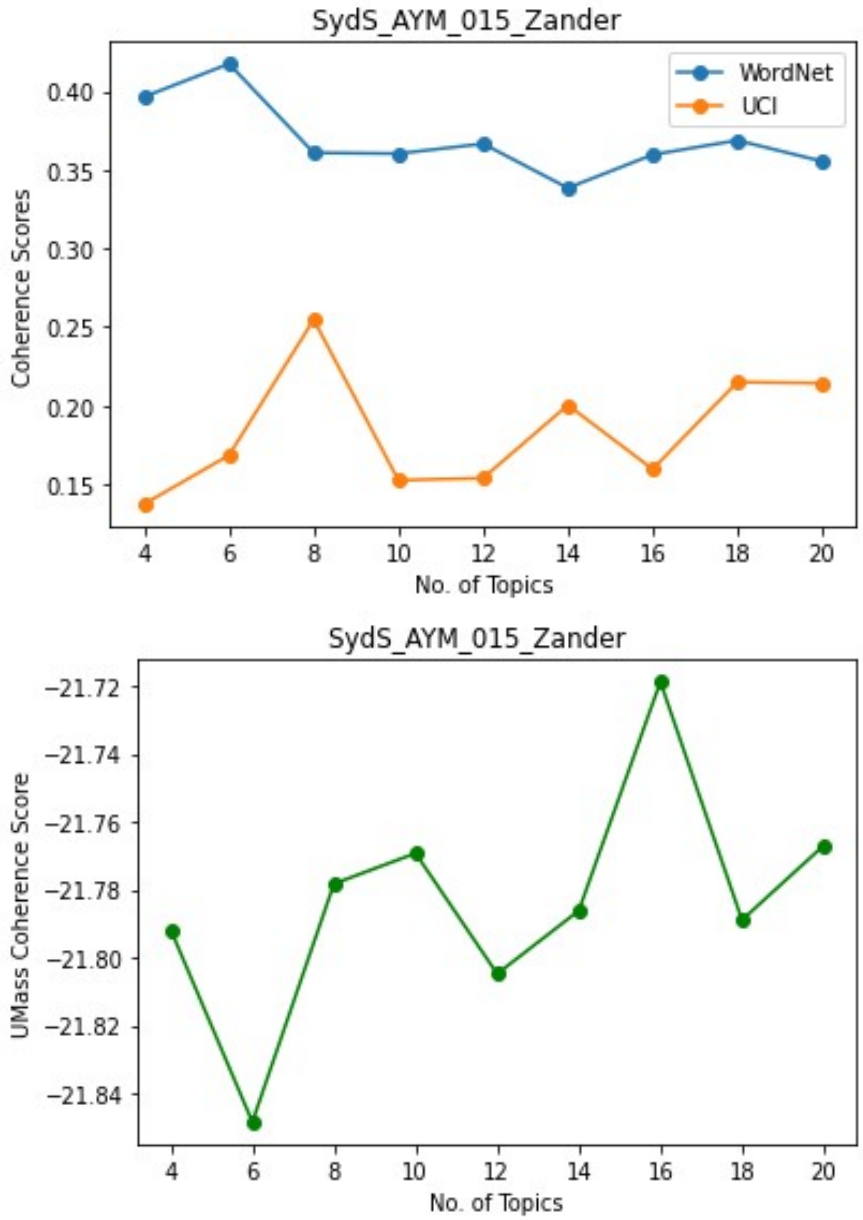


Figure 19 Coherence scores to identify number of clusters to use for Zander's transcript, as per WordNet, UCI and UMass: 6 and 8 topics look the best



A final remark comes from the consistent performance of WordNet's scores, when compared to UCI and UMass. Since we mentioned how UCI and UMass have been applied in previous literature, the non-divergent, WordNet results support its implementation in future works.

4.4 Chapter Summary

This chapter has established the tools and methodology needed to perform topic modelling, in a maximally-automated manner. The discussion began with an overview of unmodified LSI, and why partitioning is a justifiable modification, based on linguistic observations from the Sydney Speaks data. Then, each partitioning algorithm was described in detail, and their performance was compared against the manual coding for accuracy in determining topic groups. Based on these results, each partitioning algorithm is worth applying when topic-modelling. Specifically, Even and Turn partitions function better with more partitions; Pragmatic, however, is the better algorithm with fewer partitions. Then, the issue of automating the number of clusters to produce (and types per cluster) was addressed with three coherence scores – one of which is newly developed, and was explained in extra detail. Through implementing the coherence scores, their results have optimised the clusters presented in the next chapter. In particular, six types per topic is the optimal amount for all transcripts, and (excluding Janice) between four to eight topics should be extracted per transcript.

Chapter 5- Cluster Results

5.1 Introduction

This chapter presents the cluster results, with interpretative and numerical-summary analyses included for a holistic overview. The first section will detail how to interpret the semantics of clusters, from a qualitative, analyst perspective, by considering different types of clusters and their conveyance of semantic frames in the transcript. Some clusters elegantly convey a topic without needing further context, while others do require additional context from the top-40 words, or other background information. There are others, though, that still do not present any meaningful information.

We will then consider the coverage of each algorithm, inasmuch as quantifying how many manually-coded topics emerge in the topic modelling. The number of manually-coded topics which (1) were identified by the clustering algorithms, and (2) are represented by more than one individual cluster (per algorithm), will be measured, to quantify the degree of overlapping clusters per algorithm.

5.2 Interpreting Topic-Modelling Output

In order to judge clusters for their informativeness, it is important to define what constitutes “informativeness”. For this thesis, I consider a cluster to be informative if once given all available context, it conveys an intuitive concept

as a semantic frame. Figure 20 shows an example of the raw output from the Pragmatic algorithm, when run on Zander's transcript to identify six clusters, and six types per cluster. Note that the order of the clusters directly corresponds to the location of the partitions in the original transcript. The number next to each word represents the word's location in the respective cluster, such that the word with a "(+/-) 1.000" (always the first word) is the cluster's midpoint, with other words' *proximity* to the centre being directly proportional to their number's absolute value (e.g. "-0.015" would be closer to the midpoint than "0.001"). For the purposes of this analysis, I am considering all words as equal units within a cluster's semantic frame. However, when reproducing clusters in writing, I will always present them in the order shown by the algorithm.

A notably-deviant cluster is the fifth one. It shows (1) a repetition of the word "siege", (2) only four words, as opposed to six, and (3) square parentheses, as opposed to round ones. These differences are the result of the partition's contents being given directly to the user, as the partition housing this cluster was too small to run LSI (a feature mentioned in Chapter 4). Since the relevant partition had less than seven words post-stoplist-filtering (specifically, four), these words have been directly shown to the user.

By quick inspection, it is clear that a few topics are easily identifiable, including online gaming in the first cluster, language use in the fourth cluster, and the

Lindt Cafe Siege in the fifth cluster. For the other clusters, more contextual information would be needed to clarify them, such as the manually-coded topic of “socioeconomic stance” being referenced in the final cluster with the words “richer”, “different” and “hang” (as in “hang out”). Lastly, the second and third clusters are unintelligible, even with contextual information, thereby providing no use to an analyst.

Figure 20 Example of 6 clusters from Zander’s transcript, with 6 types per cluster by default

```
(0, '1.000*"zombie" + -0.000*"amaze" + -0.000*"game" + 0.000*"awesome" +
0.000*"online" + 0.000*"call"')

(0, '1.000*"streak" + 0.002*"respect" + -0.001*"physically" + -0.001*"whatnot" +
-0.001*"insight" + 0.001*"conversation"')

(0, '1.000*"australia" + 0.002*"derogatory" + 0.002*"favourite" + -0.002*"bet" +
0.002*"cruise" + -0.002*"goodness"')

(0, '1.000*"word" + -0.015*"different" + 0.000*"example" + -0.000*"happen" +
-0.000*"complicate" + 0.000*"way"')

[['lindt'], ['siege'], ['sydney'], ['siege']]

(0, '-1.000*"different" + -0.002*"hang" + -0.001*"richer" + -0.001*"across" +
-0.001*"government" + 0.001*"lucky"')
```

One area that Figure 20 does not represent, is when two or more clusters refer to the *same* topic (regardless of their intelligibility). In essence, we sometimes get *different* clusters that refer to the same topic (described as “overlap”). As an example, “*pick, Australian, difference, Queenslander, Victoria, Newcastle”* and *“ocker, America, Australian, ockerism, tend, suburb”* both refer to the topic of *Dialectal Variation and Language Use* in Janice’s Transcript, when using the Turn algorithm. Therefore, having established how to interpret informativeness, the following sections detail different clusters by this qualitative metric.

