The Seduced Speaker: Modeling of Cognitive Control

Ardi Roelofs

Max Planck Institute for Psycholinguistics, F.C. Donders Centre for Cognitive Neuroimaging, and Nijmegen Institute for Cognition and Information, Wundtlaan 1, 6525 XD, Nijmegen, the Netherlands
ardi@mpi.nl
http://www.mpi.nl/world/persons/profession/ardi.html

Abstract. Although humans are the ultimate "natural language generators", the area of psycholinguistic modeling has been somewhat underrepresented in recent approaches to Natural Language Generation in computer science. To draw attention to the area and illustrate its potential relevance to Natural Language Generation, I provide an overview of recent work on psycholinguistic modeling of language production together with some key empirical findings, state-of-the-art experimental techniques, and their historical roots. The techniques include analyses of speech-error corpora, chronometric analyses, eyetracking, and neuroimaging. The overview is built around the issue of cognitive control in natural language generation, concentrating on the production of single words, which is an essential ingredient of the generation of larger utterances. Most of the work exploited the fact that human speakers are good but not perfect at resisting temptation, which has provided some critical clues about the nature of the underlying system.

1 Introduction

Unlike most Natural Language Generation programs that run on serial, digital computers, human speakers are occasionally distracted while performing a natural language generation task. "I can resist everything except temptation" (p. 5), a play character of Oscar Wilde [1] once said, and this difficulty in resisting temptation holds for most people. Distractibility seems to be the price paid for the parallelism of the human brain. One of the key tasks of the human cognitive system is to select one appropriate action at any given moment and to focus the machinery of planning and movement on that action. Selectivity of attention is required for the coherent control of action. At the same time, the system needs to remain open to events that may happen outside the focus of attention (e.g., to detect possible danger in the background). The opposing forces of the need to focus and the need to remain open make the human system distractible. This raises the issue of cognitive control.

In speaking, the distractibility of the human cognitive system is revealed by speech errors and delays in initiating articulation. The distractibility is also evident from the eye movements that speakers make. By examining speech errors,

A. Belz et al. (Eds.): INLG 2004, LNAI 3123, pp. 1–10, 2004.

delays, and eye movements, researchers have discovered much about the cognitive foundations of speaking. Computer models have been developed that account for the kinds of speech errors that occur and their relative frequencies, and also for the eye movements and the exact duration of the delays caused by distraction. Furthermore, much has been discovered about the brain areas that are involved in speaking. Computer models can even predict the time course of the increase in blood flow to certain brain areas required for speech production. I provide an overview of work in psycholinguistics that tried to shed light on the human language generation system using evidence from speech-error corpora, chronometric experiments, eyetracking, and neuroimaging. The overview is built around the issue of cognitive control in natural language generation. It concentrates on the production of single words, which is an essential component of the generation of larger utterances. Nearly all of the work that is reviewed exploited the fact that human speakers are good but not perfect at resisting temptation, which has provided important evidence about the nature of the underlying system.

2 What Speech Errors Say About Speaking

A slip of the tongue or speech error is an unintended, nonhabitual deviation from a speech plan. Meringer and Mayer [2] were among the first to draw attention to speech errors as a data source that might illuminate the mechanisms underlying speech production. In 1895, they published a large corpus of German speech errors along with a theoretical analysis. They made several seminal observations. First, they discovered that slips of the tongue are typically meaning-based or form-based. The substitution of "dog" for "cat" is a meaning-based error and the substitution "cap" for "cat" is a form-based one. The distinction suggests that words are planned at a conceptual level and at a form level. Second, they observed that there is often a form-relation in meaning-based errors (e.g., "calf" for "cat"), suggesting that the planning levels do not operate completely independently, although this is still a hotly debated issue [3]. Third, they observed that contextual errors may be anticipations (e.g., "*l*eading list" for "reading list"), perseverations (e.g., "beef n*ee*dle" for "beef noodle"), exchanges (e.g., "*fl*ow *sn*urries" for "snow flurries"), or blends (e.g., "clear" combining "close" and "near").

