Computational approaches to the pragmatics problem

Abstract

Unlike many aspects of human language, pragmatics involves a systematic many-to-many mapping between form and meaning. This renders the computational problems of encoding and decoding meaning especially challenging, both for humans in normal conversation and for artificial dialogue systems that need to understand their users' input. A particularly striking example of this difficulty is the recognition of speech act or dialogue act types. In this review, we discuss why this is a problem, and why its solution is potentially relevant both for our understanding of human interaction and for the implementation of artificial systems. We examine some of the theoretical and practical attempts that have been made to overcome this problem, and consider how the field might develop in the near future.

Introduction

What constitutes human communication? One possible answer is to claim that it requires a sender and a recipient, and that information is encoded by the sender, transmitted, and decoded by the recipient. This concept of communication was famously formalised by Shannon (1948). However, Grice (1957) argued that communication between people was also characterised by the process of intention recognition. Specifically, he identified the notion of "non-natural meaning", in which sense a speaker "means" something if, firstly, they intend to induce a belief in the hearer as a consequence of that utterance, and secondly, they intend for this to happen as a result of the hearer recognising the intention (conveyed by the utterance) to bring about this belief. For instance, a speaker who says "Please sit down" intends for the hearer to sit down, and for this to occur because the hearer recognises that this is what the speaker wants to convey by these words. From this perspective, as Levinson (1983: 15) puts it, "communication involves the notions of intention and agency".

Grice’s view of inter-personal communication has been enormously influential in linguistic pragmatics and related fields. A striking point of contrast with the Shannon model, as Grice himself immediately noted (1957: 387), is that the intentional view of communication admits the possibility of indeterminacy. On the Gricean view, it is possible for the same signal to correspond to different intentions, in which case it is necessary to appeal to context in order to understand what the speaker actually intends on this particular occasion. Shannon, conversely, adopts a model in which encoding and decoding of a signal are one-to-one mapping processes, and in which context and the mental state of the sender are irrelevant to the recipient’s understanding of the message.

It seems undeniable that human communication does indeed have the systematic ambiguity that Grice posits, whether this is a consequence of the polysemy of words or the multi-functional nature of various actions: Grice’s own examples are the word ‘pump’ and the action of putting one’s hand in a pocket. So clearly some elaboration of the Shannon model is called for. And intuitively, it seems credible that the goal of the hearer is to understand the intention of the speaker, as Grice argues. However, given that many different intentions may be realised by the same signal, the task of recovering the speaker’s intention given a signal is logically intractable (Levinson 1995: 231) – there is not enough information in the signal to tell the hearer, precisely and unambiguously, what the intention was. In order for the Gricean, intentional analysis of communication to be tenable, we therefore need to be able to explain how hearers are so often successful in solving this ‘pragmatics problem’, and understanding what intention underlies the speaker’s choice of utterance. Given the ramifications of this model for our understanding of human interaction, foundational questions about the validity of the model are of substantial theoretical importance.

In this paper, we focus on a particular subcase of the pragmatics problem that has attracted widespread interest from philosophers of language and builders of computational systems alike: namely, the way in which we identify dialogue act types. The following section discusses why this is an important issue for both human-human interactions and for artificial spoken dialogue systems. We then outline some of the most productive linguistic and computational attempts to address this issue. We conclude by considering how these methods might usefully be synthesised into a coherent interdisciplinary approach to dialogue act type recognition.

Dialogue act recognition in interaction

As pointed out by Austin (1962), our use of language does not just consist of asserting propositions. More broadly, we perform “speech acts”. That is to say, we “do things with words” – we use utterances to achieve particular effects. We may request an action, acknowledge a request, ask for information, and so on. From this perspective, we can see language as a tool that we can use in order to accomplish things that we would not be able to accomplish by other forms of physical action. We can also analyse individual instances of language use as social actions that are performed in order to elicit specific responses, which might involve obtaining information or causing interlocutors to act upon the physical world in particular ways.

The usefulness of linguistic acts in enabling specific social accomplishments cannot easily be treated in terms of truth conditions: it doesn’t generally make sense to describe a request as “true” or “false”, for instance. Austin introduced the notion of “illocutionary act” to describe this kind of function, a notion which was later elaborated by Searle (1975). Although this research tradition is referred to as speech act theory, here we will use the term “dialogue act” rather than “speech act” to emphasise that the relevant actions may be achieved by other means than through speech (for instance, gesture, eye-gaze, and so on). There is little consensus as to what constitutes an appropriate typology of dialogue acts, but we might distinguish dialogue act types by appeal to a notion like “what kind of response is appropriate”.