To begin, there are some clusters which are easy to interpret and understand at a glance. Some of these show references to bigrams, including the names of famous characters such as Paul Hogan and the school subject of *domestic science*. Others relate to larger word co-occurrences, such as details about a *nice marble collection*. Examples of these are shown in Table 10, which presents the words in each cluster on the left-most column, followed by the cluster's transcript, the partitioning algorithm used, the corresponding manual-topic(s) and top-40 words. This final column is included due to some clusters containing topic-relevant information, despite possessing few or none of the most-frequent words.

Table 10 Six Intelligible Clusters

Cluster	**Transcript**	**Partitions**	**Manual Topic(s)**	**Top-40 Words**
Lindt, siege, Sydney, siege	Zander	Pragmatic	Lindt Cafe Siege	N/A
lesson, teach, English, Greek, speak, kid	Isa	Turn	School Life, Greek Afternoon-School	lesson, teach, English, Greek, speak, kid
Hogan, speak, try, talk, eastern, Paul	Janice	Even	Dialectal Variation and Language Use	try
Sydney, bus, issue, fighting, find	Parker	Turn	Bus (Mis)management	Sydney, bus
domestic, science, school, Sydney, cross, different	Marjorie	Pragmatic	School Life	school, Sydney, cross, different
marble, milky, appear, round, collection, nice	Janice	Pragmatic	Games	marble, nice

Other clusters present a borderline case when it comes to intelligibility, as they do require an understanding of the top-40 words and metadata context to be interpreted clearly. Four examples of these clusters are seen in Table 11.

The first cluster may seem nonsensical on first inspection, until one learns that Simmons was the manager of a Depression-era mission, which doubled as a flop-house for workers. Hence, this cluster refers to the running of a cheap motel, including its costs and services. In the second cluster, the participant is describing how he was taught the Greek insults “tenekes” (τενεκες) and “skoupidotenekes” (σκουπιδοτενεκες), meaning bin and rubbish-bin, respectively, by an extended family member (who, jokingly, would call him these words). The third cluster details how the participant would stand under a shed to avoid being soaked, during wet weather at school. Lastly, the fourth cluster presents an interesting scenario, in which the word “mum” seems to be an outlier on initial inspection (as one would think that it is merely school-related). However, the word “mum” frames that the cluster relates entirely to the school-life of the participant’s mother – particularly, the differences in subjects.

While Table 11 does contain meaningful clusters, the question of utility arises, given that the manual-coding was the easiest way to interpret some of them. This statement largely holds for the first and third clusters, as the top-40 words do not help in deciphering their true meanings. With the second and

fourth clusters, an analyst could decipher their meanings using just the top-40 words, if they correctly realise that “call” and “Greek” go together (possibly after seeing the relevant cluster), and that the presence of “mum”, “learn” and “different” refers to differences in the school life (again, after seeing the relevant cluster).

Table 11 Four clusters which need to be understood in context

Cluster	**Transcript**	**Partitions**	**Manually-Coded Topic(s)**	**Top-40 Words**
pay, experience, tea, four-pence, Simmons, seven	Marjorie	Even	Sydney City Mission	N/A
call, Greek, bin, tenekes, basically, skoupidotenekes	Parker	Turn	Family Life	call, Greek
school, area, allow, stand, shed, weather	Janice	None (Stock LSI)	School Life	school
mum, geography, mathematics, subject, learn, different	Isa	Pragmatic	Mum’s School Life	mum, learn, different

The final set of clusters, seen in Table 12, display no meaningful nor intelligible semantic-frame information. Even considering these in the context of the information obtained during the manual coding, top-40 words, or background details about the participants, very little sense can be made of these clusters.

Table 12 Four Entirely-Unintelligible Clusters

Cluster	**Transcript**	**Partitions**
speak, paper, outside, unfair, scalp, browny	Isa	None (Stock LSI)
hop, plant, describe, pole, entrance, love	Janice	Even
school, commercial, fourteen, bee, hope, somebody	Marjorie	Even
day, group, o'clock, number, middle, mother	Parker	None (Stock LSI)

Therefore, having interpreted multiple scenarios of cluster informativeness, it is reasonable to conclude that word-frequency counts and topic-modelling clusters should not be used in a mutually-exclusive manner. Rather, the clusters should work together with the most-frequent words, to "fill in gaps" when the other method misses information. This statement is quite evident for clusters which possess (1) a few of the most-frequent words, but provide their contextual use (e.g. the Greek insults example), or (2) show none of the most-frequent words, but nevertheless refer to an event in the common conscience (e.g. the Lindt Cafe Siege example).

Since the clusters have been evaluated qualitatively, it is now worth exploring the topic coverage offered by each algorithm. That is to say, how many manually-coded topics could an analyst deduce *from merely observing the clusters?* And, of these topics, how many are represented by overlapping (i.e. multiple) clusters?

5.3 Manual-Coding Coverage

To address the question of topic coverage, any clusters which are considered informative (with additional context or not) will be tallied per transcript and algorithm, with the final statistics revealing the ability of each algorithm to extract human-identified topics.

Thus, Table 13 provides a condensed overview of each algorithm's recognition of manually-coded topics. The table's columns represent each algorithm, with the first row indicating the number of topics ("Total N") which each algorithm's clusters managed to identify. If an algorithm only produces multiple, repetitive clusters about a few, select topics (even if the individual clusters are incredibly coherent), then it is not providing a holistic overview of a transcript's topics, and this metric will therefore score it poorly. The second row takes the number of identified topics (i.e. the number in the first row), and divides it by the total number of topics found through the manual coding – this value being 56. This process scales the results, for easier identification of trends. The raw data used to generate the table is visible in Appendix 3.

Table 13 Number of Manually-Identified Topics (out of total of 56), Represented by Clustering Algorithms

Manual Topics Covered by Clusters	**No Partitions**	**Even**	**Pragmatic**	**Turn**
Total N	22	29	27	27
Proportion	39%	52%	48%	48%

Table 13 reveals that the use of partitioning provides a noticeable, general improvement, when compared to the stock, non-partitioning LSI. Interestingly enough, the Pragmatic and Turn algorithms perform identically, with both showing a mere 4% difference from Even Partitions. The key idea to extrapolate from this finding, is the clear benefit that partitioning itself delivers, despite its simple concept. However, while the results support the idea that an analyst will see approximately 50% of all topics when partitioning is used – as opposed to 39% – this result does seem a bit underwhelming. After all, missing half of all topics is not ideal. But, as there is no relevant literature for judging partitioning's efficacy in spontaneous speech, it is tough to decide what percentage of coverage is "good enough".

A positive result, though, does come from the Even algorithm's flexibility, due to its non-reliance on a list of pragmatic markers or speaker-role labels. This implies that it can be easily used on any data, with minimal annotations. As ease-of-use is also the main attraction to using unmodified LSI, the Even algorithm leaves non-partitioning, stock LSI looking fairly underwhelming.

Another comparison of interest, is the difference between Even Partitions, and Pragmatic and Turn Partitions. As a quick remark from the interpretative section, it is curious that the majority of intelligible clusters came from the Pragmatic and Turn algorithms. When this observation is considered alongside the 4% difference between Even and these two algorithms, their use is

justifiable *when obtaining more detailed clusters*. In essence, Even Partitions works nicely as a topic surveying tool, to obtain a quick overview of potential topics. However, by then running Pragmatic or Turn Partitions, the resulting clusters should provide clearer information about said topics.

Another area to explore is that of cluster overlap. To address this question, Table 14 presents similar statistics to Table 13, except that it only tallies topics which are represented by more than one cluster. Additionally, the "Proportion" row lists the "Total N" value divided by the number of topics identified by the relevant algorithm in Table 13 (e.g. 22 for No Partitions, thereby giving 11/22 = 50%), to scale the result properly for the purposes of solely measuring overlap. The raw data used to generate the table is visible in Appendix 3.

