End-of-Turn Detection

Sean Leishman

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2023

Abstract

This skeleton demonstrates how to use the `infthesis` style for undergraduate dissertations in the School of Informatics. It also emphasises the page limit, and that you must not deviate from the required style. The file `skeleton.tex` generates this document and should be used as a starting point for your thesis. Replace this abstract text with a concise summary of your report.

Research Ethics Approval

Instructions: *Agree with your supervisor which statement you need to include. Then delete the statement that you are not using, and the instructions in italics.*
Either complete and include this statement:
This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: ???
Date when approval was obtained: YYYY-MM-DD
[If the project required human participants, edit as appropriate, otherwise delete:]
The participants' information sheet and a consent form are included in the appendix.
Or include this statement:
This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Sean Leishman*)

Acknowledgements

Any acknowledgements go here.

Table of Contents

Chapter 1

Introduction

The preliminary material of your report should contain:

- The title page.
- An abstract page.
- Declaration of ethics and own work.
- Optionally an acknowledgements page.
- The table of contents.

As in this example `skeleton.tex`, the above material should be included between:

```
\begin{preliminary}
    ...
\end{preliminary}
```

This style file uses roman numeral page numbers for the preliminary material.

The main content of the dissertation, starting with the first chapter, starts with page 1. ***The main content must not go beyond page 40.***

The report then contains a bibliography and any appendices, which may go beyond page 40. The appendices are only for any supporting material that's important to go on record. However, you cannot assume markers of dissertations will read them.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default single spacing). Be careful if you copy-paste packages into your document preamble from elsewhere. Some LaTeX packages, such as `fullpage` or `savetrees`, change the margins of your document. Do not include them!

Over-length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

1.1 Turn-taking: From the Conversational Analaysis Perspective

Over the last few decades, psycolinguists have been fascinated with the complexity of the mechanisms of conversation along with the apparent ease with which speaker's are able to converse in a orderly and timely manner. Sacks et al. [1974] is a widely cited paper that outlines some general observations that has gone on to inform general turn-taking literature. Primarily, that turn-taking organisation is not planned in adavance however the actions taken are still coordinated, in a flexible manner that can be decided upon by the current participants in a conversation. Typically one person speaks at a time and most transitions have a small gap or overlap but transitions do occur with no gap and no overlap. Levinson and Torreira [2015] used automatic analaysis to show that these observsations are indeed statisically valid. They note, that turns are generally short (mean 1680ms, median 1227ms) and turn transitions most commonly fall between 100ms and 200ms but the vast majority fall in the -100ms and 600ms range.

1.1.1 Models of Turn-taking Organisation

Turn-taking organisation has generally been characterised in two different ways within literature: the `reactionary` and the `predictive` approach. The former assumes that participants simply understand end-of-turn signals and react to them accordingly while the predictive approach entails the listener predicting the end of turn in advance such that responses are well timed.

The reactionary approach assumes that turn-taking organisation is regulated by both vocal and gestural signals (Yngve [1970]). This approach was pioneered by (Duncan [1972, 1973, 1974], Duncan and Fiske [2015]) who argued for a precise set of context free turn-yielding 'signals'. Duncan [1972] described phrase-final intonation, drawl on the final syllable, termination of hand gesticulation, changes in pitch and a termination of a grammatical clause as turn-yielding signals.

Later literature goes on to argue against the general model of a reactionary approach as, put simply, turn-transitions occur too quickly and turn-yielding signals occur too late within a speaker's utterance for the listener to simply react to an end-of-turn signal.

Sacks et al. [1974] pioneered the `predictive` approach and in their analysis of turn-taking argued that the observed speed of turn-transitions required some form of 'projection' with the production of language beginning prior to the end of a turn. This model of turn-taking is based off of separating speech into units, where one speaker is the speaker, called `Turn Construction Units (TCU)` and immediately after completing a TCU a `Transition Relevance Place (TRP)` occurs that signals that a turn-transition (turn-shift) can occur. It is also important to note that a TRP does not always result in a turn-shift and a turn-shift does not always occur at a TRP. Nevertheles, every TRP is governed by a set of rules determing whether or not a TRP will transpire:

1. The current speaker may select a new speaker during which the other participants act as listeners

2. If the current speaker does not select then any participant can self-select. The first to start gains the turn.
3. If no other party self-selects, the current speaker may continue.