In order for the speaker’s dialogue act to be effective, it is generally necessary (under the Gricean assumptions discussed above) for the hearer correctly to identify it, as without doing so, it is impossible for the hearer to respond in such a way as to satisfy the speaker’s goals. However, as has long been observed, this is not a straightforward matter. Consider for example the potential dialogue act of ‘asking a question’. Nearly all human languages possess the interrogative sentence-type, which is usually distinguished from the declarative by some complex of morphosyntactic and intonational factors. It is tempting to assume that the task of recognising the dialogue act ‘asking a question’ is reducible to that of recognising an interrogative sentence. But this is simply not true: a formally declarative sentence may perform a questioning function (“You’ll let me know”), and a formally interrogative sentence may function as a request (“Could you close the window?”) Indeed, interrogative forms can easily be ambiguous between various dialogue act types depending on context (“Can you come?” could be a question, a request or an invitation). Moreover, the notion of ‘asking a question’ might not even constitute a single coherent dialogue act type: it might include such distinct dialogue acts as ‘asking a polar question’, ‘asking a wh- question’, ‘asking a check question’, and so on. If these need to be distinguished, that clearly cannot rely on appeal to the sentence-type alone, which is typically the same (interrogative) in all cases.

The recognition of dialogue act types can thus be seen as a specific case of intention recognition, and one that succumbs to the pragmatics problem: given that several different intentions may be expressed by the same form, how can the hearer locate the right one? And just as we ask this question for human interactors, so we can ask it for artificial systems, and in particular spoken dialogue systems – that is, systems that are designed to converse with humans. To get computers to understand one another, we can program them to communicate unambiguously: but the ultimate goal for a spoken dialogue system is to be able to accommodate all the ambiguity and uncertainty of normal human discourse. (In practice, humans tend to adjust their choice of words to match the abilities of artificial systems (see Branigan et al. 2011), but ideally this would not be necessary.) Moreover, the system must understand what the speaker is actually trying to achieve, rather than merely formalising the content of the speaker's utterance in some way. This kind of understanding also proves useful in enabling the system correctly to identify individual words that would otherwise not have been correctly parsed (Stolcke et al. 2000, Taylor et al. 2000). In order to allow systems of this kind to approach human performance levels, it would be helpful to have a fuller and clearer account of how humans actually recognise dialogue act types.

A growing body of evidence underscores the impressive nature of human performance in this particular domain. Our own experience suggests that competent language users are able correctly to identify the intended dialogue act in the vast majority of cases, as shown by the appropriateness of their responses. For instance, a hearer asked "Could you pass the salt?" will usually do so, unless they deliberately choose to misinterpret the speaker's intention and merely say "Yes". In cases such as this, the formal ambiguity of the utterance is not necessarily noticed by the dialogue participants, unless it is pointed out by a response that is inappropriate to the speaker's actual intention.

The success of dialogic communication speaks to the accuracy of the conclusions arrived at by hearers about the speakers' intentions. Experimental work suggests that hearers are not only accurate but also remarkably fast in identifying the speaker's intention in ongoing utterances. Relevant evidence here comes from turn-taking. De Ruiter, Mitterer and Enfield (2006) demonstrated that, in spontaneous Dutch conversation, almost half of the new conversational turns started within 250ms (either way) of the end of the current turn. Stivers et al. (2009) generalised this result to a typologically mixed sample of 10 languages: for each language, the mean duration of the gap between turns was less than half a second, the "fastest" being Japanese with a mean gap of just 7ms. This supports the observation by Levinson (1995: 237) that a half-second delay in responding can (in English) be interpreted as conveying some pragmatic effect (in that case, the impossibility of the hearer responding 'yes' to a question).