Table 14 Number of Topics which were Manually- and Cluster-Identified that are Represented by Two or More Clusters

Topics Covered by Two or More Clusters	**No Partitions**	**Even**	**Pragmatic**	**Turn**
Total N	11	12	18	16
Proportion	50%	41%	67%	59%

The results are truly remarkable, as they support the idea that Even Partitions (owing to a proportion of 41%) really does *survey* a transcript across a broad scope of topics, when compared to the other partitioning algorithms. Furthermore, the results reveal that stock LSI is less repetitive than the Pragmatic and Turn algorithms, which strongly suggests that speakers use

specific constructions to introduce certain topics, while avoiding these when introducing others. This conclusion is most apparent for the Pragmatic algorithm (with 67% overlap), as the current markers do seem to favour the introduction of certain topics over others. However, the high overlap seen with the Turn algorithm is also incredible, as it suggests that long interviewer turns are (at least implicitly) favoured for soliciting certain groups of topics, when compared to shorter equivalents. This opens a new area to explore within discourse-pragmatics and conversation-analyst work, to improve the representativeness of the Turn and Pragmatic algorithms.

5.4 Summary

To summarise, this chapter has evaluated the performance of four, topic-modelling implementations, by scoring the information from their clusters against the manual coding. The results reveal multiple findings: these being (1) the objective benefit to using partitions over stock, non-partitioning LSI, (2) the noticeable benefits from applying Even Partitions as a quick *topic-surveying* approach, (3) the use of some linguistic-constructions to introduce certain topics, as supported by overlap statistics for the Turn and Pragmatic algorithms, and lastly (4) the differing degrees of qualitative accuracy from each partitioning approach, particularly more niche and detailed clusters from Turn and Pragmatic Partitions. Overall, these results show that while topic modelling provides some insightful information – including details that the

most-frequent words may miss – the coverage is limited to half of the manually-identified topics.

Chapter 6- Conclusions, Limitations and Further Research

This thesis has explored the role of topic modelling in the semantic labelling of sociolinguistic transcripts. The journey began with an overview of the latest literature, which outlined the previous applications of topic modelling, and noted a lack of work on spontaneous speech data, as well as minimal assessment of topic modelling from a linguistic perspective.

To assess the performance of topic modelling in spontaneous speech data, transcripts of sociolinguistic interviews were chosen. Due to the defined roles of an interviewer and participant, this form of data was hypothesised to provide an easier basis for detecting topics, when compared to more unstructured, conversational data. Having chosen the data type, five transcripts were manually coded to identify topics. This provided a human benchmark for the number and contents of topics. While this approach is entirely novel for the current literature, it was incredibly time-consuming and tedious, which led to only one person completing it. The manual coding also revealed the range in topic- and type-counts. Therefore, future research should compare manual-coding results from multiple people, or explore other genres of spontaneous speech.

The second step was to explore the 40 most-frequent words in a transcript, to extract topic-relevant information. However, the raw data from the transcripts contained many noise words. To address this issue, a pre-existing stoplist of

grammatical words was applied to the data, which failed to remove many function words, due to the grammaticalised use of certain words in speech. Once a conversational stoplist was developed, its application tremendously reduced the amount of noise words, thereby distilling the most-frequent words per transcript, into recognisable words from the manual coding. Future research should explore the efficacy of the conversational stoplist on other genres of spontaneous speech.

The question of how to determine the number of topics, and unique words per topic (i.e. types), was also explored from the perspective of automation. This was addressed through the use of coherence-scoring algorithms – two of which were pre-existing, and the other one developed as part of this thesis, through the use of WordNet. All three methods indicated that a minimal number of types and topics was ideal, with no more than eight topics recommend for almost all transcripts. Furthermore, no more than six word-types per topic were recommended by the three methods. Despite agreement across the three methods – including the WordNet method, with its unique architecture – the minimal amount of literature measuring the efficacy of coherence scores is concerning. Future research, therefore, should compare the numerical scores for coherency (as per coherence scores) to human evaluations of clusters.

Existing topic-modelling algorithms consider a document (i.e. transcript) as a whole and identify a specified number of topics anywhere within it. Yet, the manual-coding revealed the tendency for words from the same topic to co-occur in groups. To apply this finding to topic modelling, three methods were devised for dividing a transcript into segments, such that each segment contains one topic within it. By running a topic-modelling algorithm individually in each segment, the algorithm determines a representative cluster for each topic segment. This is in direct contrast to feeding an entire, unpartitioned transcript for a topic-modelling algorithm to analyse, without any context on linguistic features, such as pragmatic markers indicating topic shifts.

That being said, the question of efficacy arose with the cluster results, as the use of partitioning yielded a consistent benefit against no partitioning. Specifically, partitioning ensured that around half of all manually-coded topics were evident in the clusters, when compared to 39% in the stock, non-partitioning algorithm. While 50% is an improvement, it does raise the questions of (1) what is a feasible upper-limit, and (2) what would be the agreement rates across human coders, since it may be that if multiple people are asked to manually-code five transcripts, they will only agree on 50% of their identified topics. Answering these questions requires a comparison of agreement across multiple, human coders.

Given the consistent performance for all partitioning methods, it is worth exploring adjustments to the list of pragmatic, topic-shift markers. Or, if partitioning a transcript by long, interviewer turns, or evenly-sized segments (by word count), always works to a comparable standard.

Lastly, I would welcome further testing of the partitioning and conversational stoplist, with other topic-modelling algorithms.

References

Barth, D., & Schnell, S. (2021). *Understanding corpus linguistics*. Routledge: Abingdon, Oxon. https://doi.org/10.4324/9780429269035

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Bou Orm, H. (2021). *Ethnolectal variation in Lebanese Australian English* [Master of Arts Thesis, Australian National University, School of Literature, Languages and Linguistics].

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, *21*(1), 3–21. https://doi.org/10.1177/1461445618814032

Clayman, S. E. (2012). Turn-Constructional Units and the Transition-Relevance Place. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 151–166). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118325001.ch8

de Arruda, H. F., Costa, L. da F., & Amancio, D. R. (2016). Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *26*(6), 063120. https://doi.org/10.1063/1.4954215

Drew, P. (2018). Epistemics in social interaction. *Discourse Studies*, *20*(1), 163–187. https://doi.org/10.1177/1461445617734347

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S. & Paolino, D. (1993). Outline of discourse transcription. In J. Edwards and M. Lampert (Eds.), *Talking data: Transcription and coding in discourse* (pp. 45–89). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fillmore, C. J. (1982). Frame semantics. *Linguistics in the morning calm* (pp. 111–137). South Korea: Hanshin Publishing Company.

Fillmore, C. J. (1976). Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences*, *280*(1), 20–32. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

Fino, E., Hanna-Khalil, B., & Griffiths, M. D. (2021). Exploring the public's perception of gambling addiction on Twitter during the COVID-19 pandemic: Topic modelling and sentiment analysis. *Journal of Addictive*

Diseases, *39*(4), 489–503. https://doi.org/10.1080/10550887.2021.1897064

Fontana, E. (2020, October 26). Hyperparameters tuning—Topic Coherence and LSI model. *Betacom*. https://medium.com/betacom/hyperparameters-tuning-topic-coherence-and-lsi-model-d31701f8aeec

Ford, C. E., Fox, B. A., & Thompson, S. A. (2014). Social interaction and grammar. In M. Tomasello (Ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, Volume II Classic Edition* (pp. 119–144). New York: Psychology Press. https://doi.org/10.4324/9781315777443

Ford, C. E., Fox, B. A., & Thompson, S. A. (2002). *The Language of Turn and Sequence*. Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/uql/detail.action?docID=430364

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (1st ed., pp. 134–184). Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.003

Fromont, R., & Hay, J. (2008). ONZE Miner: The development of a browser-based research tool. *Corpora*, *3*(2), 173–193. https://doi.org/10.3366/E1749503208000142

Gardner, R. (1998). Between Speaking and Listening: The Vocalisation of Understandings1. *Applied Linguistics*, *19*(2), 204–224. https://doi.org/10.1093/applin/19.2.204

Ghasiya, P., & Okamura, K. (2021). Understanding the Middle East through the eyes of Japan's Newspapers: A topic modelling and sentiment analysis approach. *Digital Scholarship in the Humanities*, *36*(4), 871–885. https://doi.org/10.1093/llc/fqab019

