Applying Discourse Analysis and Data Mining Methods to Spoken OSCE Assessments

Meladel Mistica, Timothy Baldwin
The University of Melbourne
CSSE
{mmistica,tim}
@csse.unimelb.edu.au

Marisa Cordella, Simon Musgrave
Monash University
School of Languages, Cultures and Linguistics
{marisa.cordella,simon.musgrave}
@arts.monash.edu.au

Abstract

This paper looks at the transcribed data of patient-doctor consultations in an examination setting. The doctors are internationally qualified and enrolled in a bridging course as preparation for their Australian Medical Council examination. In this study, we attempt to ascertain if there are measurable linguistic features of the consultations, and to investigate whether there is any relevant information about the communicative styles of the qualifying doctors that may predict satisfactory or non-satisfactory examination outcomes. We have taken a discourse analysis approach in this study, where the core unit of analysis is a 'turn'. We approach this problem as a binary classification task and employ data mining methods to see whether the application of which to richly annotated dialogues can produce a system with an adequate predictive capacity.

1 Introduction

This paper describes our experimentation with applying data mining methods to transcribed doctor–patient consultations. It is in essence a discovery project: we apply methods to a field and task that is not ordinarily associated with such approaches in order to ascertain whether this could make for a tractable learning task.

The task involves the extraction of discourse features from doctor–patient consultations performed by international medical graduates (IMGs), and what is known as 'simulated patients' (Vu et al., 1994), respectively. The IMGs are enrolled in a bridging course in Melbourne as preparation for their Australian Medical Council (AMC) examination, the successful completion of which is one of the pre-requisites to becoming a fully accredited practitioner in Australia. This partially replicates the AMC examination by studying in detail how IMGs perform two objective structured clinical examinations (OSCEs). See Section 2 for full details of the examination environment and participants involved.

The main questions raised when initiating this study were:

- How objective is the testing?
- What is the importance placed on language skills in OSCE environments?
- What makes for a successful OSCE?

In this research, we aim to build classifiers that make reasonable predictions of the data being tested, and possibly point us in the right direction with respect to the questions above. From the classifiers we build, we also hope to ascertain which of our features best predict a successful examination.

We organise the paper as follows. In Section 2, we briefly describe the examination environment and process, the marking scheme, and the participants involved in the testing of the IMGs. We also outline some of the issues that have arisen with regard to the current methods of IMG testing. In Section 3, we present details of the data used. Section 4 describes the features we develop for the task and discusses the reasoning behind the selection of features from a discourse analysis perspective. Section 5 discusses the results of the experiments, with further examination of the data. The

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 577–584
Manchester, August 2008

last two sections, Sections 6 and 7, comprise a discussion of the results and concluding remarks.

2 Background

With Western nations becoming increasingly reliant on medical professionals trained overseas, there is, in turn, a growing need to develop a reliable means of objectively assessing IMGs. The shortage of medical doctors is a worldwide phenomenon currently affecting many Western societies such as the UK, Canada, US and New Zealand, which compete for the best medical practitioners available around the world. Australia is not immune to this global phenomenon, and in the last two decades the shortage of local medical practitioners in Australia has worsened (Birrell et al., 2004). Challenges to the healthcare system in the country are particularly evident in the areas of providing medical care for a growing elderly population and of servicing rural areas, where locally trained doctors do not feel particularly attracted to practise medicine (Han et al., 2006). Currently 35% of the rural medical workforce and 20% of the total national medical workforce consist of IMGs (Flynn, 2006). These figures may increase even further in some regions (Spike, 2006), as preparation of fully educated and trained local medical graduates takes up to thirteen years to complete.

There is considerable disparity among IMGs in their background training, clinical skills, understanding of the health system and communication skills (McGrath, 2004). In order to be registered to practice in Australia, IMGs must successfully complete the Australian Medical Council examinations and a period of supervised training. The medical knowledge of IMGs is assessed in two ways: by multiple choice examinations and by clinical examinations. This second form of examination consists of a series of simulated medical consultations in which a role-player takes the part of the patient, and the IMG's professional knowledge, lay-cultural knowledge, socio-cultural assumptions, institutional norms, and values and personal experiences are all in full display during the unfolding of the medical event (Roberts et al., 2003). Whenever cultural factors are not shared with their patients, the interpretative schema and therefore the comprehension of speech are affected by this lack of commonality in the participants' inferences and contextual cues (Gumperz, 1999). Such effects are likely to cause miscommunication in medical visits and have a potential negative effect on patients' satisfaction in the consultation. Identification of the communication difficulties faced by IMGs can therefore inform modifications to the training provided to IMGs when they prepare for the Australian Medical Council examinations, as well as suggesting more nuanced and targeted procedures for assessing communicative skills within those examinations, all with the goal of working toward a better equipped medical workforce for the future. The use of automated analytic procedures to try to establish objective criteria for communicative success is an important step in this process.

