Automatic Induction of Dialogue Structure from the Companions Dialogue Corpus

Debora Field[1] and **Simon Worgan** [2] and **Nick Webb**[3] and **Mark Hepple** [4] and **Yorick Wilks**[5]

Abstract.
We present a mechanism to learn repeated substructures of dialogue from an appropriately annotated dialogue corpus. We hypothesised that we could automatically segment a dialogue corpus into 'chunks' corresponding roughly to 'game'-like structures. In our dialogue system we manage dialogue by employing a set of hand-crafted networks (ATNs) which we call 'Dialogue Action Forms' (DAFs). DAFs deterministically guide the progress of a conversation with the user, and are individuated by conversational topic—when the conversation is moved onto a new topic (either by the system or the user), a new DAF is pushed onto the stack. We hoped to automatically induce these DAFs from a dialogue corpus. To do this, we needed to derive substantive DAFs, such as might be used to learn about someone's family (names, ages, relationships to the user, *etc.*), as well as meta-conversational DAFs that manage the dialogue at the 'exchange structure' or 'game' level, and we needed to combine the two. We hoped to find generalisations that clustered discovered topic chunks into classes of functionally similar chunks. To achieve this, our method was to devise a text segmentation technique specifically suitable for dialogue, and to use this in conjunction with a tool for analysing dialogue structure in terms of dialogue act sequences.

1 Introduction

In the core area of dialogue management, there have been attempts to produce more sophisticated but learnable structures than the finite-state rule systems, of which the best known is TRINDI.[6] Lemon's WITAS [11] at Stanford and the COMIC[7] system showed that more complex dialogue behaviours could be got from network and stack architectures that allow reentry and recovery of topics. Following in a similar vein, our dialogue system uses the stack architecture to manage dialogue, employing a set of hand-crafted networks (ATNs) which we call 'Dialogue Action Forms' (DAFs), which get pushed onto and popped off the stack. DAFs are individuated by conversational topic—when the conversation moves to a new topic, a new DAF is pushed onto the stack. However, since the DAFs are hand-crafted, we are in the position of having to manually determine what constitutes a change of topic, and what a conversation which is changing topic looks like. Since we wanted our system's dialogue to resemble natural conversation, we decided to examine some data to look for evidence of these phenomena. In this paper we describe the machine-learning strategy that we are therefore using to attempt to discover naturally occurring DAFs in our dialogue corpus.

The motivation for this work was the on-going development of a personal automated *Senior Companion*[8] that learns about its (senior citizen) owner's life story, needs, and preferences through free-ranging natural language dialogue with him/her.

1.1 Companions Dialogue Corpus

This research was carried out using a spoken dialogue corpus[26] collected under Companions. It (currently) contains 54 transcribed spoken dialogues, 24,262 turns, and 243,574 words (further additions are expected). In the dialogues, two participants (*A* and *B*) discussed the family photographs of one of the participants (*B*). The general aim of all these conversations was for *A* to gradually get to know *B*'s life story by means of perusing *B*'s family photos together with *A*, and discussing what (which significant (to *B*) events, individuals, locations, *etc.*) the photographs depicted. This particular domain for the corpus collection was carefully chosen to provide suitable empirical data that would help with the development of the Senior Companion, which incorporates an application for encouraging reminiscence while using a user's personal photos as a prompt.

A subset of the dialogues in our corpus was obtained by using the *Wizard of Oz* (WoZ) method, in which *B* had a conversation with a computer system (*A*), but the system was actually being operated remotely by a human being. *B* believed he/she was talking to an actual computer, and not to a human being. *A* was observing some restrictions in his/her conversational abilities, attempting in some carefully defined ways to converse just as it is expected the mimicked computer system will eventually converse, once it is finished. The point of using this method was to try to lead *B* into behaving more as he/she would behave if *A* really were a conversationally compromised computer system, and thus to elicit utterances from him/her that are supposedly more like the utterances that our real computer system will eventually elicit.

In using the WoZ method to collect data to help steer the design of the Senior Companion, there is an implicit assumption that we have correctly guessed the conversational prowess of our eventual system, which, of course, we cannot have done accurately, and so the data we have collected by the WoZ method cannot be trusted. For this reason, we did not restrict the corpus used for this research to the WoZ dialogues, but we also included dialogues in which *A* presented him/herself as a human, and an openly human-to-human dialogue about *B*'s photographs was conducted.

