Non-chronometric Experiments in Linguistics

Bruce L. Derwing
Department of Linguistics
University of Alberta
Edmonton, Canada

Roberto G. de Almeida
Department of Psychology
Concordia University
Montreal, Canada

Address correspondence to:

Bruce L. Derwing
Department of Linguistics
Assiniboia Hall
University of Alberta
Edmonton, Alberta, Canada T6G 2E7

Phone: 780-492-5698
Fax: 780-492-0806
E-mail: blde@ualberta.ca

Roberto G. de Almeida
Department of Psychology
Concordia University
7141 Sherbrooke Street West
Montreal, QC, Canada H4B 1R6

Phone: 514-848-2424 extension 2232
Fax: 514-848-4545
E-mail: almeida@alcor.concordia.ca

Draft: Oct 1, 2004

[Preliminary draft for D. Eddington (Ed.), *Experimental and quantitative linguistics*; please do not circulate or cite this version without authorization from the authors]

Non-chronometric Experiments in Linguistics[1]

Bruce L. Derwing
Department of Linguistics
University of Alberta

Roberto G. de Almeida
Department of Psychology
Concordia University

1. Why experiments?

A first question that might naturally be asked by the reader is why do we talk of experiments at all in a book and chapter that have "linguistics" in their titles? What, after all, do linguists have to do with experiments? The standard image of a linguist is of someone who spends his research time either in the library, reading about languages, or in the field, studying them directly, or perhaps even someone who theorizes about language from the comfort of an easy chair. Unless he or she is a phonetician, however, we tend not to picture a linguist as someone who spends much time running experiments in a laboratory.

This picture, of course, accurately reflects the history of the discipline, which arose out of philology and whose earliest practitioners were scholars well versed in many languages, either through literature and other written documents or through direct exposure, or both. And by and large this fascination with the forms and varieties of languages around the globe continues unabated today. To analyze, describe, and compare languages is as common and natural to a professional linguist as it is rare and unmotivated for most other people. As a consequence of this unique focus of attention, linguists quite naturally came to be known as the group who best knew what languages were like and were credited, by default, with the best skill at describing them accurately.

Even after the advent of experimental psychology in the late 19th century (an event most often associated with the name of Wölfgang Wundt), psychologists--who, after all, had the whole range of the human psyche as their research domain--took relatively little interest in language per se, especially during the behaviorist era in North America[2], which saw the rise of North American structuralism in linguistics, as well. It is only in the past few decades, with the development of the subfield which has come to be known as psycholinguistics (which, despite the name[3], has developed more as a branch of cognitive psychology than as a branch of linguistics), that psychology has seriously embraced questions related to the mental representation and processing of language.

It was quite natural at the outset of this new psychological endeavor, however, that psychologists' ideas about the actual "structure" of language should have been heavily influenced,

---

[1] This work was supported in part by grants from the Social Sciences and Humanities Research Council of Canada and from the Fonds québecois de la recherche sur la société et la culture.

[2] Thus Boring's 1950 *History of experimental psychology* fails to mention the term "language" in its index, and although Kling & Riggs 1971 mention it once, that reference turns out to apply to a section on "speech".

[3] Bloomfield (1933) used the more appropriate term "linguistic psychology" (p. 32) to characterize that specialty which studies language knowledge and processes, but, unfortunately, this term never caught on.

at least at first, by the views of linguists, who were at the time the only ones who had seriously looked at it. For a time, therefore, even some psychologists bought into the idea that it was the linguist's job to determine what linguistic structure was like, while the psychologist (or psycholinguist) studied how that structure was learned and used (cf. Hörmann 1971, p. 31).

An important question that is often lost in this implied collaboration, however, is the question of just what linguistic structure *is*. The typical linguist (who, as often as not, is not concerned with psychological questions at all) has tended to view a language as an abstract object (something akin to what the 19th century Swiss linguist Ferdinand de Saussure called a "sociological reality", incomplete in any one speaker and with only some parts shared by speakers of the same language community). Quite naturally, therefore, the structure of a language thus became associated by many investigators with the structure of that abstract object.

Unfortunately, however, there is no place "out there" where one might go in order to find "language" (or even "a language") under such an abstractionist conception. One may speak metaphorically about describing "the English language" as though one were talking about describing a real object or entity, but no such entity can be found that will serve as the standard against which linguistic theories can be empirically tested, unless one looks inside of the minds of its users.[4] Autonomous linguistics (i.e., a linguistics that is uninformed by psychology and psychological experimentation) can describe languages in a (perhaps endless) variety of ways, but decisions of the "best" way must always either be endlessly deferred or else resolved in terms of criteria that are, in the last analysis, arbitrary.[5]

The rise of generative grammar in the late 1950s and early 1960s did not change this, as Derwing (1973) argued in some detail.[6] New evaluation criteria were offered (such as rule naturalness, economy of lexical representation, and rule/system generality), and some new sources of data were also introduced and widely exploited (notably speakers' intuitive judgments about sentence grammaticality or well-formedness), but the former were not motivated by any known psychological facts or well-established principles, while the latter were seldom collected in a systematic way and, in any event, proved to be quite insufficient to resolve the wide range of descriptive alternatives that could still be seriously entertained to account for them. And, most importantly, language continued to be treated by most linguistic theorists as a "thing apart" (e.g., as "a set of sentences", as in Chomsky 1957, p. 13; 1965, p. 51; etc.), whose structure could somehow be described and assessed without any necessary reference to the speakers who produced it and the hearers who comprehended it.

But if linguistic structure is not "out there", then just where is it, and what kind of reality *can* be used to test theories of such structure against? Halle (1964, pp. 324-5) tried to ground it in the physical stream of speech, as when he characterized the phonemic segment as a discrete component

---

[4] In a paper entitled "Where's English?", Ross (1979) answers the question this way: "Clearly (I would imagine), English does not exist independently of those who speak it" (p. 135).

[5] Cf. Halle (1964, p. 329): "For the linguist it suffices if the attributes selected yield reasonable, elegant, and insightful descriptions of all relevant linguistic data."

[6] Cf. Hymes & Fought (1975): "[I]f the fundamental premise of structuralism is seen as the study of language as an autonomous system…, then the ways in which Chomsky's work continues preceding structuralism and completes it seem more decisive than the ways in which it does not (p. 922).

of the speech signal.[7] It is quite clear in retrospect, however, that neither the phoneme nor any other aspect of linguistic structure is intrinsic to utterances, either, but is rather something that is *attributed* to them by speakers and hearers. Thus a linguistic unit (at any level) "exists as a unit only because the language user treats it as a unit… [and] linguistic structure is not something that is 'built up' out of utterances or 'overlaid' upon [them], but is… something which receives its only possible empirical realization as part of the language process of speech production and comprehension." (Derwing 1973, p. 305).[8] In short, the only conception of linguistic structure that makes any scientific sense is one that is intrinsically psychological in its own right.

In this perspective, then, the question of "Why experiments?" is much easier to answer. For if linguistic structure is inherently psychological, then psychological reality is the *sine qua non* of any linguistic theory worth its salt. At the same time, considering the many arbitrary decisions that inevitably go into theory construction by linguists, it also seems obvious that psychological reality is not something that can be simply taken for granted or established by fiat,[9] nor does it seem likely that linguists have the skill to divine psychological reality from the regularities they observe in utterance forms alone. This by no means implies, of course, that linguistic theory is unimportant or unnecessary. In fact, it is a virtual contradiction in terms to imagine an experimental linguistics that was not driven by theories derived from the careful and detailed examination of linguistic data, and all of the experiments to be discussed in this chapter have followed from one or another construct or claim that arose initially from the theoretical linguistic literature. To be scientifically credible, however, linguistics must be more than just theory and rather than attempting to "go it alone" must subject its constructs to rigorous testing against the relevant knowledge and skills of real language users.

Furthermore, by implication, if linguistic structure is psychological structure, then the same kinds of methods that are used to answer psychological questions in general are also the appropriate ones for testing claims about linguistic structure, as well. In the sections that follow, therefore, a variety of specific experimental techniques will be described in some detail, with examples also provided to illustrate how these techniques have already been used to assess theories of linguistic structure and, in particular, questions about phonological and morphological units, which by this time have been quite extensively investigated from a psychological point of view**.** At the same time, references will also be given to work that has been done on other aspects of linguistic structure, including syntactic and semantic representations, using some of these same methods.

2. Chronometric vs non-chronometric methods

One convenient way to distinguish the experimental methods used in testing theories and models about language structure and processing is in terms of the response measures that they employ, with time being a particularly prominent factor, especially in the currently very lively research areas of the mental lexicon and sentence parsing, where chronometric methods clearly predominate (see Chapter N of this volume). The present chapter, however, deals only with non-chronometric methods, in which reaction or response times are not collected and analyzed as data.

---

[7] Cf. also Bloomfield (1933, p. 32), who also identified the "speech signal" as the domain of linguistic analysis and inquiry.

[8] On the status of linguistic structure within the autonomous view, see also the exchange between Kac (1982) and Prideaux (1980).

[9] Contrast Valian (1976, p. 66), where generative grammars are described as psychologically real "by definition".

Such methods have quite a number of practical advantages, often including a relatively simple and straightforward approach and freedom from the need to employ the kind of elaborate laboratory equipment that is often required to collect and score temporal responses (such as response boxes, eye-trackers, and ERP devices). This often makes it possible to increase efficiency by testing large numbers of subjects together in groups, rather than individually, as is required in a chronometric task. Another advantage of non-chronometric methods is that they can circumvent the many possible complications involved in measuring and interpreting response times, which can obviously be affected by many factors beyond the internal language processing times that are always the main focus of interest. As we will see in the sections below, however, the main advantages of non-chronometric methods are that they allow for the experimental examination of questions that do not readily lend themselves to chronometric tests, as well as providing means for cross-methodological corroboration (see section 4 below).

One major disadvantage of the non-chronometric approach must also be mentioned, however, and that is the fact that most of the methods involved exploit so-called "metalinguistic" judgments[10], which are by and large conscious and analytical, as opposed to the more covert and less cognitive tasks that typify most chronometric studies. Thus, while in a non-chronometric rating task, for example, an experimental participant might be explicitly required to reflect on how similar (in meaning or sound or whatever) one word is to another, this is not the case in a chronometric task such as naming, which involves no judgments about the linguistic properties of stimuli. The key issue here is what may be called the "naturalness" of the tasks involved and the extent to which they might relate to ordinary language behavior.

The distinction involved here is roughly parallel to the one between so-called on-line vs. off-line tasks. In an on-line task, data are collected while the stimuli are still being processed and, typically, normal language processing mechanisms are involved. This affords such tasks a degree of ecological validity that cannot be matched by off-line tasks, which involve making judgments about stimuli after they have been presented and answering questions that typically do not arise under ordinary conditions of language use. Compare the questions/instructions in Set A below with those in Set B:

Set A:
(A-l) Is X a real word?
(A-2) Does X sound similar to Y?
(A-3) Does X mean about the same thing as Y?
(A-4) Does utterance X make sense in the context Y?
(A-5) Press a button as soon as you can recognize the word you hear.

Set B
(B-1) What is wrong with X?
(B-2) How many speech sounds do X and Y share?
(B-3) How similar in meaning is X to Y?
(B-4) Is X a grammatical or well-formed utterance?
(B-5) Repeat the word you hear with the first syllable reduplicated at the end.

---

[10] The recall and recognition paradigms discussed in section 3.6, however, do not involve such metalinguistic judgments.

Although it is highly unlikely that any of the ten questions or commands listed above would ever be directly encountered in any familiar, everyday setting, it is also clear that the ones in Set A are more closely allied to the activities of normal language processing than are those in Set B, which are all quite analytical and derivative in character. For example, in order to answer question (A-l), a hearer need only consult his or her own mental lexicon (presumably in a manner very much the same as would be required to comprehend the word in any ordinary spoken or written utterance), and to respond "yes" if the word is there (as in the case of the word *should* for a native adult English speaker) and "no" if it is not (as in the case of a non-word like **shalt*). To answer question B-1 about a stimulus like **shalt*, however, the hearer must explicitly invoke knowledge about English grammar and morphological structure (such as this is not the standard or correct "past tense form of the modal auxiliary verb *shall*"), all information which is, at best, only implicitly and incidentally employed in ordinary speech comprehension, where the focus is almost always on the extraction of a speaker's intended meaning. By the same token (to borrow an example from Heringer 1970 *apud* McCawley 1979), it is one thing to judge the grammaticality of an utterance like "John left until 6 P.M." in the absence of any appropriate context of use (cf. question B-4 above) and quite another to view its acceptability in a context that makes it clear that the intended meaning is something like "John left and is to return at 6 P.M." (cf. question A-4).