Assessing language knowledge and competence quantitatively is not a novel concept in second language learning assessment. However, the application of data mining methods to automatically assess language proficiency in a discourse setting is novel. Levow et al. (1999) propose an architecture to automatically assess language proficiency. In their paper, they propose an architecture that employs data mining methods, but do not build classifiers over their spoken data to test this proposal. A closely related line of research is on the automatic classification of discourse elements to assess the quality of a written genre (Burstein et al., 2001). Like this work, it focuses on extracting features from the discourse as a whole. But unlike this study, the authors extract high level features, such as rhetorical structure, of written discourse. The study we present in this paper is rather unique in its approach to language assessment.

3 Data

The data is taken from transcribed recordings of examinations from students enrolled in a bridging course at Box Hill Hospital in Melbourne, Australia. Each candidate was video-recorded enacting medical consultation scenarios with what is known as a standardised or simulated patient (SP). This method of testing is known as an objective structured clinical examination (OSCE), which is an emulation of a doctor–patient consultation, much like a role-play setting.

In this study, the role of the patient (SP) is enacted by a qualified doctor who follows a script, and has well-defined ailment(s) and accompanying concerns. Even though the SP assumes the same ailment and disposition with all the candidates, the interaction between the candidate and the SP is uncued and free-form. They simply present the information in a standardised manner across all candidates and perform the role of the patient as felicitously as possible.

For this set of examinations there are 2 types of OSCE stations referred to as STD (sexually transmitted disease – genital herpes) and BC (bowel cancer). The SP for the STD station is played by a female doctor. The patient she plays has genital herpes and is concerned about how this will affect her chances of falling pregnant, and how this condition may also affect her baby. The SP for the BC station is played by an older male. A tumour is discovered in his bowel lining and he is reluctant to undergo any treatment because one of his good friends suffered a similar condition and his quality of life was severely diminished.

Even though the consultation is free to be negotiated between doctor (candidate) and patient (simulated patient), each of the OSCEs cannot exceed 8 minutes, and is terminated by the examiner if it does so.

3.1 Transcription

The recordings are transcribed in ELAN, a multimedia annotation tool developed at the Max Planck Institute, to help encode low-level linguistic features such as overlapping and timing information. The information and features extracted from the discourse are largely based on a 'turn'.

Here we consider a turn as being normally dominated by one speaker. It can be made up of multiple intonation units. When there is backchannelling, overlapping, or any interruption by the other participant, then the turn is encoded as ending at the end of the interrupted intonation unit. Otherwise, transition pauses commonly signal turn changes, unless latching occurs.

Given that the OSCE setting aims to emulate as close as possible a real medical consultation, this interaction, like all uncued spoken dialogues, also has evidence of complicated turn-taking negotiations, disfluent and unintelligible speech, interrupted speech, challenges for the floor, and the like, all of which must be encoded and noted in ELAN. Transcribing such data is not a trivial matter. In addition, transcribing the data in order to extract these features is also a demanding task in itself, which makes creating data for such tasks an involved process.

Disfluencies and repairs are encoded in a limited way, only by way of marking up truncated or unfinished words. We also do not take a fine-grained approach in encoding delaying strategies (Clark et al., 2002), that is we do not differentiate whether the *uh* or *ah* encoded represents lexical search, a wish to hold the floor, a wish to give up the floor or buying time to construct what to say next.

