Formulaic Language in Computer-supported Communication: Theory Meets Reality

Alison Wray
Centre for Language and Communication Research, Cardiff University, PO Box 94, Cardiff, CF10 3XB, UK

A recent model of language processing in normal native speakers (Wray, 2002a) proposes that speakers reap substantial benefits from storing and retrieving prefabricated utterances from memory, rather than always constructing novel ones on line. Substantial evidence from a wide range of linguistic research is consistent with the model, but independent tests of its viability would be difficult to conduct. A small number of 'natural experiments', however, provide an opportunity to gain insights into some of the parameters of the model. One such 'natural' experiment is TALK, a system developed to promote conversational fluency in non-speaking individuals. The design and efficacy of TALK are explored and evaluated as a potential working model of Wray's depiction of normal language processing. TALK is demonstrated to be a valuable tool for the pursuit of language awareness, both because it demands of its users a highly developed sensitivity about how conversation works, and because it provides researchers with glimpses of a phenomenon that is normally inaccessible – language processing in action.

Introduction

All our beliefs about language are mediated by some sort of model, explicit or implicit. Models commit us to a view about such things as the relative importance of the components and/or functions of language, how it is organised and/or created, and what the surface manifestations of the underlying mental processes signify. For our insights about language to be valid, we need to acknowledge the parameters of the models we employ, and subject them to rigorous scrutiny and testing.

This paper concerns an attempt to externally validate a psycholinguistic model of language processing. Language processing is highly complex and theoretical descriptions of it necessarily impose substantial simplification in order to tease out the feature or features central to their particular focus. A standard device for achieving this simplification is to present the account in terms of an abstract model, usually metaphorical in nature. Box models, flow diagrams, and even Connectionist networks, are all metaphorical, insofar as no direct association is made between units of processing in the model and real anatomical areas of the brain, beyond the general level.

It is a necessary corollary of psycholinguistic modelling that it creates fictions. These fictions need accurately to simulate the patterns in observed linguistic data. However, it is never going to be easy to move beyond the stage of saying 'yes, that fits what we see' to saying 'yes, that is what is really going on'. Good psycholinguistic models are an effective device for explaining, exploring and

0965-8416/02/02 0114-18 $20.00/0

predicting linguistic behaviour, but there is only limited scope for evaluating them independently.

Three testing options exist. The first is to develop computer programs that confront basic linguistic units (as defined by the model) with the dynamics of the model itself, to see if the output is consistent with (a) the model's predictions and (b) real language. This is a procedure followed in, for instance, Connectionism (e.g. McLeod *et al.*, 1998). Although an immensely powerful tool, an inevitable limitation of this approach is that the computer programs normally have to be developed as a bespoke device for testing the model, and so they are likely to dovetail with parameters of the model in ways that may obscure certain kinds of design limitations and/or assumptions.

The second option is to apply the model to the human language machine, that is, test it on more and more linguistic material. While this is in one respect the most direct kind of test, and certainly models are often severely strained by new kinds of evidence, a good model of language processing will have been designed to be robust to the surface manifestations of processing, and the part that really needs testing is the underlying dynamic. This dynamic cannot be investigated using the human brain in its normal state, because processing is too fast and too effective to give up its secrets easily. A certain amount may be gleaned from observing abnormal language – from tip of the tongue to aphasia – but it remains immensely difficult to get very far below the surface.

The third opportunity for testing a model is to look for 'natural' experiments, that is, situations in the real world in which the model is tested incidentally. The situation need not be 'natural' in the sense of created by nature, but if it is artificial, it will have been created for some purpose unrelated to the testing of the model. Such experiments may be rare for any given model, but they can offer considerable advantages if recognised and exploited. Firstly, the parameters of the situation are independent of the parameters of the model, offering the potential for a degree of independent external validation for the model. Secondly, if the situation created in the 'natural' experiment is motivated by genuine human communication, it forges a valuable link between the simplicity of the model, the simplicity of the natural experiment *per se* (part of what makes it a suitable forum for testing), and the real life complexity of the underlying focus of investigation: language in use. Thirdly, by having its own impetus, a 'natural' experiment can 'fight back' so to speak, making demands upon a model that purports to formalise its previously *ad hoc* or other-oriented logistical design. Such confrontation turns the tables on the model, challenging it either to account for the 'natural' experiment more fully, or to justify why certain characteristics of the experiment fall beyond its scope. Finally, the independent and continuing existence of the 'natural' experiment as a language phenomenon in its own right offers the potential for long-term structured symbiosis. In the present case, as will become clear, continued improvements in the design of the language-support software under examination could both influence, and be influenced by, our developing models of human language processing. Thus, there is the potential for lasting benefits in relation to, on the one hand, our general and particular understanding of complex psychological processes, and, on the other, future enhancements in the practical support available for the communicatively disadvantaged.

In this paper one such natural experiment is used to test a recent model of language processing. In the next section, the model itself is described, and a set of questions arising from it is identified. In the third section, TALK – an augmentative communication system – is introduced and accounted for as an incidental representation of the processing model. In the fourth section, TALK and the processing model are compared as a means of ascertaining whether the validity of the model extends beyond its own immediate domain. The final section briefly draws out the contribution made by this work to the TALK user's and the researcher's respective awareness of language.

Formulaic Language: A Model

In recent years, Wray (1998, 1999, 2000a,b, 2002a,b; Wray & Perkins, 2000) has explored the nature of formulaicity in language from a number of angles, with the aim of capturing its fundamental essence within a model of language processing. The model in its fullest form (Wray, 2002a) is wideranging. It offers not only an explanation for repetitive linguistic behaviour in normal and abnormal speech, but also suggests how such behaviour may enter the speaker's repertoire – both during first-language acquisition and subsequently; why it tends to be avoided – with some disadvantageous consequences – during post-childhood second-language learning; and how the lexicon may be organised to accommodate it. Wray (1998, 2000a, 2002b) extends the discussion to a possible role for formulaicity in the evolutionary origins of grammatical language. There is clearly no space here to describe, nor even justify fully, the various aspects of the account, and the interested reader is referred to the works listed for more detail. In this context, it will suffice to identify certain key features that are pertinent to the task in hand.

