`

Submitted by Geraldine Bengsch

**Project Report**

54th Essex Summer School

1K Intro to Programming with Python for Social Scientists

Dr Phillip D. Brooker

30th July 2021

**Interaction at the international hotel front desk: an alternative form of analysis**

# **Summary**

This project takes three canonical extracts from the data corpus collected for my PhD to explore if basic NLP analysis in Python could be used to confirm the pseudo-quantitative analysis carried out. The script acts as a proof of concept. It continues the exploration of the Python library NLTK that my group started to use in week two of the course and applies it my own data. I explore ways of manipulating CA transcripts in Python. I presented my aim to do such work with my existing data set during the methods showcase at the Essex Summer School. I expand on this presentation with the skills learned in the course.

Please refer to the readme file in the project to run the script.

The project is available online here[[1]](#footnote-1): <https://github.com/GeriNZ/NLPforCA>

Table of Contents

[**Summary** 2](#_Toc78553040)

[**Introduction** 2](#_Toc78553041)

[*Introduction of the context and data in the PhD project* 3](#_Toc78553042)

[**Framework and data** 3](#_Toc78553043)

[**Results from prior CA analysis** 4](#_Toc78553044)

[*RQs for this project:* 5](#_Toc78553045)

[*Data selection for this project* 5](#_Toc78553046)

[**Structure of the script** 5](#_Toc78553047)

[**Findings** 7](#_Toc78553048)

[**Limitations** 9](#_Toc78553049)

[**Future improvements** 9](#_Toc78553050)

[**Conclusion** 9](#_Toc78553051)

[**References** 11](#_Toc78553052)

# **Introduction**

Text that is treated as qualitative data often has a “wild” side (Richards, 2020), that is, no coherent structure which makes working with it far from easy (Roberts & Wilson, 2002). The data used in this project stems from transcripts I created during my PhD. While Conversation Analysis was used as the main framework, the transcripts are of varying completeness, based on which interactions were taken further in the analysis. The transcripts also reflect my journey of trying to deal with the data in various formats, including Microsoft Word and CLAN versions. Over the years, I have worked and assisted with research using CA transcription conventions and a common theme has been the development of transcripts over the course of a project (this was not just the case for CA work, but also e.g. interview data where transcriptions changed based on what was currently understood from the source) (Erickson, 2010). Extracts of data get copy-and-pasted in between documents and projects and often bear very little resemblance to the original text (Azevedo et al., 2017; Duff & Roberts, 1997; Wellard & McKenna, 2001). I aim to develop a way of engaging with text that is less destructive of the original corpus through the use of Python. In this project, I wanted to learn more about the challenges – and opportunity that working with CA transcripts poses when using a script to extract information. I worked on ways to read in text files and to analyse them individually or as a whole corpus in various ways. The analysis is currently limited to basic word frequencies, sentiment analysis and some basic visualisations.

## *Introduction of the context and data in the PhD project*

The world continues to become more and more interconnected, creating implications for research in global communicative settings (Blommaert, 2010). The hospitality industry, especially hotels as institutions provide an interesting interface for access to interaction research. Conversations between a guest and a receptionist connect every day, ordinary life with that of an institution. Staying in a hotel means that guests transpose their lives voluntarily for a limited time to an organisation, often at a considerable distance to their homes, frequently in unfamiliar surroundings, such as a foreign cultural context. Familiar social structures are reimagined in a space where “the basic requirements of life - food, drink, warmth, shelter and security - are supplied by others” (Rojek, 1993, p. 191). As a result, it may be argued that guests and staff engage in interactions that aim to emulate an everyday life version of “home away from home” (Marriott Jr. & Brown, 1997) and a “practical place for tourist interest” (Goffman, 1953, p. 30). They may be considered “natural laboratories for the investigation of stress and coping, culture learning, and social identification” (Berno & Ward, 2005, p. 598).

# **Framework and data**

Conversation Analysis (CA) and the social sciences have a long-standing tradition in investigating talk in a large variety of settings, including asymmetrical interactions, such as service encounters (Drew & Heritage, 1992). CA is used here to describe interactional structures that govern the ceremonial order of conversations at the interface of mundane and institutional. Most of modern interactions occur in organisations (Grey, 2013). Service encounters are still learned behaviours that provide insights into current social structures (Manning, 2008). Common business practices often promote prescriptive linguistic practices which remove the communicative and inherently playful nature of interpersonal interactions (Kennelly Isaacs, 2006). Ten hours of naturally occurring conversations at the hotel front desk of four hotels in three countries (Germany, England, Spain) were videotaped and analysed using CA and ethnographic notes in relation to interactional, sociological and business literature.