[1] University of Sheffield, UK, email: d.field@shef.ac.uk
[2] University of Sheffield, UK, email: s.worgan@dcs.shef.ac.uk
[3] State University of New York, USA, email: nwebb@albany.edu
[4] University of Sheffield, UK, email: m.hepple@dcs.shef.ac.uk
[5] University of Sheffield, UK, email: yorick@dcs.shef.ac.uk
[6] http://www.ling.gu.se/projekt/trindi/
[7] http://www.hcrc.ed.ac.uk/comic

[8] Being developed under *Companions*[25], European Commission Sixth Framework Programme Information Society Technologies Integrated Project IST-34434 (http://www.companions-project.org/).

2 Background and Previous Work

In order to automatically induce DAFs from our dialogue corpus, we first needed to find segments of text that revolved around the same topic—which were *about* the same thing. Having found our chunks, we then hoped to find generalisations that clustered discovered topic chunks into classes of functionally similar chunks. These strategies required us to take into consideration previous work on topic segmentation and on techniques and tools for analysing dialogue structure.

2.1 Text Segmentation

Text segmentation originated in the text summarisation field, and has largely been applied to prose rather than dialogue. Different approaches have been tried, including lexical cohesion [9] (TextTiling), [5], lexical chains [13, 2], story segmentation using lexical chains [22], topic detection [1], and others. To our knowledge, little work has been done on topic segmentation for *dialogue*, as opposed to prose. We hoped that one of the pre-existing topic segmenters for prose would perform well for our dialogue corpus, but found that it did not, and so we developed our own topic segmentation strategy that builds segmented 'topic chunks' from the discrete system-user dialogue turns (to be discussed).

2.2 Dialogue Acts

In order to carry out analysis of the dialogue structure, we decided to first assign a dialogue act to each turn. The notion of 'dialogue act' comes from the field of pragmatics, a discipline which emerged in the late 19th and early 20th century proposing a new view of language as 'meaningful action' in contrast to the then accepted view of it as a matter of truth-conditional semantics. All major dialogue theories now treat dialogue acts as, to some extent, central.

The conceptual granularity of the dialogue act labels that people have used varies considerably between alternative analyses, often being driven by demands specific to some application or domain. In 1998 the Discourse Resource Initiative finalised a task-independent set of DAs, called DAMSL (Dialogue Act Markup in Several Layers), for use across different domains [6]. DAMSL has been used to mark up several dialogue corpora, including TRAINS [7], and SWITCHBOARD [10]. This research used a small set of acts (24 acts) derived from the DAMSL set, and a pre-existing automatic dialogue act tagger [23]. The dialogue acts we used were as follows:

Full name	**Abbr.**	**Full name**	**Abbr.**
CONVENTIONAL-CLOSING	fc	CONVENTIONAL OPENING	fp
STATEMENT-NON-OPINION	sd	STATEMENT-OPINION	sv
WH-QUESTION	qw	YES-NO-QUESTION	qy
OPEN-QUESTION	qo	RHETORICAL QUESTION	qh
ACKNOWLEDGE	b	AGREE-ACCEPT	aa
MAYBE-ACCEPT-PART	aap	HEDGE	h
NO-ANSWERS	nn	SIGNAL-NON-UNDERSTANDING	br
HOLD-BEFORE-ANSWER	ĥ	REPEAT-PHRASE	bm̂
ACTION-DIRECTIVE	ad	COLLABORATIVE COMPLETION	2̂
APPRECIATION	ba	DOWNPLAYER	bd
APOLOGY	fa	THANKING	ft
OR-CLAUSE	qrr	OFFERS	oo