The nature of formulaic language

Without question, there are some strings of words that are processed holistically. How else could we explain the use of phrases like *laissez-faire*, *au fait*, *sine qua non* and *et cetera* by people who know no French and Latin? Evidently, such users memorise the phonological and graphemic forms of these phrases along with a holistic meaning, without engaging with their internal lexical or grammatical composition. Something similar must occur with idioms that are semantically opaque, such as 'kick+TENSE the bucket', 'tick+TENSE off' (in the sense of 'castigate'), 'once in a blue moon', 'cold turkey', and so on. They must be associated with a holistic meaning and thus require no construction or internal analysis beyond, as appropriate, local morphological amendments. So much is relatively uncontroversial.

Opinion divides at the point where strings constructed according to the regular rules of the language, using recognised lexical items in their customary interpretation, are also assessed for their potential to be formulaic. In such cases, the association of a holistic meaning with the form would have to be one of two ways of handling the string (see 'Dual systems processing', below), and any approach to language modelling that values parsimony is likely to reject such a possibility, since it both increases many fold the number of entries in the lexicon and also duplicates a great deal of core lexical material. There are, nevertheless, several significant explanatory advantages to preferring a model of language processing in which regular word strings can be handled in two ways: stored and retrieved as a single large unit, and also constructed from, and reduced to, their smallest components by rule.

Storing regular material holistically: Explanatory advantages

Holistic meaning

Many regular word strings have an additional level of meaning – pragmatic or situational – that resides in the string as a whole. An expression such as 'it's a small world' has a meaning derivable from its components. However, its characteristic use in particular situations, such as when A and B discover that they have a mutual acquaintance, or meet by chance in an unexpected location, is a feature of the expression as a whole, not its parts. The strings 'nice to meet you' and 'nice meeting you' are, componentially speaking, equivalent in meaning. Both are used on the occasion of a first meeting but, for many native speakers, while the former can act as both an opener and a closer, the latter is only a closer (Schmidt, 1983: 152). Analysing the strings offers no explanation of their difference, which resides in the associations made with each expression as a complete entity, within a particular speech community.

Holistic meanings, then, belong to speech communities in a way that componential meanings do not. Outsiders, whether non-native speakers or just speakers of different varieties, are in danger of understanding all of the words of an expression (and how they are grammatically associated) without accessing the most relevant part of the meaning. The role of the speech community is not, Wray (2002a) argues, purely circumstantial. Rather, interaction with in-group members transmits and forges formulaic language and maintains its formulaic status, by favouring speakers with certain advantages when their output matches what is prefabricated in the hearer's lexicon (see 'Dual systems processing' and The hearer's processing effort' below; Wray, 2002a, Chapters 4 and 5).

Accounting for patterns of usage

The proposal that regular word strings can be stored and retrieved whole enables us to explain the contrast between what *could* be said, and what *is* said. Put succinctly, our grammar and lexicon afford many different ways of expressing the same idea, but we have a very strong tendency to prefer a subset of these ways. As Pawley and Syder (1983) observe, 'native speakers do *not* exercise the creative potential of syntactic rules to anything like their full extent, and ... indeed, if they did so they would not be accepted as exhibiting nativelike control of the language' (p. 193). For 'only a small proportion of the total set of grammatical sentences are nativelike' and the remainder 'are judged to be "unidiomatic", "odd", or "foreignisms"' (Pawley & Syder, 1983: 193).

Pawley and Syder go on to raise the question of why non-native speakers might fail to pin down the set of native-like strings from the larger set of grammatical sentences, a puzzle that Wray (2000b, 2002a) addresses within the terms of her model. Her proposal is that any speech community, or complex of speech communities, establishes a set of idiomatic ways of expressing ideas, by favouring, purely through repeated use, certain complete phrases and a great many partly filled phrase-frames. As they are encountered by community members they are handled with the *least* processing necessary for comprehension within context. In other words, a new expression will be accepted with a holistic meaning, and stored as a single item *unless there is a good reason to break it down* – the principle of *needs only analysis* (Wray, 2002a: 130ff; Wray & Perkins, 2000). Second- and foreign-language learners insufficiently exposed to an appropriate native speech community, or restricted in their range of interactions within it, will fail to encounter the full range of native-like expressions, and will be obliged to create their own versions. Leaving aside the potential effects of interlanguage, the product, based on rules and atomic lexical units, is as likely to be a grammatical but unidiomatic string as a grammatical *and* idiomatic one – indeed, probably more likely, since the unidiomatic ones outnumber the idiomatic ones, and, for independent reasons (briefly outlined below) idiomatic strings may tend to drift away from full grammaticality and semantic transparency. In addition, Wray proposes that second language learners resist needs only analysis and so are likely to analyse and then discard the word strings they encounter, so that they fail to build up a native-like store of formulaic material (Wray, 2000b, 2002a: Chapter 11).

Irregularity, residual language and intuition

Several other advantages to viewing regular linguistic material as potentially formulaic are explored in Wray (2002a). Three will be briefly mentioned here. Firstly, it offers a way of explaining the tenacity of irregularity in languages. If we only create and understand utterances by applying rules to words and morphemes, it is difficult to see why irregularity should be tolerated, let alone why an item or construction should progress from regular, to marked, to antiquated, to a fossilised historical relic. Needs only analysis permits a polymorphemic word or a word string to remain unanalysed if its meaning and usage are clear, and it is thus protected from the normal policing of the regular system.