As will readily be observed in many of the examples provided below, non-chronometric methods typically involve the use of off-line tasks, which means that questions of validity are likely to be raised and will have to be dealt with. The ideal situation, of course (as implied by the discussion on cross-methodological verification in the last section of this chapter) is one in which on-line and off-line tasks complement one another in presenting a consistent picture of what is going on inside of the minds of language learners and users.

3. Some prominent examples of the non-chronometric approach

In this chapter we will focus on six general types of non-chronometric tasks, not because they are the only ones that have been successfully employed, by any means, nor even because they are necessarily the ones most likely to bear the most fruit in the long run. Instead, the techniques to be described in this chapter were selected primarily because, in addition to being both widely and productively employed, they also shared the dual virtues of being both efficient (in that they are capable of yielding relatively large amounts of new data with relatively small amounts of effort) and flexible (and hence readily adaptable to a wide range of specific research questions or problems). By way of introduction, these six general approaches are as follows:

(1) Segmentation tasks;

(2) Rating tasks;

(3) String manipulation tasks (or "experimental word games")

(4) Manipulation of miniature artificial real language subsystems;

(5) Stimulus classification (or "concept formation");

(6) Recall and recognition tasks

3.1. Segmentation Experiments

3.1.1. Unit Counting

Suppose we wanted to know how many elements (say, for example, "speech sounds") that a linguistic utterance was thought to contain. One very easy way to do this is simply to ask participants to perform a count. What could be easier? It might seem surprising to learn, though, that quite a number of interesting findings have emerged from studies using so simple and straightforward an experimental task as this.

In the domain of phonological analysis, for example, one issue that troubled descriptive linguistics for decades was the question of the proper phonemic treatment of speech sounds that involved complex articulations, such as affricates, as in a language like English. Should the stop+fricative combination that ends a word like *rich*, for example, be treated as two segments (as implied by its IPA symbol [tʃ]) or rather as a single segment (as implied by its YPA[11] symbol [č]). It was noted that simplicity arguments could be formulated that favored either position. From the standpoint of economy of phonemic inventory, for example, the first analysis seemed better, since /t/ and /š/ were required for English, anyway, so a phoneme could be "saved" by treating [tš] as a combination of /t/ + /š/. From the standpoint of economy of lexical or textual representation, however, the second treatment was advantageous, since a word like *rich* could be listed as the three-segment string /rɪč/, rather than as the four-segment strong /rɪtš/. So which "economy" ought one choose? And why should we think that economy of any kind might be the decisive factor, in any case? The approach advocated here, of course, is to ask the relevant psychological question, which is "How do English speakers *actually perceive* such sounds? Do they hear them as one segment or as two?" It is in answer to this question that a technique like unit counting has something to offer.

A pioneer in this area was Linnea Ehri, who devised clever ways to ask such questions even of young children in a completely oral context, in order to minimize the possible orthographic bias that, in English, typically treats this sound as a two-letter digraph.[12] In Ehri and Wilce (1980), for example, children were asked to repeat words aloud, placing poker chip counters down for each "speech sound" that they heard themselves saying. For words like *rich* the typical count was three, corresponding to the three-segment treatment outlined above.

Other long-standing questions of phonemic segmentation were also clarified using this very simple unit counting technique. One of these was the question of the appropriate phonemic treatment of complex vowel nuclei in English, such as the diphthongs that appear in words like *pain*/*pane* or *sole*/*soul*. Again, should such English words be analyzed in terms of four segments (e.g., as /peyn/ and /sowl/, respectively), with the glides included, or rather as three (as in /pen/ and /sol/), with the presumably predictable off-glide elements omitted? Once again, Ehri and Wilce's

---

[11] YPA is a facetious abbreviation invented by our colleague, Terry Nearey: It stands for "Yankee Phonetic Association". YPA symbols (with hacheks, [y] instead of IPA [j], etc.) are used preferentially throughout this chapter.

[12] Interestingly, some orthographic effects still did emerge, even in some of Ehri's studies, as when words like *pitch* were presented to children who already knew how to read. Though *rich* and *pitch* clearly rhyme, her subjects showed a tendency to perceive an "extra sound" in the latter, which presumably corresponded to the extra letter in its spelling, and a portion of the adult subjects reported on in Derwing, Nearey, and Dow (1986, p. 62) did the same.

children provided a pretty clear answer, showing an overwhelming preference for the latter, three-segment treatment. Furthermore, as reported in Derwing, Nearey, and Dow (1986), oral counting tasks conducted with English-speaking adults showed the same strong tendencies, even in the case of a vowel nucleus like [ay], for which the offglide element is contrastive and not automatic (cf. *sod* /sɑd/ vs. *side* /sɑ^{y}d/).

In some other cases, however, segment counting led to inconsistent results, either because of uncertainty as to what the units were that were being counted, because of confusion between orthographic and phonological units (see n. 11), or perhaps even because of individual differences between speakers. Examples of this were words with prevocalic glides (whether alone, as in *yule* vs. *wait*, or in consonant clusters, as in *cue* vs. *view* and *mute*, vs. *quote*), and in some cases where rhymes with different spellings (e.g., *cower* vs. *sour*) were treated differently (see Derwing, Nearey & Dow 1986, pp. 54-55). Such findings highlight the need for supplementary tasks to clarify these issues.

3.1.2. Slash and Pause Insertion

Another segmentation question that has long troubled linguistic theorists is the scope of syllables. For a language like English the biggest headache in this area was not so much the number of syllables contained in an utterance, as even very young children seem quite adept at syllable counting (Liberman, Schankweiler, Fischer & Carter 1974), but rather the boundaries or break points between syllables, which were highly disputed and which seemed to depend on a number of potentially independent factors, such as stress and vowel quality. As part of their first attempt to study this problem, Treiman and Danis (1988) used a simple segmentation task that involved inserting a slash (/) to mark syllabic divisions in standard orthographic representations, asking their subjects to choose between written representations like *le/mon* vs. *lem/on*. Unfortunately, this approach did not allow for the possibility that the medial consonant of this word might be included as part of both syllables, so the results of this study will be postponed until section 3.3.1, which describes a different task that Treiman and Danis also used that did allow for this possibility to emerge.

This example also highlights the main disadvantage of the slash-insertion technique, which is the fact that it relies on written letter strings as stimuli. This not only restricts use of the technique to languages that employ segment-based orthographies, such as English[13], but even in English some of the posited break points involve segments that are not consistently represented as divisible elements in the standard orthography. (How, for example, would one use a slash to mark a syllable division between the /k/ and /s/ sounds of a word like *Texas*?) In order to explore syllable division in such cases, as well as in unwritten languages or with speakers who were nonliterate, a technique was sought (like Ehri's unit counting approach described above) that could be employed in a purely oral testing situation. Derwing (1992a) introduced a "pause-break" procedure for this, in which participants were presented with a recorded series of words with pauses inserted in the middle of them, breaking the words into two clearly articulated parts. Thus, for the English word *lemon*, the following three recorded options were offered in a forced-choice task (where "…" represents a pause of approximately 500 ms in duration):

---

[13] Taking the Japanese word *kimono* as an example, note that while this romanized or *romaji* transcription of the word can be used to distinguish the segmentation *ki/mono* from the segmentation *kim/ono*, the standard *hiragan*a and *katakana* transcriptions cannot, since they both represent the sequence *mo* as a single, indivisible character.

(a) /lɛm...ən/ (where /m/ is treated as the coda of the first syllable)
(b) /lɛ...mən / (where /m/ is the onset of the second syllable)
(c) /lɛm...mən / (where /m/ is ambisyllabic)

This task not only largely replicated the main findings of Treiman and Danis (1988) for English (as reported in some detail in the section on string manipulations below), it also provided a segmentation task for languages with different orthographies (such as Arabic, whose orthographic norms for vowels do not lend themselves to the slash-insertion technique), as well as for others that are normally not written at all (such as Blackfoot and Swiss German), leading to the discovery of strong cross-linguistic tendencies to treat single intervocalic consonants as syllable onsets and to divide CC clusters across syllables.

Using a similar oral procedure in which the participants themselves inserted the pauses, Rice, Libben, and Derwing (2002) were also able to explore the question of morpheme divisions in Dene (Chipewyan) by speakers who were illiterate in the language, opening a door for the psycholinguistic investigation of languages with extremely complex (and often descriptively intractable) morphologies, many of which are unwritten and even dying out.

Notwithstanding the disadvantages of the slash-insertion technique for the investigation of segmentation in phonology, there is still much room for its effective use in many "higher level" areas. To give one recent morphological example, consider the research of Yin, Derwing, and Libben (2004) on branching-direction preferences for trimorphemic compounds in Mandarin Chinese. The ideographic characters used in standard Chinese orthography, of course, are syllable-based, which makes perfectly good sense, since the meaning units (morphemes) of the language are coextensive with syllables. While this precludes the possibility of using slashes to explore subsyllabic or submorphemic elements, most of the words in the language are actually compounds, consisting of two or more characters that are written with spaces between them, allowing for the insertion of a slash within the spaces. In the case of a trimorphemic compound like *da hui tang* ('big meeting hall'), for example, there is often also the possibility of two different constituent "groupings" within the word, one that is left- branching ('big-meeting hall', i.e., a hall for big meetings) or right-branching ('big meeting-hall', i.e., a meeting hall that is big). Since left-branching compounds predominate statistically for the language as a whole (Chao 1968), and since this preference is related to potential processing advantages generally (i.e., left-branching structures can be amalgamated by listeners as the signal comes in, while right-branching amalgamations have to be postponed until later), it was of some interest to see if this preference was real for speakers or merely a statistical artifact. One of the experiments described in this study, therefore, involved the use of three types of trimorphemic Chinese compounds: (1) real ambiguous ones, controlled for surface frequency (such as the example already used above), with overall frequencies balanced for the two branching directions in the stimuli actually used); (2) compounds with a non-branching flat structure (e.g., *zhen shan mei*: lit. 'real kind beautiful', meaning 'all kinds of virtues'); and (3) sound imitation loan words in which the meanings of the component parts play no role at all in the meaning of the whole word (e.g., *mai ke feng* : lit. 'wheat grammar wind', meaning 'microphone'). Participants were then asked to insert a slash to break these words up into two component parts. In all three categories, significantly more slashes were inserted between the second and third characters (indicating a left-branching preference) than between the first and the second (right-branching), confirming the pre-experimental expectation. (For a similar study in English involving ambiguous trimorphemic words, see de Almeida & Libben, in press.)

3.2. Rating/Scaling Experiments

One of the most widely used non-chronometric techniques in psycholinguistic research (and in psychological research generally) involves rating some phenomenon or property on a fixed scale. The earliest applications to linguistic material were actually done by psychologists, and these were primarily concerned with questions of lexical semantics. A particularly well-known example of this is the work by Charles Osgood on what he called the "semantic differential". What this research involved was the collection of a vast amount of data on the judged connotative (rather than denotative) meanings of words, using scales like good-bad, small-large, weak-strong, active-passive, etc. (see Osgood 1952 for a concise characterization, and Osgood, Suci & Tannenbaum 1957 for more details and numerous examples). As an example for denotative meaning, we can mention the study by Segalowitz and de Almeida (2002), who used a seven-point rating scale to investigate the meaning similarity of verb pairs selected from the two categories of verbs of motion (such as *walk* and *run*) and psychological verbs (such as *think* and *wonder*), based on the judgments of English and French bilinguals.

