Chapter 13:

Visualizing conversations in health care: Using Discursis to compare Cantonese and English data sets

Authors and affiliations (in order):

1. **Alice YAU**
Centre for the Applied English Studies
The University of Hong Kong, Hong Kong SAR, China
2. **Margo TURNBULL**
International Research Centre for the Advancement of Health Communication
Department of English
The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR, China
3. **Daniel ANGUS**
Digital Media Research Centre
School of Communication
Queensland University of Technology, Brisbane, Australia
4. **Bernadette WATSON**
International Research Centre for the Advancement of Health Communication
Department of English
The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR, China

Funding: The research reported in this chapter was funded by the Faculty of Humanities, Dean's Reserve, The Hong Kong Polytechnic University.

Abstract

Health care is shaped by often complex communication between multiple people such as doctors, nurses, patients and carers. Research has repeatedly shown that effective communication is key to safe and high-quality care yet improving communication remains a challenge across health systems. In recent years the field of natural language processing has developed analytic tools to supplement the study of verbal communication through visual representation of anlaysis. To date these tools have primarily been used on English data. This study used the software tool Discursis to compare visual representations of Cantonese conversational data that were analysed before and after English translation. Results indicate that some linguistic features of Cantonese that carry meaning through may be lost in translation into English. Specific concerns relate to the multidimensional issues of equivalence, ranging from cultural and social associations to semantic, lexical and conceptual differences. These results highlight the importance of developing visual analytic tools that can be used on Cantonese data. Generating visual representations of such data contributes to local and international understandings about communication in health care.

Keywords: Discursis, Cantonese, Natural language processing

Introduction

Communication in health care is complex and involves ongoing and dynamic interactions between people. Communication incorporates multiple modes of interaction (such as spoken and written formats), speakers, contexts and communicative events (such as meetings, conversations, emails, etc.: Bondi, 2017). Effective and meaningful communication involves both the transfer of information *and* the generation of shared meaning between doctors, nurses, patients and carers. Meaningful communication in health care is vital in terms of safety and quality of care, as well as empowering people to improve their own health by taking up preventative and curative advice (Goldstein, MacDonald, & Guirguis, 2015; Street, Makoul, Arora, & Epstein, 2009).

Although health-related information can be shared across a range of modalities, such as electronically or in print, *conversations* between people remain a key element of health communication. Conversation involves multiple interactants and is dynamic, purpose driven and reflective (Bondi, 2017). As conversation unfolds meaning is generated and shared, as the interactants come to discuss a shared topic. This can be described as semantic alignment (Tolston, Riley, Mancuso, Finomore, & Funke, 2018). Semantic alignment can be achieved within a conversation even if participants do not agree or reach consensus.

Human conversation and dialogue has been studied extensively using a range of well-established theoretical and methodological approaches (see Weigand, 2017). Recent developments in Natural Language Processing (NLP), a sub-field of computer science concerned with analyzing and understanding human language (Hirschberg & Manning, 2015), have seen the increasing use of computer programs to process conversational data and supplement other methods of qualitative and quantitative analysis. Discursis, one such NLP tool, has been used successfully in recent studies of healthcare in both clinical and managerial contexts (Atay et al.,

2015; Chevalier, Watson, Barras, Cottrell, & Angus, 2018; Watson, Angus, Gore, & Farmer, 2015). Although Discursis and other NLP tools may be theoretically able to process non-alphabetic input, most published work to date has reported on the analysis of data sourced in or translated to alphabetic languages (in the case of the related software Leximancer: English, Italian and Danish: Evers, Marroun & Young, 2017; Franzoni & Bonera, 2019).

This chapter reports on a novel study that compared the analytic outputs produced by Discursis following the analysis of conversational data uploaded to the program written both in Cantonese Chinese characters and English. Analytic outputs were compared in terms of the representation of inter-speaker engagement, conceptual alignment and levels of interaction. The comparison of data outputs based on Cantonese and translated data highlighted important conversational markers that are at risk of being lost through either translation or inadequate customization of software for use with a logographic language.

This chapter begins by briefly examining the importance of conversation in health care, as both a tool for information exchange and as a process through which meaning is generated. We then introduce Discursis as an analytic tool that can supplement other qualitative and quantitative approaches to conversational analysis. We pay particular attention to Discursis's visualisation plots which represent data-grounded, time-ordered exchanges between conversational participants (Chevalier et al 2018, p. 3). These plots show map patterns of conversational features such as turn taking, concept sharing and topic maintenance, which contribute to the generation of meaning through ongoing exchanges. We draw on the analytic outputs from Discursis to discuss three key features of Cantonese that influence analysis in Discursis. These features include the logographic nature of written Cantonese, the significance of the 'word' units used in the software analysis and the need for software adaptation to

accommodate the lexical and semantic role of tone in Cantonese. We conclude this chapter by examining how the results of this research underscore the need for continued investment in the development of analytic tools like Discursis that can be used for the analysis of alphabetic and logographic data in health communication research.

Conversations in Healthcare

Health care is shaped by a diverse range of interactions between people. Recent research has noted the impact of spoken communication and conversation on knowledge, motivation, diagnosis, treatment and management of health conditions (Nouri & Rudd, 2015). It has also been argued that as doctors spend *decreasing* amounts of time with patients yet need to exchange *increasingly* large and detailed amounts of information associated with treatment and diagnosis, the need for effective and efficient communication has grown (Nouri & Rudd, 2015). Communication about health in a broader sense involves discussion about a wide range of information including physical conditions, lifestyle factors and the broader context in which people live.