Although speech error analyses continued to be carried out during the next half century, there was a real revival of interest in the late 1960s. In 1973, Fromkin [4] edited an influential book on speech errors that included an appendix with part of her own speech error corpus. Another important corpus was collected during the early 1970s at MIT by Garrett and colleagues. Garrett [5] discovered that word exchanges such as the exchange of *roof* and *list* in "we completely forgot to add the *list* to the *roof*" tend to involve elements from different phrases and of the same syntactic category, here noun. By contrast, segment exchanges such as "she is a real *r*ack *p*at" for "pack rat" are likely to involve elements from the same phrase and they do not respect lexical category. Garrett explained this finding by assuming a level of syntactic planning (at which the lexical exchanges

occur) that is different from the level of form planning (at which the segment exchanges occur). Garrett argued that the speech errors also provide support for a distinct morphological planning level.

Some morphemic errors appear to happen at the syntactic level, whereas others arise at the form level. For example, in "how many *pie*s does it take to make an *apple*?", the interacting stems (i.e., *pie* and *apple*) belong to the same syntactic category and come from distinct phrases. Note that the plurality of *apple* is realized on *pie*, which suggests that a number parameter is set. The distributional properties of these morpheme exchanges are similar to those of whole-word exchanges. This suggests that these morpheme errors and whole-word errors occur at the same level of planning. They seem to occur when lexical items in a developing syntactic structure trade places. Similarly, errors such as "I'd *hear* one if I *knew* it" for "I'd *know* one if I *heard* it" suggest that syntactically specified lexical representations may trade places independently of their concrete morphophonological specifications. By contrast, the exchanging morphemes in an error such as "*slice*ly *thinn*ed" for "thinly sliced" belong to different syntactic categories and come from the same phrase, which is also characteristic of segment exchanges. This suggests that this second type of morpheme error and segment errors occur at the same level of planning, namely the level at which morphemes and segments are retrieved and the morphophonological form of the utterance is constructed.

On the basis of his speech error analyses, Garrett [5] proposed an unimplemented model of speech production that distinguished between conceptual, syntactic, morphological, phonological, and phonetic levels of speech planning. Ten years later, Dell [6] developed the first computer model of memory retrieval in sentence production, instantiating several of Garrett's insights. Following a long associationist tradition that began with Aristotle [7], Dell convincingly argued that our word memory is organized as an associative network that is accessed by spreading activation. The network contains nodes for conceptual, syntactic, morphological, phonological, and phonetic information about words. In retrieving information for concepts to be verbally expressed, activation spreads from the corresponding concept nodes to associated nodes in the network. After fixed periods of time, the highest activated lexical, morpheme, phoneme, and phonetic nodes are selected. Dell's associative network model of word memory provided quantitative accounts of the major facts about speech errors: the kinds of errors that occur, their relative frequencies, and the constraints on their form and occurrence. On the account, errors occur when, because of noise in the system or influences outside the intended utterance (distraction), another node in the network than the target one is the most highly activated node and becomes erroneously selected.

3 What Response Times Say About Speaking

The first person to measure (in milliseconds) speech production latencies – the time between stimulus onset and the initiation of a verbal response – was Don-

ders [8]. Until Donders' work in the 1860s, most scientists had assumed that the mental operations involved in responding to a stimulus occur instantaneously. Donders designed a subtraction technique to time the different mental processes that the brain goes through when faced with different tasks. His chronometric work demonstrated a simple principle: The time it takes to perform a task depends on the number and types of mental stages involved. With this observation, he laid the foundation of a research programme that is still extremely productive today: the componential processing analysis of human task performance. At the end of his seminal article on the measurement of mental processing times [8], Donders reports that "distraction during the appearance of the stimulus is always punished with prolongation of the process" (p. 428). This observation is interesting in the light of later research developments exploiting distraction, in particular, the work of Stroop in the 1930s. Surprisingly, it was only in the 1990s that speech error and chronometric analyses became equal partners in the study of speaking. Most of the chronometric work that has been done addressed the production of single words or simple phrases. This seems to be due to the fact that it is awfully difficult to investigate the generation of more complex utterances (sentences and discourse) in controlled experimental settings. Still, the investigation of word production has provided some key insights into the algorithms that underlie human language generation. The first computer model of word production built on chronometric evidence is WEAVER++ [9] [10] [11]. This model recognizes the key insights from the speech error analyses, but it was specifically designed to provide a unifying account of the increasing body of chronometric data.