The first issues to be faced in any ratings study are the choice of the features or dimensions to be measured (and the labels to be used for them) and the particular scale that is to be used. While the dimensions will vary widely, of course, depending on the particular theoretical question of interest (see below for several examples from the psycholinguistic literature), the descriptions and labels used for these have to be chosen with some care, in order that non-specialists (the usual participants of choice [see section 4 below]) will be able to understand and use them**.** For this reason, labels involving technical terms from linguistics (such as "grammaticality", "morphological relatedness" and "phonemic similarity") are avoided, and often considerable ingenuity and careful pilot testing are required to come up with (near) equivalents that can be expressed in ordinary language (see the discussion below of the problem of testing for "shared morphemes").

A second key choice to be made concerns the type of measure to be used (continuous or discrete) and, if the latter, as is usually the case, the number of points to be used on the scale. Both of the studies already mentioned, for example, used a discrete seven-point scale, with descriptive labels supplied only for the poles, while other researchers have opted for the convenience of the kind of ten-point scale that is provided on many standard optical scoring sheets. Most common, however, is the so-called "Likert" scale (after Likert 1932), which involves five alternatives, with a label provided for each. While in many cases the number of points (and labels) may not be critical, the ideal is to have a sufficient number to reveal real differences in the scores, yet not too many to unduly tax the memory and discrimination skills of the typical participant, and some pilot testing may also be required to assess these factors for any given phenomenon of interest.

To illustrate the use of a Likert-type scale with linguistic material, consider the five-point (0-4) scale that was used by Derwing (1976) in his attempt to elicit semantic similarity judgments as part of a larger study on morphological relatedness:

4 – a clear and unmistakable connection in meaning between the two words

3 – probably a connection in meaning between the two words
2 – unable to decide whether there is a connection in meaning or not
1 – probably *not* a connection in meaning between the two words
0 – no connection in meaning whatsoever between the two words

At this point we may proceed to illustrate some of the specific linguistic questions, in addition to similarity in meaning, that have proven to be amenable to investigation through the use of rating scales. At the opposite extreme from meaning, in a sense, is sound, which is the other end dimension of the language encoding problem.[14] Unsurprisingly, this dimension, too, has been extensively studied, beginning with the work of Greenberg and Jenkins (1964, 1966), who used rating scales to assess sound similarities among words and syllables.

Extending this research and focusing on CVC-CVC word pairs, Nelson and Nelson (1970) systematically varied the number and the position of shared phonemes and found that there was a strong correlation between these formal counts and the global sound similarity ratings that their participants provided. This was then codified as a model of "predicted phonemic distance" (PPD) by Vitz and Winkler (1973), based on words that contained up to three syllables. Derwing and Nearey (1986) showed that sound similarity ratings were also sensitive enough to detect the effects of both subphonemic (i.e., feature) differences and differences between initial vs. final consonants**.** (Thus the pair *pit-bit* was judged to be more similar than *pit-fit*, while the latter pair was judged to be less similar than *cap-calf*.) Interestingly, although different scales were used by all of these researchers (7-points by Nelson and Nelson, 5-points by Vitz and Winkler, and 10-points by Derwing and Nearey[15]), and while a mixture of visual and aural presentations was also employed, the same basic findings emerged throughout, attesting to the extreme (and actually rather surprising!) reliability of the task, even with linguistically naïve participants.

Emboldened by both the sensitivity and reliability of global sound similarity ratings in English, other investigators collaborated in applying the same technique to a variety of other languages, including Arabic (Beinert & Derwing 1993), Taiwanese (Wang & Derwing 1993), Japanese (Derwing & Wiebe 1994), and Korean (Yoon & Derwing 1994, 1995, 2001), reinforcing the idea that the "basic" phonological units can vary from language to language and that the phonemic segment does not have the same status across these language types. While the procedures and analytic techniques employed in these cross-linguistic studies are a bit too varied and complex to relate here, what they showed, in essence, was that, while the segment (i.e., C or V) was basic for English, the syllable (e.g., a whole CVC) was a more fundamental phonological unit in Korean (and presumably also Taiwanese, by default), while the others showed a preference for units that were intermediate between these extremes, such as the mora (e.g., CV) in Japanese[16] and the discontinuous consonantal "tier" (e.g., C…C…C) in Arabic.

To move on to higher levels of application, Derwing (1976) used rating scales in an effort to assess a capacity for "morpheme recognition" in English speakers, which he saw as primarily a function of the two independent factors of similarity in meaning and similarity in sound. This

---

[14] What language encoding involves, ultimately, is the conversion of private, covert meanings into overt sound symbols that can be transmitted from speaker to hearer.
[15] Bendrien (1992) even used a sliding scale labeled for percentages and was thus using a scale that was essentially continuous, but the same results still followed as in the other studies, looking at comparable items.
[16] For more on the mora, see section 3.5 below.

followed from the discussion in Derwing (1973, pp. 122-126), which pointed out that word-pairs that most clearly exhibited shared root morphemes (such as the derivationally related *ride*-rider, *joy-joyous*, and *sad-sadly*) were also pairs for which both semantic and sound similarity seemed high. However, other pairs (such as *plumber-lead*, *holy-Halloween*, and *moon-menstrual*) seemed like poor candidates for morphological relatedness (despite their historical etymological connections), since they showed little in the way of similarity in either meaning or sound. Even in cases where similarity on one of these dimensions seemed quite high, morphological transparency seemed to be destroyed if similarity on the other dimension was very low, as illustrated by the pairs *irrigate-ear* (where sound similarity is high but semantic similarity low) and *milk-lactate* (the reverse). The most interesting cases, however, were those in the "middle" (such as *fable-fabulous*, *tame-timed*, and *moon-month*), where the similarity on both dimensions seems to be middling and where the morphemic status of the words was correspondingly quite unclear.

Measuring meaning and sound similarities was a relatively straightforward matter in this study, as rating scales had already been successfully used for both of these tasks, as noted above. The main practical question was thus how one might go about using a rating scale to assess the ability of ordinary speakers to recognize morphological connections between words (e.g., to determine whether or not words like *fable* and *fabulous* were thought to contain the same root by speakers who were unfamiliar with either formal morphological analysis or with the history of the language). Seeking to avoid the use of technical terms like "morpheme" or "morphological similarity" as labels for such a scale (for obvious reasons), the idea that Derwing (1976) opted for was to play on the "historical" connection and address the morpheme awareness issue by asking if one word "came from" another (e.g., "Does the word *fabulous* come from the word *fable*?"), with response ratings provided on the basis of the following five-point confidence scale:

4 -- No doubt about it
3 -- Probably
2 -- Can't decide
1 -- Probably not
0 -- No way

The result was a task with a good deal of face validity, in that it gave the expected results for items at the extremes (e.g., pairs like *teacher-teach* were rated very high, while pairs like *carpenter-wagon* were rated very low), as well as for items where only one of the two input dimensions was high or low, but not both. (Thus both *eerie-ear* and *puppy-dog* were rated quite low for morphological relatedness by the "comes from" test, despite the high sound similarity rating in the first case and the high meaning similarity rating in the second.)

For all of this, a number of obvious flaws were also noted in this study. For one thing, the results in some cases (e.g., *handkerchief-hand* and *cupboard-cup*) seemed to reflect the factor of spelling, which was not controlled in the experiment (but see Derwing, Smith & Wiebe 1995 for a follow-up study that confronted the orthographic knowledge issue directly). For another, the "pseudo-etymological" character of Derwing's "comes from" test raised questions in many minds as to whether this was the best way to go. Schirmeier, Derwing, and Libben (in press) explored this by comparing the "comes from" version of the morpheme recognition test with other formulations that seemed to bring out the formal/morphological and semantic connections in a more direct way. In a study of German derived verbs and their presumed roots and stems, therefore, these investigators supplemented a "comes from" task ("Does the word X come from the word Y?")

with one that focused on inter-word relationships that might be involved ("Does the word X contain a form of the word Y?") and also by one that focused on the meanings of the supposedly shared morphemic elements ("Does the word X contain the meaning of the word Y?"). Interestingly, all three tasks gave essentially the same pattern of results for the 64 German verbs tested, which spoke to the fact that all three versions of the test were likely measuring pretty much the same thing (i.e., morphemic relatedness).

But the possible applications of rating scales need by no means end here. To our minds, a particularly fascinating and promising application to syntax is provided by the study of inter-word "connectedness", as pioneered by Levelt (1970) in Dutch. In this technique, participants are presented with a sentence (e.g., *Carla took the book and went to school*) and are then presented with all of the "content" words from the sentence in pairs: *Carla-took*, *Carla-book*, *took-book*, etc. For each pair they have to rate (on a 5-point scale) how closely related the two words are in the context of the particular sentence provided. Levelt created a scale of relations and showed that the final hierarchy that emerged was similar to what was then taken to be the "deep structure" of the sentence. So, for instance, *Carla-took* and *Carla-went* were equally rated, showing that ratings were not based on proximity (surface structure) but rather on perceived relations between words in some "deeper" syntactic or semantic representation of the sentence.

This technique was used again by Fodor, Garrett, Walker, and Parkes (1980), Gergely and Bever (1986), and by de Almeida (1999a) to investigate not syntactic but rather the semantic structure of sentences containing different types of verbs and how they are affected by the perceived "distance" between their subject and object nouns, as in the following examples:

a. The engineer expected the public to leave the station.
b. The engineer persuaded the public to leave the station.

It is also worth mentioning that the last three studies described all used two types of off-line tasks to investigate similar issues (thus, one "corroborating" the other): the "scaling" one, as above, identified with Levelt's work, and a "forced choice" version, in which subjects had to select a sentence in which they considered two underlined words to be the more closely related, as in the examples above. These two sentences differ in syntactic structure, since the verb *expect* does not take a direct object, while *persuade* does. Naïve subjects (psychology undergraduates) correctly judged the two underlined words to be more "closely related" in the second sentence than in the first.

Finally, rating scales have also been used to assess grammaticality judgments, though their number is disappointingly (even shockingly) few, considering the major role that such judgments have played in the development of influential theories such as generative grammar (in all of its various manifestations). Perhaps the most interesting and influential example is provided by Ross (1979), who used the following four-point scale (but ultimately wished that he had tried to cut things a little finer) to assess grammatically judgments for about a dozen carefully selected sentences of medium to high syntactic complexity:

1. The sentence sounds perfect. You would use it without hesitation.
2. The sentence is less than perfect--something in it just doesn't feel comfortable. Maybe lots of people could say it, but you never feel quite comfortable with it.

3. Worse than 2, but not completely impossible. Maybe somebody might use the sentence, but certainly not you. The sentence is almost beyond hope.
4. The sentence is absolutely out. Impossible to understand, nobody would say it. Un-English.

These assessments were also supplemented by three levels of confidence judgments ("Pretty sure/Middling/Pretty unsure"), as well as by some supplementary questions designed to assess each participant's general level of liberality/conservatism with regard to usage.

Interestingly, the main finding of this study was to provide support for the now quite widespread realization of "the staggering extent of interspeaker variation [in grammaticality judgments] on any given set of sentences" (p. 128), together with the view that "the sentences of a language seem to be viewed by speakers as falling into three groups: a core, a bog, and a fringe." (See also McCawley 1979 on the shaky status of the whole notion of grammaticality, judged apart from situations of actual language use.)

Perhaps better than any other, this latter case illustrates the truism that an experiment is basically nothing more than a systematic way to go about collecting data (cf. Ohala, 1986), and, given the choice, which (and whose) data ought we to value and trust more: Those provided willy-nilly by a single linguist, exploring his or her own inner best judgments, which have likely also been biased by years of very specialized training and/or by a particular theoretical orientation? Or rather those that have been provided by large numbers of more representative, ordinary speakers, whose language facility is the one we all truly most want to understand in the end, anyway? Simple experiments involving rating scales (and other techniques as herein described) not only make that choice possible, but they also enable comparisons between trained and untrained observers, in order to determine if linguistic background does indeed have a significant effect on the kind of judgments they provide (see, in fact, Spencer 1973 for a concrete example of this, indicating that there may actually be more consistency in the judgments of the naïve speakers than among linguistically trained ones[17]).