Despite technological advances reflected in the growth of tools associated with e-health and tele-health, conversation remains a core element of health care. Conversations are made up of linguistic interactions and exchanges between people which result in the transmission of information and generation of shared meaning even when speakers do not agree about the subject being discussed. These exchanges are shaped by the use of semantic and lexical features including vocabulary and grammar as well as contextual and relational markers that connect words, phrases and spoken sentences in meaningful ways. Importantly, the language used in

conversation also reflects the relationships that connect people in terms of familiarity, authority and professional expertise.

As conversation unfolds between participants, meaning is shared and generated through various linguistic devices such as turning taking, asking questions, making statements or repeating what another speaker has said. Through this process, interpersonal and semantic alignment (Tolston et al., 2018) is achieved. Semantic alignment is not predetermined but is generated as a conversation progresses and meaning is negotiated and debated. Spevack et al (2018) described conversation as marked with ambiguities that are progressively addressed and resolved through dynamic and unpredictable interaction. This dynamic nature of conversation can be contrasted with comparatively static types of interactions such as letters or emails and other verbal interactions such as giving instructions which feature a predominantly one-way flow of information.

Tolston et al (2018) argued that semantic and interpersonal alignment are important features of conversation that can be quantitatively and qualitatively measured. Markers such as turn taking, cross-referencing of ideas and links across time periods are features that can be identified, measured and described. Similarly, Chevalier et al (2018) noted that effective communication between participants can be reflected in the mapping of the extent of engagement between speakers, the relative contributions each speaker makes and how consistently topics are maintained within a conversation.

Natural Language Processing and Using Discursis to Produce Visual Analysis

Various methodologies have been developed to study and describe how people interact through conversation including Communication Accommodation Theory (CAT), Conversational

Analysis (CA) and Discourse Analysis (DA) (see chapters by Harrison & Lam; Yip & Zhang; Schoeb & Yip in this edition). These approaches focus in different ways on the analysis of the content and process of linguistic communication – that is, the study of what is said or understood by participants in a conversation and how these messages are conveyed, modified, accepted or rejected. These approaches can be described as *process analysis* which focus on in-depth coding of data and the *microanalysis* of conversational features (Heritage & Maynard, 2006). Various elements of conversation are described in terms such as syntax, semantics and phonetics (Spevack et al., 2018).

In recent years, the field of NLP has developed a range of computer-based tools which can be used to provide additional and supplementary data to the study of conversational content and process by focusing on multiple dimensions of conversation. Analysis of these multiple dimensions aims to identify patterns, structures and the orderliness of exchanges and interactions and to present these findings visually (Angus, Smith, & Wiles, 2012). Examples of such machine-based or automatic analytic methods include latent semantic analysis and word2vec (Tolston et al., 2018). These tools map semantic alignment through the analysis of large amounts of data and by quantifying, "linguistic and communicative co-ordination…the alignment of the meaning of the content of the utterances rather than the syntactical, morphological, or lexical alignment" (Tolston et al., 2018, p. 3).

In contrast to these tools which draw on large quantities of data, the Discursis software described in this chapter can analyse relatively small data sets. The software is straightforward to use as the analyst simply uploads a transcript or text as a comma separated value (.csv) file, makes a few parameter selections, then analyses the resulting visual and metric outputs. A

transcript of as few as 10 conversational turns can be used as the basis for Discursis' language corpus, thereby enabling analysis of very brief conversational exchanges (Tolston et al., 2018).

Once a transcript is selected, Discursis automatically builds a data-grounded natural language model from this same input text (and only this input text) by using Leximancer's conceptual modelling algorithm (Smith & Humphreys, 2006). The Leximancer concept model uses Bayesian statistics to identify groups of words that co-occur within an input text. This bottom-up approach is based on a bag-of-words assumption which draws on the idea that words which collocate often have an associated meaning, or in the words of John Firth: "You shall know a word by the company it keeps!" (Firth, 1962, p.11). Leximancer and Discursis refer to these bags-of-words as concepts. By identifying and grouping words in this way, input text can be reduced in length from thousands of unique words, to a smaller number of concepts. This process is preceded by the removal of words included in a stop word list. This stop word list consists of words that occur frequently but carry little or no semantic meaning. Examples of such words in English include *a, an, and, the*. A default stop word list is used by Discursis and this can also be manually generated or customized by the analyst depending on the text and communicative interaction being analysed.

After building a language model from the entire input transcript, Discursis then codes the transcript to indicate which concepts are present per turn. This coding process involves the automatic identification and labelling of concepts within each conversational turn. Discursis performs this coding by looking for words within a single turn that can be considered as evidence for the presence of specific concepts, and if found, uses these words as markers that this concept is present in this turn.

Once a transcript is coded, Discursis uses a graphical interface to present this coding visually as a recurrence plot. A Discursis recurrence plot features vertically and horizontally adjacent coloured squares (recurrence elements) that highlight instances of conceptual repetition between pairs of conversation turns (see Figure 13.1 for an example). These plots are a useful tool for the analyst who can then click on squares to see detail about the concepts identified as well as their actual location within a text. Discursis can also produce other statistical and visual data to support quantitative and qualitative analysis (Angus, Rintel, & Wiles, 2013; Angus et al., 2012).