Secondly, damage to certain areas of the adult brain can result in disruption to, or complete loss of, the ability to construct and/or understand grammatical sentences. Yet it has long been recognised that individuals with this acquired disability remain able to produce and understand some quite complex grammatical strings. Wray's proposal is that these strings are being drawn from the lexicon in prefabricated form, so no grammatical processing is taking place. However, to accommodate the particular patterns of loss and retention observed in acquired aphasia, the lexicon needs to be modelled in a particular way (Wray, 2002a: Chapters 13 and 14).

Thirdly, the storage of complex units in the lexicon offers an explanation for the non-alignment between a native speaker's intuition about the meaning and usage of words, and the patterns of their actual manifestation in large corpora. Wray suggests that holistically stored units are not easily subjected to the inspection that underpins an intuitive judgement. In short, if you want to know what a word means, it is easier to rely on its entry as a separate unit than to look for it as one component of larger strings. In this way, our intuitive judgement of the word 'large' ignores its occurrence in 'at large' and 'by and large', in exactly the same way as our understanding of the word 'pet' quite reasonably ignores its apparent occurrence within 'carpet'.

Dual systems processing

As already mentioned, a construction that is grammatically regular and semantically transparent must be capable of generation using the grammatical rules, even if it is also stored holistically. This means that there is a choice of processing routes. In line with others, including Sinclair (1991), Wray proposes that formulaic processing is the default, and that construction out of, and reduction into, smaller units by rule occurs only as necessary. In actual fact, because a great deal of the formulaic material is only partially specified – it contains gaps for the insertion of closed class items (e.g. bound morphemes) or open class items (e.g. nouns) – the two strategies work hand in hand. The key feature of the combined system, however, is that language input is assumed to be familiar and holistically processable until found otherwise, and language output takes the processing route of least resistance compatible with the interactional goal.

Like others before her, Wray identifies direct processing benefits for the speaker in using larger rather than smaller chunks when creating output. In addition, however, she shows how the speaker derives other kinds of benefits from producing material that is holistic for the hearer. A substantial subset of formulaic language is manipulative in nature (e.g. greetings, commands, requests, warnings, bargains, hedges), and she argues that the less processing entailed in the decoding process, the more likely the hearer is to react appropriately to the message. There are two reasons. First, there is less danger of misunderstanding a familiar form, even if it is subject to distortion or disruption. Second, the hearer associates the particular expression with a pre-agreed meaning, the situational and pragmatic elements of which may not be derivable easily from the composition. In short, by delivering an expression that is associated, within the speech community, with a holistic layer of pragmatic meaning, the speaker reduces the danger of the hearer failing to construe the implicatures accurately. A significant corollary is that it is in the speaker's interests to anticipate what formulaic expressions perform the desired function for the hearer. This will encourage individuals to accommodate their speech patterns to those of the groups that they prioritise for interaction, thus generally promoting cohesion in the linguistic behaviour of speech communities.

Issues arising

How plausible are the dual systems processing scenario and its accoutrements? A great many questions may legitimately be posed regarding the assumptions and priorities within the model. Not all will be addressed here. However, we shall explore later four issues:

- Are the human brain and the priorities of human communication well suited to dual systems processing?
- At what points will fully analytic processing be obliged to cut in, and what features of prefabricated material will tend to increase and decrease the necessity for this?
- What role can formulaic language play in making conversation operate successfully?
- How robust is the account of how formulaic language contributes to language processing?

First, however, the following section describes the testing ground for the model.

Computer-supported Conversation

Augmentative communication (AC) systems are a means by which individuals with physical difficulties that prevent them from articulating speech clearly are furnished with a synthetic voice. All systems entail the entering of text into a computer, and, as a result, have to overcome the compound effects of three practical difficulties. The first is that many target users (e.g. those with cerebral palsy) lack precise motor control. This makes the entering and selection of text a slow and cumbersome process, whether it be done via a keyboard, a mouse or a pointer. The second is the inherent slowness of typing, which simply takes longer to execute than speech. A third difficulty arises specifically with some aphasic users of AC systems (e.g. Waller *et al.*, 1998), whose language processing may be damaged in such a way that they are unable to construct the text that they want to express, even though they can understand it once constructed.

Any one of these difficulties will render AC output slow relative to normal speech – 2–15 words per minute, versus up to 180 (Todman *et al.*, 1999a: 325) – and combinations of them seriously undermine progress beyond the basic conveyance of simple information. Although communication may be achieved, it lacks the spontaneity and detail that are characteristic of real conversation, and can fail to deliver a sense of enjoyment or satisfaction (Todman & Lewins, 1996: 285). One solution to these problems is to build in mechanisms for word recognition and predictive selection. These anticipate likely completions for words and phrases that have been started, and can be sensitive to the user's personal lexicon and style. However, this approach only chips at the edges of the problems inherent in producing original text on line.

The design of TALK

A different solution to the problem of production speed and accuracy is to by-pass the lion's share of on-line processing demands by using a database of pre-constructed utterances, entered at leisure in advance of the interactional event(s) in which they are used. In the case of TALKsBAC, an AC system for people with non-fluent aphasia, these utterances are entered by a carer (Waller *et al.*, 1998). In TALK, designed primarily for people with cerebral palsy and motor neurone disease (e.g. Todman & Lewins, 1996; Todman *et al.*, 1994a,b, 1999a,b), the user enters his or her own material. In both cases, the skill lies in anticipating accurately the utterances that will be needed on a future occasion.

TALK was developed at Dundee University during the 1990s, with the aim of enhancing the speed of communication sufficiently to transform sterile information exchange into genuine conversational interaction. Using TALK, production rates increase to an average of 60 words per minute (Todman *et al.*, 1999a: 325) and even a cursory examination of the data from its trials indicates that greater speeds could be achieved with a faster computer processor and differently designed peripherals such as the touch pad. The increased speed of output is possible because the TALK-user is primarily engaged, during the conversation, not in constructing utterances but in retrieving previously anticipated and stored ones. These utterances can be accessed quickly and easily using icons on the screen. They are spoken fluently by a synthesised voice.