Recent work on dialogue act recognition (Gisladottir et al. 2012) demonstrates directly that hearers are able accurately to identify dialogue acts off-line. Hence, given the content of a speaker's turn (and awareness of the contrast), it should not be a problem for the hearer to identify the speaker's dialogue act type. However, it seems profoundly implausible that this could happen in the gaps between turns documented by Stivers et al. (2009). In the first place, many of the languages they test exhibit frequent overlap in turn transitions, which indicates that hearers cannot be waiting for the speaker's turn to be complete before they start planning their own conversational response. In the second place, research on utterance planning (for instance, Brown-Schmidt and Tanenhaus 2006) appears to indicate that even a latency of 500ms would not be enough for the hearer even to formulate a response *ab initio*. Given that the responses are usually faster than this, usually pertinent, and usually conform to the dialogic strictures laid down by the speaker (for instance, a question will be met with an answer), this strongly suggests that the hearer must often be aware of the nature of the speaker's dialogue act before it is complete.

In a similar vein, we might interpret the nature of back-channel responses (Yngve 1970) as evidence that the hearer can identify aspects of the speaker's communicative intention incrementally and on-line. Back-channel responses are utterances by the hearer that are not attempts to initiate a turn. Schegloff (1982) refers to a subset of these as "continuers", on the basis that they serve to assure the speaker of the hearer's attention and indicate that the turn can continue. Various utterances can fulfil this function, among them "uh-huh" and "yeah".

However, it appears likely that the appropriate choice of back-channel response depends to a certain extent upon the dialogue act being performed by the speaker – for instance, “yeah” would not be an appropriate back-channel if the speaker is formulating a request, unless the hearer intends to comply (cf. Schegloff 1993: 107). If this intuition is correct, it further suggests that hearers may be able to access information about the speaker’s dialogue act type from relatively early in the utterance.

In sum, there appears to be quite convincing evidence that human dialogue participants are able to draw rich inferences about dialogue act types from very early on in a dialogue turn. In the following section, we examine some approaches to explaining how this process might take place.

Approaches to dialogue act recognition

A linguistic approach to dialogue act recognition was offered by Gazdar (1981), who formulated the Literal Meaning Hypothesis. According to this account, every utterance possesses some kind of illocutionary force that is built into its surface form. Declaratives are used to make statements, interrogatives to question, imperatives to order or request, and verbs such as “promise”, “deny” and so on (performatives, in Austin’s terms) are used to accomplish whichever function their verb specifies. However, as discussed earlier, utterances are frequently used to accomplish other discourse functions than their surface form would suggest, and the same utterance may be used for multiple functions. So at the very least we need to supplement the Literal Meaning Hypothesis with some mechanism that enables hearers to calculate the alternative non-literal or “indirect” meanings that may arise.

One possibility is to appeal to traditional pragmatic notions of cooperativity and, in particular, relevance. Gordon and Lakoff (1971) suggest that reanalysis occurs when the hearer realises that the surface meaning of the utterance is inappropriate given the context.

For instance, a speaker asking “Could you pass the salt?” typically knows that the hearer is able to do so, and the hearer can infer from this that the purpose of the utterance is not to enquire as to their salt-passing capabilities. For the utterance not to be a waste of effort, therefore, there must be some other purpose to it. Searle (1975) tells a slightly different story: on his account, the ‘natural’ answer to the question “Could you pass the salt?” (namely: yes, the hearer could do so) must be relevant to the speaker. A possible reason for this is that the speaker wants the salt; and the hearer, being cooperative, should therefore pass the salt to the speaker, without an explicit request being necessary.

Can we, however, reconcile this kind of account with the data on turn-taking discussed above? Timing presents a serious problem. Both versions of the pragmatic account take as their starting point the realisation that the literal meaning of the utterance is in some way inadequate given the conversational context, and has to be enriched. However, if the reasoning in the previous section is correct, this process has to begin before the utterance is complete. The problem is, how can the hearer determine that the literal meaning of the utterance is inadequate before knowing what the utterance is? A sentence beginning “Could you…”, or even “Could you pass…”, could certainly be a genuine question that was not a request (“Could you pass for 21?”). More generally, we might observe that almost any sentence beginning “Could you…” might conceivably be used either as a question or as a request, and for many such cases, it is easy to imagine contexts in which either use might be intended (“Could you teach a course in psycholinguistics?”) In order to know that “Could you pass the salt?” cannot (normally) be intended as a question about the hearer’s capabilities, the hearer must identify the meaning of the sentence and realise that the speaker knows the answer to the question that is ostensibly being posed. This is completely reasonable *post hoc*, but as an account of online reasoning it doesn’t appear to give the hearer enough time to formulate their response.