Like Dell's model, WEAVER++ assumes that word planning involves the retrieval of information from an associative network through spreading activation. In addition, WEAVER++ assumes that the associations are labeled, because a mere associative link between two nodes in a network tells nothing about the relation between the entities represented. For example, the concept RED(X) is strongly associated with both GREEN(X) and FIRE(X), but the relationship between RED(X) and GREEN(X) is very different from the relationship between RED(X) and FIRE(X). The importance of representing the relation between entities symbolically was first recognized by Selz [7] in the early 1900s. Labeled links have become a central part of semantic networks in computer science since the seminal work of Quillian in the late 1960s.

To explain even the simplest forms of language generation, like single word production, it is not enough to assume an associative memory and spreading activation. Natural language is a very flexible tool that can be used to achieve various goals. Around 1900, Watt, Ach, and Külpe of the Würzburg school [7], as well as Selz [7], called attention to the importance of understanding the directedness of action in general and verbal action in particular (the problem of cognitive control is also referred to in the literature as the problem of attentional, executive, or willed control.) They convincingly argued that the various associative models that had been developed during the past two millennia failed to explain the directedness of thought and action. Plato already drew attention

to the problem, which he characterized in *Phaedrus* as the problem of a charioteer attempting to manage a number of horses pulling in different directions. Until recently, this aspect of natural language generation was neglected in psycholinguistic research on word production. The directedness of natural language generation has, in its simplest form, perhaps been most intensively studied by using the "gold standard" of attentional measures, the color-word Stroop task [12] and picture-word analogs of it. The Stroop task is one of the most widely used tasks in academic and applied psychology, reviewed by MacLeod [13]. In the classic, color-word version of the task, speakers name the ink color of color words (one basic task variant) or they read the color words aloud (another basic task variant). In performing the Stroop task, speakers are slower naming the ink colors of incongruent color words (e.g., the word BLUE in red ink) than of a series of Xs. Word reading times are unaffected by incongruent ink colors. The correct naming of the colors of incongruent color words shows that goals keep verbal actions on track in the face of distraction, albeit with a temporal cost.

Issues of cognitive control were already explored in the early days of experimental psychology (between 1870-1920) by Cattell, Donders, James, and Wundt, who saw all his work on response times as studies of volition [14]. However, no progress was made in understanding the mechanisms of control. Associationist and behaviorist theories accounted for action selection by postulating associations between stimuli and responses (e.g., Müller in the early 1900s [7], and later Watson and Skinner). However, if all our actions were determined exclusively by stimulus-response associations, goals could not determine which action to make because the strongest association would automatically determine the response. Watt and Ach of the Würzburg school [7] extended the idea of stimulus-response associations to associations between stimuli and an internally represented task goal ("Aufgabe"), on the one hand, and responses, on the other. Later theoretical developments are descendants of this idea.

On the view that currently dominates the attention and performance literature, which was anticipated by Müller [7] and recently implemented in the GRAIN computer model by Botvinick and colleagues [15], goals associatively bias the activation of one response pathway (e.g., for color naming) rather than another (e.g., for oral reading). On another view, implemented in WEAVER++ [16], attentional control arises from explicit, symbolic reference to goals, accomplished by condition-action rules. WEAVER++'s associative network is accessed by spreading activation while the condition-action rules determine what is done with the activated lexical information depending on the task. When a goal symbol (e.g., indicating to name the color) is placed in working memory, attention is focused on those rules that include the goal among their conditions (e.g., those for color naming rather than reading). Words are planned by incrementally extending verbal goals. Lexical nodes are selected for concept nodes, morpheme nodes for lexical nodes, segment nodes for morpheme nodes, and phonetic syllable program nodes for syllabified segment nodes, whereby the syllabification of segments proceeds incrementally from the beginning of a word to its end.

The idea of incrementality in natural language generation was first proposed by Wundt [14].