3.2 OSCE scoring

In an OSCE setting, candidates are given an overall pass or fail rating for each station by an OSCE examiner observing the interaction. This overall evaluation can be based on a number of performance criteria which tests the candidates medical, clinical and communication skills (Grand'Maison et al., 1992). The OSCE marking scheme used for this study consists of 5 assessable categories, as follows:

APPROACH: the ability of the candidate to communicate with the patient;

HISTORY: the ability of the candidate to collect medical history;

INTERPRETATION: how well does the candidate interpret his or her investigation in order to formulate an appropriate diagnosis;

MANAGEMENT: how well does the candidate formulate a management plan for the diagnosis;

COUNSELLING: is the candidate able to give appropriate counselling to the patient.

The first category tests language knowledge and competency both at the lexical and discourse level, while the remaining four categories test medical knowledge and clinical competency.

4 Feature Engineering

We extracted a total of 38 features from the transcribed data. Some of these features are based on what is marked up according to the transcription scheme, while others are based on timing information or lexical information as encoded in ELAN. These include features such as signals for delaying speaking or hesitation (Clark et al., 2002), features of conversational dominance (Itakura, 2000), the manner in which turn-taking is negotiated (Sacks et al., 1974), temporal features such as pausing (ten Bosch et al., 2005), as well as our own features, which include 'lexical introduction', and 'lexical repeat'.

In encoding features of conversational dominance, we focus on *participatory dominance* (Itakura, 2000), which looks at which speaker contributes most to the dialogue in terms of content.

Lexical introduction refers to a non-stop word that is introduced by the doctor (IMG) or the patient (SP), while *lexical repeat* encodes how many times a word introduced by the other interlocutor is repeated by the speaker.

Almost all of the features developed are continuous, based on timing information or word counts. The only binary feature used encodes whether the doctor initiates the consultation or not.

As mentioned in the previous section, the features developed were largely based on turns. This is to capture, along with other features such as overlapping and pauses, the interactional aspect of the communication. For example, conversational cooperation and speaker reassurance can be captured with these features. Another aspect to the development of these features, particularly for the lexical-based features, is whether the IMG has a suitable vocabularly and if they employ it appropriately in the interaction.

We arrive at 11 feature sets from which we build our classifiers, as described in Table 1.

Not all features are exclusive to any one feature set, that is, it is possible for a single feature to belong to a number of feature sets.

The sets were designed to isolate possible characteristics of not only the discourse as a whole, but how the participants negotiated their interaction. These features sets were developed from observing each of the consultations with the expectation that these were salient and determining features of a successful examination.

5 Experiments

There was a total of 11 OSCE candidates, all of whom performed an STD and a BC station, giving us in total 22 instances for this binary classification task to predict a pass or fail examination result. Of the 22 instances, we had 5 failures and 17 passes. Given the small number of instances, we maximised our dataset by employing 10-fold stratified cross-validation, as well as leave-one-out cross-validation which uses all but one instance in training and the held-out instance for testing.

Feature set	Example features
all	- all 38 features
cooperation	- overall word count - length of interaction - number of turns
hesitation	- number of *uh* and *ah* - number of unfinished words
overlap	- number of overlapping words - length of overlap (time)
pause	- transition pauses - within turn pauses
timeBased	- all time-based features
turns	- all turn-based features - number of turns - longest turn - single word responses
uniqNrepeat	- number of introduced content words by each speaker - number of times speaker uses word introduced by other
wordBased	- number of words in dialogue - longest number of words in a turn
patient	- all SP-based features
doctor	- all IMG-based features

Table 1: The 11 feature sets developed

The baseline system we use for comparison is zero-R, or majority vote. For our supervised classifier, we employ a lazy learner in the form of the IB1 algorithm implemented in WEKA.

5.1 Results for Feature Sets

Our initial classifiers held some promise. The classifier built from all of the features was equivalent to the baseline system, and the combination of the word-based features surpassed the baseline's results, as shown in Table 2.

To evaluate our system, we employ simple classification accuracy, in addition to precision, recall and F-score. Classification accuracy is the proportion of correct predictions by the classifier, irrespective of class. Precision gauges how successful the pass predictions of a given classifier are, while recall gives us an indication of how successful a given classifier is at identifying the candidates who actually passed. Finally, F-score is a composite of precision and recall, and gives us an overall performance rating relative to passed candidates.