Studying A Logographic Language: Contrasting Features of Cantonese and English Data

As previously discussed, Discursis has been used to analyse a range of health related data. However, at the time of writing, this growing body of published work described the analysis of alphabetic data (i.e. data sourced or translated into English or other alphabetic languages). Cantonese, the version of Chinese spoken in Hong Kong and the broader Guangdong province including Macau, is a logographic language and is written in Chinese (hanzi) characters (Ho & Bryant, 1997). From a historical perspective, Cantonese developed as a predominantly spoken, rather than written, language. This is reflected in Cantonese's linguistic subtlety and complexity (Snow, 2004). Cantonese has been the focus of significant linguistic analysis but discussion of this is beyond the scope of this chapter (see Matthews & Yip, 2011; Snow, 2004). There are, however, three key features of Cantonese that can be contrasted with English and are relevant to this discussion of Discursis and the ongoing development of NLP models.

Firstly, written Chinese characters and symbols used in other logographic systems do not correspond directly to spoken components (phonemes or morphemes) as they do in alphabetic languages like English, Spanish or Russian. In contrast, Chinese characters map onto spoken syllables and carry semantic and lexical significance (Liu & Hsiao, 2014; Wong, Juang, & Chen, 2012). The semantic and lexical information that is encoded within the characters is often carried within English sentences. For example, the Cantonese 多謝 (do1ze6) and 唔該 (m4goi1) can both be translated into the English form 'thank you.' However, 多謝 (do1ze6) is used when someone offers you a gift while 唔該 (m4goi1) is used when someone offers you a service or help. In English, these differences and the additional information would be set out in the context of the phrase or sentence in which 'thank you' is embedded. For example, 'thank you for the gift' or 'thank you for helping me.'

Secondly, Cantonese is a *tone* language and each spoken unit, "has a lexical tonal pattern" (Fok 1972, p. 1 in Chan & Li, 2000, p. 76). The significance of tone in Cantonese is not the same as intonation in English. A change in intonation in English usually suggests a difference in attitude or significance rather than a change in meaning (Chan & Li, 2000). For example, changing emphasis on the italicised words below does not change the underlying meaning of the sentence [i.e. the doctor (subject) walked (verb) to the hospital (object)]:-

The doctor *walked* to the hospital.

The doctor walked to the *hospital*.

In contrast, tone in Cantonese carries fundamental lexical and semantic significance. Changing the tone in a Cantonese word changes the meaning of the word. There are six distinctive tones in Cantonese (Bauer & Benedict, 1997; Matthews & Yip, 2011) which are

indicated by numbers when transcribed in the Romanized form of Jyutping[1] (as shown in the example below). The tone used with the morpheme or word partly determines lexical and semantic meaning and additional meaning is determined by the broader context of the utterance. This can be seen in the examples given below:

- First tone (high level): *maa1* 媽 (mother), 孖 (twin)
- Second tone: not possible
- Third tone (mid level): *maa3* 嗎 (question particle)
- Fourth tone (low falling): *maa4* 麻 (hemp)
- Fifth tone (low rising): *maa5* 馬 (horse)
- Sixth tone (low level): *maa6* 罵 (scold, abuse)

However, this tonal system restricts a speaker's ability to manipulate pitch which usually conveys a range of communicative information about speaker's interactions including attitudes (surprise, doubt, hesitation, reluctance, etc.) as well as speech-acts such as asking, requesting, refusing or persuading (Bauer & Benedict, 1997). While English largely expresses this information through intonation, Cantonese relies on a rich variety of sentence-final particles (SFPs) (also referred to as utterance particles (Gibbons, 1989; Luke, 1990)) to compensate for this limitation (see Matthews and Yip (2011) and Luke (1990) for discourse-

[1] Jyutping is a Romanised written version of Cantonese, introduced by the Linguistic Society of Hong Kong in 1993 and is the most widely used system of Romanization (https://www.lshk.org/). The numbers used in the in the Romanized examples indicate the relevant tone for that word/utterance. When written Chinese characters are used tone is embedded within the character itself and is not indicated separately.

related functions and meanings of these particles). For instance, 咩 (me1) can change a statement into a question while 啩 (gwaa3) conveys the speaker's uncertainty about the truth of the statement.

The third contrasting feature relevant to this discussion relates to the identification of semantic 'word' units and how these are embedded within larger 'sentences'. The majority of alphabetic writing systems segment words by using spaces and punctuation as word delimiters. Logographic Asian languages, such as Cantonese, Japanese and Korean, do not delimit words by whitespace (Bai, Yan, Zang, Liversedge, & Rayner, 2008) (see Taylor & Taylor (2014) for an overview of literacy and writing in these languages). This is illustrated in the following example:

你老人著得少衫呢

You old people are wearing so few

This Cantonese sentence contains eight characters but these are not separated by spaces. Each character in this sentences can be considered to be the semantic equivalent of a 'word' unit. However, this is not always the case. Cantonese words can be made up of multiple characters. Segmentation, the process of the separation of characters into semantic word units, is subjective and open to interpretation by the analyst rather than being governed by rules (Fung & Bigi, 2015; Luke & Wong, 2015).

There are also significant differences in the way in which Cantonese sentences are structured. For example, in English verb tenses are used to indicate temporal features such as time. In Chinese, adverbs and contextual information are used. When translating written Chinese into English, the translator has to make word choices based on contextual understanding rather than relying on word for word translation.

These contrasting features between languages raise issues in terms of translation and the potential loss of important semantic, social and cultural meaning all of which are relevant in the study of health communication. For example, it is quite common in Cantonese to address a person using his or her family role especially in a medical consultation, for example 媽咪 (maa1 mi5), 哥哥 (go1 go1), 爹哋 (de1 dei2) and 阿仔 (aa3 zai2). These terms could be accurately translated into 'you', the common form of address in English, or their literal translation of 'mummy', 'elder brother', 'daddy' and 'son'. Although these translations correspond to the original meaning, the *relational* aspects of the speakers involved in that conversation are difficult to translate into another language. The loss of this relational information could have implications for researchers in interpreting and analyzing data.