At first glance, only being able to say things that you or someone else thought of on some previous occasion appears highly restrictive, and quite unlikely to compare favourably with normal conversation. How could interaction conducted by drawing on pre-empted halves of a conversation ever be more than a very blunt tool for genuine information exchange or social engagement? Given these natural reservations about the capacity of this type of AC system to deliver a useful outcome, it is something of a curiosity that both TALKsBAC and TALK are evidently successful as conversation tools (Todman *et al.*, 1999b; Waller *et al.*, 1998).

The successful operation of a system like TALK relies on the ease with which the next utterance can be located and selected: the longer it takes, the more detrimental to conversational fluency and to both participants' enjoyment (Todman *et al.*, 1999b: 154). Therefore, it is of great importance that the material is stored in a way that is both memorable and logical. TALK utterances are stored according to three semantic principles. The first principle operates on the intersection of three dimensions of perspective: person (me or you); orientation (where, what, how, when, who or why) and time (past, present or future) (Todman *et al.*, 1999a: 324). Examples, from Sylvia's computer,[1] are in Table 1.

Table 1 Examples of utterances stored and accessed by perspective combination

Perspective combination	*Example utterances*
you, why, future	'why shouldn't you?'; 'why will you do that?'
me, how, present	'just now I use a number of different ways to communicate'; 'I've been helping John with the TALK system for about four years now'
me, what, present	'I generally keep myself busy most of the time'; 'Eastenders is too depressing for me to watch'; 'I like all kinds of pizza, but I think my favourite kind is Hawaiian'

The second accessing mechanism is topic/function-based. There is, for instance, a set of options for greeting, another for finishing a conversation, and a large set containing stories. The story screens often contain extended texts, but Sylvia, at least, tends to split a story across a series of entries, so that there is flexibility about how much is said and when. Some of the effects of this, both deliberate and unintended, are considered later. Table 2 exemplifies the topic/function-based storage.

The third type of storage relates to short interjections. The categories include a set of expressions for 'don't know' (*dunno*), apologising (*sorry*), hedging (*hedge*), interrupting (*intrup*), requesting time to reply (*wait*), indicating that an incorrect selection has been made (*oops*), agreeing (*agree*), and giving minimal responses (*uhhuh*). Table 3 illustrates the types. In order to simulate the style of natural conversation, the computer randomly selects one of the utterances stored for that function (Todman *et al.*, 1999a: 325), though some sub-types also offer a controlled option (Sylvia Grant, personal communication).

As the examples for 'wait' in Table 3 indicate, the TALK-user has two further

Table 2 Examples of utterances stored and accessed by topic/function

Topic	*Example utterances*
Greet	'Hello, how are you?'; Hi there, it's a long time since I last saw you'; 'What's your name?'; 'Hello, I'm Sylvia'
Finish	'Thank you for taking the time to talk to me, I enjoyed our chat very much'; 'Bye-bye for now. Hope to see you again some time'; 'Good night'; 'Cheerio. See you'
Stories	'I went to Vancouver two years ago'. 'It was for a conference for people who use and work with speech aids'. 'And John and I did a presentation at the conference'. 'It went really well' …

Table 3 Examples of utterances stored and accessed by interactional function

Function	**Example utterances**
dunno	'I'm afraid I don't know'; 'I haven't got a clue'
sorry	'Sorry, go ahead. I interrupted you'; 'Sorry, I didn't quite catch that'; 'Sorry but I'm having a bit of trouble here'
hedge	'That's a good question'; 'Well, I suppose you could say that'; 'I'll have to think about that'
intrup	'Could I say something there?'; 'Excuse me, may I interrupt there'; 'I'd like to butt in here'
oops	'I'm sorry about that, it was a mistake'; 'Sorry for repeating myself there, but my hand slipped'; 'Sorry, that wasn't what I meant to say there'
wait	'I haven't got an answer in the computer to that just now, so would you please wait until I type out an answer'; 'Please wait a second while I find the next thing I want to say'; 'I'll need to edit something here, so please hang on a sec'
agree	'Yes, I suppose so'; 'Yes, that's very true'
uhhuh[2]	'Right'; 'Uh-huh'; 'yeah yeah'

facilities, aimed at honing the appropriacy of utterances to their context. One is the on-line composition of a completely new utterance. The other is the on-line editing of an existing utterance. Both of these require the entry of individual words, letter by letter. On-line creation is not only time consuming but subject to error. The user may mistype words, so that they are incomprehensible when synthesised, or inadvertently erase a prepared string instead of forwarding it to the synthesiser. The stakes are high, since the purpose of using TALK is to enable conversation, and if the pauses between utterances are too long the momentum of the exchange is lost.[3]

Talk as a Model of Real Language Processing

Superficially it is clear that the configuration of TALK resembles in certain regards Wray's model of language processing. The resemblance is unplanned and serendipitous. How far do the similarities extend beneath the surface, and what, if anything, do they signify? As a means of judging this, we shall focus on the questions raised under 'Issues arising', above.

Two systems of processing

The first question posed was 'Are the human brain and the priorities of human communication well suited to dual systems processing?'

The ease with which TALK-users and their speaking partners adapt to the TALK system, and the capacity of TALK to facilitate prolonged exchanges with the fundamental characteristics of conversation, suggest that dual systems processing is compatible with the way we think and communicate.