3.3. String Manipulations ("Experimental Word Games")

A third general approach to the study of linguistic elements, apart from crude segmentation/counting and the use of rating scales, is to perform some type of manipulation on them. Such string manipulation tasks are sometimes referred to as "artificial language games", since the manipulations involved are reminiscent of (and often directly inspired by) those performed spontaneously in any of a number of "secret languages" or "ludlings" that have been documented in the literature (for copious examples, see Laycock 1972 and Bullard 1979). Perhaps the most familiar such natural "language game" in English is the one called Pig Latin, which (in one familiar dialect) involves moving the onset of the first syllable of a word to the end and adding the vowel /e/. Thus the ordinary word *secret* is converted into its "disguised" counterpart *eekrut-say*. (Day 1973)

3.3.1. Unit Inversion

[17] Ross (1979) gives very little information about his 30 participants, at least some of whom may well have been trained linguists, given the situation under which his data were collected..

To illustrate the application of a string manipulation task to the experimental study of phonological units, we may return to the case of Treiman and Danis (1988), who supplemented their slash-insertion task with written representations (as described in section 3.1 above) with an oral production task of unit inversion, in order to investigate the scope of syllables in bisyllabic words. The specific theoretical question at issue was the controversial one involving the syllable break-point in words containing a single intervocalic consonant, such as *melon*, *lemon*, *seven*, and *radish*. Taking the word *lemon* as our example again here, should such a word be syllabified as *lem-on*, with the critical consonant included as the end (coda) of the first syllable, or as *le-mon*, with the medial consonant as the onset of the second syllable, in accord with the theoretical notion of the universal obligatory onset principle (e.g., Hooper 1972)? Or, perhaps, might the consonant even be treated as both, as in the ambisyllabic representation *lem-mon*, as other theorists had suggested (e.g., Kahn 1976)?

Avoiding explicit use of the term "syllable" in their instructions, Treiman and Danis trained their participants to move the first part of a word to the end, using examples like *grandfather* > *fathergrand* and *catfood* > *foodcat*, in which the first syllable also corresponded to a morpheme. For a test stimulus like *lemon*, therefore, the experted inversions were to *monle*, *onlem*, or *monlem*, depending on which of the syllabications was preferred for the word. Using a carefully selected set of stimuli that incorporated systematic factors that were expected to affect the results, Treiman and Danis found indeed that there was no single, simple answer, but that the results tended to vary as a function of at least four different factors, with the preferred syllable breaks shown with a hyphen here: (1) the position of stress (thus *lem-on* tended to be treated differently from *de-mand*); (2) the quality of the preceding vowel (thus *lem-on* also tended to be treated differently from *mo-ment*) (3) the relative 'sonority' or 'strength' of the intervocal consonant itself (thus both *mel-on* and *lem-on* tended to be treated differently from *se-ven* and *ra-dish*), and (4) by the spelling of the word, as only words with doublet spellings led to significant numbers of ambisyllabic responses, distinguishing *lem-on* from both *com-mon* and *com-mand*.[18]

Finally, in a particularly fascinating application of the unit-inversion technique, Hombert (1986) used it in a study of the status of tone in three African languages (Bakwiri, Dschang, and Kru). He discovered that the tone pattern of disyllabic words was never affected by inversion of the syllables (e.g., kwélì > líkwè in Bakwiri), indicating that tone was a feature of the word in these languages, rather than a feature of the individual syllables.[19]

3.3.2. Word-blending

Another and even more widely used string manipulation task is word-blending, which is also modeled after a naturally occurring linguistic phenomenon. Natural word blends in English include the word *smog*, which was created from the onset *sm-* of the word *smoke* and the rhyme *-og* of the word *fog*. (cf. also *brunch*, which comes from the onset *br-* of the first syllable of *breakfast* and the rhyme *-unch* of *lunch*.) Treiman (1983) exploited this idea to test whether the onset (everything before the vowel) and the rhyme (the vowel plus everything after it) were in fact significant

---

[18] As already noted in section 3.1 above, Derwing (1992) replicated these basic findings using a forced-choice oral segmentation task, as well as adding morpheme integrity to the list of relevant factors (cf. *oil-y* vs. *doi-ly*) and extending the testing to other languages.

[19] Hombert's results were less consistent in a replication with three Asian tone languages..

intrasyllabic units in English (as suggested by the natural blends illustrated). Taking pairs of meaningless but well-formed English monosyllables (such as *krint* and *glupth*), Treiman trained individual participants to combine them into a new monosyllable that contained parts of both, and she found that responses like *krupth* (which consisted of the onset *kr-* of the first syllable and the rhyme *-upth* of the second) were produced far more often than any other possible combination, suggesting that the natural break point within English syllables occurred immediately before the vowel.

Treiman also reasoned that if syllables were composed of onset and rhyme constituents, then "games" that kept these units intact should be easier to learn than games that broke the syllables up in a different way. She thus also taught her participants a word game in which the onset of a nonsense CCVCC syllable was blended with the rhyme of another (e.g., *fl-irz + gr-uns > fl-uns*), as well as other games in which the onsets and rhymes were broken up (as in the blends *f-runs*, *fli-ns*, and *flir-s*). She found that the game that kept both onset and rhyme intact was learned with fewer errors than were the other three games. (see also Treiman, 1985, 1986, 1988 for other, similar studies, and Treiman & Kessler, 1995, for a defense of the onset-rhyme interpretation of this research). Wiebe & Derwing (1994) also describe a forced-choice version of Treiman's word-blending task that can be efficiently used for group-testing in English and other languages.

3.3.3. Unit Substitution

Another useful manipulation that has been used in research in this area is unit substitution (or substitution-by-analogy). As the name suggests, in this type of task participants are trained to replace some portion of a string with something else, as indicated by modeled examples. In Dow and Derwing (1989), for instance, participants were trained to replace various substrings of words (e.g., to change *might* to *plight* and *drank* to *plank*, where the onsets *m-* and *dr-* have both been replaced by the onset *pl-*). These two exemplars were then immediately followed by a test item, as illustrated below for each of the substitution types that were investigated.[20]

Substitution	Modeled examples	Stimulus:(target)
1. Replace onset by /pl-/	might-plight drank-plank	scum:(plum)
2. Replace rhyme by /-old/	baste-bold strict-strolled	scant:(scold)
3. Replace body by /kræ-/	floss-crass drift-craft	blush:(crash)
4. Replace coda by /-m/	tote-tome clasp-clam	prince:(prim)
5. Replace vowel by /-u-/	bath-booth crown-croon	stowed:(stewed)

[20] Controls were imposed across all of these categories on the total number of phonemes exchanged, number of spelling matches vs. mismatches in the unchanged portions of stimuli and targets (cf. *scant-scold* vs. *kept-cold*), and number of real vs. nonsense targets (cf. *monk-mold* vs. *chomp-*chold*).

6. Replace margins[21] by /b…t/	strife-bite tense-bet	gloom:(boot)

The results (both in terms of accuracy and response time) were quite clear and consistent, indicating that onsets and rhymes were the easiest units to manipulate, and that bodies (called "heads" in those studies) and margins were the most difficult, with vowels and codas falling in between (see also Derwing, Dow & Nearey 1989). An analysis of error types also showed a strong tendency for incorrect substitutions to involve changes in the onsets or rhymes, with very few errors of the other types.

Derwing, Dow, and Nearey (1989) also used a combination of substitution and deletion (= substitution by null) tasks to investigate the role of sonority and vowel quality on intrasyllabic break points (cf. the Treiman study already described above), but the details are too complex to recapitulate here (see also Derwing & Nearey 1991, which provides a brief summary of this research under the rubric of the "vowel stickiness" phenomenon). Suffice it to say that the hierarchy that consistently emerged from these studies was much the same as the one that had been proposed in many theoretical accounts on the subject (e.g., Hooper 1976, Selkirk 1984): Glides > R > L > Nasals > Obstruents (where > denotes a increase in "strength" or a decrease in sonority, which corresponds in the experiments to a decreasing tendency for the consonants involved to adjoin with or "stick to" a preceding vowel).

In short, while only a small sampling specific studies are described here, the range of manipulation types and their applications would seem to be virtually endless, limited only by the energy and imagination of the investigator involved.[22]

3.4. Miniature Artificial Real Language Subsystems (MARLS)/ "Berko-Type Studies"

Our title for this section is adapted from the term "miniature artificial languages" (or MALs), which have been quite widely used by psychologists to investigate learnability and the course of learning development. In the typical case, novel (nonsense) names are assigned to geometrical figures on the basis of variation in some of their physical properties (size, shape, color, etc.) and the relative ease of learning various kinds of artificial systems is explored. Esper (1925) describes perhaps the first English experiment of this type, whose stated goal was to help achieve a better understanding of the phenomenon of analogic language change.[23] However, as this research paradigm developed (largely in the hands of psychologists), it became clear that the focus was not on any particular aspects of the linguistic material itself (i.e., linguistic structure) but rather on some aspect of general learning theory, with the artificial language-like material used merely as convenient

---

[21] The discontinuous pseudo-unit "margins" consists of the onset (everything before the vowel) and the coda (everything after it).

[22] A fourth methodological variant in this category that was excluded here only because of space limitations is the method of inducing speech errors, patterned on the kind of natural errors that frequently occur in ordinary speech. For a good example of the experimental use of a task of this kind to answer a linguistic question, see Stemberger and Lewis (1986).

[23] It is also of some historic interest that this first monograph published by the Linguistic Society of America was experimental in character.

stimuli (see, e.g., Foss 1968). Moreover, as Schlesinger (1977) noted in his critique, to show that something was "learnable" or that some suggested mechanisms "might *possibly* account for language learning", did not at all demonstrate that such things were indeed learned or that such mechanisms were indeed employed in normal language acquisition--any more than (we might add) describing an abstract "language" in a particular way, however elegantly, means that the description in question has any necessary claim to psychological reality for the speakers of the language. In both cases we are faced with empirical issues, and such issues must be tested empirically.

In our view, however, there is one particular variant of the general "artificial language" approach that does seem to show a great potential for revealing some of the details of what it is that speakers know and do when they learn and use a language. The production experiment of Berko (1958) epitomizes the approach--so much so, in fact, that we have used the term "Berko-type studies" as a clarifying addendum to the title of this section. What Berko did was to make up a set of novel English-like vocabulary items, and then direct her participants to create inflectional and derived forms of those "words" , just as if they were part of their own language (hence the use of the phrase "real language subsystems" above). The possible benefits of this approach can be readily seen if we look at Berko's (1958) so-called "wug" experiment in more detail.

The main research question that motivated Berko's experiment was whether or not the English nominal and verbal inflections (such as the plural forms of nouns and the past tense form of verbs) were productive, rather than learned merely by rote, and, if so, to discover what kind of psychological mechanism (such as a rule) might be responsible for this productivity. For example, when a native speaker of English said something like "two dogs", was the word *dogs* necessarily dredged up out of long-term memory ("the plural of *dog* is *dogs*"), or could it have been constructed instead on the basis of some general principle that took the singular form of the word into account (such as a subrule that specified, "if a word ends in the /g/ sound, its plural must be formed by adding the sound /z/").

Note first of all that this is formulated as a psychological question, and not as a purely "linguistic" one. Clearly, since the number of (head) nouns in English is finite, their plural forms could in principle be described in the form of a list (although this would not seem to be a very parsimonious approach). Alternatively, a set of rules (in fact, many alternative sets, as noted below) could be formulated to describe them. Or some (say, the most frequent ones) could be listed and some rule(s) or analogical principles used to generate the others. But Berko was not interested in linguistic descriptions *per se*. What she was interested in was what speakers of English actually knew, and what they actually did when they engaged in pluralization behavior. She therefore performed a psychological test that was disarmingly simple in its conception and realization, yet which did something that no abstract linguistic description or theory could ever do: It informed us about the linguistic abilities of real speakers of English.

There was, of course, much anecdotal evidence available at the time to suggest that even preschool children were able to create novel inflected forms, especially in cases of so-called "overgeneralization" errors, where irregular plural and past tense forms like *sheep* and *ran* were replaced by their "regularized" counterparts *sheeps* and *runned*, respectively. However, rather than waiting around and cataloguing children's spontaneous errors in the (likely) vain hope of revealing the precise nature of the productive mechanism involved and its pattern of development, Berko decided instead to create a controlled semi-artificial situation (in other words, an experiment) that would answer her specific research questions immediately and directly.