As will be discussed further, these issues of equivalence of language as well as cultural and social associations and semantic, lexical and conceptual differences (Al-Amer, Ramjan, Glew, Darwish, & Salamonson, 2015; Hilton & Skrutkowsky, 2002; Twin, 1997) raise important questions about how NLP tools such as Discursis can be customized for use on first or native languages. These questions will be addressed through a comparative analysis of the same texts in Cantonese characters and in English translation.

Data

Conversational data used in the research[2] discussed in this chapter were collected in 2008 and 2009 in a number of government-run health settings in Hong Kong S.A.R. Data collection was approved by The Hong Kong Polytechnic University Human Research Ethics Committee

[2] This research was funded by a grant from the Faculty of Humanities, The Dean's Reserve, The Hong Kong Polytechnic University.

(HREC), and relevant hospital bodies in accordance with local requirements. Research participants provided written and informed consent prior to the audio-recording of conversations between them and a research assistant. Research participants were native speakers of Cantonese and Hong Kong residents. Conversations were audio-recorded, de-identified and transcribed verbatim into Cantonese (written in Chinese Hanzi characters) following the standard of the Hong Kong Supplementary Character Set (HKSCS) (https://www.ogcio.gov.hk/en/). Three extracts from the de-identified Cantonese transcripts were shared with the authors of this chapter in May 2018.[3] A total of 155 conversational turns were analysed in the research described in this chapter.

Method

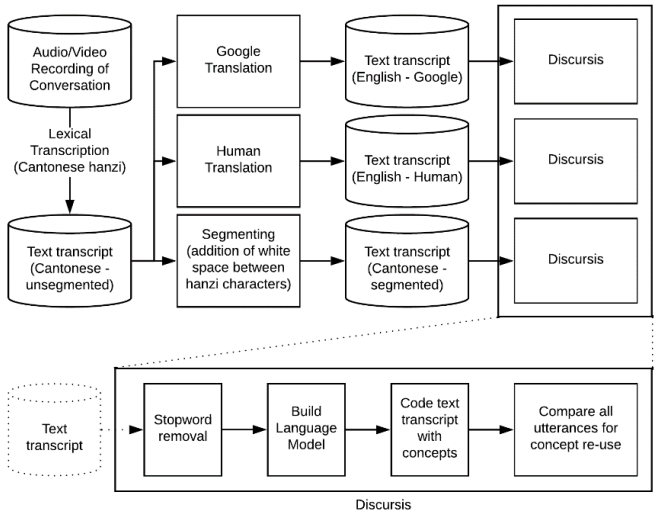
Data were entered into Discursis in three formats – Cantonese, English translated from Cantonese by the first author and English translated from Cantonese using Google Translate. Each set of data was entered into Discursis twice and analysed using (i) a default stop word list; and (ii) a customized stop word list generated by the first and second authors. These workflows are shown in Figure 13.1. On this basis, six Discursis visualisation plots were produced per transcript and there were 18 plots in total. Although Google Translate draws on a corpus of simplified Mandarin Chinese characters rather than Cantonese, as this is the most widely used

[3] The authors would like to acknowledge and thank the original research team at The Hong Kong Polytechnic University and its' representatives who allowed the data to be used in this research. Details of that research are recorded under approval code HSEARS20131104001.

and freely available tool for translation, it was included in the initial analysis. The low quality of the translation produced by Google Translate, however, limited its usefulness for analysis in Discursis. Therefore, the two data sets (12 plots) discussed in this chapter are based on the Cantonese (written in traditional Chinese characters) transcript and the transcript translated into English by the first author.

Comparison of the plots was based on a combination of visual qualitative examination and quantitative assessment of differences in outputs. For the quantitative comparison, each plot was directly compared via the number of recurrence elements (and shared absence of recurrence) each plot has in common. In Discursis, there is a unique recurrence element for each pair of utterances. For example if turns 2 and 34 of a transcript share concepts, then there will be a recurrence element visible at position (2, 34) that contains a value more than zero and less than or equal to one that indicates the strength of this conceptual overlap. These recurrence elements combine to create the recurrence plot and can also be compared directly in the case here where we have two versions of a transcript in different languages, but whose utterances should be equivalent. Differences between corresponding recurrence elements in the case of this study are an indicator of disparity between how transcripts are coded and processed by Discursis' language model. To measure the overall similarity of any two Discursis plots we simply sum the similarity (measured as 1 - difference between two corresponding recurrence elements) of all paired recurrence elements across two plots, and divide this sum by the total number of recurrence elements.

Figure 13.1. Workflows Used in the Analysis of Cantonese and Translated Data: Workflow Begins with the Transcription of the Conversation



Findings & Analysis

The results of the analysis of these three transcripts using the customised stop word lists are shown in Figures 13.2-13.4. The letter ‘a’ refers to Cantonese transcripts and letter ‘b’ refers to transcripts translated into English prior to analysis in Discursis. Analysis of the comparison of the (a) and (b) figures will be discussed in turn in order to highlight changes in visual representation of speaker engagement, interaction and the sharing of meaning and concepts through the conversation. Quantitative results (shown in Table 13.1 at the end of this chapter) suggest that, at least in the case of the customised stop word list, the workflows were mostly comparable (with an average 92% similarity according to direct comparison of recurrence elements between plots), however the differences that do exist are enough to affect and potentially alter qualitative interpretations as discussed below. Note also that for the quantitative comparison, the presence of corresponding white space on two plots (recurrence elements with a

value of zero) will count towards overall similarity, making it easy to qualitatively assess two plots as being more dissimilar than they are.