The rules predict that intra-speaker silent gaps are longer than inter-speaker gaps. Ten Bosch et al. [2005] reports that intra-speaker gaps are, on average, 25% larger than inter-speaker gaps. This can be explained by rule (3) as for a speaker to continue (an intra-speaker gap) they have to first go though the other selection criteria and then continue speaking. Sacks et al. [1974] note that in order for a listener to project the end-of-turn than the speaker would have to construct their turns, with successive TCUs, in such a way that a turn transition is foreshadowed, showing that the turn is, in effect, winding down. Some effort has been taken by Heldner and Edlund [2010] to critique this `predictive` approach. They argued that the systematic properties outlined within Sacks et al. [1974] are consistent or that they exist at all. Most interesting, is their dissmissal of projection as a central principle of turn-taking, where instead they argue that for inter-speaker gaps longer than 200ms, the listener simply reacts to silence and further on, argue that listeners react to end-of-turn prosodic information. Levinson and Torreira [2015] provide a systematc rebuttal to these claims. Firstly, they argue that, for gaps longer than 200ms, participants cannot simply react to silence as the time taken for silence to become recognisable, react to said silence and produce a response is at minimum 550ms. Riest et al. [2015] point out that the presense of longer gaps could be explained by a speaker intetionally delaying a response when producing a 'dispreferred' response (Levinson [1983], Kendrick and Torreira [2015]). (Second objection?)

1.1.2 Turn-taking Cues

The question remains, what features of speech are relevant when projecting TCU completion and as such an end-of-turn? Prior research related to turn-yielding signals (Duncan [1972]), pointed out prosodic, syntactic and gestural features coincide with turn-completion at an end-of-turn. Later work focussed on expanding these turn-yielding signals for use in projecting a turn completion, and discussing features found eariler in an utterance than some of the phrase-final signals outlined by Duncan [1972]. Most work has focussed on three aspects of conversation: syntactic, prosodic and pragmatic features. Gestural features Duncan [1972] and gaze Kendon [1967] have shown to be a useful part of turn-taking but findings in gaze have suggested these features are action dependent and as such more context-sensitive than other features Clayman [2012]

Although **?** left solving the question of how projection occurs they suggested that syntax provides a main projection cue allthough they also point out that intonation could also be used to differentiate between syntacitcally complete phrases. Sacks et al. [1974] argued that a relationship between syntax and projection can be illustrated by a listener's behviour of overlapping the final-phrase of a sentence while the speakers 'drawls' on the final syllable. As such the listener was projecting the turn-end based on the unfolding syntactic form and the current tempo of the utterance and as such an overlap occurs due to the slowing down of tempo by the speaker.

In their study Ford and Thompson [1996], attempts to characterise TCUs using syntactic,

intonational and pragmatic features and to quantify these features' role in TRPs. To do so they operantialised syntactic, intontation and pragmatic completness and use these to define points in an utterance that are complete with respect to each of these features. An utterance is `syntactically complete` according to Ford and Thompson [1996] if "in its discourse context, it could be interpreted as a complete clause, that is, with an overt or directly recoverable predicate, without considering intonation or interactional import.". They go on to describe syntactic completeness is "judged incremenatally within its previous context". `Intonational completeness` follows from other literature **?** and characterised as "a stretch of speech uttered undera single coherent intonation contour". `Pragmatic completeness` is based on the notion of conversational action and in this instance the action completion, or pragmatic completeness, is based on whether an utterance is a part of a greater conversational action. Ford and Thompson [1996] found most points of turn transitions can be accurately predicted by a combination of all three measures of completeness, known as a `Complex Transition Relevance Place (CRTP)`, where they report that CRTPs predict 71% of actual turn-shifts.