The least successful classifier was built on the

	10-fold cross validation				Leave-one-out cross validation			
Feature set	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
baseline	.773	.773	1.00	.872	.773	.773	1.00	.872
all	**.773**	**.773**	**1.00**	**.872**	**.773**	**.773**	**1.00**	**.872**
cooperation	.682	**.789**	.824	.806	.682	**.778**	.824	.800
hesitation	.636	.737	.824	.778	.636	.737	.824	.778
overlap	**.773**	**.833**	.882	.857	**.773**	**.833**	.882	.857
pause	.682	**.778**	.824	.800	.727	**.789**	.882	.833
timeBased	.500	.647	.688	.667	.545	.706	.706	.706
turns	.727	**.789**	.882	.833	.727	**.789**	.882	.833
uniqNrepeat	.636	.765	.765	.765	.682	**.778**	.824	.800
wordBased	**.864**	**.850**	**1.00**	**.919**	**.864**	**.850**	**1.00**	**.919**
patient	.727	**.867**	.765	.813	.727	**.867**	.765	.813
doctor	.733	**.800**	.941	.865	.727	**.789**	.882	.833

Table 2: Classification results for STD and BC

feature set based on timing, which contains information such as the overall length of the dialogue, the overall length of transition pauses, in-turn pauses and other time-based features. This was most surprising because as a general observation, candidates who allowed extended pauses and uncomfortable silences were those who seemed to perform poorly, and those who did not leave too many silences, and could maintain the flow of the dialogue, seemed to perform well.

Given the small number of training instances each classifier is based on, these first results were somewhat encouraging. With respect to the baseline, the overall performance of two of the systems equalled or surpassed the baseline in terms of F-score. Most of the classifiers performed well in terms of precision but less well in terms of recall, i.e. when the classifiers predicted a pass they were generally correct, but there were significant numbers of candidates who were predicted to have failed but passed in practice.

5.2 Data Introspection Retrospectively

Although the results show promise, it was expected that more of the feature sets would return more favourable results. The possible reasons why the time-based features, and many of the other feature sets developed, did not perform as well as expected may have been because the features used in building the classifiers could have been combined in a better way, or because the data itself had too many anomalies or was too disparate. We would expect that extra data could iron out such anomalies, but developing additional data is expensive and more recordings are not always available. The advantage of having a small dataset is that we are able to do fine-grained annotation of the data, but the obvious disadvantage is that we cannot easily generate extra amounts of training data.

One very noticeable feature of the OSCE stations was that the STD SP had a very different communicative style to that of the the BC SP. Based on this observation we conducted tests given the hypothesis that the possible bias in the data could have stemmed from having two very different testing approaches from the two SPs. In general, the BC SP was more leading and in a sense more forgiving with the candidates. In contrast to this, the STD SP tended to be more felicitous in her role as a patient, allowing awkward silences and not prompting the candidates for further exploration.

We conduct the Mann-Whitney test, a rank sum test, over the data in order to diagnose whether the poor results were due to the distribution of the data or whether the classifiers built with the selected features were simply poor predictors. The Mann-Whitney test ascertains whether there is a difference in the population mean of the two samples given, without making any assumptions about the distribution of the data.

We sub-sample the data in two ways in examining its homogeneity: (a) FAIL juxtaposed with PASS candidates; and (b) BC juxtaposed with STD stations. Test (a) essentially tests which examinable category contributes the most to a pass or fail outcome, whilst test (b) examines whether there is an inherent difference in the way the test-

Category	OVERALL	APPROACH	HISTORY	INTERPRETATION	MANAGEMENT	COUNSELLING
z-score	1.84	1.21	-0.03	2.53	0.85	1.64

Table 3: Mann-Whitney z-score for BC and STD samples (OVERALL is the cumulative total of all 5 categories)

Category	OVERALL	APPROACH	HISTORY	INTERPRETATION	MANAGEMENT	COUNSELLING
z-score	-3.29	-3.13	-2.43	-2.31	-2.31	-1.57

Table 4: Mann-Whitney z-score for failed and passed samples

ing was conducted between the BC and STD stations.

BC VS. STD

We use the ranking from the 5 assessable categories outlined in Section 3 and obtain the Mann-Whitney z-score for each category. The z-score gives us an indication of how disparate the two separated datasets, BC and STD, are. The further away from 0 the z-score is, the greater the evidence that BC and STD data are not from the same population, and should be treated as such. The results of this test, as seen in Table 3, show that these two groups differ quite markedly: the candidates were consistently marked differently for all assessable categories except HISTORY. This is a striking peculiarity because each candidate was tested in both the STD and BC stations.