Figure 13.2a

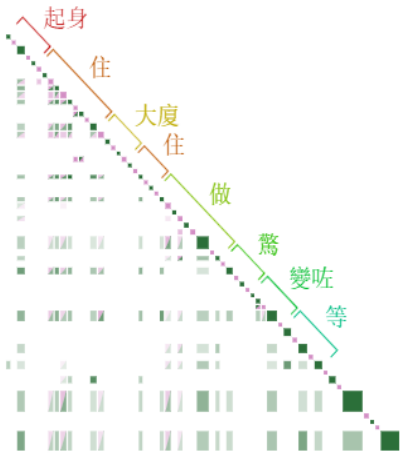


Figure 13.2b

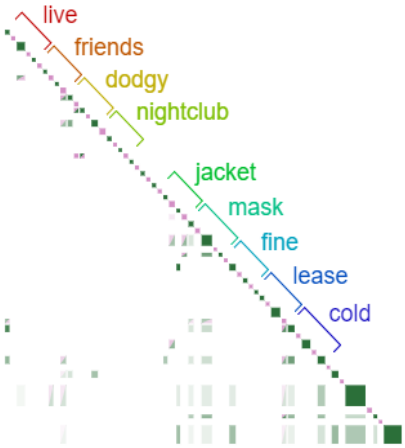


Figure 13.3a

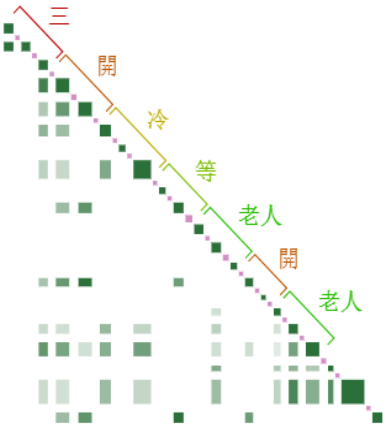


Figure 13.3b

Figure 13.4a

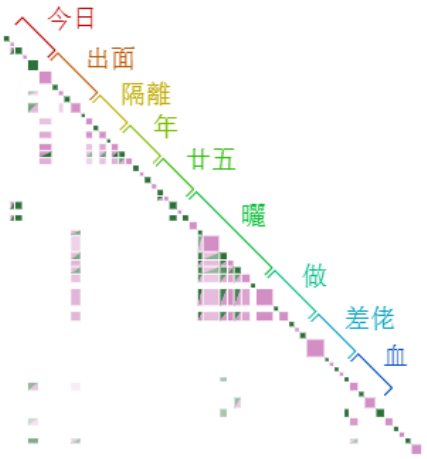


Figure 13.4b

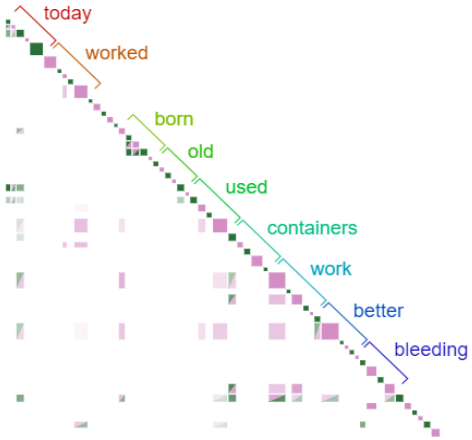


Figure 13.2. Discursis Plots Showing Analysis of the Cantonese and English Transcripts

Visual comparison of figures 13.2a and 13.2b highlights differences in colours and sizes of the recurrence elements (shown in the plots as coloured squares). Figure 13.2b shows a large white space in the first half of the plot. The white space indicates no engagement between the speakers. This suggests that the speakers were neither repeating each other's nor their own concepts. However, Figure 13.2a shows a higher level of speaker engagement reflected by the two-colour off-diagonal blocks in the same area of Figure 13.2b. By clicking on the coloured block within the plot the analyst can take a closer look and identify recurring concepts. In this case, the recurring concept is '住' (zyu6) (live) which occurred in five turns. The moderate-high level speaker engagement was also attributed to the association between this concept and three other words '大廈' (daai6haa6) (building), '雜' (zaap6) (dodgy) and '長大' (zoeng2daai6) (grew up). That means when the utterance includes either of these words, they will be categorized into

the concept of ‘住’. However, in Figure 13.2b, all these words were identified as separate concepts, showing no connection between each other. One possible explanation of this difference is that ‘住’ (zyu6) was not always translated as live/living in English. For example, ‘哦！住 邊頭 呀？’ was translated as ‘Oh, where?’ as a question raised following the response “I have some friends who used to live there.” In another instance 住宅 was translated as residential building, instead of living home. Although the translation into English maintained the semantic and syntactic relations within the utterance, some of the nuance and implied meaning of the original communication was lost.