2.3 Dialogue Structure

Having marked up dialogue acts, we then looked for repeated patterns of consecutive sequences of dialogue acts (while also taking into account patterns of conversational topic). This approach was in line with the tradition that has been concerned with the functional structures of interactions [19]—turn-taking, when and how interruptions occur, how repairs are signalled and detected, *etc.* Under such analysis, much dialogue appears to be constructed from adjacency pairs, *e.g.*, question/answer, complaint/apology, greeting/greeting, accusation/denial. Dispreferred responses are also detailed—responses which are appropriate, but dispreferred by the hearer, *e.g.*, invitation/refusal (versus invitation/acceptance), and assessment/disagreement (versus assessment/agreement)[12]. Similar to these approaches is the notion of *exchanges* [20], considered to occur in situations where there is inherent structure in the interaction, but little exchange of initiative, *e.g.*, in classroom and doctor/patient interchanges. The highest level of exchange representation is captured by conversational 'games' [17], which comprise transactions that have some correlation with subtasks, or subtopics. [4] describe *dialogue games* as exchanges between speakers that fulfil some limited goal, embodying our expectation of natural patterns within dialogue, *e.g.*, that questions usually precede answers, and requests precede either an acceptance or a refusal.

The goals in each of the conversations in our corpus are less clearly defined than the goal in, say, a tutor/pupil dialogue in which the tutor is teaching the pupil how to solve some problem or other. The overarching goal is for the system (or 'interviewer') to encourage the user (or 'interviewee') to talk about his/her life, while using personal photographs as a prompt. (This on the assumption that older people enjoy looking at their photos, and being reminded of events and people from their lives.) A subgoal of this overarching goal is for the system to learn the details of significant events in the user's life, so as to eventually be able to construct a 'story' of the user's life (probably in the form of a timeline), and so as to be able to share its knowledge with other family members who are likely to be interested in the family stories (grandchildren, for example).

The subtasks of these two goals are at present somewhat opaque to us, although no doubt there are practical strategies for how to go about achieving these goals outlined in the *reminiscence* literature, which we intend to investigate. We hypothesised, however, that we might be able to empirically discover (at least clues to) what the 'game' is like that the participants of our photo-centric conversations were playing.

3 Method and Results

In order to discover repeated patterns of consecutive sequences of dialogue acts, while also taking into account patterns of conversational topic, we used a two-stage clustering process. First, a novel hierarchical clustering technique was used to build segmented topic chunks from the discrete system–user dialogue turns. From these chunks a number of Hidden Markov Models (HMM) were derived to capture common structures [21], *i.e.*, the probability that one particular dialogue act follows another. These models were then ranked according to their descriptive capabilities. Those that describe a greater proportion of the corpus are being ranked highly and returned in the results.

3.1 Topic Chunks

In order to identify topic chunks, we needed to form an indeterminate number of clusters from an unlabelled corpus. After trying some well-respected text-segmentation software without success (to be discussed) we developed our own clustering method tailored specifically for dialogue. We adapted a standard additive hierarchical clustering method [14] and used the WordNet similarity measure [15, 3] to establish the 'semantic distance' between the *nouns* in dialogue turns.[9] We established the average similarity between nouns within a cluster (the *intra-similarity*), and the similarity between a noun and the nouns in the cluster at the next time step (the *inter-similarity*).

Given this measure of inter- and intra-similarity we then joined together two adjoining clusters if they increased the validity of the set of topics [18] (*i.e.*, if the inter-similarity was greater than the intra-similarity) and if they were next to each other on the same level of the hierarchy. This joining formed a cluster at the next level of the hierarchy which continued to build up until a single cluster was reached, encompassing the whole document, or until no further gains in validity could be made. The advantages of this approach are that the clusters remain unified in time, and change at distinct time points in the discourse. We also discover the optimum number of clusters (topics) without having to set any variables in advance.

3.1.1 Results

Having applied our topic clustering algorithm, we found that each dialogue had been divided into multi-turn chunks which appeared on manual inspection to revolve around one or more conversational topics. Here are two topic chunks from the same dialogue. Each chunk shows the speaker (U: user, S: system), the turn number, the dialogue act (automatically tagged) and the nouns being considered by the clustering algorithm (which is using a semantic similarity measure derived from WordNet):

1.
```
U  218   sd  [uncle, aunt, lily, face]
S  219   qy  [face]
U  220   sd  [secret, take, war, kind, saving,
             little, being, nature]
S  221   sd  []
U  222   sd  [singing, voice, people]
S  223   b   []
U  224   sd  [day, routine, bit, go, ilford]
S  225   sd  []
U  226   sd  [concert, friend]
S  227   ba  [great]
U  228   sd  [concert, friend, little, bit, life,
             ilford]
S  229   sd  []
U  230   sd  [aunt, lily, bit]
S  231   aa  []
U  232   sd  [english, breakfast, egg, lily, cut,
             bit, instinct, time]
```