The key to the success of Berko's experiment was the use of nonce forms as stimuli, and these are what gave the "artificial language" character to her study.. These novel forms were word-like in phonological form but which lacked any meanings for English speakers; critically, the children she tested could not have known the inflected forms of these words in advance, as Berko herself had made the words up. To deal with the meaning problem, Berko used pictures of make-believe creatures (her "wug" looked much like a small, nondescript bird) for her nouns and stick drawings of hitherto unnamed actions for her verbs (her "bing" referred to the action of standing on the ceiling). In the experiment itself, then, she provided oral syntactic frames (along with the pictures) to elicit the desired inflectional endings, such as "There are two _____" for the plural forms of nouns and "Yesterday he _____" for the past tense of verbs. Her results demonstrated not only productivity but also a good deal of systematicity (such as adding /-z/ to mark the plural forms of stems ending in /g,n,r,ɔ/ but /-əz/ to mark those ending in /s, z, č, ž/), strongly suggesting the involvement of some kind of linguistic rules or other generalizations that were sensitive to the phonological properties that she had chosen to manipulate in her nonce stimuli.

However, although Berko looked at a variety of both inflectional and derivational forms in her experiment, more than half of her total stimuli involved the two inflections just noted, and even for these she had only nine nonce items for plurals and six for the past tense. She thus obtained only a very partial picture of what her child subjects could do, as these items represented only a very small portion of the range of possible stem types involved; moreover, the children she tested ranged from only four to seven years, with three-quarters of them aged six or over, representing only a portion of the full range of development, as later studies showed. Furthermore, she considered only one formulation of the rules that she had set out to test, namely, the one that was familiar from the linguistics texts of that period. Thus, although this study certainly opened exciting new ground, it was far too small in scope to address the question of the precise nature of the mechanism that her child participants were actually using to create their novel inflected and derivational forms, or to reveal the developmental details.

Surprisingly, there were very few studies in the first two decades or so after Berko's work that followed up on this very promising line of experimental linguistic inquiry (e.g., Anisfeld & Tucker, 1968; Anisfeld & Gordon, 1968; Gray & Cameron, 1980; and the papers listed in fn. 22). However, interest in morphological issues took a big leap forward as part of the rule vs. connectionist debate of the late 1980s (see especially Rumelhart & McClelland 1986 and Pinker & Prince 1988, as well as chapter **M** of this volume)--in the context of more general debates about the proper cognitive architecture--and with the rise of interest in analogical or exemplar modeling since that time (see Chapter **Q** here). Even during the heydey of implicit rule accounts, however, it was clear that many formulations could be readily conceived to account for Berko's findings (see, for example, Derwing 1980), and some attempts were made to psychologically interpret and test these alternatives in a systematic way using a Berko-style approach, mostly by greatly expanding both the variety of nonce-word stimulus types employed and the age range of the children tested.[24]

---

[24] See Derwing & Baker 1980 for a detailed summary of the main results of this work, and Derwing & Baker 1979 for a small extension to derivational morphology. For more on the English plural inflection, in particular, see Innes 1974 and Baker & Derwing 1982, who utilize Innes' data to present a clear picture of several discrete "stages" in the development of pluralization skill.

Perhaps the most ambitious of the Berko-style attempts at rule evaluation, however, was made by Dennis (1988), whose goal was to test the psychological reality of the classical "generative phonological" analysis of plural and past tense formation in English.[25] Key to that account was the postulation of an invariant "base form" in the lexicon for each of these affixes (/z/ for the plural and /d/ for the past tense) and two general phonotactic rules of voicing assimilation and vowel insertion to account for all regular variations from these shapes. Thus, if the addition of a plural marker /z/ to a noun created a phonotactically unacceptable word-final sequence like /tz/, the /z/ was postulated to be automatically and unconsciously devoiced to /s/ (as in the plural form *cats*). To complete the picture for the regular plurals, this analysis also postulated that a schwa-like lax vowel was automatically inserted to break up impossible word-final clusters consisting of two consecutive sibilants (hence the plural form of *church* is *churches*, with the vocalic suffix /-əz/ not **churchs,* with either a simple /-s/ or a /-z/*)*. For ease of reference, the lexical representations and rules presumed to be involved in such an account are given in prose form as (R1) and (R2) below:

Lexical Representation: (Plural) =-z

Rules:
(R1) Insert a lax vowel to break up word-final sibilant+sibilant clusters.
(For example, /čz/# converts to / čəz/#)
(R2) For consonant clusters that remain intact, the voicing of a final
obstruent must agree with that of the consonant immediately before it.
(Thus, /tz/ converts to /ts/.)

What Dennis did next was to create a vocabulary of 16 monosyllabic nonce words, having the following properties:

(1) Four ended in the voiced obstruents /b/ (*zabe*, *tib*) or /v/ (*lev, seave*);
(2) Four ended in the voiceless obstruents /p/ (*lep*, *kip*) or /f/ (*paff*, *giff*);
(3) Two ended in the voiced sibilants /z/ (*muzz*) or /ǰ/ (*sadge*);
(4) Two ended in the voiceless sibilants /s/ (*biss*) or /č/ (*gutch*);
(5) Two ended in the voiced velar stop /g/ (*boog*, *deg*);
(6) Two ended in the voiceless velar stop /k/ (*gack*, *beck*).

For the sub-experiment to be described here, she also created a training set with the same properties as stimulus set (1) above (*feb*, *gib*, *sav*, *tev*).

Each of these 16 test stimuli was then associated with a picture card of a nonsense creature (colored pink) and presented as depictions of infant creatures from another planet. Participants were 45 third and fourth graders, who practiced these names (in a shuffled order) until they could correctly recall all of them from the pictures.

At this point the children were introduced to the first of the training stimuli, which came in two colors, pink and blue. For the first pair it was also explained that when the creatures on this planet started to walk, their colors changed to blue and their names also changed slightly. Thus the baby (pink) example shown was called a *feb*, while the older (blue) example was called a *febzh* (/fɛbž/).

---

[25] Since this experiment was part of a master's thesis that has never been published and is not readily accessible, it will be described here in more detail than would otherwise be the case.

Each child then practiced both forms of the word and was introduced to the other three pink and blue training pairs in a similar manner. In the testing phase, then, these training items were randomly mixed with the original 16 stimulus items, which now also came in both pink and blue variants. The responses of interest, of course, were those to the blue variants that the children had never seen or named before.

At this point we can systematically view what these children were being required to do for the test items that were colored blue:

(a) For the test items in set (1), the task was merely to add the suffix -/ž/ after /b/ or /v/, just as they had done with the training items;

(b) For the test items in set (5), the task was quite similar, which was to add the suffix after the stem-final consonant /g/, something that was not done in the training session but which did not create any particular phonotactic difficulties[26];

(c) For the items in sets (2) and (6), however, the task was complicated by the fact that the children were now being asked, in effect, to create a word-final consonant cluster that violated presumed rule R2. Specifically, in each case they were being asked to place a voiced obstruent after a voiceless one. For example, the straightforward answer for the name of the blue *lep* would be *lepzh* (*/lɛpž/), which violates the premise of the voicing agreement rule. The idea, of course, was that, if a rule of this kind had actually been internalized as part of the children's acquisition of English, we should expect them to automatically devoice this final segment and give the name as *lepsh* (/lɛpš/), in strict accordance with the rule.

(d) Finally, for the items in sets (3) and (4), a second complication arose, which involved the problem of placing one sibilant after another at the end of a word, in violation of presumed rule R1. An example of such a violation would be the name *bisszh* (*/bɪsž/), and if the premise behind rule R2 was correct, we would naturally expect that the problem would be resolved by giving the name as *bissezh* (/bɪsəž), with the lax vowel automatically inserted, as specified by the rule.

In short, in the MARLS described, not only was some new vocabulary introduced, but also a new suffix /-ž/, which was, of course, modeled after the real plural suffix (supposedly /-z/ in its underlying form), with which it shared all of its properties except point of articulation and meaning. By requiring participants to juxtapose these elements in the experiment, the expectation was that, if the voicing assimilation and vowel insertion rules were truly operative for speakers, then they should automatically come into play with synthetic stimuli, as well, under the appropriate circumstances.

In fact, for the cases described by R2, that is precisely what happened nearly all of the time, as fully 92% of the responses involved devoicing the /-ž/ suffix to /-š/ after the stem-final voiceless items, with only 3% preserving the voiced variant that was actually taught. (There were also a few anomalous responses.) This supplies a measure of experimental support for the psychological validity of the phonotactic explanation for the alternation between /-z/ and /-s/ for the real English (plural) forms, as well.

The same, however, cannot be said for rule R1, as a mere 8% of the responses involved the insertion of a lax vowel to produce the variant /-əž/ in the appropriate cases. In 21% of these cases

---

[26]An initial screening test was performed to insure that all participants included for testing were able to correctly articulate all of the critical word-final clusters required for this experiment (e.g., /gž/#), with the exception of those that violated either of the presumed two phonotactic rules that were the focus of the study.

the voiced variant /-ž / was also used here (e.g., /bɪsž/), in direct violation of the supposed phonotactic constraint that motivated the rule, and 19% involved null responses, where no suffix at all was used (e.g., /bɪs/). (The remaining 46% of the responses were anomalous, which often involved deleting the stem-final consonant and replacing it with the inflection, e.g., /bɪž/), This suggests that the presumed phonotactic explanation for the /-əž// variant of the regular plural is not psychologically correct and that some other theory ought to be devised to account for it.[27] (Dennis herself proposed a hybrid solution, whereby the choice of the /-əž/ variant after sibilants is lexically determined.)

It is worth noting, finally, that it is perhaps just as important to the interpretation of this study that the devoicing rule "worked" as it is that the vowel insertion rule did not, for without the positive result in the first case, Dennis would have been faced with the possible conclusion that the whole experiment was simply too "artificial" to yield valid results at all. The fact that the devoicing phenomenon came through loud and clear, however, while the vowel insertion one did not, suggests instead that it is the two linguistic phenomena that really differ, and that the experiment tapped the first of them rather well, while at the same time shedding some light on the correct interpretation of the other, and this result ought to give encouragement to those who might be interested in extending this same approach to other subsystems of English grammar, as well as to other languages.

3.5. Concept Formation

The so-called "concept formation" (CF) technique has been widely used in general psychological research and has appeared in many versions (see Deese and Hulse 1967; Dominowsky 1970; Bolton 1977). Perhaps better viewed as "category identification", the version used in psycholinguistic research involves defining a target category on the basis of some common property or properties. Experimental stimuli (usually words or pictures) are then selected which represent both positive and negative instances of the category. Participants are instructed to respond "yes" (usually by pressing a designated computer key) to items that they think belong to the target category and "no" (another key) to other items, with each response immediately reinforced for correctness. Whereas participants typically have no idea at the outset what the target category is and must resort to guessing for the first few trials, the reinforcement provided is intended to gradually make them aware of how the positive and negative instances differ from each other and eventually to "learn" or "master" the target category, as evidenced by a variety of criteria that are discussed below. In the typical case, a number of clear positive and negative instances of the target category are first presented in training, and then tests are carried out to see how participants categorize other, less clear instances, whose category membership is in question. Another popular approach is to investigate the ease with which the mastery of a given category can be achieved during the training session itself, particular vis-à-vis some other category that is defined in a different way.

To illustrate how this procedure works with a concrete example, we can create a hypothetical case that captures the flavor of some of the early CF studies in psychology (e.g., Bruner, Goodnow, & Austin 1956), which were often concerned with the identification of categories defined in terms of one or more attributes that make up complex geometrical objects, such as shape (e.g., circle vs.

---

[27]In a second phrase of the study not reported here, Dennis also introduced the mock suffix -/g/ (as an artificial analog of the real past tense suffix -/d/), and tested whether vowel insertion would occur under the appropriate conditions with it (i.e., insertion between two velar stop consonants, rather than between two alveolar ones, as with the real suffix), with very similar negative results.

square vs. triangle), size (large vs. small), and color (blue vs. green, vs. red). For example, suppose that the simple concept "circle" was selected as the target category, and then a series of pictures could be presented that showed variation along all three dimensions of shape, size, and color, such as "large blue circle", "small green square", "large red triangle", "small green circle", etc. In this scenario, a "yes" response would only be reinforced as correct whenever a circle was pictured (regardless of size or color), with the "no" response reinforced for all squares and triangles. While participants would naturally make guessing errors during the first few trials, even in a simple case like this, they would very soon learn to isolate the concept/feature of interest, after which time their performance would dramatically improve and the conclusion could be drawn that the target category has been correctly identified. Of course, a more complex target category (such as a "large red" item) might take a little longer to figure out, but so long as simple and familiar categories were all that were involved, the task would likely never be particularly challenging or onerous, as it might have been had some less familiar or more complex category been designated as the target (such as a "small vertically aligned green amorphous shape").