Figure 13.3

Both figures 13.3a and 13.3b shared similar recurrence in the initial opening and final sections, but there was a significant difference in the intersection of the first half and second half of these conversations. Recurring concepts in Figure 13.3a are ‘開’ (hoi1) (start / turn on) and ‘冷’ (laang5) (cold) yet these are absent in figure 13.3b. Other words associated with these two concepts were also identified. For example, 開 (start / turn on) - 車 (ce1) (car) and 冷氣 (laang5hei3) (air-conditioner); 冷 (cold) - 衫 (saam1) (clothes) and 脾氣 (pei4hei3) (temper). The Cantonese word ‘開’ (open/start/turn on) can collocate with car and air conditioner while different verbs were needed to collocate with those nouns in the English transcript, for example, *start* the car, *turn on* the air conditioner. This contextual information was not added into the English transcript when it was translated and thus the relationship between these words was lost and not shown in figure 13.3b. This suggests that the process of translation, while technically correct from a linguistic perspective, has altered the representation of topic recurrence.

Figure 13.4

The visual differences between figures 13.4a and 13.4b in the first and middle sections of the plots indicate different levels of speaker engagement. Figure 13.4a shows a higher engagement level with more blocks of darker colours, indicating a higher concept similarity compared with Figure 13.4b. It is a more accurate representation of the original conversation as the recurring concepts '做' (zou6) (do) and '曬' (saai3) (expose in the sun) appeared consecutively in conversational turns. The differences between these two figures can be attributed to some translation issues similar to those discussed in relation to Figures 13.2 and 13.3. For example, '年' (nin4) (year) occurred five times in the initial opening of the conversation which was accurately represented by the moderate-high level of speaker engagement in Figure 13.4a. The reason why this level of speaker engagement was not shown in Figure 13.4b is because '年' was not always translated as 'year', for example, in the case of '七九年' (79 years) which should be 1979 in the translation. Another translation issue is related to verb inflection. Tense in English is often reflected by verb inflection such as *-ed*. In Cantonese, however, this is expressed lexically with the help of, for example, temporal phrases such as ('而家' (ji4gaa1) (now) and '之前' (zi1cin4) (before)) (Lin, 2006). In this case, '做' was translated into three different words (*worked, working* and *work*) and then three separate concepts in the English plot. This affected the level of speaker engagement that was shown in the plots. As was discussed earlier, the inherent differences between Chinese and English make transcription and translation challenging and even high quality translations may have discrepancies or inconsistencies in meaning.

Discussion

The process of translating the Cantonese data into English and preparing it for analysis in Discursis involved modification of syntactic and lexical markers. This included the initial translation from Cantonese to English as well as Segmentation of the written Chinese characters and development of stop word lists for both languages. Within the context of studies of health communication, the complexity of these processes and the risk of losing important relational data supports the argument for ongoing development of software that can facilitate bottom-up, first language analysis. Each of these stages of data preparation are discussed below.

Transcription and Translation of Conversation into Chinese Characters

Many Cantonese words can have identical sounds and tones and share similar meaning yet can be written differently. For example:

Cantonese words	**Jyutping (Cantonese Romanization)**	**Translation**
噉/咁	(gam2)	Like this / that; in this way
喺/係	(hai6)	To be; yes; right
哋/地	(dei6)	To indicate plurality after a personal pronoun
返/番	(fann1)	Return; come / go back
冇/無	(mo5)	Do not; not

畀/俾	(bei2)	Give
晒/曬	(saai3)	Completely; show off; bask in the sun

When preparing a transcript for analysis, consistent and accurate use of Chinese characters becomes crucial as it can affect the level of speaker engagement shown in the analysis. This characteristic of Cantonese is also the reason why the Cantonese transcripts were transcribed using Chinese characters, rather than Jyutping. Jyutping can show six lexical tones; however, many Cantonese words can have identical sounds and tones yet be written differently and carry different meaning, e.g. soeng2 - 想 (want, hope), 相 (photo); coi3 - 菜 (vegetables), 蔡 (surname), 賽 (race, competition). This complicates the transcription and process of segmentation and requires verification of the accuracy of the representation of tone. Additionally, Discursis can only identify different characters or words for data coding. Showing identical sounds and tones will affect how Discursis identifies concepts and depicts relationships as well as the level of speaker engagement in the plot. This consideration also impacts upon the creation of the stop word list for a transcript, for example, in the case of 喎 (wo5) (a model particle included in the stop word list) shares the same sound and tone as 禍 (calamity) (a content word which should be excluded from the stop word list). As demonstrated in this research, Discursis can analyse *logographic* characters. This suggests that transcribing Cantonese health-related conversations in Chinese characters and then segmenting the utterances prior to analysis in Discursis is preferable to using the Romanized Jyutping format. As written Chinese characters convey significant semantic and relational information and more accurate Discursis plots can be produced, the additional preparation time is warranted.

Segmentation of Chinese Characters

As noted earlier in this chapter one of the most important differences between alphabetic writing systems such as English and the Chinese logographic system from the perspective of text analytics is that the latter is written without spaces between characters (Bai et al., 2008). However, NLP tools such as Discursis which aim to work with logographic and alphabetic languages, need to draw on semantic word units as essential parts of their analysis. The Segmentation of a Cantonese transcript is therefore a fundamental but complicated step in the preparation of data for analysis. Segmentation not only delimits the words into meaningful semantic units, but also provides a unit (i.e. the word) which can be analysed. The analyst can then identify parts of speech or grammatical functions of words or units in order to generate the stop word list.

Generation and Modification of Stop Word Lists

As has been discussed, stop word lists generated for the analysis of English transcripts usually include words which do not have specific meaning but rather fulfil a predominantly grammatical function such as particles, auxiliaries, connectives (also referred to as conjunctions, prepositions, adverbs). Common examples of these words include *the, and, of, a, be*. Cantonese stop word lists also include words that have similar functions as the English ones. However, there are two main issues in relation to translating stop word lists across languages which have also been well-documented in previous studies in English-Cantonese translation (e.g. Lin, 2006; Yip & Matthews, 2017).