For this project, three canonical examples from the data corpus were selected. However, the script is set up so that there is no limit to the number of transcripts that can be used and analysed.

# **Results from prior CA analysis**

The data shows that interactions at the hotel front desk contain elements of both mundane and institutional interactions which are navigated and oriented to by interactants through the use of role ascriptions, rituals, ceremonial behavioural patterns and personae (Bente, Leuschner, Al Issa, & Blascovich, 2010). In addition to authentic behaviour, it is a notion of creating a sincere interaction (Taylor, 2001). In a hotel environment, quality assurance remains at the centre of interest for researchers and businesses (Xiang, Schwartz, Gerdes Jr, & Uysal, 2015).

While service encounters traditionally are goal focused, the data corpus shows that interactants at the front desk show concern for more than simple task completion. Both parties negotiate their interests and demonstrate understanding for the concerns of the other. As such, the interactions function to role play traditional hospitality, such as exists at home, outside of a paid for service to recreate a sense of belonging (Blue & Harun, 2003; Doorne & Ateljevic, 2005; Pearce, Moscardo, & Ross, 1996).

On one hand, the hospitality industry has been striving to increase automation and standardisation in service (Bitner, Ostrom, & Morgan, 2008; Palmer, 1998); on the other, the industry shows an interest in providing experience and personalisation for its guests (e.g. Hemmington, 2007). Commercial hospitality may thus be said to be performing a paid service while also emulating aspects of traditional hospitality designed for friends and family (Blue & Harun, 2003), which are negotiated explicitly in conversations at the hotel front desk.

One main observation from the data analysis was the organisation of interaction into a larger interactional project with both intra-sequential and inter-sequential organization (Orthaber & Márquez-Reiter, 2011; Robinson, 2006): Interactions produce salient information for subsequent encounters; Service encounters, like mundane interactions occur in larger episodic structures which follow sequential ordering. This creates distinctive phases (arrival, stay, departure) which, as I claim in my thesis, are observable and available to non-participants.

* solution oriented and driven by the individual’s need to complete a routine task not only according to the company’s standard, but to fit with the individual’s personal understanding
* environment provides strong incentive to develop a suitable communicative style to minimise work effort
* jargon, or specific vocabularic terms, are employed to successfully minimise miscommunication
* immersive atmosphere provides opportunities to develop pragmatic and conversational competence in action.

However, the analysis generated in CA is done through the reliance of hand coding, despite the use of increasingly large corpora. The abundance of text in the present day is present in many fields, which has led researchers to call for viewing text differently (Benoit, 2020; Gentzkow, Kelly, & Taddy, 2019).

This leads to the question I aim to answer through the use of a Python script:

## *RQs for this project:*

Can a Python script be useful when dealing with CA transcripts?

Can exploratory data analysis in NLP with Python be used to confirm elements of a pseudo-quantitative, hand-coded analysis?

**Data selection**

## *Data selection for this project*

As a proof of concept, I wanted to choose examples of data that I labelled as “canonical” in the corpus. I chose one extract from each of the “distinctive phases” that I claim to exist in interactions at the hotel front desk.

# **Structure of the script**

Steps:

1. Reading in txt files
2. Tokenize text, lower case, remove punctuation
3. Remove stop words, add own stop words
4. Frequency distribution using matplotlib (should be different in the three phases)
5. Wordcloud for frequency distribution
6. Bigrams of words
7. Sentiment analysis
8. Working with transcripts as a corpus, repeating tokenizing and frequency distribution as above using pandas and seaborn for visualisation

The main aim of this project was to begin creating a Python script that allows a better management of CA transcripts[[2]](#footnote-2).

The script is divided into two parts: Part one iterates over the individual transcripts and analyses them separately, part two creates a corpus of all the transcripts and analyses them.