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports*, *7*(1), 45141. https://doi.org/10.1038/srep45141

Gordon, M. J. (2013). *Labov: A Guide for the Perplexed: A Guide for the Perplexed*. Bloomsbury Publishing Plc. http://ebookcentral.proquest.com/lib/anu/detail.action?docID=1106794

Grama, J., Travis, C. E., & Gonzalez, S. (2020). Ethnolectal and community change ov(er) time: Word-final (er) in Australian English. *Australian Journal of Linguistics*, *40*(3), 346–368. https://doi.org/10.1080/07268602.2020.1823818

Grzech, K. (2021). Using discourse markers to negotiate epistemic stance: A view from situated language use. *Journal of Pragmatics*, *177*, 208–223. https://doi.org/10.1016/j.pragma.2021.02.003

Guo, C., Lu, M., & Wei, W. (2021). An Improved LDA Topic Modeling Method Based on Partition for Medium and Long Texts. *Annals of Data Science*, *8*(2), 331–344. https://doi.org/10.1007/s40745-019-00218-3

Guy, G., Horvath, B., Vonwiller, J., Daisley, E., & Rogers, I. (1986). An Intonational Change in Progress in Australian English. *Language in Society*, *15*(1), 23–51.

Hart, G. W. (1994). To decode short cryptograms. *Communications of the ACM*, *37*(9), 102–108. https://doi.org/10.1145/182987.184078

Horvath, B., & Sankoff, D. (1987). Delimiting the Sydney Speech Community. *Language in Society*, *16*(2), 179–204.

Horvath, B. (1985). *Variation in Australian English: The sociolects of Sydney*. Cambridge: Cambridge University Press.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science*, *64*(6), 2833–2855. https://doi.org/10.1287/mnsc.2017.2751

Jaworska, S., & Nanda, A. (2016). Doing Well by Talking Good: A Topic Modelling-Assisted Discourse Study of Corporate Social Responsibility. *Applied Linguistics*, amw014. https://doi.org/10.1093/applin/amw014

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018). *Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey* (arXiv:1711.04305). arXiv. http://arxiv.org/abs/1711.04305

Joty, S., Carenini, G., & Ng, R. T. (2013). Topic Segmentation and Labeling in Asynchronous Conversations. *Journal of Artificial Intelligence Research*, *47*, 521–573. https://doi.org/10.1613/jair.3940

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, *42*(1), 56–73. https://doi.org/10.1016/j.ipm.2004.11.007

Labov, W. (1984). The Sociolinguistic Interview. In *Field Methods of the Project on Linguistic Change and Variation* (pp. 32–34).

Labov, W. (1963). *The social motivation of a sound change*. Word 19: 273-309.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Language Data Commons of Australia (LDaCA). (2021 – 2023). ARDC. Retrieved 2 July 2022, from https://ardc.edu.au/project/language-data-commons-of-australia-ldaca/

Lee, E. (2020). *Quotatives in Australian English: Ethnicity and change over time* [Honours Thesis, Australian National University, School of Literature, Languages and Linguistics].

Lee, J.-B., Lee, T., & In, H. P. (2019). Automatic Stop Word Generation for Mining Software Artifact Using Topic Model with Pointwise Mutual Information. *IEICE Transactions on Information and Systems*, *E102.D*(9), 1761–1772. https://doi.org/10.1587/transinf.2018EDP7390

Lee-Goldman, R. (2011). No as a discourse marker. *Journal of Pragmatics*, *43*(10), 2627–2649. https://doi.org/10.1016/j.pragma.2011.03.011

Lin, H., Zuo, Y., Liu, G., Li, H., Wu, J., & Wu, Z. (2020). A Pseudo-document-based Topical N-grams model for short texts. *World Wide Web*, *23*(6), 3001–3023. https://doi.org/10.1007/s11280-020-00814-x

Lonergan, J., & Cox, F. (2010). Is there any evidence of rhoticity in historical Australian English? In: L. de Beuzeville & P. Peters (eds.), *From the Southern hemisphere: Parameters of language variation, E-Proceedings of the 2008 Conference of the Australian Linguistics Society,* https://ses.library.usyd.edu.au/handle/2123/6860

Miller, G. A., & Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, *41*(2), 209–214. https://doi.org/10.1007/s10579-007-9044-6

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272. https://aclanthology.org/D11-1024

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*, *12*(2), 243–277. https://doi.org/10.3366/cor.2017.0118

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. https://aclanthology.org/N10-1012

NLTK. (2022a). *NLTK :: nltk.stem.wordnet*. Retrieved 11 June 2022, from https://www.nltk.org/_modules/nltk/stem/wordnet.html

NLTK. (2022b). *NLTK :: Sample usage for wordnet*. https://www.nltk.org/howto/wordnet.html

NSW Bicentennial Oral History Project. (1987). *NSW Bicentennial oral history collection:* Council on the Ageing NSW Branch and NSW Oral History Association of Australia. Housed at the National Library of Australia.

Onyimadu, O., Nakata, K., Wang, Y., Wilson, T., & Liu, K. (2013). Entity-Based Semantic Search on Conversational Transcripts Semantic. In H. Takeda, Y. Qu, R. Mizoguchi, & Y. Kitamura (Eds.), *Semantic Technology* (pp. 344–349). Springer. https://doi.org/10.1007/978-3-642-37996-3_27

Pleple, Q. (2013). *Topic Coherence To Evaluate Topic Models*. http://qpleple.com/topic-coherence-to-evaluate-topic-models/

Princeton University. (2022). *WordNet: A Lexical Database for English*. Retrieved 20 March 2022, from https://wordnet.princeton.edu/

Purser, B., Grama, J., & Travis, C. E. (2020). Australian English over Time: Using Sociolinguistic Analysis to Inform Dialect Coaching. *Voice and Speech Review*, *14*(3), 269–291. https://doi.org/10.1080/23268263.2020.1750791

Qiao, F., & Williams, J. (2022). Topic Modelling and Sentiment Analysis of Global Warming Tweets: Evidence From Big Data Analysis. *Journal of Organizational and End User Computing*, *34*(3), 1–18. https://doi.org/10.4018/JOEUC.294901

Rehurek, R. (2022). *Gensim: Topic modelling for humans*. https://radimrehurek.com/gensim/models/coherencemodel.html

Rosner, F., Hinneburg, A., Roder, M., Nettling, M., & Both, A. (2013). Evaluating topic coherence measures. *NeurIPS*, *7*, 4.

Shadrova, A. (2021). Topic models do not model topics: Epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, *2021*. https://doi.org/10.46298/jdmdh.7595

Singular value decomposition. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Singular_value_decomposition&oldid=1113838604

Sinka, M. P., & Corne, D. W. (2003). Towards modernised and Web-specific stoplists for Web document analysis. *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, 396–402. https://doi.org/10.1109/WI.2003.1241221

Skaggs, B. (2014). *Topic Modeling for Wikipedia Link Disambiguation* [M.S, University of Maryland, College Park, Department of Computer Science]. https://dl.acm.org/doi/10.1145/2633044

Stegner, W. (2021). *Context-Aware Malware Detection Using Topic Modeling* [M.S., University of Cincinnati, Department of Electrical Engineering and Computer Science]. https://www.proquest.com/docview/2595956127/abstract/53C053175C71435EPQ/1

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961. https://aclanthology.org/D12-1087

Tengi, R. (1998). Chapter 4 Design and Implementation of the WordNet Lexical Database and Searching Software. In C. Fellbaum (Ed.), *WordNet: An*

Electronic Lexical Database (p. 24). The MIT Press. https://doi-org.virtual.anu.edu.au/10.7551/mitpress/7287.003.0009

Thompson, S. A. (2002). "Object complements" and conversation towards a realistic account. *Studies in Language*, *26*(1), 125–163. https://doi.org/10.1075/sl.26.1.05tho

Travis, C. E. (2014-2022). *Sydney Speaks*. Australian Research Council Centre of Excellence for the Dynamics of Language, Australian National University: http://www.dynamicsoflanguage.edu.au/sydney-speaks/