Though obviously more complex (both in design and implementation) than any of the other approaches discussed in this chapter so far, the CF technique is almost ideally suited to the investigation of categories about which participants might have only very limited overt awareness, which is typically the case for the internalized categories of language (cf. Lakoff 1982). Consequently, like the simple counting and rating techniques already discussed, the CF technique has proven to be extremely flexible in its application to a wide range of linguistic notions, ranging from semantic concepts to phonological units (which, as in earlier sections, will be the primary emphasis in the present discussion). A high level of controllability can also be achieved with the technique, either by manipulating the explicitness of the instructions or the reinforcement schedule, as discussed further below.

A typical CF experiment involves the following four components, ordered as indicated: (1) instructions, (2) learning session, (3) test session (optional, depending on the nature of the problem, as discussed below), and (4) post-experimental interview (also optional, but highly recommended). The instructions, of course, are designed to explain the basic nature of the task, as well as to help direct participants' attention to the relevant attributes that define the target class. If they have difficulty understanding the task from simple verbal instructions and a few examples, a practice session can also be included, using categories that are far removed from those that are involved in the experiment itself.

The purpose of the learning session is to gradually teach a concept to participants by reinforcing their responses for correctness. It is here that participants are provided with both positive (target) and negative (distractor) examples of the particular concept or category that they are being trained to learn. In order not to bias participants to respond in one way or the other, the total number of positive ("yes") and negative ("no") items is normally kept the same over the whole course of training and testing. In the case of complex targets, it is advantageous to begin with the clearest or most typical examples of the targets and distractors, moving only later to less typical targets and to distractors that are more similar to the targets and thus more likely to be confused with them. Furthermore, if the focus is on comparing the learnability of one target category with another, it is critical that the reinforcement schedule (including the target and distractor types presented at each stage, as well as the pattern of "yes" and "no" items throughout) be kept the same across the two experiments. (See Jaeger 1986 and Yoon & Derwing 2001 for detailed examples of these procedures.)

Several response measures have been used to determine whether a given target category has been mastered by participants during the learning session. Most common among these is the "trials to criterion" measure, which looks for the first sequence of correct "yes" and "no' responses that conform to some fixed standard and then tabulates the trial number of the last correct response in such a sequence. While the criterion "ten consecutive correct responses" was often used in early work, this rather strict standard can often lead to misleadingly high scores, in that it does not allow for the possibility that an occasional incorrect response might occur inadvertently. Jaeger (1986), therefore, recommended the more liberal criterion of "15 trials in a row with 2 or fewer errors", which can be justified on statistical grounds (see Yoon & Derwing 2001, p. 208). Other (supplementary) measures include the number (or percentage) of participants who reach the mastery criterion by the end of the learning session, as well as the total number of correct responses overall. As Jaeger (1986) also emphasized, an analysis of the specific errors that participants make can also be very informative, as well as their ability to name (or even correctly describe) the target category during the post-experimental interview (see below).

The main purpose of the test session (if included) is to observe participants' categorization of various ambiguous or theoretically controversial stimuli (i.e., test words), whose category membership is the main focus of the investigation. Unlike the learning session, no feedback is provided about the correctness of the participants' responses during the test session. Separate instructions are thus necessary before the beginning of the test session to bring this point out, and the (new) test words are intermixed with examples of targets and distractors previously used in the learning session, in order to provide a means for testing that the participants are still on track (Jaeger 1986 also recommends the use of new "control tokens" during the test session, which are designed to confirm that that participants have identified the intended target category, rather than some extraneous one.)

In the post-experimental interview, finally, participants are asked if they can name or describe the category that they had been trained to learn and to explain their decision-making strategy, as well as note any particular problems that they may have encountered along the way. Such observations not only help in the interpretation of the experimental results, they can also contribute to improved experimental designs in the future. Although participants may form a concept without being able to name it, there seems to be a definite correlation between participants' naming ability and the ease with which they identify the concept or category involved (see Deese & Hulse 1967 on "codability").

.

Although the details of procedure vary greatly in CF experiments, depending on the particular categories being tested, the basic assumption remains the same: The easier it is to identify a category or concept on the basis of its membership, the more psychologically salient that category or concept is. Thus, as Jaeger and Ohala (1984) anticipated, previously existing or natural categories ought to be more readily be brought to conscious awareness than non-existent or unnatural ones. By the same token, Rosch (1973ab, 1978) discovered that her "basic level " categories (e.g., *cat*) were easier to learn and faster to recognize than superordinate (*animal*) or subordinate (*Siamese*) ones, and that semantic categories were better represented by certain "prototype" members than by other "peripheral" members.

Jaeger (1980b, 1986) and Ohala (1986) replicated these effects for phonological categories in a series of experiments designed to test the psychological status of the phoneme in English. Using

the CF technique, Jaeger (1980b) found evidence that the phonemic segment was a "basic level" category in English, as demonstrated by the fact that her participants learned a phoneme-sized category readily and spontaneously identified various allophones of a phoneme as belonging to the same category. Her target set consisted of words containing the phoneme /k/ in English, and the target examples she used in her training session were words containing the aspirated allophone [k^h] (e.g., *kind*), including orthographic variants such as *clear*, *chrome*, *acclaim*, and *queen*. Her distractors were words which caused different kinds of interference with her target words: no interference (e.g., *left*), orthographic interference (*knit*), and phonetic interference (*gift*), where a similar but contrasting category (the phoneme /g/) was involved. In her test session, then, new words were included that introduced the unaspirated [k] (e.g., *skin*) and unreleased [k^-] (e.g., *take*) allophones, which were readily included in the same target category. Ohala (1986) completed the picture by showing that participants had much more difficulty learning a category that included [k] with [g] than one that joined [k] with [k^h].

In her CF experiments in Japanese, however, where the mora[28] is a more salient unit than the segment, Jaeger (1980a) found that her participants were much less successful in identifying a concept based on a single phoneme. It remained unclear, however, the extent to which orthographic differences might have been responsible for these results (see section 4 below for further discussion of this problem).

Jaeger and Ohala (1984) also explored the conceptual salience of phonological features. In this study, the five consonantal features [+anterior], [-anterior], [+voice], [+sonorant], and [+continuant][29] were tested in English, and the target categories that their participants had to identify all involved words beginning with a phoneme that could be characterized by one of the these. Overall, the concept of the feature (or the "natural class of phonemes" that each feature was designed to represent) was found to be much more difficult to form than the concept of phoneme (as a class of allophones). The results also provided evidence that some members of a category were more prototypical than others. In the category [+anterior], for instance, bilabial consonants were more frequently and readily included than alveolar consonants. In addition, the status of /w/ proved to be ambiguous, since it was identified as a member of both the target sets [+anterior] and [-anterior].

Other phonological units that have been explored by means of the CF technique are the CVC syllable (Yoon and Derwing, 1995) and some of its presumed subcomponents, such as the onset (initial C) and rime (VC), or the body (CV) and coda (final C), as reported in Derwing and Wang (1995) for Taiwanese (Min) and in Yoon & Derwing (2001) for Korean. The great flexibility of this technique is also revealed by its successful adaptation to the study of notions as diverse as sentence types (Baker, Prideaux & Derwing 1973) and phonological rules (Wang & Derwing 1986).

3.6. Recall and Recognition Experiments

Many psycholinguistic experiments have manipulated what we will call memory-dependent variables. These have used some form of recall or recognition task, including proactive interference, false memory for words, and sentence or propositional recall. All these tasks have been employed in

---

[28] See Vance (1987) for an explication of the mora notion, particularly as it applies to Japanese.
[29] For an explication of these feature names, see Chomsky & Halle (1968).

the study of different levels of linguistic representation—including phonological, morphological, syntactic, and semantic—with varying degrees of success. In this section we will discuss only a fraction of these studies, focusing on those particular levels which we believe have been most effectively manipulated in off-line tasks involving memory.

3.6.1. A historical perspective on memory and psycholinguistics

It is interesting to note that the very first experimental studies of memory—those of Ebbinghaus' (1885/1964)— can also be considered perhaps the first "psycholinguistic" studies, judging by the nature of the linguistic materials employed and some of the later studies they motivated (see Jenkins 1985). It is clear that Ebbinghaus can be considered a pioneer in the cognitive sciences, using scientific methods that were far ahead of his time. It would be something of a stretch, however, to think of him as the first psycholinguist, as he was not interested in language *per se*. What Ebbinghaus did do was to develop a method for the study of recall—and forgetting—employing lists of nonsense CVC syllables as stimuli. His studies relied on his own learning and relearning of lists of those syllables, followed by series of recall sessions at different time periods after list learning. These studies led to the now famous "forgetting curve", showing the dramatic decrease in retention as a function of the time between study and recall. His goal was to try to understand the "pure" aspects of memory research (or what later became known as "verbal learning"). His deliberate use of nonsense materials was an attempt to avoid mnemonic strategies (such as associations and elaborations) that a list of words would trigger**.**

Though harshly criticized by Kintsch (1985) and others, Ebbinghaus' work with nonsense syllables strongly influenced Peterson and Peterson (1959) on the role of rehearsal on short-term memory, while the latter served in turn as the basis of the work of Wickens (1970; Wickens, Dalezman, & Eggemeier 1976) that will play such a prominent role in the body of our discussion below. It should also be noted that Jenkins (1985) found applications of Ebbinghaus' work in the study of phonological units, such as syllables (see also Greenberg & Jenkins 1966).

Another pioneer of memory research also with direct links with language research was Bartlett (1932), who, as a reaction to Ebbinghaus' approach, focused on how semantic representations from sentences and texts are retained over time. As we will see later in this section, his work led to the development of many modern studies on the nature of semantic representations.

3.6.2. Normative studies and lexical associations

Some now classical experiments in memory research are illustrative of the period of approximation of the study of memory to the psychological study of language. The studies and techniques we discuss in this section deal primarily with short-term (STM) or working memory but also draw on (and inform about) long-term (LTM) memory representations.[30]

---

[30] For the present purposes we will not make a distinction between working memory and short-term memory. We will assume a standard "memory systems" approach with interacting but functionally distinct short-term (STM) and long-term (LTM) memory systems.

Off-line techniques that probe mnemonic representations have been heavily employed in normative studies. Among these we can mention word-association tasks, which are taken to reflect the interconnections between words in the mental lexicon; categorization tasks, which serve to determine the strength of the semantic or conceptual relations between words; and scaling techniques, as discussed in section 3.1 above. These types of tasks tap the long-term memory representation for words and their meanings and thus serve largely as practical aids in the preparation of linguistic materials for experiments that involve more stringent experimental protocols, such as lexical priming (e.g., Swinney 1979).

Word association tasks have also proved to be revealing of the nature of the semantic, morphological, and phonological relations that are established between words, beyond their use in normative studies. In fact, one of the very first studies on semantic associations, done by Kent and Rosanoff (1910), was devised to show the types of semantic relations obtained between words in LTM. They presented over a thousand participants with a list of words, one by one, and asked them to say the first word that came to mind (with the exception of the word stimulus itself). Although there was an enormous variability in the responses, Kent and Rosanoff noticed that the majority of responses clustered on only a few words. They also noted that many word-association responses could be classified according to what they called "logical relations" between the items, which could be revealing of the strength of their semantic (e.g., *health-sicknesss*) or phonological (e.g., *health-wealth*) connections. Numerous word association norms have been created following the same paradigm used by Kent and Rosanoff (see, e.g., Nelson, McEvoy & Schreiber 1998; the norms themselves are available on the web at http://www.usf.edu/FreeAssociation/).