Firstly, no single English words can be translated as equivalent to a Cantonese sentence-final particle. Sentence-final particles do not have any semantic content and their meaning comes

from the sentence, clause, phrase or word they are attached to in a specific discourse context (Luke, 1990). For example, the accompanying response 金國大廈 “Kam Kwok Building” with the sentence final particles 吖嘛 (aa1maa3) attached can be translated as “Kam Kwok Building of course. Don’t you know?” The additional words “of course. Don’t you know?” in the translation express the intonation and speaker’s intention as naturally and closely as in that context where the speaker might have assumed the listener should have known this building prior to this discourse. Also, these particles are used primarily in relatively informal colloquial speech and are rarely found in written Chinese (Luke & Nancarrow, 1997). In contrast, many of the words in the English stop word lists can appear in both informal and formal English writing. This suggests that these sentence-final particles might only understood through additional words or punctuation at sentence level rather than at the word level as may be possible in English. This distinctive difference between Cantonese and English makes the application and translation of the English stop word list to Cantonese almost impossible.

Secondly, in English verbs are conjugated to express tense, e.g. *-ed* for past events. However, this kind of inflection is absent in Cantonese as tense is realized through aspect and verbal particles. For example, the possible Cantonese equivalents of ‘went’ (which is a verb in the default English stop word list) could be 有去, 去咗, 去過, 去完. If a conjugated verb is translated, all the possibilities may also have to be included in the Cantonese stop word list. However, in some cases these particles may not be necessary as tense can also be indicated through temporal phrases as discussed in the analysis of Figure 13.4. The tense of 去 (go/went) can already be clearly understood if the contextual information is sufficient.

In view of these two major issues, the translation and application of English stop word lists to Cantonese transcripts is not a feasible option for generating a meaningful analysis of a

Cantonese plot in Discursis. Results of the analysis in this chapter therefore suggest that as Discursis provides a dynamic analysis of language the generation of a customized stop word list for different languages is fundamental. This is a time consuming step in the process but is important for the validity of the analysis.

Conclusion

This chapter has detailed the findings of a unique research study which compared the visual analysis of Cantonese and English data using Discursis. As was discussed earlier in this chapter, the analytic value of using Discursis has been demonstrated in other research. Although the program can be theoretically used on non-alphabetic data at the time of writing this is the first such published study using logographic transcripts. The results described in this chapter have highlighted the relational information which can be lost through the translation of Cantonese into English. This emphasizes the importance of developing visual analytic tools that can be used on Cantonese data sets particularly in health-related research in which relational information embedded within the semantic and lexical features of a language is important. Generating visual representations of such data has benefits in terms of contributing to local and international understandings about how health and health care are discussed in different communities and cultures.

Research has consistently shown that communication about health and care is more effective when conversational features such as turn taking and semantic alignment are balanced between participants rather than dominated by experts such as doctors. Complex yet critical interpersonal components of health care such as building trust and rapport, decision-making, managing medication and explaining risk and uncertainty unfold through dynamic processes of

interaction and communication and often involve talking about broad lifestyle-related information that goes beyond description of symptoms or treatments. Language is a fundamental data source in research into this area yet relational data can be lost through the process of translation from one language to another (Squires, 2009).

Conversational transcripts recorded in first or native languages, therefore, provide unique insights in to cultural and social perceptions of health. Throughout this chapter we have argued that the complex processes of both the transcription of spoken language and translation between languages present unique challenges to language and communication researchers. Developing analytic tools that can be used with logographic *and* alphabetic languages without requiring translation will help to preserve the subtle and relational aspects of language that shape communication about health and health care. Expanding the field of health communication in Asia will be supported through the continued development of analytic tools which can be used with first-language, logographic data. Discursis outputs have been used to inform the development of training programs for a variety of professions that aim to increase awareness of communication between people. Such insights could be of benefit across multiple language groups if the software can be appropriately customised. Expanding this work with Asian languages will also make significant contributions to the fields of NLP and machine translation.

References

Al-Amer, R., Ramjan, L., Glew, P., Darwish, M., & Salamonson, Y. (2015). Translation of interviews from a source language to a target language: examining issues in cross-cultural health care research. *Journal of Clinical Nursing, 24*(9-10), 1151-1162.

Angus, D., Rintel, S., & Wiles, J. (2013). Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis. *International Journal of Social Research Methodology, 16*(3), 261-267.

Angus, D., Smith, A., & Wiles, J. (2012). Conceptual recurrence plots: revealing patterns in human discourse. *IEEE Transactions on Visualizations and Computer Graphics, 18*(6), 988-997.

Atay, C., Conway, E., Angus, D., Wiles, J., Baker, R., & Chenery, H. (2015). An automated approach to examining conversational dynamics between people with dementia and their carers. *PLos ONE, 10*(12), e0144327.

Bai, X., Yan, G., Zang, C., Liversedge, S. P., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: evidence from eye movements. *Journal of Experimental Psychology Human Perception and Performance, 34*(5), 1277-1287.

Bauer, R., & Benedict, P. (Eds.). (1997). *Modern Cantonese phonology*. Berlin and New York: Mouton de Gruyter.

Bondi, M. (2017). Corpus linguistics. In *The Routledge Handbook of Language and Dialogue* (pp. 46-61). New York: Routledge.