In both Wray's model and TALK, one system is low in processing requirements but offers only restricted forms, while the other offers infinite flexibility but makes on-line processing demands that are prohibitive. The two systems can operate independently or can interact, when the analytic system is applied to the holistic for the purpose of on-line editing. The real-time processing limitations on the on-line construction of novel configurations so constrain the speaker/user that holistic processing is established as the default. In addition, the driving principle in both cases is successful communication, and the retention of fluency is a high priority. Todman *et al.* (1994a,b) report that TALK-users preferred, when they had no suitable response stored, to keep the conversation going with a filler and/or to instigate a topic shift, rather than stop to generate a more accurate response letter by letter. TALK conversations suggest that the user will even compromise on truth and/or grammatical accuracy (relative to the form of the previous utterance) in order to take the turn and sustain fluency. For instance, the following exchange occurs in a conversation about TALK, transcribed in Grant (1995):

Question: Do you have any special ways of coping with emotions. Say you're really annoyed with someone. Does that handle it at all?
Sylvia: I haven't thought about that much.

Post hoc interpretations are, of course, unwise. However, Sylvia's response is a typical 'hedge' in the TALK database. Furthermore, it is unlikely that Sylvia had never considered the relative difficulty of expressing her current emotional state through pre-stored utterances. A reasonable interpretation of this exchange, therefore, is that she had no easy way to answer the question truthfully, so, rather than disrupt the fluency of the conversation by constructing a reply, she dodged the question by telling a white lie. This subjugation of accuracy to fluency is reminiscent of Tannen's (1984) observation that the drive for rapport and fluency in normal conversation can lead to irrelevance (p.95) including, in one example, the confirmatory repetition of what another speaker has said, even when it was known to be incorrect (p. 76).

The relative involvement of analytic and holistic processing

The second question posed under 'Issues arising' was 'At what points will fully analytic processing be obliged to cut in, and what features of prefabricated material will tend to increase and decrease the necessity for this?'

Storing the right things

The level of intervention of on-line analytic processing may be taken as an indicator of several independent but potentially co-occurring variables. One is the match between what is stored and what is needed. In the case of normal language, this will reflect such things as the novelty of the situation and the speaker's previous level of exposure to relevant linguistic input. For TALK, it is more a matter of the user's ability accurately to translate passively observed linguistic material to the dynamic medium, and to anticipate accurately what it is most valuable to store. It is clear from comparing TALK conversation transcripts with the contents of Sylvia's personal computer database, that certain word-strings are used many times, while others may never be used at all. More precise details of just what, in terms of form and meaning, undermines the usability of a string that has, presumably, only been stored in TALK because it was anticipated to be useful, will only become obvious when current detailed analyses of the data are completed.

Levels of specificity

A second variable operating on the level of activation of the marked, analytic, processing option, is appropriacy to the local context. In normal speech this largely relates to pragmatic appropriacy, where a prefabricated form may be deemed likely to be unacceptable in a given, probably unaccustomed, context, even though it is semantically appropriate. For example, an individual whose friend or acquaintance is diagnosed with a life-threatening illness may struggle for the right thing to say, not trusting that any of the consolatory expressions holistically stored are quite appropriate. The TALK-user will be subject to the same pragmatic limitations on the appropriacy of pre-stored items. But in addition, formal limitations may be observed.

These are limitations that could exist in the normal context too, but which would be difficult to spot there, since editing, the manifestation of analytic processing in this situation, is accomplished in the normal speaker adeptly and with little if any external evidence of its occurrence. For a TALK-user, it could be that a particular string has been stored, and is in every way appropriate to the current requirements except for one detail. Take, for instance, one of the strings in Table 2, 'And John and I did a presentation at the conference'. In the context of the story, the word 'and' contributes to textual coherence. But it compromises the range of use for the string as an independent item. Another example is the pre-stored utterance 'Did you have a nice weekend, Sian?'. Sylvia created this utterance for use in a series of 19 conversations with Sian (reported in Todman *et al.*, 1999a,b), in which context it may be considered useful. However, the tagging of the addressee's name clearly invalidated it for use with any other addressee. The string could be edited on-line, but the cost incurred, measured in delay to the output, would be considerable.

Wray proposes that, in normal language, slots arise where there is paradigmatic variation in a formulaic string. These slots require the insertion of a variable item, something that entails analytic processing, but only at a minimal level. Matching this procedure in TALK would, for the present example, entail a two-tier structure, in which the name, as a single stored item, was embedded into the slot in 'Did you have a nice weekend, __ [name]?' This is somewhat different from a sequential selection of 'Did you have a nice weekend' and 'Sian', which TALK would produce unintegrated, each encased in its own tone unit. A better solution would have been to store the string without the name tag. In other words, parsimony, in aiming to maximise the applicability of a stored utterance across situations and addressees, will tend to impose constraints on the form of stored strings. In TALK the best solution is fully complete strings devoid of compromising specificity. In normal language the possibility of underspecification arises, in the form of slots.

Operational stress

A third determiner of analytic involvement is the level of operator stress and the effect that this has on the ability to access prefabricated material. In normal language processing, certain kinds of stress can affect the ability of the speaker to produce fluent, idiomatic output and to recall memorised material. Wray (1992, 2002a) goes so far as to suggest that the heightened level of self-consciousness induced by the experimental conditions in much laboratory-based linguistic research could force processing into analytic mode, removing the possibility of observing the holistic processing operational in unobserved conditions.

In TALK, however, things are rather different. Although there is no clear evidence of such stress in the data from the various TALK trials examined so far, it may reasonably be inferred that, since access to any given prefabricated string relies on, first, working out and/or remembering where it has been lodged and, second, using physical movements to select the necessary sequence of screen icons in the correct order, a high level of stress could inhibit efficiency and make access slower. Yet, in contrast to the normal speaker, it would not necessarily help the stressed TALK-user to shift to the more cumbersome analytic method, whereby there was specific control, letter by letter and word by word, of the output, because every additional action carries a risk of error. On the other hand, if an error were made during on-line construction, the hearer might at least be able to repair it, whereas the wrong selection of a holistic string results in entirely the wrong message, with little hope of successful hearer-initiated repair.

To summarise, TALK convincingly illustrates some aspects of the fundamental conflict between holistic and analytic processing as described by Wray, though the resolution of the conflict is not always the same.