Travis, C. E. (2021, December 6). *Sydney Speaks: Examining language variation and change through the stories people tell – Sydney Corpus Lab*. https://sydneycorpuslab.com/sydney-speaks-examining-language-variation-andchange-through-the-stories-people-tell/

Travis, C. E., & Johnstone, C. (In Progress). *Introduction to the Sydney Speaks project – compiling sub-corpora for a historical view of Australian English*.

van Dijk, T. A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Longman.

van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Winter, J. (2002). Discourse Quotatives in Australian English: Adolescents Performing Voices. *Australian Journal of Linguistics*, *22*(1), 5–21. https://doi.org/10.1080/07268600120122535

Xia, L., Gu, P., Li, B., Tang, T., Yin, X., Huangfu, W., Yu, S., Cao, Y., Wang, Y., & Yang, H. (2016). Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication. *Journal of Computer Science and Technology*, *31*(1), 3–19. https://doi.org/10.1007/s11390-016-1608-8

Xu, Z., & Zhang, J. (2021). Extracting Keywords from Texts based on Word Frequency and Association Features. *Procedia Computer Science*, *187*, 77–82. https://doi.org/10.1016/j.procs.2021.04.035

Zipf, G. K. (1935). *The psycho-biology of language; an introduction to dynamic philology*. Houghton Mifflin company. https://catalog.hathitrust.org/Record/000359461

Appendices

Appendix 1 – Sydney Speaks Notation

Items removed for computational process (cf. Du Bois et al., 1993)

TOKENS	**NOTES**
[2	
2]	
[3	
3]	
[	
]	
<CRYING	
CRYING>	
<MRC	
MRC>	
<SNG	
SNG>	
<VOX	
VOX>	
<crk	
crk>	
<rct	
rct>	
<sgh	
sgh>	
<ywn	
ywn>	
<WH	
WH>	
<RP	
RP>	
<hi	
hi>	
<rh	
rh>	
<sm	
sm>	
<al	
al>	

<br	
br>	
<sh	
sh>	
<et	
et>	
<@	
@>	
<X	
X>	
<F	
F>	
<P	
P>	
<Q	
Q>	
<a	
a>	
<h	
h>	
<	
>	
th-	common truncation in Sydney Speaks data for the
@	
.	
,	
?	
!	
=	
-	
~	
%	

Appendix 2 – Conversational Stoplist

Conversational stoplist to complement the NLTK stoplist

TOKENS	CATEGORY	BROADER_CATEG	ORIGINS	NOTES
'd	contraction	grammatical	Highly_Frequent	
'll	contraction	grammatical	Highly_Frequent	
'm	contraction	grammatical	Highly_Frequent	

're	contraction	grammatical	Highly_Frequent	
's	contraction	grammatical	Highly_Frequent	
've	contraction	grammatical	Highly_Frequent	
actually	adverb	grammatical	Highly_Frequent	
ah	backchannel	interactional	Highly_Frequent	
along	preposition	grammatical	Highly_Frequent	
alright	backchannel	interactional	Highly_Frequent	
also	adverb	grammatical	Highly_Frequent	
always	adverb	grammatical	Highly_Frequent	
among	preposition	grammatical	Highly_Frequent	
another	vacuous_adj	topic_modelling	Hart_Article	
anyway	adverb	grammatical	Highly_Frequent	
around	preposition	grammatical	Highly_Frequent	
b	letter	other	Highly_Frequent	
back	adverb	grammatical	Highly_Frequent	
bit	functional	conversational	Highly_Frequent	bit of
blah	vacuous	conversational	Highly_Frequent	
c	letter	other	Highly_Frequent	
ca	truncated	other	Highly_Frequent	occasional residue of can’t
can’t	modal	grammatical	Highly_Frequent	
cause	conjunction	grammatical	Highly_Frequent	convo variant of because
certain	vacuous_adj	topic_modelling	Highly_Frequent	
certainly	adverb	grammatical	Highly_Frequent	
cetera	truncated	other	Highly_Frequent	error correction
come	light_verb	conversational	Highly_Frequent	
cool	disc_marker	conversational	Highly_Frequent	
could	modal	grammatical	Highly_Frequent	
duh	vacuous	conversational	Highly_Frequent	
e	letter	other	Highly_Frequent	
em	backchannel	interactional	LING4106_Corpus_Creation	
et	truncated	other	Highly_Frequent	error correction
etcetera	vacuous	conversational	Highly_Frequent	
even	adverb	grammatical	Highly_Frequent	
ever	adverb	grammatical	Highly_Frequent	
every	vacuous_adj	topic_modelling	Highly_Frequent	
everyone	vacuous_pron	grammatical	Highly_Frequent	

everything	vacuous_pron	grammatical	Highly_Frequent	
exactly	adverb	grammatical	Highly_Frequent	
f	letter	other	Highly_Frequent	
feel	epistemic	conversational	Heba	
first	vacuous_adj	numbers	Hart_Article	
g	letter	other	Highly_Frequent	
gather	epistemic	conversational	Highly_Frequent	
get	light_verb	conversational	Highly_Frequent	
give	light_verb	conversational	Highly_Frequent	
go	light_verb	conversational	Highly_Frequent	
gonna	functional	conversational	LING4106_Corpus_Creation	convo variant of going to
good	vacuous_adj	topic_modelling	Hart_Article	
got	light_verb	conversational	Highly_Frequent	
guess	epistemic	conversational	Highly_Frequent	
h	letter	other	Highly_Frequent	
hey	disc_marker	conversational	LING4106_Corpus_Creation	
hm	backchannel	interactional	Highly_Frequent	
hmhm	backchannel	interactional	Highly_Frequent	
hmhmhm	backchannel	interactional	Highly_Frequent	
hmoh	backchannel	interactional	Highly_Frequent	
hmyeah	backchannel	interactional	Highly_Frequent	
hmyeahyeah	backchannel	interactional	Highly_Frequent	
k	letter	other	Highly_Frequent	
kind	functional	conversational	Highly_Frequent	kind of
know	epistemic	conversational	Highly_Frequent	
last	vacuous_adj	numbers	Hart_Article	
let	light_verb	conversational	Highly_Frequent	
like	disc_marker	conversational	Highly_Frequent	
little	vacuous_adj	topic_modelling	Hart_Article	
long	vacuous_adj	topic_modelling	Hart_Article	
look	light_verb	conversational	LING4106_Corpus_Creation	
lot	functional	conversational	Highly_Frequent	lot of
make	light_verb	conversational	Highly_Frequent	
man	vacuous_abstract	topic_modelling	Hart_Article	
many	adverb	grammatical	Hart_Article	
may	modal	grammatical	Hart_Article	
maybe	adverb	grammatical	Highly_Frequent	

mean	epistemic	conversational	Highly_Frequent	
men	vacuous_abstract	topic_modelling	Hart_Article	
mhm	backchannel	interactional	Highly_Frequent	
mhmhm	backchannel	interactional	Highly_Frequent	
might	modal	grammatical	Highly_Frequent	
mostly	adverb	grammatical	Highly_Frequent	
mr	vacuous	conversational	Hart_Article	
much	adverb	grammatical	Highly_Frequent	
must	modal	grammatical	Highly_Frequent	
n	letter	other	Highly_Frequent	
n't	contraction	grammatical	Highly_Frequent	
never	adverb	grammatical	Highly_Frequent	
new	vacuous_adj	topic_modelling	Hart_Article	
oh	backchannel	interactional	Highly_Frequent	
okay	backchannel	interactional	Highly_Frequent	
okayoh	backchannel	interactional	Highly_Frequent	
okayso	backchannel	interactional	Highly_Frequent	
okayyeah	backchannel	interactional	Highly_Frequent	
one	vacuous_adj	numbers	Highly_Frequent	
p	letter	other	Highly_Frequent	
people	vacuous_abstract	topic_modelling	Hart_Article	
perhaps	adverb	grammatical	Highly_Frequent	
probably	adverb	grammatical	Highly_Frequent	
put	light_verb	conversational	Highly_Frequent	
quite	adverb	grammatical	Highly_Frequent	
r	letter	other	Highly_Frequent	
rather	adverb	grammatical	Highly_Frequent	
really	adverb	grammatical	Highly_Frequent	
reckon	epistemic	conversational	Highly_Frequent	
remember	epistemic	conversational	Highly_Frequent	
right	backchannel	interactional	Highly_Frequent	
say	light_verb	conversational	Highly_Frequent	
see	light_verb	conversational	Highly_Frequent	
seem	epistemic	conversational	Highly_Frequent	
slightly	adverb	grammatical	Highly_Frequent	
soandso	vacuous	conversational	Highly_Frequent	
someone	vacuous_pron	grammatical	Highly_Frequent	
something	vacuous_pron	grammatical	Highly_Frequent	
sometimes	vacuous_pron	grammatical	Highly_Frequent	
sort	functional	conversational	Highly_Frequent	sort of
still	adverb	grammatical	Highly_Frequent	