2.
```
U  sd    322 [system, stage, english, thought,
             material]
S  aa    323 []
U  sd    324 [material, sums, thought, entrance]
S  b     325 []
U  sd    326 [course, school, thought, idea,
             show, going, help, paper, sort]
S  aa    327 []
U  sd    328 [material, paper, clue]
```

[9] We included person names and location names in these calculations (identified by using the named entity recognition tool from the ANNIE system within the GATE framework [8]), but did nothing fancy concerning semantic distances between these proper nouns (such as establishing whether a particular location was geographically close to or part of another). We just looked for repetition of the proper nouns themselves.

3.1.2 Topic Troughs

We also found that a common phenomenon was for the semantically content-ful topic chunks such as these to be separated from one another by sequences of one-turn chunks that contained comparatively little semantic content, and had therefore not clustered together into multi-turn clusters. We believe these 'topic troughs' may be a reflection of the fact that while all conversational turns have pragmatic force, much of what is uttered in conversation has little semantic content. It is not hard to find semantically spartan sequences like the following, which is an extract from the corpus:

S: Yes, it's this couple here in this photograph.
U: Yes, you've got the ...
S: This one ... erm
U: You've spotted them, have you?
S: It's this couple?
U: Yes, absolutely.
S: Yes.

It is of note that if thorough syntactic analysis of each turn were being carried out, it would be possible for our program to spot and resolve definite noun phrases, which would give us a clearer indication that these seven turns are in fact about the same thing: *this couple*. However, we are currently not attempting any syntactic analysis at all. One tactic we may try in the next round of experiments is to observe the presence of (but not to attempt to resolve) third person pronouns (with the exception of *it*, which is often a dummy), and possibly other definite NPs. We would then embody a heuristic in the topic clustering algorithm which says that if pronouns *etc.* are present, it is highly likely that the conversational topic has not changed.

Before designing our own algorithm for finding topic chunks, we tried a program written by Choi,[10] which implements the work discussed in [5]. We found, however, that the topic chunks identified by this program were very large (scores of turns long), and that where a topic boundary had been drawn, it was clear on manual inspection that there were semantically closely related nouns on either immediate side of many topic boundaries. We believe that Choi's segmenter may have been confused by the presence of the topic troughs that we later found after having applied our own algorithm. However, this bears further investigation.

3.2 Deriving Dialogue Structure

After topic clustering we wanted to compare each topic chunk we had found with every other topic chunk in the corpus, to look for recurring patterns of dialogue act sequences and conversational topic, in the hope of finding classes of functionally similar chunks that we could use to inform the design of our dialogue action forms (DAFs, see section 1). Before doing this, we substituted placeholders for the nouns within each chunk, so that we would not be looking for the recurrence of a particular noun across different chunks, rather we would be looking for a pattern of *how* the same noun was discussed within each chunk. So, for example, here again is chunk 2:

[10] Available for download at www.freddychoi.co.uk

```
2.
U  sd    322 [system, stage, english, thought,
              material]
S  aa    323 []
U  sd    324 [material, sums, thought, entrance]
S  b     325 []
U  sd    326 [course, school, thought, idea,
              show, going, help, paper, sort]
S  aa    327 []
U  sd    328 [material, paper, clue]
```

In this (and every) chunk, we replaced each noun with a placeholder, using the same placeholder name for the same noun. So chunk **2** became:

```
2.
U  sd    322 [1, 2, 3, 4, 5]
S  aa    323 []
U  sd    324 [5, 6, 4, 7]
S  b     325 []
U  sd    326 [8, 9, 4, 10, 11, 12, 13, 14, 15]
S  aa    327 []
U  sd    328 [5, 14, 16]
```

The fact that I have used numbers as markers here is not significant. The important point is that the numbers show where the same noun occurs within the topic chunk's turns.

Having mapped every chunk into structures like this, in order to compare each chunk with every other, we were faced with the challenge of extracting patterns from discrete data of variable lengths. Accordingly, we described each sequence of variable length data as a HMM [16], avoiding the degradation inherent when sampling or interpolating in an attempt to normalise the data. Having reformulated each topic chunk as a HMM we then established the extent to which a given HMM (h) describes a given topic chunk (t) by calculating the probability of t given h ($P(t|h)$). By iterating over each HMM in turn we established which models best described certain sections of the corpus. The most descriptive models are being returned in the results.