One conceivable way of rescuing this account is to propose that the hearer in fact guesses how the sentence will end, and reasons on the basis of that guess, thus being able to draw the inferences discussed above before the end of the speaker's turn. After all, Sacks, Schegloff and Jefferson (1974) proposed that hearers anticipate the end of speakers' turns in order to achieve smooth transitions; and Magyari and De Ruiter (2012) provide evidence that the accuracy of this anticipation is correlated with the rapidity of turn transition. However, as an account of dialogue act type recognition, this explanation is in danger of becoming circular: a hearer may well guess that the sentence "Could you pass…" concludes with the words "the salt", but this continuation only makes sense if the utterance is a request, whereas by hypothesis the hearer currently takes the utterance to be a question. To put it another way: intuitively, we might expect the words "the salt" because we guess that the speaker wants the salt passed to them. But how did we guess that the speaker wanted something passed to them? Presumably because "Could you pass…" tends to signal that this is the case, notwithstanding that it is formally part of an interrogative sentence-form.

An alternative approach, foreshadowed by Levinson (1983), is to dispense with the Literal Meaning Hypothesis, and instead treat the identification of dialogue act type as a puzzle to be solved by any means available. That is not to propose that the hearer ignores the sentence-type: that might be a valuable clue to the dialogue act type. However, according to Levinson, most speech acts are indirect, in the sense that they do not correspond to the surface form of the sentence. Fortunately, there are many other forms of information that might be helpful to the hearer. Within the speech signal itself, other indications of the likely dialogue act type are present. These include the prosody, as discussed by Bolinger (1964) and extensively explored by Shriberg et al. (1998) among many others. It is also likely that specific lexical choices are strongly associated with particular dialogue acts. For instance, "I want you to…" strongly suggests that the current sentence has the character of a request, even though the sentence-type is purely declarative. Even more generally, the use of “please” seems typically to mark a request whether it is appended to a declarative (“The door should be closed, please”), imperative (“Close the door, please”) or interrogative (“Could you close the door, please?”) sentence-type.

At a higher level, there are considerations deriving from the structure of dialogue, as studied within the research tradition of conversation analysis: for instance, the idea of adjacency pairs (Schegloff and Sacks 1973). If the preceding dialogue turn was a question, the current turn is likely to be an answer, even if its form suggests otherwise. If the previous turn was an offer, the current turn is likely to involve accepting or declining that offer. Thus, when we encounter the first turn of an adjacency pair, we might (with some degree of confidence) expect that the second turn of that pair will follow. Adjacency pairs can also have non-linguistic constituents, as argued by Schegloff (1968). Clark (2004) originates the notion of ‘projective pair’ to cover cases where a non-linguistic communicative act such as a gesture serves to trigger a particular kind of communicative act in response. He later argues (Clark 2012) that we can identify wordless exchanges that are analysable as question-answer sequences. At a still higher level of discourse organisation, an awareness of the overarching purpose of the dialogue and of the participants’ roles in it might help a hearer disambiguate dialogue act types. In a restaurant, for instance, if a customer states the names of dishes, this is likely to be a request; if a waiter does so, it is more likely to be an offer (or effectively a multiple-choice question).

Computational implementations of dialogue act recognition have predominantly adopted this kind of permissive, inclusive approach, in which all available forms of information are used to make the relevant decisions. This cue-based approach essentially dispenses with the assumption of literal meaning elaborated by the kind of stepwise inference discussed earlier, although that approach has also been explored computationally (from Perrault and Allen 1980 to Allen et al. 2007). The role of the cue-based model is simply to identify which dialogue act is instantiated by a given utterance, appealing as necessary to lexical, syntactic, prosodic and conversational-structural factors, among others.

It would perhaps be fair to say that cue-based implementations are primarily focused on improving the performance of systems, rather than necessarily providing insights into the process of dialogue act recognition *per se*. However, the models are linguistically informed, in important respects. They are trained on labelled corpora, from which they can learn the strengths of association between specific signals and specific dialogue acts. The choice of signals may, and typically does, reflect empirically-determined findings as to which aspects of the utterance are likely to constitute informative cues. Identifying potentially useful signals is a non-trivial problem in domains such as prosody, where it is unclear precisely what properties of the acoustic pattern have informational value (see for example Rangarajan Sridhar, Bangalore and Narayanan 2009).