Ford and Thompson [1996] theorised that TCUs and their partnering TRPs are a complex notion and as such multiple factors should be considered for predicting a turn completion. There has been some debate, however, about which feature is most important for projection.

An experimental study De Ruiter et al. [2006], showed that listeners can predict turn-completion equally accurately even when intonational contours are flattened. However turn-completion prediction was heavily effected by the removal of lexicosyntactic data. This showed that listeners use content of an utterance to predict turn-endings. Magyari and De Ruiter [2012] extended De Ruiter et al. [2006] by demonstrating that listeners' accuracy in end-of-turn prediction was correlated with the listener's anticipation of the last words in a turn. As such, they theorise that people make predictions about the remaining content of a turn in order to, or in parallel to, predicting the time left within a turn. Pickering and Garrod [2013] produces an improvement on the previous findings and propose and backup that the listener predicts a speaker's utterance and as such discern the intention of the speaker and in combination with the speaker's current speaker rate to predict the end-of-turn of the current speaker

Bögels and Torreira [2015] have pointed out that the findings within De Ruiter et al. [2006] could be explained by a lack of controlling of other prosodic features aside from intontation, namely final syllable lengthening, which has been pointed out by Duncan [1972] as a end-of-turn feature. Bögels and Torreira [2015] used long and short questions that contained equivalent syntactically equivalent completion points and suggest that since the same syntactic completion points were treated differently (short turn transitions far more prevelant in long questions) than lexicosyntactic information is not sufficient for turn projection. Another experiment carried out by Bögels and Torreira [2015] found that it was late prosodic cues, close to turn boundary, rather than other cues that allowed for accurate turn-detection. They conclude that both lexicosyntactic and intonational cues are used by listeners to time their response. It may appear that these findings are contradictory to results pointed out earlier related to 600ms required for planning the production of content-word turns Indefrey and Levelt [2004] in that these turn-final cues are too late for the speed of a turn transitions as pointed out by

Sacks et al. [1974]

1.1.3 A Concrete Model of Turn-Taking

Levinson and Torreira [2015], from the experimental results listed above, derived a psycholinguistic model in order to account for the observations of Sacks et al. [1974] and with additional temporal considerations. A few results were mentioned previously but as suggested by Sacks et al. [1974] the latencies of speech production means that turn-taking is predictive in nature. However, this could purely relate to the production processes as turn-final cues are used in order for the production to be released. This is inline with arguments of speech production as the speech has already been prepared and the only action required is articulation. These cues have been identified as prosodic: phrase-final syllable lengthening, intonation; syntactic: syntactic completness and the overall conversational action of the contained utterance.

1.2 Models for End-of-Turn Detection and Prediction

The tradition around conversational systems' turn-taking ability is based on the existence of a silence threshold. In these models a turn is assumed to have been yielded by the current spekaer once some threshold has been past (around 650ms). However, as it is to be expected, this approach yields sluggish or possibly, misstaken interruptions. As discussed above, human-human turn-taking organisation is complicated and nuanced and as such the models generated should aim to be able to utilise the signals available in conversation. Further research brought this idea into fruition with what could be interpreted as 'IPU-based' models. An IPU in this instance, is a `Interpausal Unit`, which is a segmented part of continuous speech without silence exceeding a certain threshold (200ms). IPU-based models still undertake some form of silence detection, just with a shorter threshold than a pure-silence model, and after the sufficient silence has been detected the model predicts whether the silence is a TRP or a non-TRP and as such whether the turn has been yielded by the speaker.

Naturally, these models resembled the natural progression of state-of-the art machine learning models moving from rule-based classifier **?**, to a decision-tree classifier **?**Ferrer et al. [2002], **?**], **?**], **?** and then now onto deep learning architectures including the use of the LSTM RNN architecture **?**. Each model uses a different set of features and found varying results on the effectiveness of various prosodic, lexicosyntactic and pragmatic features. Specifically **?** and **?** found that prosody did not contribute significantly to a decision while Ferrer et al. [2002] and **?** found that syntactic and prosodic features both contribute to turn-taking accuracy. Models such as **???** specifically use silence thresholds that are fixed in size. As such, if the speaker yields their turn and if the model does not detect an end of turn then a state of silence may continue. As such other models such as Ferrer et al. [2002], **?** incorporate silence length in order to continouly condition a response based on the time of silence and as such the longer after a pause the more likely that the turn has in fact been yielded. **?**, took this step further by also using turn-holding cues in order to condition the silence threshold so when more turn-holding cues are detected, the system will wait longer before considering a