3.2.1 Results

At the time of writing this the final version of this paper, we do not yet have a full set of results from the final HMM process to present, owing to the large amount of processing power needed, and the consequent extreme slowness of processing speed. However, we sincerely hope that the results of the final stage of the analysis will have been produced in time for presentation at the workshop. The results will tell us what is the probability of one particular dialogue act being followed by another (according to the entire corpus). More significantly, we hope that they will reveal patterns of discourse structure which show relationships between *what* is being talked and the pragmatic *way* this is being done. Our intention is to use any such patterns that emerge to inform the design of the dialogue action frames which are the topic-individuated base components of the dialogue manager of our dialogue system.

4 Further Discussion

This paper was originally submitted as a position paper. The work described in this the final version of the paper has been carried out between acceptance of the submitted paper, and submission of this the final version. For this reason the pressure of shortage of time has meant that we still do not have the final results we were hoping for by now, and in our methodology we have had to cut some corners that we would otherwise have liked more time to consider carefully, and develop.

We chose to use a small number of dialogue acts, with the intention that if we get some interesting patterns emerging, we will then try making the dialogue acts less coarse-grained, and see whether this affects the results adversely or beneficially.

It is of note that many turns in our corpus are multi-'sentence' turns. Some turns are, in effect, monologues during which the user describes some situation or event at some length. When the corpus was dialogue-act-tagged (DA-tagged), we calculated and assigned only one DA-tag per turn. We did this because we wanted to observe system-user interaction patterns, rather than within-turn patterns (though this does not mean that within-turn patterns are not of interest to us in this research). The way we calculated the single DA-tag for each turn was to treat the whole turn (regardless of how many sentences) as a single sentence. There are clearly more sophisticated ways of assigning a single DA tag to a multi-sentence turn, for example: to DA-tag only the first sentence of the turn; DA-tag only the last sentence of the turn; DA-tag all the sentences of a turn and devise some heuristic for deciding which one is the 'most relevant' one for our purposes. And so on. This matter bears more and careful consideration.

The DA-tagger that we used has not been trained on our own corpus. It had been previously trained on a database of 220,000 utterances of the SWITCHBOARD-DAMSL data set. Since SWITCHBOARD dialogues and the photo-centric Companions dialogues are so functionally different, the performance of the DA-tagger will have been degraded when applied to our corpus. (The applicability of using out-of-domain data to train a DA tagger is explored in detail in [24]). However, the smaller number of dialogue acts that we have used will have mitigated against this in some measure. Of course, in order for the results of this research to be significant and/or useful, we need to have confidence in the performance of the DA-tagger we are using for our corpus. However, we do not yet have accuracy figures for this, although we do intend to obtain them. Training the DA-tagger on our own corpus (which, of course, requires a significant manual annotation effort) is something we have been considering doing anyway in order to improve the performance of the DA-tagger in the implemented Senior Companion dialogue system, and so it is likely to be just a matter of time before this is done.

The final stage of processing (which is taking place as I write) is comparing each topic chunk with every other topic chunk, and is looking for the presence of nouns *in particular places* within each topic chunk (as well as the sequence of dialogue acts within each chunk). If when the results are returned they are poor, meaning that we do not observe any (or we observe very few) functionally similar topic chunks, during the next round of testing we may try looking simply for the *presence* of nouns within a turn, rather than for their positions within a turn. We are also considering again invoking the WordNet semantic distance measure during this final stage, so that we are not limited to looking for functional patterns relying purely on the *repetition* of particular nouns within a topic chunk.

We are considering applying a similar methodology as described in this paper to a very much larger corpus. The *Oral History* movement in the UK has over many years recorded and transcribed two-party conversations during which one person (the 'interviewer') tries to encourage the other (the 'interviewee') to discuss his/her life, using various objects as prompts. The functional similarity between these kinds of conversations and the Senior Companion's photo-centric conversations may well mean that Oral History transcripts can provide us with valuable domain-specific data. We are currently investigating getting permission to use some of this data.

We are aware that in the literature there are a number of different proposed ways of measuring semantic similarity/distance. Our decision to use the WordNet similarity measure [15, 3] was a quick decision based on the reputation of this work, and of WordNet, and on the availability of software to do this work for us. We acknowledge that there may be other measures of semantic similarity that are more suitable for our purposes, and this bears further investigation.