stuff	vacuous	conversational	Highly_Frequent	
super	adverb	grammatical	Highly_Frequent	
suppose	epistemic	conversational	Highly_Frequent	
sure	backchannel	interactional	Highly_Frequent	
take	light_verb	conversational	Highly_Frequent	
thanks	vacuous	conversational	LING4106_Corpus_Creation	
thing	vacuous_abstract	topic_modelling	Highly_Frequent	
think	epistemic	conversational	Highly_Frequent	
though	conjunction	grammatical	Highly_Frequent	
three	vacuous_adj	numbers	Highly_Frequent	
time	vacuous_abstract	topic_modelling	Highly_Frequent	
two	vacuous_adj	numbers	Highly_Frequent	
u	letter	other	Highly_Frequent	
uh	backchannel	interactional	Highly_Frequent	
um	backchannel	interactional	Highly_Frequent	
use	functional	conversational	Highly_Frequent	use to
usually	adverb	grammatical	Highly_Frequent	
w	letter	other	Highly_Frequent	
wanna	functional	conversational	LING4106_Corpus_Creation	convo variant of want to
want	light_verb	conversational	Highly_Frequent	
well	disc_marker	conversational	Highly_Frequent	
went	light_verb	conversational	Highly_Frequent	
whatever	vacuous_pron	grammatical	Highly_Frequent	
whereas	conjunction	grammatical	Highly_Frequent	
whether	conjunction	grammatical	Highly_Frequent	
within	preposition	grammatical	Highly_Frequent	
work	light_verb	conversational	Hart_Article	
would	modal	grammatical	Highly_Frequent	
wow	backchannel	interactional	Highly_Frequent	
yeah	backchannel	interactional	Highly_Frequent	
yeahhm	backchannel	interactional	Highly_Frequent	
yeahhmyeah	backchannel	interactional	Highly_Frequent	
yeahi	backchannel	interactional	Highly_Frequent	
yeahit	backchannel	interactional	Highly_Frequent	
yeahoh	backchannel	interactional	Highly_Frequent	
yeahso	backchannel	interactional	Highly_Frequent	
yeahyeah	backchannel	interactional	Highly_Frequent	

yeahyeahyeah	backchannel	interactional	Highly_Frequent	
yep	backchannel	interactional	Highly_Frequent	
yepyeah	backchannel	interactional	Highly_Frequent	
yes	backchannel	interactional	Highly_Frequent	
yet	adverb	grammatical	Highly_Frequent	

Appendix 3 – Clusters vs. Manual Coding

When reading Table 15 to Table 19, it is important to note that two cluster quantities were run per transcript, using the coherence scores in Chapter 4 to decide these values. For each quantity of clusters, the respective columns list the specific cluster which corresponded to the relevant topics. For example, "#8" under 8 clusters for Even Partitions in the topic of Alcohol Use, indicates that the 8th cluster in the total set of eight corresponded to the topic of Alcohol Use, when the Even algorithm was run. This structure allows for the relevant cluster to be located in the Appendix for verification.

Table 15 Marjorie's Results with 4 and 8 Clusters obtained per Algorithm

Topics		**4 clusters**	**8 clusters**	**Total**
Alcohol Use	No Partitions			0
	Even Partitions		#8	1
	Pragmatic Partitions			0
	Turn Partitions	#3	#6, 7, 8	4
Clothing and Sewing	No Partitions			0
	Even Partitions		#6	1
	Pragmatic Partitions			0
	Turn Partitions			0
Familial Relations	No Partitions	#1	#7	2
	Even Partitions		#5	1
	Pragmatic Partitions	#4	#8	2

	Turn Partitions	#2	#5	2
Heart Problems	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Language Attitudes and Use	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Miscellaneous Charity	No Partitions			0
	Even Partitions	#1		1
	Pragmatic Partitions			0
	Turn Partitions			0
Radios and Broadcasts	No Partitions			0
	Even Partitions		#7	1
	Pragmatic Partitions			0
	Turn Partitions			0
School Life	No Partitions	#2, 4	#4, 6	4
	Even Partitions	#2	#3, 4	3
	Pragmatic Partitions	#2, 3	#2, 3, 5	5
	Turn Partitions		#2, 3	2
Singing	No Partitions		#2	1
	Even Partitions			0
	Pragmatic Partitions		#7	1
	Turn Partitions		#4	1
Stores	No Partitions		#1	1
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Sydney City Mission	No Partitions	#3		1
	Even Partitions		#1	1
	Pragmatic Partitions	#1	#1	2
	Turn Partitions		#1	1
The Depression	No Partitions			0
	Even Partitions		#2	1
	Pragmatic Partitions			0

	Turn Partitions	0

Table 16 Isa's Results with 4 and 6 Clusters obtained per Algorithm

Topics		**4 clusters**	**6 clusters**	**Total**
Childhood in Greece	No Partitions	#1		1
	Even Partitions			0
	Pragmatic Partitions	#1		1
	Turn Partitions			0
Games	No Partitions			0
	Even Partitions		#1	1
	Pragmatic Partitions			0
	Turn Partitions			0
Greek Afternoon-School	No Partitions		#1	1
	Even Partitions		#5	1
	Pragmatic Partitions		#5	1
	Turn Partitions	#2, 3	#2, 3	4
Language Use and Attitudes	No Partitions		#2	1
	Even Partitions	#4	#6	2
	Pragmatic Partitions	#4		1
	Turn Partitions		#6	1
Mum's School-Life	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions	#2, 3		2
	Turn Partitions			0
School Life	No Partitions	#3, 4	#3, 4, 5	5
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions		#1	1

Table 17 Janice's Results with 4 and 20 Clusters obtained per Algorithm

Topics		**4 clusters**	**20 clusters**	**Total**
Appreciating Glass	No Partitions			0
	Even Partitions		#8	1
	Pragmatic Partitions			0

	Turn Partitions		#6	1
Attitudes about Education	No Partitions		#14	1
	Even Partitions		#12, 13, 14	3
	Pragmatic Partitions	#4	#15	2
	Turn Partitions	#3	#13, 14	3
Class Post-Interview	No Partitions			0
	Even Partitions		#15	1
	Pragmatic Partitions			0
	Turn Partitions		#16	1
Dialectal Variation and Language Use	No Partitions		#17	1
	Even Partitions	#4	#16, 17, 18, 19, 20	6
	Pragmatic Partitions		#16, 17, 18, 19, 20	5
	Turn Partitions		#17, 18, 19, 20	4
Games	No Partitions	#4	#5, 6, 7, 8, 9, 10	7
	Even Partitions		#5, 6, 7, 9	4
	Pragmatic Partitions	#3	#5, 6, 7, 8, 9, 10, 11	8
	Turn Partitions	#2	#4, 5, 7, 8	5
Janice's Mother	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions		#12, 13	2
	Turn Partitions		#9, 10, 11	3
School Life	No Partitions	#2	#2, 12, 13, 14, 15, 16	7
	Even Partitions		#1, 2, 3, 4, 11	5
	Pragmatic Partitions	#1, 2	#1, 2, 3, 4	6
	Turn Partitions	#1	#1, 2, 3, 12	5