* The script begins by defining the transcript location so that they can be accessed by the script.
* The remainder of the first part of the script is wrapped into a for-loop, so that the output is produced for every single transcript.
* The basic transcript is read in and printed to the shell to show its current state and highlight some issues, e.g. the amount of whitespace within CA transcripts
* The transcript is converted to lower case (this may need to occur at a later time when the script becomes more sophisticated)
* Duplicate whitespaces in the strings are removed
* NLTK retains whitespace as a token. To remove those extra tokens from the corpus, NLTK’s WhitespaceTokenizer() was used.
* The resulting, unfiltered tokens are printed to the screen
* Stop words are removed from the corpus and the list of filtered tokens is printed on the screen
* Punctuation is removed from the transcripts and the list is printed to the screen
* Frequency distribution is generated from the token list that contain no punctuation and ploted using matplotlib
* A word cloud is generated using the raw text from the transcript
* A list of unordered bigrams is generated and printed to the screen
* An ordered list of bigrams with frequency is generated and printed
* Sentiment analysis is generated (based on what we did in the group project)

Part 2:

* Corpus is generated from all transcripts using NLTK’s PlaintextCorpusReader()
* Prints the list of files that form the corpus onto the screen
* Tokenising works differently in the PlaintextCorpusReader(): .words() is used to generate the tokens and print them to the screen
* .sents() is used to extract sentences from the corpus and printed to the screen - moderately useful for the way that CA transcripts are often constructed
* .paras() is used to extract paragraphs that are then printed to the screen
* Frequency distribution is generated from the token list
* Tokens are converted to lower case
* Stop words are removed from the tokens and filtered tokens are printed to the screen
* 20 most common words are generated from all the frequency distribution list
* Tabular representation of 10 most common tokens in filtered list is generated
* 20 most common terms are converted to a Pandas series to make plotting easier. This is done via a Python dictionary and plotted using seaborn
* A second plot is generated using the frequency distribution of the 20 most common terms in the filtered token list

# **Findings**

Working with data I am very familiar with has been an interesting experience in Python. It allows to view and compare files rather quickly, without being overwhelmed by them, e.g. by having them open in multiple windows in Word, or multiple tabs in a software like NVivo, thus aiding in query and retrieval of data (Gephart & Wolfe, 1989; Wolfe, Gephart, & Johnson, 1993). It has also shown that there may be more accurate ways of working with CA notation if it is to be understood by a machine, not only by a human. Whitespace is used extensively in CA, especially when marking overlaps. However, the print out of the data with whitespaces removed, still holds those CA markers for overlap, so that a later iteration of this script could use regex to help map speakers and overlaps, without relying on the use of whitespace, which is often heavily manipulated as the font may not line up properly in word (even when using an appropriate font).

CA can also rely on capitalisation to mark e.g. loudness of an utterance. In this project, I wanted to practice with methods available, thus chose to generate a version of the text that only uses lower case letters to allow for the removal of stop words. However, it means that Python is able to deal with these intricacies of transcription and can offer ways of engaging with them further. This is in line with the notion of constant comparison in working with quasi-statistical data (Drass, 1980; Gephart & Wolfe, 1989).

While bag of word methods, such as tokenising may not be of immediate interest to CA analysis, it can help in understanding what is consistently foregrounded by speakers (Ellis & Ferreira–Junior, 2009). This could be done in a more sophisticated way than the script currently does. It could be used to find common elements in specific actions, e.g. requests, complaints, greetings, etc to generate more robust understandings of how those are constructed (Housley, Albert, & Stokoe, 2019). Since the speaker tags are currently present in the frequency distribution, it shows the distribution of turns by individual speakers (Stasser & Taylor, 1991). The common assumption in CA literature is that service encounters are dominated and led by the service provider. However, the distribution in the project here shows that the conversations still very much rely on the turn-response structure, e.g. through the use of minimal response tokens to create a supportive conversational environment (Bodie et al., 2021; Wang, Yaman, Precoda, & Richey, 2011).

Counting things, e.g. frequency distribution, would also help in helping to deal with the frequency and distribution of silence in conversation (Çetin & Shriberg, 2006). Python would make it easy to compute time of silence in conversation for an interaction, or for individual speakers – a notion that has been addressed as concept in traditional CA literature (Wilson & Zimmerman, 1986). Individuals interactionally produce a perceived and accountable competence to receive and adequately process information (Morrison & Bellack, 1981).

Chart, line chart

Description automatically generatedA picture containing text

Description automatically generatedChart, line chart

Description automatically generatedText

Description automatically generatedChart, line chart

Description automatically generatedA picture containing text

Description automatically generated

The basic frequency distribution of words for the transcripts confirm my CA analysis. Arrival sequences contain the most minimal response tokens. Notably, arriving in a hotel does not contain goodbyes, even at the end of the interaction. Thanking is far more prevalent in Departure sequences. Stay sequences are governed by request sequences in which individuals work on agreements, e.g. through repetition.

Ngrams in the analysis are not typically part of CA analysis, but may be useful in e.g. showing how common ground is constructed through the use of particular expressions which has generated some interest to researchers working at the intersection of corpus linguistics and CA (Bednarek, 2014; Chaudhuri & Raj, 2011; Morton, Walsh, & O'Keeffe, 2011).