We now turn to studies that have employed other types of mnemonic techniques—focusing in particular on recall of words and sentences—which have played a more central role in experimental psycholinguistics.

3.6.3. Lexical Units and Memory Representation

Studies on lexical representation—beyond lexical associations—relying on memory techniques abound. For this reason we deliberately impose some constraints on the scope of our discussion. We will focus on patterns of recall (and forgetting) and how we can advance our understanding of linguistic representation by relying on well-established paradigms in memory research and by manipulating particular linguistic variables. We will discuss three general types of experimental paradigms, providing a few examples of experimental studies and findings. Our goal is twofold: to offer the reader information about key experimental procedures and show their usefulness for the investigation of particular types of linguistic issues.

3.6.3.1. Proactive Interference

The "release from proactive interference" (PI) paradigm is perhaps one of the best-known techniques used in memory research to probe the nature of lexical-semantic relations. The development of this technique is due mostly to Wickens (1970).

The technique relies on four major findings of STM research: (1) that STM has severe capacity limitations, holding an estimated seven items or "chunks" of similar material at any given

point in time (Miller 1956); (2) that items or chunks need to be rehearsed to be kept "alive" in STM (Peterson & Peterson 1959); (3) that recall of items decays rapidly after a few seconds (with varying durations depending on the chunking properties of the material) (Peterson & Peterson, 1959); (4) and that memory retention is subject to proactive interference, i.e., that material learned in one session or trial affects one's capacity to learn or retain new material belonging to the same general category in successive trials (Wickens 1970).

In the original PI test developed by Wickens, participants are presented with triads of words, one by one. Following the presentation of the three words comes a distractor task (usually a three-digit number from which participants have counted backwards in threes). After the distractor task, participants have to try to recall the words presented in the beginning of the trial. The task sequence—word triads, distractor, and recall—is repeated three more times. What Wickens found is that when words belong to the same semantic category (e.g., when all words name fruits), recall accuracy declines dramatically from trial to trial, going from close to 100% in the first trial to about 30% in the fourth trial, thus constituting a case of proactive interference. However, when items in the fourth trial belong to a different category (e.g., animals instead of fruits), recall is as accurate as in the first trial—a case of "release" from proactive interference.

In some studies, Wickens used words from opposite ends of Osgood's semantic differentials (as discussed above) and obtained high release values (i.e., high recall in the fourth trial as compared to the third trial), when the shift was from words at one polar end of a scale (e.g., "negative" words like *hate, fire, kill*) to the other (e.g., "positive" words like *able, mother, wise*). Wickens and his colleagues also studied a variety of "shifts" with more linguistic (i.e., structural) variables such as grammatical category (e.g., from verbs to adjectives, or from verbs to nouns), gender (masculine to feminine), tense (infinitive to past), number of syllables, and number of phonemes. In most of these studies, at least small amounts of proactive interference and release were observed (see Wickens et al. 1976 for a review).

This test has been used more recently to study further semantic properties of lexical items, beyond the similarity of their referential attributes. Marques (2000, 2002) used a variant of Wickens' paradigm to investigate the nature of semantic memory. His innovations included the use of both words and pictures and the introduction of a cue before the triad of items within each trial to probe for particular aspects of the semantic relation between items (e.g.,"means of transportation"), with a triad like *bicycle, airplane, bus* (in the build-up trials) followed by another like *horse*, *elephant, camel* (in the release trial). What he found was no release occurred when the cue was used, unlike the control condition, which did not involve a cue to indicate that the animals could also be construed as belonging to the same category as the vehicles. What is important to notice from a methodological point of view is that the PI-release technique can be used to investigate very specific linguistic hypotheses, such as the possible role of features in semantic memory.

In a recent series of experiments investigating the nature of verb-conceptual representation, de Almeida and Mobayyen (2004) and Mobayyen and de Almeida (in press) employed a paradigm similar to Wickens' and Marques' but with different classes of verbs, such as lexical causatives (e.g., *bend*, *crack, grow*) and morphological causatives (e.g., *thicken*, *darken, fertilize*), and perception verbs (*see, hear, smell*). What de Almeida and Mobayyen investigated was whether verbs are represented in semantic memory in terms of their decompositional template features (as proposed by Jackendoff 1990; see also Rappaport Hovav & Levin 1998) or whether they are represented in terms of their categorical relations (de Almeida 1999b). Notice that the main

difference between these verb classes is in their predicted semantic complexity. While lexical and morphological causatives are thought to be represented by the same semantic template, as shown in (1), morphologically simple and morphologically complex perception verbs are taken to have different templates, as shown in (2) and (3).

(1) a. [*x* ACT [CAUSE [*Y*]]]
b. The gardener grew the plants
c. The gardener fertilized the plants

(2) a. [*x* PERCEIVE *y*]
b. The gardener smelled the plants

(3) a. [AGAIN [*x* PERCEIVE *y*]]
b. The gardener re-smelled the plants

Notice that the lexical causatives in (1b) and the morphological causatives in (1c), where causation is overtly marked with the suffix *-ize*, are posited to have the same template structure, unlike their morphologically simple and complex perceptual counterparts in (2) and (3).

In the de Almeida and Mobayyan study, groups of participants received four triads of verbs, with the three triads of the build-up phase corresponding to one class of verbs (e.g., lexical causatives) and the fourth triad corresponding to common semantic categories (e.g., fruits). In this study, proactive interference was obtained for all classes of verbs that were related semantically, but the effect was stronger in the case of morphologically complex verbs. This result suggests that lexical and morphological causative concepts may not be represented by the same type of complex semantic structure, but possibly by structures that correspond to the morphological complexity of their lexical items.

Another finding that may be revealing of the use of this technique for the investigation of semantic properties of words is that the only verb class that did not produce PI build-up was that of intransitives (e.g., *sneeze*, *sweat*), which have similar argument (syntactic) structure but are not usually clustered in terms of any semantic dimension.

In a related study but with sentences and auditory presentation, Mobayyen and de Almeida (in press) found no significant effects of PI build-up, but found differences in memorability for sentences containing different verb classes: Sentences containing lexical and morphological causatives (as (1b) and (1c)) were recalled better than sentences with simple and complex perception verbs (as in (2b) and (3b)). This suggests that causative verbs create longer-lasting memory codes, perhaps because of the greater number of inferences that they trigger (see also Breedin, Saffran & Schwartz 1997 for an alternative explanation).

From a methodological standpoint, these last results cast doubts on the usability of the PI technique with materials more complex than simple lists of words. Another possibility is that the PI technique is not suitable for auditory presentation. For all of this, the technique is certainly established well enough to be used with a variety of more fine-tuned linguistic variables, but in particular for the manipulation of semantic aspects of word encoding.

3.6.3.2. False Memories

The proactive interference technique presented above relies on the notion that related lexical items (as represented in LTM) produce confusions when activated in STM, leading to false recall within a trial of items that were presented in a previous trial. Semantic interference of this type can also be obtained with paradigms that attempt to induce source errors. In a false recognition task originally developed by Deese (1959), Buchanan and colleagues (e.g., Buchanan, Brown, Cabeza, & Maitson 1999) investigated whether lexical items related by association (e.g., *pie, core, tree*) or by feature sets (e.g., *banana, grapefruit, peach*) would produce "better" false recall of a foil-target such as *apple*.

In this paradigm, during the "study" phase, words within a list (usually about 10) are presented one-by-one for two seconds. After this phase, participants are given a distractor task for a short period of time (2 min in Buchanan et al.'s study), followed by a test phase. In the test phase a list of words is presented, again one-by-one, and participants have to indicate whether the word was in the original study list or whether it is a new word. The critical condition is the presentation of the foil target (the word that is associated or categorically-related to the words in the study lists). The degree of false positives (i.e., the proportion of false recalls of the target foil) is revealing of the nature of the relation obtained between the list words and the target.

In their study, Buchanan et al. (1999) obtained more false recalls of the target *apple* when subjects were presented a list of associated words than when a categorical/featural list was used. Their suggestion was that lexical associations in the mental lexicon are stronger than semantic relations obtained via semantic features. Regardless of the strength or generalizibility of their results, what is important for our present purposes is that false memory techniques can be quite useful in clarifying the nature of semantic relations. Westbury, Buchanan, and Brown (2002) also report on using this technique for the exploration of phonological properties of lexical items.

3.6.3. Memory for sentences: Probing the nature of semantic structure

We now turn to paradigms used in the study of the semantic properties of sentences. We will focus on recall and recognition techniques, all of which rely on the assumption that remembering a sentence is largely a function of the complexity of its semantic representation in memory.

One of the most influential studies on the nature of semantic encoding of sentences was developed by Sachs (1967). She presented participants with recorded stories. At different points during the presentation of each story, subjects were presented with a sentence and asked to verify whether or not the sentence came verbatim from the story. She manipulated two main variables: lag between the original sentence in the story and the sentence presented for verification (with 0, 80, and 160 intervening syllables between original sentence presentation and the probe), and type of sentence probe presented for verification. For this latter variable, Sachs presented participants with sentences such as those shown in (4), where (4a) is identical to the sentence in the story, (4b) is similar in meaning but with a change in the syntactic/clausal structure, (4c) is similar in meaning but with a change from active to passive voice, and (4d) is different in meaning.

(4) a. He sent a letter about it to Galileo, the great Italian scientist
b. He sent Galileo, the great Italian scientist, a letter about it

c. A letter about it was sent to Galileo, the great Italian scientist
d. Galileo, the great Italian scientist, sent him a letter about it

Results showed that, at 0 delay, subjects correctly detected (in about 90% of the trials) which sentence was originally presented. Later, 180 syllables after the original sentence presentation, however, they could not distinguish the original from the sentences with syntactic and voice changes (about 55% accuracy), but they were quite sure that the sentence with a meaning change was not the original one (80% accuracy). [31]

The main methodological lesson from the study—and its main psycholinguistic import—is that memory for syntactic structures (and verbatim information) is short-lived, but that our memory for semantic content remains relatively intact over time.

The study of semantic representation—and the nature of what is retained in memory—was also advanced by some ingenious sentence recall experiments done by Kintsch (1974). In the technique that Kintsch used, participants were presented aurally with blocks of five sentences, after which they had to recall in writing as much as possible from each sentence. It is important to notice that of the five sentences that participants heard in each block, only the three in the middle were experimental sentences: the first and the last were used as foils to avoid a "serial position" effect (since items at the beginning and the end of list are generally recalled better than those in the middle). This looks like a simple off-line paradigm, but the cleverness of Kintsch's study was in the manipulation of the types of sentences he used, which involved systematically controlling the number of content words and varying the number of propositions. For instance, sentences such as those in (5) have the same number of content words (three) but convey (or are represented by) a different number of propositions.

(5) a. The policeman issued a summons
b. The crowded passengers complained

While (5a) conveys one proposition, as represented in (6a), (5b) conveys two propositions, as shown in (6b).

(6) a. [ISSUE [POLICEMAN, SUMMONS]]
b. [[[CROWDED [PASSENGERS]] & [COMPLAINED [PASSENGERS]]]

Thus although both sentences are the same in terms of the number of content words they contain, they are different in terms of their semantic or propositional complexity. Using sentences containing as many as four content words and up to three propositions, Kintsch found that the more propositionally complex a sentence was, the worse it was recalled in full. However, more sentence subparts were recalled in the propositionally complex sentences than in the propositionally simple ones.

Kintsch's paradigm can be adapted to investigate a range of questions on semantic representation of words and sentences. The issue of propositional complexity engendered by

[31] In a replication of Sachs' study with deaf-mute participants, Hanson and Bellugi (1982) found similar effects using American Sign Language, thus showing that the paradigm is valid across populations and modalities.

different verb classes was investigated in a study by de Almeida and Turbide (2004). Like Kintsch, they employed sentences with varying number of content words and propositions, but also manipulated the semantic complexity of verbs, using the same verb types shown in examples (1) - (3) above, as well as other factors. Participants recalled full sentences containing lexical causatives and simple perception verbs better than they recalled sentences with morphological causatives. Also, there was no difference in recall between lexical causatives and morphologically simple perception verbs. This corroborates the results found earlier using the PI technique, supporting the "atomic" theory of lexical-conceptual representation elaborated in Fodor, 1998 (see also de Almeida 1999ab; de Almeida & Fodor, 2004).