Although traditional linguistics and computational modelling approaches find common cause when it comes to identifying signals, the customary meaning of ‘dialogue act’ varies significantly between the two traditions. As Thomson (2010: 10) puts it, “In the traditional definitions of both speech and dialogue acts, the semantic information is completely separated from the act”. That is to say, the utterance “Could you pass the salt?” is an instance of a dialogue act type like REQUEST rather than one like REQUEST-SALT. From a linguistic point of view, the motivation for this is fairly clear: the notion of dialogue act type captures the idea that there are commonalities between all forms of REQUEST, regardless of what is being requested. However, from a dialogue systems standpoint, this is not necessarily an advantage. If the goal of the system is to fulfil the user’s request, then merely identifying the utterance as ‘some kind of request’ is not helpful: it does not enable the system to formulate a response, as this response will depend upon what is being requested. Unless the system has an abstract understanding of how to fulfil generic requests, the ‘type’ level of dialogue acts is not useful here.

Moreover, by dispensing with the ‘type’ level, it may be possible for a system to identify dialogue acts more efficiently than a human could. Consider the case of a robot receptionist (as implemented, for example, by Paek and Horvitz 2000). Suppose that John Smith is an employee at the company and that the robot is programmed with only one action that relates to John Smith, namely putting a call through to him. Confronted with the input “Could you call John Smith?”, the robot can use the words “John Smith” as a cue to the action it should take, and thus use the name as evidence that it should put a call through. A more capable robot, just like a human, would be disadvantaged here, because if it could take various different actions with respect to John Smith, recognising the name would not suffice to identify which one should be performed. Of course, the simple robot may misidentify dialogue acts that are outside its knowledge base (“My name is John Smith”), but it has no problem using lexical cues to choose among its limited repertoire of abilities.

The question arises of whether the traditional notion of dialogue act type is at all helpful for implementations of spoken dialogue systems. Traum (1999) considers this point, coming to the conclusion that dialogue act types may not be strictly necessary but are potentially useful as an intermediate step in communication planning. The practice of identifying dialogue acts at a finer level of granularity (REQUEST-SALT, CALL-JOHN-SMITH) certainly has implications for the scalability of dialogue systems, as the number of distinct dialogue acts increases drastically as the coverage of the system expands to multiple conversational domains (whereas, by hypothesis, the number of dialogue act types is relatively small even for the whole of human interaction). This becomes especially pertinent when we consider statistically-driven dialogue systems of the kind surveyed by Young et al. (2013). These models use the approach named POMDP (partially observable Markov decision processes) and treat dialogue as a Markov process, in which transitions between dialogue states are modelled probabilistically. Even within a small domain, it is impractical to track dialogue state fully in such a model; for a general spoken dialogue system, the resulting state space would be intractably large (Young et al. 2010: 152).

In particular, a domain-general system that identified highly specific dialogue acts would necessarily have to incorporate thousands of distinct dialogue acts. Consider the receptionist scenario: a person entering the building might request the receptionist to make a call to any individual in the building, using the form of words "Could you call X?" A system that treats every such request completely separately, depending on the identity of X, could not make useful generalisations across this set of requests. For instance, if the name of X is mumbled or unfamiliar, it will not be able to respond "Sorry, who?" unless it identifies the utterance as a request: it could only announce its inability to respond to the request as a whole, which might prompt futile reformulations ("I would like to talk to X"). That is, although such a system might be very efficient at learning the mappings between specific strings and specific tasks, it will struggle to generalise these mappings in any remotely human-like way. Similarly, if it is possible to make generalisations about dialogue act sequences (e.g. question-answer, apology-acceptance, check-confirmation, and so on), these generalisations will not be as evident when the coarse-grained dialogue act types are broken down into fine-grained ones.[i] If each particular kind of apology must be separately associated with a kind of acceptance, a large volume of data may be required for the pattern to be learnt by the system across all pertinent occasions.