In the arrival transcript, response tokens are prevalent in the bigrams. In the departure sequence, there are thanking and farewell that occur together. The stay sequence features descriptive elements.

# **Limitations**

Exploratory data analysis that is limited to “bag-of-words method”, that does not care about the notion of larger contextual structures, e.g. sentences or sequences. More sophisticated methods are needed to work with those (e.g. a different base library, such as spaCy, genism,language/topic modelling, machine learning etc) (Baburajan, e Silva, & Pereira, 2018).

# **Future improvements**

Working in Python opens up a much more fine-grained command of working with texts. Python makes it possible to compare individual transcripts, examine the full corpus, and allows much greater control for maintaining collections.

Here are some points that I hope to incorporate at a later stage.

* Use regex to separate speakers in the corpus: the literature claims that service encounters are led by the service provider. Is there more than the CA “hunch” that this is the case?
* Use regex to work with CA specific notation, e.g. pauses, but also overlaps
* Use results from EDA to identify areas of interest that may have been missed in hand-coded analysis
* Generate corpus not only with bag of words, but also with sentences (at a later stage, consider “utterances”, “sequences”, “actions” and “activities”
* Introduce normalisation, e.g. lemmatisation, morphological parsing, etc., if useful (for the purpose of this project, normalisation was seen as more damaging rather than helpful)
* Create relevant data frames to work with libraries like pandas
* Convert bag of words to wordnets
* Find more ways to visualise the data to gain novel insights not available through CA
* Incorporate batch processing for organising, cleaning and converting transcript files
* Working with audio files
* Working with images
* And other things

# **Conclusion**

The majority of time working on this script has been spent on “data wrangling”, rather than on analysing it. While somewhat frustrating, it has given me the opportunity to re-engage with the transcripts I created during my PhD and assess what I needed them to do back then and to contemplate what those decisions mean for the future of the data. Looking at the work done with the data and the revision of the literature, it seems that my ideas are supported by researchers to some degree or other. As such, what I am proposing is nothing new as such.

Goffman (1961, p. 312) ascribes “encompassing tendencies” to organisations. Behaviours found within them describe as much as promote the overarching structures within them and creates the social frame. Creating a script to work with the data has allowed me to work with the data on a different level. While the current output is somewhat underwhelming, I believe that it has taught me ways in which I can engage with the data in future. It also has provided some pointers for future learning (e.g. regex).

# **References**

Azevedo, V., Carvalho, M., Fernandes-Costa, F., Mesquita, S., Soares, J., Teixeira, F., & Maia, Â. (2017). Interview transcription: conceptual issues, practical guidelines, and challenges. *Revista de Enfermagem Referência, 4*(14), 159-167.

Baburajan, V., e Silva, J. D. A., & Pereira, F. C. (2018). *Opening up the conversation: Topic modeling for automated text analysis in travel surveys.* Paper presented at the 2018 21st International Conference on Intelligent Transportation Systems (ITSC).

Bednarek, M. (2014). Involvement in Australian talkback radio—a corpus linguistic investigation. *Australian Journal of Linguistics, 34*(1), 4-23.

Benoit, K. (2020). Text as data: An overview. *The SAGE Handbook of Research Methods in Political Science and International Relations, SAGE Publishing, London (forthcoming)*.

Bitner, M. J., Ostrom, A. L., & Morgan, F. N. (2008). Service blueprinting: A practical technique for service innovation. *California Management Review, 50*(3), 66-90.

Blommaert, J. (2010). *The sociolinguistics of globalization*: Cambridge University Press.

Blue, G. M., & Harun, M. (2003). Hospitality language as a professional skill. *English for Specific Purposes, 22*(1), 73-91. Retrieved from <http://www.sciencedirect.com/science/article/pii/S088949060100031X>

Bodie, G. D., Jones, S. M., Brinberg, M., Joyer, A. M., Solomon, D. H., & Ram, N. (2021). Discovering the Fabric of Supportive Conversations: A Typology of Speaking Turns and Their Contingencies. *Journal of Language and Social Psychology, 40*(2), 214-237.

Çetin, Ö., & Shriberg, E. (2006). *Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition.* Paper presented at the Ninth international conference on spoken language processing.

Chaudhuri, S., & Raj, B. (2011). *A comparison of latent variable models for conversation analysis.* Paper presented at the Proceedings of the SIGDIAL 2011 Conference.