WEAVER++'s combination of a spreading activation network with a parallel system of goal-factored condition-action rules yields a simple but powerful and efficient device for selecting one line of action among the available options. The crucial role of spreading activation in the model is to provide a relevance heuristic. Spreading activation serves to solve the "frame problem" that confronts any cognitive system. In making decisions, a cognitive system can, in principle, draw on all the information available, but the amount may be indefinitely large in that everything may potentially be relevant. The frame problem is how to get at the relevant information and when to stop thinking and start acting. Spreading activation is a parallel mechanism for making relevant information available and triggering relevant computations, following the heuristic that information associated with the current information is likely of direct relevance too. Triggering condition-action rules by spreading activation prevents the problem of all rules having to test their conditions at any one moment in time. Only the rules that are linked to a sufficiently activated piece of associative memory become evaluated. For example, in naming a color, no more than a dozen or so condition-action rules test their conditions rather than all rules in a speaking lexicon of some 30,000 words. Moreover, because condition-action rules may be triggered by the activation of elements outside the current focus of attention, the system remains open to what happens in the background.

The idea of goal-referenced control originated with Selz [7] and it flourished in the work of Newell and Simon, Anderson and colleagues [17], and others, on higher-level cognitive processes like problem solving (e.g., playing chess, proving logic theorems, and solving puzzles such as the Tower of Hanoi), where associative models generally failed. However, due to the traditional partitioning of experimental psychology into cognition, perception, and action, with little communication across the boundaries, goal-referenced control models have had little impact in the perception-action literature, because they generally did not aim at fine-grained modeling of the temporal structure of human information processing in the attention and performance tradition. Only recently, goal-referenced control made successful strides into the attention and performance literature. For example, there are now successful models for fine-grained aspects of visual attention, dual-task performance [17], and Stroop [16]. It seems that we are on the verge of a unified account of the control of cognition, perception, and action.

4 What Eye Movements Say About Speaking

In the second half of the nineteenth century, well before the modern era of cognitive and brain sciences, Donders and Wundt studied eye movements and constructed mechanical models for them. Whereas before those days the eyes used to be poetically called a window to the soul, Wundt took gazes to be a window into the operation of the attention system. As Wundt [14] reasoned in his *Outlines of psychology*, visual acuity is best at the point of fixation. Therefore,

to bring aspects of the visual world in the focus of attention, eye fixations are directed to those visual aspects that are of most interest. This makes a shift of gaze an overt sign of the allocation of attention, as later studies confirmed, although at times visual attention and eye movements can be dissociated. According to Wundt [14] "the successive movement of attention over a number of objects is a discontinuous process made up of a number of separate acts of apperception following one another" (p. 212).

Whereas it has long been assumed that we look at aspects of the visual world just as long as is needed to identify them and that response factors play no role, recent research suggested that when we want to verbally describe the visual aspects, the gaze durations depend on the time to plan the corresponding words [18]. In naming objects, Wundt's "successive movement of attention over a number of objects" has been shown to be determined by word planning. For example, when speakers are asked to name two objects in a row, they look longer at first-to-be-named objects with disyllabic than with monosyllabic names even when the object recognition times do not differ [18]. The effect of the number of syllables suggests that the shift of gaze from one object to another is initiated only after the phonological form of the name for the object has been encoded.

Dissociations between vocal response latencies and gaze shifts suggest that the signal to move the eyes is the completion of (a critical part of) planning the phonological form of the vocal response rather than a flag that a signal to begin a vocal response has been sent out to the articulatory system. Response latencies and gaze durations can be dissociated in that gaze durations may reflect the phonological length (e.g., number of syllables) of the utterance even when response latencies do not [19]. Speakers were instructed to describe colored left and right objects (e.g., a big red scooter and a ball) in a simple or in a complex way. They either had to respond with "the scooter and the ball" or "the *big red* scooter and the ball". The gaze durations for the left object (the scooter) were much shorter for the simple utterances than for the complex utterances. However, the vocal response latencies did not differ between the two utterance types. Furthermore, the shift of gaze to the right object was initiated before articulation onset for the simple utterances, but after articulation onset for the complex utterances. This suggests that the shift of gaze, but not the onset of articulation, is triggered by the completion of phonological encoding of the first object name. It seems that the attention required for planning the vocal response prevents the eyes to move before the object name has been planned. Because the planning takes longer with disyllabic than with monosyllabic names, the attention shift, and consequently the gaze shift, occurs later with two syllables than with one syllable.