When we were clustering the corpus into topic chunks, we decided in the first instance to look only at the *nouns* in each turn. This seemed sensible as a first step, however, we do not preclude adding other parts of speech into the analyses in future.

We are considering developing probabilistic DAFs (as opposed to our current fully deterministic DAFs) based upon the transition probabilities of the HMMs.

ACKNOWLEDGEMENTS

Our thanks go to Christopher Brewster, WeiWei Cheng, Adam Funk, Ting Liu and Diana Maynard for their practical help or guidance in carrying out the work done in preparation for this paper.

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report, 1998. In *Proc. DARPA Broadcasting News Workshop*, pp. 194–218.

[2] R. Barzilay and M. Elhadad. Using lexical chains for text summarization, 1999. In I. Mani and M. T. Maybury (eds), *Advances in Automatic Text Summarization*, Chapter 10, MIT Press.

[3] A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of lexical semantic relatedness, 2006. *Computational Linguistics* **32**(1):13–47.

[4] J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialogue structure coding scheme, 1997. *Computational Linguistics* **23**(1): 13-31.

[5] F. Choi. Advances in domain independent linear text segmentation, 2000. *Proc. 1st Meeting of North American Chapter of the ACL*, pp. 26–33.

[6] M. Core, M. Ishizaki, J. Moore, C. Nakatani, N. Reithinger, D. Traum, and S. Tutiya. *The report of the third workshop of the discourse resource initiative*, 1999.

[7] M. G. Core and J. Allen. Coding dialogs with the DAMSL annotation scheme, 1997. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, MA.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications, 2002. In *Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*.

[9] M. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages, 1997. *Computational Linguistics* **23**(1): 33–64.

[10] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. Switchboard discourse language modeling project final report., 1998. *Research Note 30*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore.

[11] O. Lemon, A. Bracey, A. Gruenstein, and S. Peters. The WITAS multi-modal dialogue system, 2003. In Proc. *Eurospeech 2003*.

[12] S. C. Levinson. *Pragmatics*, 1983. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

[13] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text, 1991. *Computational Linguistics* **17**(1): 21–48.

[14] B. Pavel. Survey of clustering data mining techniques, 2002. Technical Report, Accrue Software.

[15] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet: Similarity-measuring the relatedness of concepts, 2004. In *Proc. 19th National Conference On Artificial Intelligence*, San Jose, CA, pp. 1024–1025.

[16] F. Porikli. Clustering variable length sequences by Eigenvector Decomposition using HMM, 2008. In *Lecture Notes in Computer Science* 3138: 352–360.

[17] R. Power. The organisation of purposeful dialogues, 1979. *Linguistics* **17**:107–152.

[18] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation, 1999. In *Proc. 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, Calcutta, India, pp. 137–143.

[19] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation, 1974. *Language* **50**: 696–735.

[20] J. Sinclair and M. Coulthard. *Toward an Analysis of Discourse: the English Used by Teachers and Pupils*, 1975. Oxford, England: Oxford University Press.

[21] P. Smyth. Clustering sequences with Hidden Markov Models, 1997. In M.C. Mozer and M.I. Jordan and T. Petsche (eds) *Advances in Neural Information Processing Systems*, MIT Press.

[22] N. Stokes, J. Carthy, and A. F. Smeaton. Select: A lexical cohesion based news story segmentation system, 2004. *Journal of AI Communications* **17**(1): 3–12.

[23] N. Webb, M. Hepple, and Y. Wilks. Dialogue act classification based on intra-utterance features, 2005. In *Proc. AAAI workshop on spoken language understanding*, Pittsburgh.

[24] N. Webb and T. Liu. Investigating the portability of corpus derived cue phrases for dialogue act classification, 2008. In *Proc. 22nd International Conference on Computational Linguistics (COLING2008)*, Manchester, UK.

[25] Y. Wilks. Artificial companions, 2005. *Interdisciplinary Science Reviews*, June, Vol. 30, pp. 145–152.

[26] Y. Wilks, D. Benyon, C. Brewster, P. Ircing, and O. Mival. Dialogue, speech and images: The Companions Project data set, 2008. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.