Doorne, S., & Ateljevic, I. (2005). Tourism performance as metaphor: Enacting backpacker travel in the Fiji islands. In A. Jaworski & A. Pritchard (Eds.), *Discourse, communication and tourism* (pp. 173-198). Clevedon: Channel View Publications.

Drass, K. A. (1980). The analysis of qualitative data: A computer program. *Urban life, 9*(3), 332-353.

Drew, P., & Heritage, J. (1992). Analyzing talk at work: An introduction. In P. Drew & J. Heritage (Eds.), *Talk at Work* (pp. 3-65). Cambridge Cambridge University Press.

Duff, P. A., & Roberts, C. (1997). The politics of transcription; Transcribing talk: issues of representation. *Tesol Quarterly, 31*(1), 167-172.

Ellis, N. C., & Ferreira–Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal, 93*(3), 370-385.

Erickson, F. (2010). The neglected listener issues of theory and practice in transcription. *New adventures in language and interaction*, 243-256.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535-574.

Gephart, R. P., & Wolfe, R. A. (1989). *Qualitative Data Analysis: Three Microcomputer-Supported Approaches.* Paper presented at the Academy of Management Proceedings.

Goffman, E. (1953). *Communication conduct in an island community.* (Doctor of Philosophy). University of Chicago, Retrieved from <http://cdclv.unlv.edu//ega/documents/eg_phd.pdf>

Grey, C. (2013). *A fairly short, fairly interesting and reasonably cheap book about studying organizations* (3 ed.). Los Angeles, CA: Sage.

Housley, W., Albert, S., & Stokoe, E. (2019). *Natural action processing.* Paper presented at the Proceedings of the Halfway to the Future Symposium 2019.

Kennelly Isaacs, C. (2006). How do you like your cereal? A linguistic analysis of a service encounter. *Working Papers in Educational Linguistics, 21*(2), 87-102.

Manning, P. (2008). Barista rants about stupid customers at Starbucks: What imaginary conversations can teach us about real ones. *Language & Communication, 28*(2), 101-126. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0271530908000177>

Marriott Jr., J. W., & Brown, K. A. (1997). *The spirit to serve: Marriott’s way*. New York, NY: Harper Business.

Morton, T., Walsh, M., & O'Keeffe, A. (2011). Analyzing university spoken interaction: a corpus linguistics/conversation analysis approach.

Orthaber, S., & Márquez-Reiter, R. (2011). “Talk to the hand”. Complaints to a public transport company. *Journal of Pragmatics, 43*(15), 3860-3876.

Palmer, C. (1998). From theory to practice experiencing the nation in everyday life. *Journal of material culture, 3*(2), 175-199.

Pearce, P. L., Moscardo, G., & Ross, G. F. (1996). *Tourism community relationships*. Oxford: Pergamon.

Richards, L. (2020). *Handling qualitative data: A practical guide*: Sage.

Roberts, K. A., & Wilson, R. W. (2002). *ICT and the research process: Issues around the compatibility of technology with qualitative data analysis.* Paper presented at the Forum Qualitative Sozialforschung/Forum: Qualitative Social Research.

Robinson, J. D. (2006). An interactional structure of medical activities during acute visits and its implications for patients' participation. In P. Drew & J. Heritage (Eds.), *Conversation analysis: Institutional interactions* (Vol. IV, pp. 49-84). London: Sage.

Rojek, C. (1993). *Ways of escape: Modern transformations in leisure and travel*. Houndmills: Macmillan Press Ltd.

Stasser, G., & Taylor, L. A. (1991). Speaking turns in face-to-face discussions. *Journal of personality and social psychology, 60*(5), 675.

Wang, W., Yaman, S., Precoda, K., & Richey, C. (2011). *Automatic identification of speaker role and agreement/disagreement in broadcast conversation.* Paper presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Wellard, S., & McKenna, L. (2001). Turning tapes into text: Issues surrounding the transcription of interviews. *Contemporary Nurse, 11*(2-3), 180-186.

Wilson, T. P., & Zimmerman, D. H. (1986). The structure of silence between turns in two‐party conversation. *Discourse processes, 9*(4), 375-390.

Wolfe, R. A., Gephart, R. P., & Johnson, T. E. (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. *Journal of Management, 19*(3), 637-660.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120-130.

1. Transcripts are not available on github [↑](#footnote-ref-1)
2. As a separate project, I aim to automate converting transcripts into useful file formats. The state of my own transcripts from my PhD corpus has certainly degraded over the years, and cleaning those up was out of the scope for this project. [↑](#footnote-ref-2)