Recent research from my own laboratory showed that Stroop-like interference is reflected in the gaze durations of speakers during object naming. Speakers were presented with picture-word stimuli. They either named the picture, named the word, or categorized the word, and shifted their gaze to a left- or right-pointing arrow to manually indicate its direction. Eye movements were monitored. Overall, there was a close correspondence between the magnitude of the distractor

effects on the latencies of vocal responding, the gaze shifts, and the manual responding. This further supports the idea that the eyes are only free to move elsewhere when the verbal action goal is achieved.

The evidence from eyetracking suggests that in generating multiple-word utterances, speakers are not operating in a maximally incremental way. For example, in naming two objects in a row, they do not perform a lexical selection for the second object before they have planned the phonological form of the first object name. Moreover, the evidence suggests that a major reason why speakers fixate objects until having planned their names is that word planning requires attention. This conclusion agrees with recent evidence from dual-task performance by Ferreira and Pashler [20], which suggests that individuals cannot select a word for production and select a manual response at exactly the same moment in time. Ferreira and Pashler argued that all selections in word planning require attention. However, that does not need to be the case. If only one task goal (e.g., vocal responding or manual responding) can be in the focus of attention at any one moment in time, goal-referenced control predicts that one cannot perform selections for two tasks concurrently, even when they are automatized. This would explain the available data without assuming that attention is required for all individual selections in word planning. Instead, if selections are made with explicit reference to the task goal, word planning requires attention until (a critical aspect of) the word form is planned, as empirically observed.

5 What Brain Activity Says About Speaking

Currently, Donders' [8] subtraction technique developed for the temporal aspects of mental processes is widely applied to their spatial aspects – their correlates in the human brain. In his seminal article on response times, Donders [8] remarked that "as in all organs, the blood undergoes a change as a consequence of the nourishment of the brain" (p. 412). One "discovers in comparing the incoming and outflowing blood that oxygen has been consumed" (p. 412). This insight, together with the subtractive method designed by Donders, constitutes the basis of the two most widely used modern functional neuroimaging techniques, PET (positron emission tomography) and fMRI (functional magnetic resonance imaging).

Recent neuroimaging and electrophysiological studies have shed light on the neural correlates of speaking. The techniques include PET, fMRI, MEG (magnetoencephalography), and LRP (lateralized readiness potential) analyses. Indefrey and Levelt [21] performed a meta-analysis of 82 neuroimaging studies in the literature, which anatomically localized the word planning system in the brain. As can be expected from the classic neuropsychology literature and most later studies, the system is basically located in the left hemisphere (for most people). Visual and conceptual processing appear to involve the occipital, ventrotemporal, and anterior frontal regions of the brain. The middle part of the left middle temporal gyrus seems to be involved in lexical selection. Next, activation spreads to Wernicke's area, where the phonological code of the word seems to be

retrieved. Activation is then transmitted to Broca's area for post-lexical phonological processing such as syllabification. Finally, phonetic encoding takes place, with possible contributions of the supplementary motor area and the cerebellum, while the sensorimotor areas are involved in articulation.

Recent neuroimaging studies also revealed that an extensive network of brain areas is involved in the attentional control of word planning. For example, color-word Stroop performance engages the anterior cingulate and dorsolateral prefrontal cortices, both subserving attentional control, the left lingual gyrus for color processing, the left extrastriate cortex for visual word-form processing, and the left-perisylvian language areas including the areas of Broca and Wernicke [22]. Whereas evidence suggests that the dorsolateral prefrontal cortex serves to maintain the goals in working memory, no consensus exists as to whether anterior cingulate activation reflects the presence of conflict [15] or goal-referenced control [16] [22]. The latter view is in line with Paus' [23] characterization of the anterior cingulate cortex as the brain area where "motor control, drive and cognition interface" and Simon's [24] characterization of attention as the principal link between cognition and motivation. For action control, it is not enough to have goals in working memory, but one should be motivated to attain them. Extensive projections from the thalamus and brainstem nuclei to the anterior cingulate suggest a role for drive and arousal. Extensive reciprocal connections between the anterior cingulate and dorsolateral prefrontal cortices suggest a role for working memory. The motor areas of the cingulate sulcus densely project to the spinal cord and motor cortex, which suggests a role of the anterior cingulate in motor control. The motor areas seem to contain subregions controlling vocal responses, manual responses, and eye movements [23]. Goal-referenced control was supported by successful WEAVER++ simulations of the hemodynamic response in the anterior cingulate during Stroop task performance [22].