Table 18 Parker's Results with 4 and 6 Clusters obtained per Algorithm

Topics		**4 clusters**	**6 clusters**	**Total**
Assault	No Partitions		#4	1

	Even Partitions	#1		1
	Pragmatic Partitions			0
	Turn Partitions		#2	1
Bus (Mis)Management	No Partitions	#2	#1	2
	Even Partitions		#3, 4	2
	Pragmatic Partitions			0
	Turn Partitions	#2	#3	2
Bus Life	No Partitions		#2	1
	Even Partitions	#2		1
	Pragmatic Partitions	#2	#3	2
	Turn Partitions	#1		1
Driving Instructor Work	No Partitions		#5, 6	2
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Family Life	No Partitions			0
	Even Partitions		#2	1
	Pragmatic Partitions	#3	#5	2
	Turn Partitions		#5, 6	2
Foxing	No Partitions			0
	Even Partitions		#5	1
	Pragmatic Partitions		#4	1
	Turn Partitions	#3	#4	2
Funding Cuts	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Interactions with Bus CEO	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Magazine Distributing	No Partitions			0
	Even Partitions	#1	#1	2
	Pragmatic Partitions	#1	#1	2
	Turn Partitions		#1	1
Public Service Work	No Partitions			0

	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Race Relations	No Partitions	#4		1
	Even Partitions			0
	Pragmatic Partitions	#4	#6	2
	Turn Partitions	#4		1
Retail Work	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Sam's Auto Parts	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Train Driver Work	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Warehouse Job	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions		#2	1
	Turn Partitions			0

Table 19 Zander's Results with 6 and 8 Clusters obtained per Algorithm

Topics		**6 clusters**	**8 clusters**	**Total**
Attitudes on American Terms	No Partitions	#1, 5		2
	Even Partitions		#5	1
	Pragmatic Partitions		#4	1
	Turn Partitions	#3	#3	2
Bogan Characteristics	No Partitions			0
	Even Partitions	#3	#4	2
	Pragmatic Partitions			0
	Turn Partitions			0
Dialectal Variation	No Partitions	#3, 4	#3, 4, 5	5

	Even Partitions	#2, 6	#3, 8	4
	Pragmatic Partitions	#4	#6	2
	Turn Partitions	#5, 6	#7, 8	4
Donald Trump	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Food	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Gaming	No Partitions			0
	Even Partitions	#1	#1	2
	Pragmatic Partitions	#1	#1, 3	3
	Turn Partitions	#1	#1	2
International Views about Australians	No Partitions	#2	#1	2
	Even Partitions	#6		1
	Pragmatic Partitions	#3		1
	Turn Partitions		#4	1
Lindt Cafe Siege	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions	#5	#7	2
	Turn Partitions			0
Movies	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions		#2	1
	Turn Partitions			0
Morning Shows	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Occupations	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Overseas Travel	No Partitions			0

	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
Race Relations	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions			0
	Turn Partitions			0
School and Uni Life	No Partitions		#6, 7	2
	Even Partitions	#5	#6	2
	Pragmatic Partitions			0
	Turn Partitions	#4	#5	2
Social Media	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions	#2	#4	2
	Turn Partitions			0
Socioeconomic Stance	No Partitions			0
	Even Partitions			0
	Pragmatic Partitions	#6	#8	2
	Turn Partitions			0

Appendix 4 – Sydney Speaks 2010s Clusters

Table 20 Parker's Clusters per Algorithm when 4 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Guy, short, thousand, find, aboriginal, drive	Service, magazine, pop, Bexley, Kingsford, company	Magazine, bank, markup, cash, believe, owner	Bus, na, live, kill, find, aboriginal
Bus, away, delegate, obviously, lock, thirty	Bus, story, throw, end, drive, four	Guy, head, Stanmore, gon, prostitute, passenger	Sydney, bus, issue, fighting, find
Day, group, o'clock, number, middle, mother	Group, interesting, wo, tennis, hundred, expense	Learn, slang, anything	Guy, roster, single, show, na, fulltimer
Group, day,	Guy, place, attack,	Greek, young,	Greek, hell, un,

speak, road, night, aboriginal	na, hard, six	block, change, skin, others	Eyetal, home, next

Table 21 Parker's Clusters per Algorithm when 6 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Guy, room, exam, delegate, fight, access	Magazine, timeframe, wo, ask, stack, level	Magazine, bank, markup, cash, believe, owner	Service, magazine, drive, Australia, level, warehouse
Bus, prostitute, passenger, trip, charge, aboriginal	Course, ex-wife, step, melt, mind, wife	Tell, sell, buy, dollar, packet	Bus, kill, complaint, smash, seriously, minor
Day, group, gon, tenekes, journal, basically	Bus, eight, roundabout, harass, ban, suburb	Bus, listen, human, send, story, drive	Sydney, bus, issue, fighting, find
Group, day, compensation, journal, fight, access	Depot, shift, launch, scar, live, ninety-one	Guy, Sydney, locked, pull, Wentworth, fox	Guy, roster, single, show, na, full-timer
Job, call, driver, ring, room, street	Guy, steal, hold, fox, xyz, woman	Learn, slang, anything	Call, Greek, bin, tenekes, basically, skoupidotenekes
Call, job, ring, tenekes, beat, woman	Day, guy, skoupidia, Campbelltown, sack, girl	Greek, st, racist, aboriginal, door, nine	Cake, school, across, play, mother, chess

Table 22 Zander's Clusters per Algorithm when 6 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Australian, different, form, America, rare, absolutely	Game, pair, world, feel, need, find	Zombie, amaze, game, awesome, online, call	Game, pair, next, doe, change, issue
Different, Australian, reading, fact, hang, next	Australia, talk, unite, felt, surrounding, lately	Streak, respect, physically, whatnot, insight, conversation	Australia, near, middle, bet, element, veggie

Australia, call, word, extra, interesting, Carnes	Bogan, throw, pf, full, killer, double	Australia, derogatory, favourite, bet, cruise, goodness	Call, American, thong, flop, Australian, deny
Word, Australia, call, complicated, keep, absolutely	Different, school, wor, ba, fair, forget	Word, different, example, happen, complicate, way	Year, exception, literature, fai, master, flair
Call, word, Australia, America, class, Carnes	Year, master, devote, fact, hopefully, understanding	Lindt, siege, Sydney, siege	Accent, diversity, accentuated, speak, emphasis, Australia
Love, way, flop, hill, denham, friend	Accent, mind, ti, eng, Jackman, lose	Different, hang, richer, across, government, lucky	Speech, change, normal, low, full, completely

Table 23 Zander's Clusters per Algorithm when 8 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Different, Australian agree, emphasis, hang, friend	Game, find, feel, outside, world, actual	Zombie, amaze, game, awesome, online, call	Game, pair, next, doe, change, issue
Australian, different, Carnes, form, fact, sunrise	World, successful, find, ye, j, race	movie	Australia, near, middle, bet, element, veggie
Call, word, Australia, middle, funny, send	Australia, hear, mingle, fearfu, cash, street	Game, watch, read, angry, PS-Three, difference	Call, American, thong, flop, Australian, deny
Australia, word, call, follow, white, send	Bogan, age, throughout, teeth, play, cheer	World, streak, whatnot, respect, sorry, America	Australia, live, area, young, anyone, way
Australia, call, word, person, fact, emphasis	School, American, call, taught, notice, scenario	Australia, shift, year, vegetable, embrace, agree	School, mum, beautiful, stressed, university, age

Way, love, follow, bad, stressed, family	Unit, far, beware, thoroughly, friend, happy	Different, word, reason, way, enjoy, complicate	Year, plain, clashing, amazing, day, fund
Love, way, middle, li, class, middle	Year, therefore, Michel, lose, seme, policy	Lindt, siege, Sydney, siege	Accent, diversity, speak, emphasis, Australia, change
Sorry, PS-Four, person, emphasis, need, keep	Speech, cou, Queensland, try, else, sp	Different, richer, friend, movie, nobody, person	Speech, food, simple, idea, anywhere, Queensland

Appendix 5 – SSDS Clusters

Table 24 Janice's Clusters per Algorithm when 4 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Fact, term, wall, able, interest, name	Hop, plant, describe, pole, entrance, love	City, important, play, Longueville, kid, public	Square, room, sail, facility, flag, transfer
School, northern, forget, allow, mathematics, anyway	Marble, pull, al, thi, base, father	Longueville, public school, dirt, playground	Marble, colour, play, circle, glass, big
Find, term, second, highly, kid, window	Fact, learn, school, loop, detention, university	Square, marble, wood, swirl, throw, self-expression	Miss, old, opportunity, ye, father, generation
Marble, square, difficult, system, French, enjoy	Australian, ocker, suburb, hill, particularly, sentence	Fact, age, head, sense, schooling, heart	Fact, friend, Newcastle, together, football, bring