However, this observation, like Traum's (1999) discussion, relates primarily to the operation of relatively complex dialogue agents with sophisticated 'mental' states. For simpler systems, dialogue act type recognition in the traditional sense is clearly less useful: in the limiting case, if a system does nothing but (attempt to) satisfy requests, coding a module to identify every input as a REQUEST is clearly not going to add anything to the system's efficacy. What the system needs to do is to identify what is being requested: only then can it initiate the appropriate response behaviour. Unless the system has a generic handling procedure for requests, it cannot benefit from the inclusion of this additional level of analysis. By contrast, systems that actually attempt to emulate human behaviour have the potential to benefit from including a dialogue act level. A recent example of such a system the virtual agent implemented by DeVault, Sagae and Traum (2011), designed to help soldiers practice negotiation skills. The agent uses a natural language understanding module to convert the content of the human user's utterance into a semantic frame representation. One of the attributes within this semantic frame is 'speech act type', so the artificial agent could be said to be calculating and exploiting information about the human speaker's purpose. Moreover, the agent can be configured to guess the content of the semantic frame based on partial utterances, thus effectively engaging in incremental identification of dialogue act type.

The catch, however, is that semantic frames are treated as atomic within DeVault et al.'s model, even though they are decomposable in principle. That is, their model postulates a finite set of semantic frames and aims to identify, based on the user's utterance, which one is currently being instantiated by the speaker. Each semantic frame happens to have an attribute that is called 'speech act type', but this specific attribute is not exploited in any way: responses are selected based upon the entire semantic frame that is identified. There is, in effect, no commonality between semantic frames that contain the same speech act type. The decision to treat semantic frames as atomic reflects a deliberate simplification, justified on the basis that it does not impair performance on the constrained domain in which the model operates. However, for the model to be scalable, some form of non-atomic approach would be necessary, which might involve the exploitation of dialogue act types in a more traditional way.

Towards an interdisciplinary perspective on dialogue act recognition

As the above discussion indicates, insights from theoretical linguistics have already been brought to bear productively upon the implementation of artificial spoken dialogue systems. However, our psycholinguistic questions about the process of dialogue act recognition and behaviours such as turn-taking are not directly addressed by this practical computational work. Most of the computational work has so far taken place in highly constrained domains, while we are interested in the full sweep of human communicative interaction. Moreover, computational approaches have predominantly attempted to achieve effective behaviour by any means necessary, but this may involve means that are not available to, or not exploited by, human interactors. For instance, computational models do not have the working memory limitations of humans, and can in principle use probabilistic cues that are outside of humans' knowledge (for instance, because they involve relations over too long a distance, or patterns that humans are not disposed to spot). They do not have the experiential limitations of humans: they can be trained on larger corpora than a human would ever experience. And they typically do not operate under the same time pressure as humans, assuming that they can initiate responses faster than humans can program their own motor functions.

Nevertheless, the application of these methods already gives us a useful insight into what might work, and which theoretical ideas add value in a practical context. For instance, Young et al. (2010) use a bigram model of dialogue act type, which is informed by the work of Schegloff and Sacks (1973) on adjacency pairs, to help identify the user's response to their artificial agent's questions. DeVault, Sagae and Traum (2011) use a rich array of lexical cues from the input string to support the semantic classification of the user's utterances. As discussed earlier, this latter model can also be made to operate incrementally, while the bigram approach of Young et al. also informs us about the likely nature of the current dialogue act before it is complete. It would seem quite conceivable to take these mechanisms, and others like them, equip them with a notion of dialogue act type, and use them to classify utterances in natural human-human interactions.

Furthermore, if we are interested in learning about how humans treat dialogue acts, we can calibrate such a model against experimentally verified human behaviour. That is, we can eliminate factors that do not appear to influence human performance, just as we can introduce additional factors that are posited to play a role in humans' classification of dialogue act types. And we can similarly adjust the candidate set of dialogue act types, in accordance with competing theoretical proposals. The ultimate goal of such a programme might be to establish a set of dialogue acts that are descriptively adequate as a characterisation of the components of human dialogic interaction, and which are identifiable sufficiently quickly by appeal only to utterance and contextual properties that humans are known to respond to.

Working in the opposite direction, it is also conceivable that a fully adequate theory of dialogue acts could be very useful in the development of domain-general spoken dialogue systems. It is, of course, clear that this is not a substitute for a comprehensive system of semantics – a system that reliably gives 'answers' that don't relate to the question will not survive scrutiny – but it may turn out to be a necessary component if dialogue systems are to behave in a credibly human-like fashion (and thus to allow their human users to behave normally with them). It may also transpire that the use of dialogue acts results in systems being more compact and efficient than would otherwise be the case, just as the analysis of dialogue reveals order in what might otherwise be the limitless variety of human-human interaction.