turn-shift event. This process of monitoring speaker cues, to determine a turn-holding intention **?**, has introduced the concept of a continuous model to monitor turn-taking. An approach which has been all the more feasilble with advances in both deep learning architectures and more powerful feature extractors or pretrained features. Rather than taking on a traditional approach of classifying an utterance, the continuous model processes an utterance incrementally so that at any point the model is able to predict the liklihood of a turn-shift. The system bares more symmetry with our human-human interactions as the system could be able to project turn-completions, determine intent or action and generate an appropriate response. Another issue with previous approaches to turn-taking, namaly the classification approach, is the availability of data that is accurately annotated. As well as this, speech data can be noisy as noted by Sacks et al. [1974] overlapping speech is common but brief, and these sections of speech should not constitue a turn-shift and so this has to be annotated well in data. Recognising this issue, Skantze [2017] proposed a general, continuous turn-taking model, that was trained in a self-supervised manner. Self-supervised as the model is predicting the voice activity of separate speakers over the next two secondsand so it is able to predict a turn-shift based on this speech activity data. The model is also continuous in that it makes these predictions in 50ms intervals. Others have also adopted the general LSTM approach **??** to investigate the effectiveness of certain features. , Roddy et al. [2018] introduced a multiscale approach where lexicosyntactic and prosodic features are processed with different temporal speeds,

Chapter 2

Your next chapter

A dissertation usually contains several chapters.

Chapter 3

Conclusions

3.1 Final Reminder

The body of your dissertation, before the references and any appendices, *must* finish by page 40. The introduction, after preliminary material, should have started on page 1.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default single spacing). Be careful if you copy-paste packages into your document preamble from elsewhere. Some LaTeX packages, such as `fullpage` or `savetrees`, change the margins of your document. Do not include them!

Over-length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

Bibliography

Sara Bögels and Francisco Torreira. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57, 2015.

Steven E Clayman. Turn-constructional units and the transition-relevance place. *The handbook of conversation analysis*, pages 151–166, 2012.

Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535, 2006.

JR Duncan, STARKEY. Toward a grammar for dyadic conversation. 1973.

Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

Starkey Duncan. On the structure of speaker–auditor interaction during speaking turns1. *Language in society*, 3(2):161–180, 1974.

Starkey Duncan and Donald W Fiske. *Face-to-face interaction: Research, methods, and theory*. Routledge, 2015.

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Seventh international conference on spoken language processing*, 2002.

Cecilia E Ford and Sandra A Thompson. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184, 1996.

Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

Peter Indefrey and Willem JM Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144, 2004.

Adam Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 1967.

Kobin H Kendrick and Francisco Torreira. The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):255–289, 2015.

Stephen C Levinson. *Pragmatics*. Cambridge university press, 1983.

Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015.

Lilla Magyari and Jan P De Ruiter. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, 3:376, 2012.

Martin J Pickering and Simon Garrod. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347, 2013.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89, 2015.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. Investigating speech features for continuous turn-taking prediction using lstms. *arXiv preprint arXiv:1806.11461*, 2018.

H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. volume 50, page 696 – 735. 1974.

Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, 2017.

Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86, 2005.

Victor H Yngve. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578, 1970.

Appendix A

First appendix

A.1 First section

Any appendices, including any required ethics information, should be included after the references.

Markers do not have to consider appendices. Make sure that your contributions are made clear in the main body of the dissertation (within the page limit).

Appendix B

Participants' information sheet

If you had human participants, include key information that they were given in an appendix, and point to it from the ethics declaration.

Appendix C

Participants' consent form

If you had human participants, include information about how consent was gathered in an appendix, and point to it from the ethics declaration. This information is often a copy of a consent form.