Table 25 Janice's Clusters per Algorithm when 20 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Fact, often, system, away, age, Longueville	Start, school, actual, Longueville, bang, self-expression	City, important, play, Longueville, kid, public	School, public, city, Longueville, kid, Cinderella-dressed-in-

			yellow
School, area, allow, stand, shed, weather	Wooden, arrive, class, nice, building, roof	Longueville, public, school, dirt, playground	Start, enjoy, class, Australia, school, draw
Find, self-expression, shop, particular, mathematics, fun	Nice, side, brick, able, weather, garden	Describe, primary, school, surprised	Square, step, grow, bannister, room, curve
Square, difficult, marble, age, name, week	Playground, way, dirt, rain, lunchtime, allow	Nice, class, toilet, circle, single, rounder	Marble, play, circle, game, big
Marble, square, difficult, museum, allow, away	Hop, square, allow, throw, leave, double	End, throw, taw, square	Big, become, rare, favourite, mainly, base
Difficult, marble, square, area, self-expression, single	Play, marble, square, bag, al	Hop, allow, turn, leave, square, foot	Glass, become, love, colour, piece, transparent
Nice, play, hop, start, day, spell	Marble, circle, able, main, colour, swirl	Complicated, allow, become, jump, play, couple	Br, milky, opaque, interest, prize, call
Start, day, play, nice, hop, hand	Glass, big, somehow, effect, love, belong	Square, become, anymore, away, miss, throw	Marble, nice, find, play, collection, favourite
Start, day, hop, nice, play, museum	Marble, br, nice, mother, milky, play	Marble, result, earth, able, reasonable, ball	Mother, magically, appear, dread, forget, ask
Day, play, hop, nice, start, Sydney	Shop, day, job, jeweller, pretty, keep	Opaque, br	Forget, ask, mother, nice, kitchen, war
Hop, nice, start, play, day, shop	Teacher, school, name, nickname, friend, babysitter	Marble, milky, appear, round, collection, nice	Day, shop, home, result, pretty, housework
Learn, class, great, big, way, hand	Spell, way, group, fact, miss, skill	Mother	Miss, teacher, school, name, amongst, nickname

Big, way, great, class, learn, tend	Fact, find, admire, composition, course, emphasis	Day, shop, clothes, job, jeweller, babysitter	Way, education, better, school, today, different
Great, learn, class, big, way, emphasis	Learn, advanced, heart, self-discipline, bring, wonder	Nickname, teacher	Learn, discrepancy, discipline, case, band, longer
Way, big, class, great, learn, twenty	Great, age, learn, year, twenty, education	Fact, together, literacy, moment, bright, past	Conclusion
Class, great, learn, big, way, throw	Pick, Australian, past, speak, Newcastle, start	Ocker, exist	Term, start, second, past, quarter, six
Suburb, Australian, year, ocker, become, different	Find, museum, different, art, accent, suburb	Paul, Hogan, dry, million, advertising, sense	Ask, question, language
Australian, become, miss, suburb, ocker, year	Hogan, speak, try, talk, eastern, Paul	Ocker, find, particular, Sydney, speaker, place	Pick, Australian, difference, Queenslander, Victoria, Newcastle
Different, suburb, miss, become, ocker, year	Ocker, suburb, call, Australian, tend, find	Suburb, consensus, talk-back, radio, con, loathe	Eastern, different, speak, suburb, pie, neutral
Different, year, become, suburb, miss, ocker	Accent, ocker, ockerism, cringe, fact, splendid	Ockerism, Australian, ocker, film, tend, cringe	Ocker, America, Australian, ockerism, tend, suburb

Table 26 Isa's Clusters per Algorithm when 4 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Greek, farm, fun, study, else, na	Play, live, saucer, building, fast, difference	School, play, macos, ru, fish, eighteen	School, somebody, indoor, plaster, eighty, na
Speak, paper, outside, unfair, scalp, browny	Teacher, fall, per, care, mathema, anything	Learn, different	Lesson, English, Greek, teach, kid, forget

School, gymnastics, live, funny, month, monkey	Greek, find, group, morning, menstruating, unfairly	Mum, geography, mathematics, subject, learn, differen	Greek, lesson, English, wonder, teach
Teacher, easy, Israel, Darwin, child, repeat	Speak, sound, particular, mark, sharp, understand	Greek, speak, abbey, jus, keep, white	Speak, trouble, al, small, Sydney, eat

Table 27 Isa's Clusters per Algorithm when 6 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Greek, venue, chance, old, paint, study	Play, tree, Nate, night, tell, name	School, play, macos, ru, fish, eighteen	School, somebody, indoor, plaster, eighty, na
Speak, Australia, house, repeat, else, loud	Teacher, matter, tell, fish, public, Greek	Learn, different	Lesson, teach, English, Greek, speak, kid
School, study, live, fun, keep, sgh	Call, story, live, nose, sea, depend	Mum, geography, mathematics, subject, learn, differen	Lesson, Greek, lesson, Greek, English, kid
Teacher, ten, class, voice, loud, paint	Greek, flaky, question, country, instance, tha	Call, mathema, month, baldy, person, past	Infant, language, question, forget
English, loud, night, read, class, forty	Greek, speak, learn, teacher, language, afternoon	Greek, next, year, pronunciation, period, already	English, lesson, Greek, wonder, teach
Real, loud, keep, paper, anymore, uniform	Speak, way, accent, try, English, talk	Speak, stand, Chinese, enough, test, trouble	Speak, particular, except, fo, tend, half

Appendix 6 – Bicentennial Clusters

Table 28 Marjorie's Clusters per Algorithm when 4 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Mother, dry,	Pay, experience,	Pay, Sydney, four-	Mother, dumbbell,

family, live, crown, movement	tea, four-pence, Simmons, seven	pence, Simmons, service, experience	win, sm, beautiful, live
Sing, science, room, domestic, neither, blanket	School, commercial, fourteen, bee, hope, somebody	Qualifying, certificate, qualif, pass, qualifying, certificate	Mother, forget, Peter, hunter, machine, later
Sydney, person, tennis, leave, blanket, forget	Sing, trouble, ai, sa, stay, company	Domestic, school, science, girl, cross, behind	Dad, pub, Mitch, call, drink, year
School, day, grandfather, floor, dye, science	Race, start, portion, exciting, orphan, bottle	Mother, shout, movement, nice, important, family	Please, day, early, mother, health, suit

Table 29 Marjorie's Clusters per Algorithm when 8 were Extracted

No Partitions (Stock LSI)	**Even**	**Pragmatic**	**Turn**
Mother, couple, room, business, forget, neither	Mission, Sydney, anyone, floor, sm, big	Pay, Sydney, four-pence, Simmons, service, experience	Pay, tea, four-pence, experience, Simmons, service
Sing, food, prize, able, tennis, class	Depression, pay, relief, money, food, mother	Qualifying, certificate, qualif, pass, qualifying, certificate	School, street, girl, nineteen, Sydney, domestic
Sydney, movement, great, love, keep, child	School, neither, great, talk, laneway, routine	Domestic, science, school, Sydney, cross, different	Wand, physical, movement, foot, bah, club
Day, grandfather, school, far, live, physical	Teacher, day, way, march, hard, round	Laneway, call, year	Sing, stay, disappointment, eyesight, forget, beautiful
Grandfather, day, school, die, dye, portion	Sing, ace, affection, hear, mother, regret	Year, leave, school	Mother, forget, Peter, hunter, machine, later
School, day, grandfather,	Boot, glove, archbishop, Ireland,	Learn, ed, disappointment,	Dad, pub, Mitch, call, drink, year

dry, far, play	mother, allow	commercial[4], pas, yard	
Home, saint, set, dye, family, Ireland	Choir, music, cat, whisker, aerial, saint	Sing, saint, difficult, ne, clock, Bourke	Drinker, pub, mother, early, heavy, day
Saint, home, crown, mind, Vincent, keep	Race, jack, run, tha, drinker, radi	Grandfather, pair, twenty-eight, late, granddaughter, talk	Suit, health

4 Appears as “commerical” in the output, which I assume to be a transcription typo of “commercial”