Fluency

The third question posed earlier was 'What role can formulaic language play in making conversation operate successfully?'

In TALK, fluency plays a key role in the satisfaction level of conversations (Todman *et al.*, 1999b). The production constraints are such that success can, in effect, be measured in seconds of speech, and failure in seconds of silence.[4] In what follows, therefore, we shall focus on the pursuit of fluency.

Planning

To protect fluency, a measure of forward planning is required. Wray (2002a: Chapter 5) proposes that, in normal conversation, the optimum level of fluency is achieved when the progression of the talk is mapped out in advance. The processing demands of proposition formulation are best offset by the concurrent delivery of previously selected formulaic material, which further serves to protect the turn during the development of the next idea. In TALK, Sylvia, as a successful, highly skilled user, can be observed to make her progression through icons and screens to the next desired utterance, while the computer is still producing the present one, or while the speaking partner has the turn. Furthermore, in a parallel to the conjectured semantic-associative accessing routes in the brain, TALK supports fluent progression between associated ideas by enabling strings that are likely to be needed in sequence to be accessed from adjacent locations.

Turn, topic and power

There are two ways in which the TALK-user can increase the quantity of speech-filled seconds during a conversation. One is by occupying them, whether using a sequence of short strings, each separately selected, or fewer very long strings. The other is to provide opportunities for the speaking partner to contribute extended turns. The simplest way to get the other party to talk is to instigate a series of one-off open questions, each requiring an extended answer.

More skilful is the ability to produce follow-on questions. Normally this entails picking up on something that the other party has just said and developing it. For the TALK speaker, techniques must be found to compensate for the absence of such information when the utterances are planned. One is to store generic follow-on prompts (e.g. Really, why's that?; Do you know when?). Another is to store a set of related questions that build up a picture. For instance, Sylvia leads her conversational partners through a succession of questions about their favourite television programmes, films, film stars, live shows and live performers. A third technique is to develop a theme aired on a previous occasion, since this occupies more familiar territory than a virgin subject.

In normal conversation it is judged a desirable conversational art to draw out the other party into extended talk, while one listens. For the therapist the listening may be 'active', though for others, such as the dutiful visitor of an elderly relative, the elicitation technique may be a device for operating with a minimum level of committed engagement. In either case, as also for Sylvia, this elicitation behaviour institutes a particular, and peculiar, kind of 'power' relationship: control of the conversation is maintained by relinquishing the turn, with no guarantee that it will necessarily be easy to regain. In some conversations, Sylvia's open questions result in excessively long monologues, which Sylvia appears powerless to terminate politely.

It seems, then, that TALK supports Wray's proposal that formulaic language can be used dynamically to determine the shape of conversation, particularly in relation to fluency.

Straining the analogy

The final question challenged the validity of the TALK analogy by asking, 'How robust is the account of how formulaic language contributes to language processing?'

TALK was not designed as a means of testing any psycholinguistic model of language, and we should not be surprised to find that the analogy between Wray's account and the design of TALK breaks down in some regards. However, non-alignments do not necessarily indicate a poor correspondence at the fundamental level, because the differences between the communicative contexts may bring about different responses to procedural demands.

The provenance of formulaic material

In Wray's model, *needs only analysis* accounts for how most formulaic material enters the native lexicon. The remainder comes about through 'fusion' – the 'glueing together' of a novel sequence of items into a useful whole. In contrast, TALK is fully based on the principle of fusion, and there is no operational equivalent of needs only analysis. Is this difference fundamental or secondary? It comes about because the TALK-user's capabilities in language assembly and literacy are harnessed in the design of the system. Insofar as TALK is imposed onto these pre-existing abilities, it shares characteristics with the system of the adult second-language learner, where, Wray (2000b, 2002a: Chapters 8–10) suggests, the role of fusion is also much greater. Needs only analysis predominates in the native speaker because language acquisition begins before linguistic, or general, analytic skills are established. But would TALK-users display the characteristics of needs only analysis if their analytic skills were not needed, or not available? The answer appears to be yes. In TALK, a small quantity of expressions are stored on the system before the user adopts it. Sylvia states that she has kept some of these expressions because they are useful (Grant, 1995: 24). Obviously, these strings were entered componentially in the first place, but for the user, they begin whole, in the same way as *au fait* and *laissez-faire* begin whole for the non-French speaker, even though they do have a compositional history. And in TALKsBAC, the user is unable to enter the material at all, and so all of it is holistic at first encounter. Just as with needs only analysis, it is possible to open up a string and edit it, but the processing costs reduce this to last resort, so if the string carries the desired message it will probably be left as it is. In a parallel to the irregularity that thereby survives in normal formulaic language, the TALK- or TALKsBAC-user may have to tolerate features in a string that they would never have generated themselves, such as vocabulary or constructions that are not part of their normal productive inventory.

Minimal responses

As noted above, TALK is equipped with a set of minimal responses, such as 'uh-huh', 'right!' and 'yeah yeah'. Minimal responses are used by the hearer to offer non-invasive feedback to the speaker. However, at least two practical difficulties have meant that minimal responses are considerably under-used in TALK relative to normal conversation.

Firstly, the physical movement entailed in selecting an icon is very visible to the speaking partner, and the manifestation of the synthesised voice, whatever it is saying, clearly marks a speech event on the part of the TALK-user. As a result, the speaking partner tends to treat all TALK production as invasive (Sylvia Grant, personal communication). This is exacerbated by the very fact that minimal responses coincide with the speaking partner's turn, rather than waiting for the end of it. Furthermore, there is a delay between selection and production of any TALK item, meaning that the token cannot be accurately timed, increasing further the invasiveness.

Secondly, even the simple movement of the cursor to a box at the edge of the screen is preventing the TALK-user from doing something else. As we have seen, fluency relies heavily on forward planning, which includes physically locating the next utterance ready for its selection. It seems likely that in normal speech we engage in a large measure of parallel multi-tasking, so that producing a minimal response would not affect the planning of the next major utterance. For the

TALK-user, mental planning of what to say next can, of course, operate in parallel with cursor movement, but the location and selection of items for production in TALK is serial.