Chan, A., & Li, D. (2000). English and Cantonese phonology in contrast: explaining Cantonese ESL learners' English pronunciation problems. *Language, Culture and Curriculum, 13*(1), 67-85.

Chevalier, B., Watson, B., Barras, M., Cottrell, W., & Angus, D. (2018). Using Discursis to enhance the qualitative analysis of hospital pharmacist-patient interactions. *PLOSone* (May).

Evers, W., Marroum, S., & Young, L. (2017). A pluralistic, longitudinal method: Using participatory workshops, interviews and lexicographic analysis to investigate relational evolution. *Industrial Marketing Management*, 61(February), 182–193.

Firth, J. (1962). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*.

Franzoni, S., & Bonera, M. (2019). How DMO Can Measure the Experiences of a Large Territory. *Sustainability*, 11(2), 492.

Fung, R., & Bigi, B. (2015). *Automatic word segmentation for spoken Cantonese.* Paper presented at the International Conference Oriental and Conference on Asian Spoken Language Research and Evaluation, Shanghai, China.

Gibbons, J. (1980). A tentative framework for speech act description of the utterance particles in conversational Cantonese. *Linguistics*, 18, 763-775.

Goldstein, S., MacDonald, N. E., & Guirguis, S. (2015). Health communication and vaccine hesitancy. *Vaccine, 33*(34), 4212-4214.

Heritage, J., & Maynard, D. (Eds.). (2006). *Communication in Medical Care*. Cambridge: Cambridge University Press.

Hirschberg, J., & Manning, C. (2015). Advances in natural language processing. *Science, 349*(6245), 261-266.

Hilton, A., & Skrutkowsky, M. (2002). Translating instruments into other languages: development and testing process. *Cancer Nursing, 25*(1), 1-7.

Ho, C. S.-H., & Bryant, P. (1997). Development of phonological awareness of Chinese children in Hong Kong. *Journal of Psycholinguistic Research, 26*(1), 109-126.

Lin, J-W. (2006). Time in a language without tense: The case of Chinese. *Journal of Semantics, 23*(1), 1–53.

Liu, T., & Hsiao, J., H-W. (2014). Holistic processing in speech perception: experts' and novices' processing of isolated Cantonese syllables. *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*, 869-874.

Luke, K., & Nancarrow, O. (1997). *Sentence particles in Cantonese: A corpus-based study*. Presented at The Yuen Ren Society Meeting, University of Washington.

Luke, K., & Wong, M. (2015). The Hong Kong Cantonese corpus: design and uses. *Journal of Chinese Linguistics, 25*, 309-330.

Matthews, S., & Yip, V. (2011). *Cantonese: A Comprehensive Grammar* (2nd ed.). London and New York: Routledge.

Nouri, S., & Rudd, R. (2015). Health literacy in the "oral exchange": An important element of patient–provider communication - ScienceDirect. *Patient Education and Counseling, 98*(5), 565-571.

Smith, A., & Humphreys, M. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behaviour Research Methods, 38*(2), 262-279.

Snow, D. (2004). *Cantonese as Written Language: The Growth of a Written Chinese Vernacular*. Hong Kong S.A.R.: Hong Kong University Press.

Spevack, S., Falandays, J., Batzloff, B., & Spivey, M. (2018). Interactivity of language. *Language and Linguistics Compass, 12*(e12282), 1-18.

Street, R., Makoul, G., Arora, N. K., & Epstein, R. (2009). How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ Couns, 74*(3), 295-301.

Taylor, I., & Taylor, M. (2014). *Writing and literacy in Chinese, Korean and Japanese: Revised Edition*. Amsterdam: John Benjamins.

Twinn, S. (1997). An exploratory study examining the influence of translation on the validity and reliability of qualitative data in nursing research. Journal of Advanced Nursing, 26(2), 418-423.

Tolston, M., Riley, M., Mancuso, V., Finomore, V., & Funke, G. (2018). Beyond frequency counts: Novel conceptual recurrence analysis metrics to index semantic coordination in team communications *Behaviour Research Methods*, (Oct.), 1-19.

Watson, B., Angus, D., Gore, L., & Farmer, J. (2015). Communication in open disclosure conversations about adverse events in hospitals. *Language and Communication, 41*, 57-70.

Weigand, E. (Ed.) (2017). *The Routledge Handbook of Language and Dialogue*. New York: Routledge.

Wong, A., Juang, J., & Chen, H.-C. (2012). Phonological units in spoken word production: insights from Cantonese. *PLOSone, 7*(11), e48776.

Yip, V., & Matthews, S. (2017). *Intermediate Cantonese: a grammar and workbook* (2nd ed.). Routledge.

Table 13.1. Results of pairwise quantitative comparison between recurrence plots generated using the original Cantonese (a), and hand-translate English (b) transcripts, using both the default (i) and modified custom (ii) stop word lists (similarity between plots: 1.0 = identical, 0.0 = opposite).

	Dataset1ai	Dataset1aii	Dataset1bi
Dataset1aii	0.77		
Dataset1bi	0.75	0.87	
Dataset1bii	0.76	0.92	0.92

	Dataset2ai	Dataset2aii	Dataset2bi
Dataset2aii	0.95		
Dataset2bi	0.90	0.95	
Dataset2bii	0.90	0.9	0.99

	Dataset3ai	Dataset3aii	Dataset3bi
Dataset3aii	0.76		
Dataset3bi	0.75	0.92	
Dataset3bii	0.73	0.94	0.97