Based on the above, we can posit that the distinct testing styles of the STD and BC SPs were the reason for our original lacklustre results, and that the two data samples need to be treated separately for the classifiers to perform consistently.

FAIL VS. PASS

In addition to the BC vs. STD test, we also test how the failing candidates differ from the passing candidates across the evaluation criteria.

The main idea behind this test is to see which of the assessable categories contributed the most in the overall outcome of the examination. For this test, we would not expect the absolute z-score of any of the assessment components to exceed the absolute z-score of the OVERALL category given that it is the cumulative scores of all categories.

The results in Table 4 suggest that APPROACH correlates most highly with the pass/fail divide in the OSCE assessments, followed by HISTORY, then INTERPRETATION and MANAGEMENT, and finally COUNSELLING. Recall that APPROACH is the component that assesses language and communication skills. In particular, it assesses the style and appropriateness of the way candidates convey information, from lexical choice to displaying empathy through communication style. Given that APPROACH correlates most strongly with the assessment result, the decision to focus our feature engineering efforts on linguistic aspects of the doctor–patient interaction would appear justified.

5.3 Results for STD & BC Data

Given the results from the Mann-Whitney tests reported in the previous section, we separate the data into two lots: those from the STD station, and those from the BC station.

Even though there were very few instances in the original dataset, we aim to see in these experiments whether this separation improves the performance of the classifiers. We build classifiers over each dataset using the same features as before.

The results of the tests performed over the separated datasets, as shown in Table 5, show a big improvement over the baseline for STD, while the BC dataset is more problematic.

In the STD group, we see that four feature sets, *all*, *turns*, *wordBased* and *patient* equal or surpass the baseline F-score.

In contrast to this, upon examination of the performance of the classifiers built over the BC dataset, we do not observe any improvements over the baseline and the results are markedly worse than those for the combined dataset. Having said this, when we combine the outputs of the two component classifiers, the F-score for *all* features is 0.882, an improvement over the original combined system.

6 Discussion

The OSCE assessment does not merely examine the language skills of the candidates, but it also as-

	BC				STD			
Feature Set	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
baseline	.818	.818	1.00	.900	.727	.727	1.00	.842
all	.727	**.875**	.778	.824	**.909**	**.889**	**1.00**	**.941**
cooperation	.727	.800	.889	.842	.364	.571	.500	.533
hesitation	.636	.778	.778	.778	.636	.700	.875	.778
overlap	.727	.800	.889	.842	.636	.700	.875	.778
pause	**.818**	**.889**	.889	.889	.545	.667	.750	.706
timeBased	.636	**.857**	.667	.750	.545	.667	.750	.706
turns	.636	.778	.778	.778	**.818**	**.800**	**1.00**	**.889**
uniqNrepeat	.727	.800	.889	.842	**.727**	**.778**	.875	.824
wordBased	.636	.778	.778	.778	**.909**	**.889**	**1.00**	**.941**
patient	.727	**.875**	.778	.824	**.818**	**.875**	.875	**.875**
doctor	**.818**	**.818**	**1.00**	**.900**	.455	.625	.625	.625

Table 5: Results for separated BC and STD datasets (leave-one-out)

sesses the efficacy of their communication skills in conveying correct and accurate medical information within a clinical setting. It can be seen from Table 4 that there is a high correlation between the overall pass or fail and the assessable category APPROACH.

The examiners' subjectivity of overall performance is minimised by the highly structured examination setup and well-defined assessment criteria. However, as shown in Table 3, the communicative style of the SP is a contributing factor to the perception of successful clinical and communication skills. The Mann-Whitney tests suggest that an SP's approach and their apparent satisfaction during the clinical encounter can affect the judgement of the examiner.

Additional inspection of the data revealed that the assessment criteria which focused on language and communication skills correlated highly with an overall pass grade, moreso than the other criteria. This seems to suggest that more emphasis should be placed on language skills and communication style in the assessment of the candidates.

Assessing language competency is no trivial matter, and capturing the linguistic features of dialogues in an attempt to define competence, as we have done here, is a demanding task in itself. Although many of our features were focused on turn-taking, speaker response and interaction, we did not develop features that encompass the information structure of the communicative event.