Condition-action rules are sometimes criticized for being not brain-like in their computation. However, it is important to realize that the rules mean nothing more than the operations that they specify. Crucial for the issue of neural plausibility is whether we can exclude that the brain performs such if-then operations, and the criticisms do not bring forward evidence for that. On the contrary, there is increasing evidence that the human brain, in particular prefrontal cortex, supports the use of abstract rules [25].

6 Final Remarks

I provided an overview of recent work on psycholinguistic modeling of language production together with some key empirical findings and state-of-the-art experimental techniques, including analyses of speech-error corpora, chronometric analyses, eyetracking, and neuroimaging. Most of the work examined the production of single words. Sentence and discourse generation has received much less attention. We are still far away from a complete understanding of the ultimate natural language generator.

References

1. Wilde, O.: Lady Windermere's fan. Dover Publications, Mineola (1893/1998)
2. Meringer, R., Mayer, K.: Versprechen und Verlesen. Goschenscher-Verlag, Stuttgart (1895)
3. Roelofs, A.: Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). Psych. Rev. 111 (2004) 561–572
4. Fromkin, V.A. (ed.): Speech errors as linguistic evidence. Mouton, The Hague (1973)
5. Garrett, M.F.: The analysis of sentence production. In: Bower, G.H. (ed.): The psychology of learning and motivation. Academic Press, New York (1975) 133–177
6. Dell, G.S.: A spreading-activation theory of retrieval in sentence production. Psych. Rev. 93 (1986) 283–321
7. Mandler, J.M., Mandler, G. (eds.): Thinking: From association to Gestalt. Wiley, New York (1964)
8. Donders, F.C.: On the speed of mental processes. Acta Psych. 30 (1868/1969) 412–431
9. Roelofs, A.: A spreading-activation theory of lemma retrieval in speaking. Cogn. 42 (1992) 107–142
10. Roelofs, A.: The WEAVER model of word-form encoding in speech production. Cogn. 64 (1997) 249–284
11. Levelt, W.J.M, Roelofs, A., Meyer, A.S.: A theory of lexical access in speech production. Behav. Brain Sci. 22 (1999) 1–38
12. Stroop, J.R.: Studies of interference in serial verbal reactions. J. Exp. Psych. 18 (1935) 643–662
13. MacLeod, C.M.: Half a century of research on the Stroop effect: An integrative review. Psych. Bull. 109 (1991) 163–203
14. Wundt, W.: Outlines of psychology. Engelmann, Leipzig (1897)
15. Botvinick, M.M., Braver, T.S. , Barch, D.M., Carter, C.S., Cohen, J.D.: Conflict monitoring and cognitive control. Psych. Rev. 108 (2001) 624–652
16. Roelofs, A.: Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. Psych. Rev. 110 (2003) 88–125
17. Anderson, J.R., Lebiere, C. (eds.): The atomic components of thought. Erlbaum, London (1998)
18. Meyer, A.S., Roelofs, A., Levelt, W.J.M.: Word length effects in object naming: The role of a response criterion. J. Mem. Lang. 48 (2003) 131–147
19. Levelt, W.J.M., Meyer, A. S.: Word for word: Multiple lexical access in speech production. Europ. J. Cogn. Psych. 12 (2000) 433–452
20. Ferreira, V., Pashler, H.: Central bottleneck influences on the processing stages of word production. J. Exp. Psych.: Learn. Mem. Cogn. 28 (2002) 1187–1199
21. Indefrey, P. Levelt, W.J.M.: The spatial and temporal signatures of word production components. Cogn. 92 (2004) 101–144
22. Roelofs, A., Hagoort, P.: Control of language use: Cognitive modeling of the hemodynamics of Stroop task performance. Cogn. Brain Res. 15 (2002) 85–97
23. Paus, T.: Primate anterior cingulate cortex: Where motor control, drive and cognition interface. Nature Rev. Neurosci. 2 (2001) 417–424
24. Simon, H.A.: Motivational and emotional controls of cognition. Psych. Rev. 74 (1967) 29–39
25. Miller, E.K.: The prefrontal cortex and cognitive control. Nature Rev. Neurosci. 1 (2000) 59–65