Endnotes

[1] The potential to draw useful generalisations will depend not only on defining dialogue act types at the right level of granularity, but actually choosing an appropriate set of specific dialogue act types

with which to populate the model. For reasons of space we cannot substantively address this issue here. See Král and Cerisara (2010) for a discussion of some specific candidate 'tag-sets' for dialogue acts.

References

Allen, J. F., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Swift, M., & Taysom, W. (2007). PLOW: A collaborative task learning agent. National Conference on Artificial Intelligence (AAAI), Vancouver, BC.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford: Clarendon Press.

Bolinger, D. L. (1964). Intonation across languages. In J. P. Greenberg, C. A. Ferguson & E. A. Moravcsik (eds.), *Universals of Human Language Phonology, vol. 2*. Stanford: Stanford University Press. 471-524.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. (2011). The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition*, 121: 41-57.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: an investigation of message formulation and utterance planning. *Journal of Memory and Language*, 54: 592-609.

Clark, H. H. (2004). Pragmatics of language performance. In L. R. Horn & G. Ward (eds.), *Handbook of Pragmatics*. Oxford: Blackwell. 365-382.

Clark, H. H. (2012). Wordless questions, wordless answers. In J. P. de Ruiter (ed.), *Questions: Formal, functional and interactional perspectives*. Cambridge: Cambridge University Press. 81-100.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Predicting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language*, 82: 515-535.

DeVault, D., Sagae, K., & Traum, D. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue and Discourse* 2: 143-170.

Gazdar, G. (1981). Speech act assignment. In A. K. Joshi, B. L. Webber & I. A. Sag (eds.), *Elements of Discourse Understanding*. Cambridge: Cambridge University Press. 64-83.

Gisladottir, R. S., Chwilla, D. J., Schriefers, H., & Levinson, S. C. (2012). Speech act recognition in conversation: experimental evidence. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society. 1596-1601.

Gordon, D., & Lakoff, G. (1971). Conversational postulates. *Papers from the Seventh Regional Meeting of the Chicago Linguistic Society*, 63-84.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 67: 377-388.

Král, P., & Cerisara, C. (2010). Dialogue act recognition approaches. *Computing and Informatics*, 29: 227-250.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S. C. (1995). Interactional biases in human thinking. In E. N. Goody (ed.), *Social intelligence and interaction*. Cambridge: Cambridge University Press. 221-260.

Magyari, L., & De Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3: 376.

Paek, T., & Horvitz, E. (2000). Conversation as action under uncertainty. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann. 455-464.

Perrault, C. R., & Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6: 167-182.

Rangarajan Sridhar, V. K., Bangalore, S., & Narayanan, S. (2009). Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech and Language*, 23: 407-422.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50: 696-735.

Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist*, 70: 1075-1095.

Schegloff, E. A. (1982). Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In D. Tannen (ed.), *Georgetown University Roundtable on Languages and Linguistics*. Washington DC: Georgetown University Press. 71-92.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26: 99-128.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica* VIII, 4: 289-327.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (eds.), *Syntax and Semantics, Vol. 3: Speech Acts*. New York: Academic Press. 59-82.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27: 379-423.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41: 439-487.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K. E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106: 10587-10592.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26: 339-373.

Taylor, P., King, S., Isard, S., & Wright, H. (2000). Intonation and dialogue context as constraints for speech recognition. *Language and Speech*, 41: 493-512.

Thomson, B. (2010). Statistical methods for spoken dialogue management. PhD thesis, University of Cambridge.

Traum, D. R. (1999). Speech acts for dialogue agents. In M. Wooldridge & A. Rao (eds.), *Foundations of Rational Agency*. Dordrecht: Kluwer Academic Publishers. 169-201.

Yngve, V. (1970). On getting a word in edgewise. In M. A. Campbell (ed.), *Papers from the Sixth Regional Meeting, Chicago Linguistics Society*. Chicago: University of Chicago Press. 567-578.

Young, S., Gasic, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language,* 24: 150-174.

Young, S., Gasic, M., Thomson, B., & Williams, J. (2013). POMDP-based statistical spoken dialogue systems: a review. To appear in *Proceedings of the IEEE.*