It is assumed that miscommunication between non-native and native speakers of a language is due to a lack of language knowledge pertaining to syntax, morphology or lexical semantics. However many of these communication difficulties arise not because of this lack of grammatical knowledge, but through a difference in discourse styles or information structure as governed by different cultures (Wiberg, 2003; Li, 1999).

Given that the word-based feature sets were the most successful predictors of an OSCE outcome, future work of this kind could make use of medical-based lexicons to gauge whether technical or non-technical word usage in such environments is judged favourably. In addition, further work should be done to test the hypothesis that information structure or rhetorical structure does impact on overall perception of a successful communication, such as a variation on the methods employed by Burstein et al. (2001).

One obvious improvement to this study would be to reduce the expense in producing the annotated data. Future work could also be done in automatically extracting features from non-transcribed data, such as timing information based on pause length and the turn length of each speaker.

7 Conclusions

In this research, we have built classifiers over transcribed doctor–patient consultations in an attempt to predict OSCE outcomes. We achieved encouraging results based on a range of lexical and discourse-oriented features.

In our first experiments, we combined the data from two discrete stations in an attempt to maximise training data, and achieved modest results. Subsequent analysis with the Mann-Whitney test indicated both that success in the APPROACH category correlates strongly with an overall successful OSCE, and that the data for the two stations is markedly different in nature. Based on this finding, we conduct tests over the data for the individual stations with noticeable improvements to the results.

The results of this exploratory study have been quite encouraging, given the novel domain and limited data. We have shown that a data mining approach to OSCE assessment is feasible, which we hope will open the way to increased interest in automated medical assessment based on linguistic analysis.

References

Birrell, Bob, Lesleyanne Hawthorne. 2004. Medicare Plus and overseas trained doctors. *People and Place*, 12(2):83–99.

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

Burstein, Jill, Daniel Marcu, Slava Andreyev, Martin Chodorow. 2001. Towards Automatic Classification of Discourse Elements in Essays. *ACL*, 90-97.

Clark, Herber H., Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(2002):73–111.

Flynn, Joanna. 2006. Medical Release. *Australian Medical Council*, 17(August 2006).

Grand'Maison, Paul, Joëlle Lescop, Paul Rainsberry, Carlos A. Brailovsky. 1992. Large-scale use of an objective, structured clinical examination for licensing family physicians. *Canadian Medical Association*, 146(10):1735–1740.

Gumperz, John. 1999. On Interactional Sociolinguistic Method. In *Talk, Work and Institutional Order. Discourse in Medical, Mediation and Management Settings* S. Sarangi and C. Robers (*eds*), 453–471.

Han, Gil-Soo, John .S Humphreys. 2006. Integratoin and retention of international medical graduates in rural communities. A typological analysis. *The Australian Sociological Association*, 42(2):189–207.

Itakura, Hiroko. 2000. Describing conversational dominance. *Journal of Pragmatics*, 33(2001):1859–1880.

Levow, Gina-Anne, Mari Broman Olsen. 1999. Modeling the language assessment process and result: Proposed architecture for an automatic oral proficiency assessment. *Workshop On Computer Mediated Language Assessment And Evaluation In Natural Language Processing*.

Li, Han Zao. 1999. Comunication Information in Conversations: A Cross-cultural Comparison. *International Journal of Intercultural Relations*, 23(3):387–409.

McGrath, Barry. 2004. Overseas-trained doctors. Integration of overseas-trained doctors in the Australian medical workforce. *The Medical Journal of Australia*, 181(11/12):640–642.

Roberts, Celia, Val Wass, Roger Jones, Srikant Sarangi, Annie Gillett. 2003. A discourse analysis study of 'good' and 'poor' communication in an OSCE: a proposed new framework for teaching students. *Medical Education*, 50:192–201.

Sacks, Harvey, Emanuel A. Schegloff, Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Spike, Neil. 2006. International Medical Graduates: The Australian perspective. *Acad Med*, 81):842–846.

Vu, Nu Viet, Howard S. Barrows. 1994. Use of Standardized Patients in Clinical Assessments: Recents Developments and Measurement Findings. *Educational Researcher*, 23(3):23–30.

Wiberg, Eva. 2003. Interactional context in L2 dialogues. *Journal of Pragmatics*, 35(2003):389–407.