What do these difficulties with minimal responses tell us? Should we, for instance, infer that minimal responses are not formulaic? This seems excessive, since the problems seem more attributable to the limitations of the computer hardware than the nature or function of the utterances. However, it would be reasonable to propose that the difficulty is seen here because minimal responses are a particular *type* of formulaic language, as Wray (2002a: Chapters 13 and 14) proposes for normal language, assigning them to a sub-lexicon operated by subcortical reflex. Support for this view comes from the way that Sylvia, at least, resolves the operational difficulties with minimal responses by shifting into another holistic reflex: 'I prefer to nod my head to let them know that I'm listening' (personal communication).

The hearer's processing effort

As mentioned under 'Dual systems processing', Wray proposes that there are certain tangible benefits for the speaker in providing linguistic forms that are easy for the *hearer* to decode. She views this as a lynchpin in reconciling the key features of language as a psychological phenomenon on the one hand and as a sociological phenomenon on the other. Speakers are driven, by the pursuit of self-interest, to approximate their speech patterns to those of the individuals with whom they interact. By speaking in a way that is familiar to the hearer, and thus easily decoded, a speaker increases the chances of successfully influencing that hearer towards particular behaviours or perceptions (Wray, 1998, 2000a, 2002a: Chapter 5; Wray & Perkins, 2000). This is one of two driving forces behind formulaicity. The other is the reduction of the speaker's own processing.

TALK aims only to reduce the speaker's on-line processing. There is no design equivalent of the quest to reduce the hearer's processing, particularly not as a means of serving the speaker's interests. Should this be interpreted as indicative that supporting the hearer's processing is not, after all, a driving force in the selection of formulaic language? There are good reasons for not taking this line. They reside in an examination of the underlying motivations of the TALK-speaker.

In Wray's account, all speaker behaviour can be accounted for as a projection of self-interest. In normal interaction, the physical, emotional and perceptual manipulation of the hearer offer a means of achieving goals beyond one's direct control, and the adoption of language that is formulaic for the hearer is a means to that end. The TALK-user is also pursuing a personal agenda, but the nature of the communication tool, and its context of use, carve out different priorities. These priorities do not necessarily involve moulding the output to the hearer's formulaic inventory. For the TALK-user, TALK is the only means available for holding extended conversations, though it is not the only way of getting information from another person, nor, indeed, of presenting information to one. The letter-by-letter method may be painfully slow, but it can be used to express any desired message. And e-mail is an excellent medium, in which the benefits of prefabrication and of on-line flexibility are combined, and the output of a non-speaking individual is indistinguishable from that of a speaking person. The *particular* value of TALK, then, is not as a means of communicating messages – though it clearly is that too – but as a means of communicating *fluently and quickly* face to face. It follows – and we saw as much earlier – that the TALK-user is at least as interested, if not more so, in adopting strategies that support fluency, as in trying to persuade the conversational partner to think, feel, or do any particular thing. The two are not mutually exclusive, but the ordering of the priorities is paramount for understanding the strategy choices. The preferential support of the conversation as an *activity*, as opposed to the *content* of the conversation, is further intensified by the research context. The TALK-user and partner have not selected each other as friends, but have been introduced in order to have a conversation, and it is their work to sustain it. So, how does this affect the behaviour of each party in the observed conversations?

We have already seen that the TALK-user has strategies that promote fluency, from planning turns, to sequencing questions and answers, to offering extended turns to the hearer. Earlier, the take-up of these turns by the speaking partner was flagged as a potential problem, since there are instances of monologues up to 100 lines long. However, this behaviour on the part of the speaking partner is consistent with TALK-user's priorities, since fluency is more important than who says what. The elicitation behaviour that leads to these extended turns is, as we saw, the wielding of a kind of power, and as such it is a manipulation of the hearer (i.e. the speaking partner). But since the hearer's manipulability is not dependent on his or her processing ease (though see below), the TALK-user does not have to use language that is formulaic for the hearer in order to increase its success.

In one respect, however, the hearer's manipulability *is* dependent on processing. Here, we do see strategies for alleviating the hearer's difficulties, though, again, formulaic language is not the appropriate tool for the job. The major priority for both partners is to ensure that they are not responsible for a breakdown in the conversational flow. One potential danger with TALK is that the hearer will fail to understand the synthesised voice. A strategy of hearers is to cut off their own speech as soon as the synthesised voice begins. We saw earlier than this exacerbates the intrusiveness of minimal responses. It also reduces substantially the instances of simultaneous talk, making these conversations less like natural ones. But, overall, it sustains the flow of the conversation. A strategy of the TALK-user is to head off known pronunciation difficulties. For instance, certain words might be avoided altogether, if they are known always to be incomprehensible, though this kind of negative evidence is difficult to track from the data. Meanwhile, proper names, especially, are spelled in a pseudo-phonetic way that will render the intended pronunciation, e.g. Harare entered as 'Har-ar-ay', Edinburgh as 'Ed-in-burra', Glasgow as 'Glass-go'. In addition, hyphens are added to force the desired stress pattern, e.g. 'Bul-away-o', 'Em-ma'. These deliberate attempts by the TALK user to mould the output to a closer approximation of what the hearer can easily decode are, ultimately, carried out in pursuit of the TALK user's own interests.

Contributing to Language Awareness

It has been proposed in this paper that TALK provides a testing ground for Wray's (1998, 1999, 2000a,b, 2002a) model of normal language processing, in which analytic and holistic operations work together to maximise fluency and flexibility. A key similarity is the establishment of operational patterns driven by communicative expediency. Where TALK and Wray's model fail to correspond, it is reasonable to point to different communicative agendas, supporting the suggestion that prefabricated language offers a particular kind of linguistic solution to a much wider, non-linguistic problem, namely, the promotion of the speaker's priorities.

The paper exemplifies language awareness at two levels. Firstly, the TALK-user develops, and demonstrates, conscious insights about the mechanics of language in use. They include knowledge of how particular forms of words support fluency, while others disrupt it; how a long narrative can best be divided into smaller units to provide greater flexibility without losing textual coherence; the role and manifestation of minimal responses; and the range of behaviours that different interlocutors engage in and how they are best managed, if satisfactory conversation is to be achieved. Such insights are usually taken for granted.

Secondly, TALK in action offers the observer an opportunity to increase his or her own awareness of how language works. TALK slows down production operations to observable speeds, and provides peepholes into the 'black box' of normal language processing. Certainly, those peepholes offer only particular and restricted views, but even a side-on glimpse can contribute something to our understanding.

The linguist's quest for ever greater awareness about language does not usually align with an increase in language awareness on the part of the *user*. Indeed, much of the linguist's interest lies in precisely those operations that are least open to conscious scrutiny. It is for this reason that TALK is a particularly valuable investigative tool. It raises awareness in both parties, without turning the speaker's task into a metalinguistic one. That is, the TALK-user is still fundamentally engaged in *talk*. Yet the heightened understanding required to use TALK successfully provides tangible evidence of processes that are not normally observable.

Acknowledgements

This paper is based on research supported by a grant from the Nuffield Foundation Social Science Small Grants Scheme. The author is indebted to John Todman and his colleagues for providing access to the tapes and transcriptions of the studies, to Janet Cowper for transcribing additional material, and Sylvia Grant for her personal insights into TALK.

Correspondence

Any correspondence should be directed to Alison Wray, Senior Research Fellow, Centre for Language and Communication Research, Cardiff University, PO Box 94, Cardiff, CF10 3XB, UK (wraya@cf.ac.uk).

Notes

1. The examples here, and elsewhere in the paper, are drawn from a database assembled by Todman and his team during eight separate simulation and fully operational studies of TALK. The material consists of tapes and transcriptions of conversations, and a set of printouts from the personal database of one TALK-user, Sylvia Grant, the most

researched, and most experienced user of TALK. For insights into her own experience see Grant (1995) and Todman *et al*. (1997).

2. Examples for the 'Uhhuh' category are taken from Todman *et al*. (1999a: 325).
3. In addition, some speaking partners find it extremely difficult to resist intervening during periods when the TALK-user is creating a novel utterance. If they change the subject, or even just the orientation of the question that triggered the need for the novel string, they may invalidate it as a suitable response just as it reaches completion.
4. This is not to deny the potential for silence to play a positive role in conversation. Absence of speech in TALK conversations, however, has different implications, and should be separately assessed.

References

Grant, S. (1995) Using TALK. *Communication Matters* 9 (3), 22–25.

McLeod, P.M., Plunkett, K. and Rolls, E.T. (1998) *Introduction to Connectionist Modelling of Cognitive Processes.* Oxford: Oxford University Press.

Pawley, A. and Syder, F.H. (1983) Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards and R.W. Schmidt (eds) *Language and Communication* (pp. 191–226). New York: Longman.

Schmidt, R.W. (1983) Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. In N. Wolfson and E. Judd (eds) *Sociolinguistics and Language Acquisition* (pp. 137–174). Rowley, MA: Newbury House.

Sinclair, J. McH. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Tannen, D. (1984) *Conversational Style: Analysing Talk Among Friends.* Norwood, NJ: Ablex.

Todman, J., Alm, N. and Elder, L. (1994a) Computer-aided conversation: A prototype system for nonspeaking people with physical disabilities. *Applied Psycholinguistics* 15, 45–73.

Todman, J., Elder, L., Alm, N. and File, P. (1994b) Sequential dependencies in computer-aided conversation. *Journal of Pragmatics* 21, 141–169.

Todman, J., File, P. and Grant, S. (1997) 'TALK' in different contexts. *Communication Matters* 11 (1), 17–19.

Todman, J. and Lewins, E. (1996) Conversation rate of a non-vocal person with motor neurone disease using the 'TALK' system. *International Journal of Rehabilitation Research* 19, 285–287.

Todman, J., Rankin, D. and File, P. (1999a) The use of stored text in computer-aided conversation: A single-case experiment. *Journal of Language and Social Psychology* 18 (3), 320–342.

Todman, J., Rankin, D. and File, P. (1999b) Enjoyment and perceived competence in computer-aided conversations with new and familiar partners. *International Journal of Rehabilitation Research* 22, 153–154.

Waller, A., Dennis, F., Brodie, J. and Cairns, A.Y. (1998) Evaluating the use of TalksBac, a predictive communication device for non-fluent adults with aphasia. *International Journal of Language and Communication Disorders* 33 (1), 45–70.

Wray, A. (1992) *The Focusing Hypothesis: The Theory of Left Hemisphere Lateralized Language Re-examined*. Amsterdam: John Benjamins.

Wray, A. (1998) Protolanguage as a holistic system for social interaction. *Language & Communication* 18, 47–67.

Wray, A. (1999) Formulaic language in learners and native speakers. *Language Teaching* 32 (1), 213–231.

Wray, A. (2000a) Holistic utterances in protolanguage: The link from primates to humans. In C. Knight, M. Studdert-Kennedy and J. Hurford (eds) *The Evolutionary Emergence of Language* (pp. 285–302). Stanford, CA: Cambridge University Press.

Wray, A. (2000b) Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics* 21 (4), 463–489.

Wray, A. (2002a) *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2002b) Dual processing in protolanguage: performance without competence. In A. Wray (ed.) *The Transition to Language* (pp. 113–137). Oxford: Oxford University Press.

Wray, A. and Perkins, M.R. (2000) The functions of formulaic language: An integrated model. *Language & Communication* 20 (1), 1–28.