Although many studies have shown that participants retain the propositional content of sentences (which in fact can be equated with the notion that some form of compositional semantic structure is retained), a study conducted by Bransford, Barclay, and Franks (1972) added a new dimension to the nature of semantic memory for sentences. They presented participants with sentences containing different locative prepositions, as illustrated in (7).

(7) a.Three turtles rested beside a floating log, and a fish swam beneath them
b.Three turtles rested on a floating log, and a fish swam beneath them

They argued that although the two sentences have the same syntactic and semantic structures, the types of conceptual information they trigger are different. In (7b) but not in (7a) listeners can infer that the fish swam beneath the log. Thus they predicted that if participants construct more than a simple compositional meaning of the sentences they hear, they should recognize the difference between (7) and similar sentences with a final pronoun change, as in (8).

(8) a. Three turtles rested beside a floating log, and a fish swam beneath it
b. Three turtles rested on a floating log, and a fish swam beneath it

In their procedure, participants were presented aurally with a series of 21 study sentences. They were then given a short break and presented with another series of sentences for recognition. Among the sentences presented in the recognition phase of the experiment were sentences identical to those in the study phase, and with sentences such as those in (8) with a pronoun change. They predicted that if sentence recall was a function of the semantic or propositional representation of the sentence, then participants would be inclined to falsely accept both sentences in (8) as having been presented in the study phase. However, if participants "construct" a model of the state of affairs described by the original sentence, then they would only falsely recognize (8b), because only that sentence allows for the interpretation that the fish swam beneath the log. They found support for their view, giving rise to the theory "mental models" (Johnson-Laird 1983).

In the beginning of the present section on memory, we alluded to the reactions to Ebbinghaus' research tradition brought about by Bartlett (1932) and his followers. The focus of the "new" research tradition was on the nature of semantic representations, mostly on what is retained in memory from linguistic tokens (sentences, paragraphs, stories, etc.). It is interesting to note that this line of investigation—with its focus on the meaning of learned or stored material—has led to converging evidence that relied on quite disparate experimental paradigms.

4. On Validating Experimental Techniques

In this chapter we have described quite a number of different experimental techniques, illustrating each with just a few specific examples, often selected, for both convenience and familiarity, from our own laboratories. One issue that we have not yet discussed at all, however, is the question of why we need so large a variety of tasks, rather than focusing our attention on one best "all-purpose" approach. There are two main reasons for this. One reason, of course, is that some tasks may simply better lend themselves to the investigation of a particular problem than others. Thus segmentation and concept formation tasks seem more naturally suited to answering questions about units and categories, while rating and recall tasks seem more appropriate when questions on matters of global similarity or "distance" are involved, and no one "ideal" technique is obvious for both.[32]

A second important reason for maintaining a large inventory of experimental techniques, however, is that, inescapably, the technique or task is itself a factor in each and every experiment in which it is employed, and there is always the danger that this factor might have a distorting or even profound influence on the results obtained. Sometimes this effect can occur expectedly and thus relatively innocuously, as when (as in the case of the proverbial blind men and the elephant) a given approach reveals only a small part of the total picture, leaving the other parts to be revealed by other studies. In other cases, however, the influence can be quite insidious, as in the classic case of the infamous "experimental artifact". This is a situation in which either the task itself so influences the results of an experiment as to render them completely invalid, or else where some extraneous factor not considered in the original research design is actually the operative or most influential one. Since either alternative can easily fail to be noted by the experimenter and can thus lead to the misinterpretation of any set of experimental results, such artifacts must always be guarded against in research that puts a premium on control over naturalness, as experimental research typically does (cf. the earlier discussion on on-line vs. off-line tasks). In either situation, the best safeguard against such extraneous influences is by means of cross-methodological verification[33], which seeks to eliminate the influence of a particular task or set of stimuli by exploring a given phenomenon through a range of qualitatively different experimental approaches (see Derwing 1997 and de Almeida 1999a for extensive discussions of this issue.).

In some cases, however, the validity issue can be dealt with much more directly, by comparing the results of an experiment with those obtained under more realistic, natural conditions. One criticism of the early Berko-type experiments, for example, was that the task of inflecting made-up words produced a "strangeness effect" and could not therefore be trusted to yield valid results about real-language knowledge and abilities (Kiparsky & Menn 1977). In order to counter this objection, Rollins (1980) carried out a "validity check" study in which pairs of stuffed animals were given made-up names (like Berko's original "wug") and were presented to pairs of children in a play situation that involved stuffed counterparts of familiar animals, as well. Inevitably, the children asked what the unfamiliar animals were called, soon learning and using their names just as naturally as the previously known ones; and sooner or later they invented games that also involved using the plural forms of these names, which the experimenter/observer carefully noted down for

---

[32] Notwithstanding this general observation, rating tasks have been used to assess categorical issues (e.g., Derwing & Nearey 1986, pp. 48-54) and concept formation could certainly be adapted to the study of semantic similarity (see Derwing 1973, p. 318 for a hypothetical example that illustrates the logic of such an approach).

[33] In his discussion of criteria for good measuring instruments, Osgood (1952) actually defines validity in such practical, utilitarian terms, as follows: "The data obtained should be demonstrably covariant with those obtained with some other, independent [measure]" (p. 183).

each child. These observations were then compared with the results of two controlled Berko-type experiments (one presented prior to the naturalistic "play" study and one after), using the standard two-dimensional pictures rather than the stuffed animals. The results were that the three data sets corresponded almost perfectly, showing that the results of the experiment were just as good as those observed in the naturalistic situation, where the children had no idea that they were being tested at all. While not all experiments or tasks so readily lend themselves to validation as this one does, of course, considerations of "ecological validity" (cf. Neisser 1976; Libben & Libben in press) must continue to be addressed in psycholinguistic work as much as circumstances allow, if laboratory findings are to be taken seriously as discoveries about the real (internal) world of the mind.

Finally, we have the case of the experimental artifact, which is a situation in which some uncontrolled factor may have so distorted the results of an experiment as to render them potentially invalid. As Ohala (1986) points out, however, a great strength of the experimental approach is that a new experiment can always be designed that controls for any factor that might be suspected of having such an effect. Clearly, as illustrated by several comments already made in previous sections of this chapter (beginning with the *rich-pitch* example noted in fn. 11 above), orthographic knowledge has played the role of "prime suspect" in almost all experimental research on phonological units to date (see Derwing, Nearey & Dow 1986, Derwing & Dow 1987, and Derwing 1992b, for other examples and extensive further discussion). In the recent study of Yoon and Derwing (2001), therefore, quite despite the fact that several qualitatively different experimental approaches had already been used to explore the status of the rhyme vs. body units, and all with the same basic findings, it was felt that the question of the role of orthographic knowledge had to be dealt with before the research could be considered in any sense complete.

For this purpose therefore, an entirely new variant of the general memory/recall approach was invented that could be administered both to literate and to preliterate children, in order to explore the factor of orthographic knowledge itself. This new technique involved assigning CVC nonce names to pictures of imaginary animals and comparing the ability of children to recall small sets of such names which rhymed (i.e., which all shared a common -VC element) with their ability to recall those that shared a common body (or CV-element). The results showed that, while a repeated rhyme element proved to be a better help in recalling such names for English-speaking children, Korean children found the latter, body-sharing names to be easier to remember, whether they could read or not, thus effectively removing the orthographic factor from consideration as the critical one. Since this finding was consistent with all of the other studies reported in the paper, which involved techniques as varied as word-blending, global ratings for sound similarities, unit inversion, and even concept formation, it seemed to cement the case that the body rather than the rhyme was the intrasyllabic unit of consequence for speakers of Korean.

Needless to say, none of the techniques described here are sacrosanct, and none lead magically to incontrovertible findings, just as no one experiment can ever hope to yield absolutely definitive results on even one small question. The truths about the inner workings of the human language mechanism are not displayed for ready observation and will only be arrived at through the painstaking accumulation of evidence, using a variety of experimental and other empirical techniques, many of which no doubt have yet to be invented. For all of this, however, thanks to the small repertoire of techniques described and illustrated here (and elsewhere in this volume), real progress has begun to be made, as even some of the simplest and most basic experimental techniques have proven capable of providing a glimpse into the world of psychological reality that

no purely descriptive or theoretical account alone, however sophisticated, could ever hope to match. In short, we are at last beginning to replace the endless rounds of theory construction and deconstruction in linguistics with real answers, and surely many more are soon to follow, as theoretical and experimental linguistics come to be understood as simply two sides of the same coin, with neither of much real value without the other.

5. References

Anisfeld, M., & Gordon, M. (1968). On the psychophonological structure of English inflectional rules. *Journal of Verbal Learning and Verbal Behavior, 7*(6), 973-979.

Anisfeld, M., & Tucker, G. R. (1968). The English pluralization rules of six-year-old children. *Child Development, 38*(4), 1201-1217.

Baker, W. J., & Derwing, B. L. (1982). Response coincidence analysis as evidence for language acquisition strategies. *Applied Psycholinguistics, 3*, 193-221.

Baker, W. J., Prideaux, G. D., & Derwing, B. L. (1973). Grammatical properties of sentences as a basis for concept formation. *Journal of Psycholinguistic Research, 2*(3), 201-220.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: The University Press.

Beinert, R. A., & Derwing, B. L. (1993). Segment, rime, syllable, tier or root? Evidence from global sound similarity judgements in Arabic. *Proceedings of the Ninth Eastern States Conference on Linguistics (ESCOL '92)*; pp. 1-10). Ithaca, NY: Cornell University.

Bendrien, T. A. (1992). *Sound similarity judgements in English CVCs*. Unpublished undergraduate honors thesis, University of Alberta, Edmonton, Canada.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150-177.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart, & Winston.

Bolton, N. (1977). *Concept formation*: Oxford: Pergamon Press.

Boring, E. G. (1950). *A history of experimental psychology*. New York: Appleton-Century-Crofts.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*, 193-209.

Breedin, S. D., Saffran, E. M., & Schwartz, M. F. (1998). Semantic factors in verb retrieval: An effect of complexity. *Brain and Language, 63*, 1-31.

Bruner, J.S., Goodnow, J.J., and Austin, G.A. (1956). *A study of thinking*. New York: Wiley.

Buchanan, L., Brown, N. R., Cabeza, R., & Maitson, C. (1999). False memories and semantic lexicon arrangement. *Brain and Language, 68*(1-2), 172-177.

Bullard, J. D. (1979). *Mechanisms of speech disguise*. Unpublished undergraduate honors thesis, University of Alberta, Edmonton, Canada.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Day, R. S. (1973). On learning "secret languages." *Haskins Laboratories Status Report on Speech Research, 34,* 141-150

de Almeida, R. G. & Fodor, J. A. (2004). *Against lexical decomposition again: Some psycholinguistic evidence*. Manuscript in preparation.

de Almeida, R. G. & Libben, G. (in press). Changing morphological structures: The effect of sentence context on the interpretation of structurally ambiguous English trimorphemic words. *Language and Cognitive Processes*.

de Almeida, R. G. & Mobayyen, F. (2004). *Semantic memory organization for verb concepts: Proactive interference as a function of content and structure*. Manuscript submitted for publication.

de Almeida, R. G. & Turbide, J. E. (2004). *Recall of sentences with propositionally-complex verbs: Evidence for the atomicity of causatives*. Manuscript in preparation.

de Almeida, R. G. (1999a). *The representation of lexical concepts: A psycholinguistic inquiry*. Unpublished doctoral dissertation, Rutgers University, New Jersey.

de Almeida, R. G. (1999b). What do category-specific semantic deficits tell us about the representation of lexical concepts? *Brain and Language, 68* (1-2), 241-248.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.

Deese, J., & Hulse, S. (1967). *The psychology of learning*. New York: McGraw-Hill.

Dennis, D. (1988). *Rule governed behaviour in English inflectional morphology*. Unpublished master's thesis, University of Alberta, Edmonton, Canada.

Derwing, B. L. (1973). *Transformational grammar as a theory of language acquisition: A study in the empirical, conceptual and methodological foundations of contemporary linguistics*. London & New York: Cambridge University Press.

Derwing, B. L. (1976). Morpheme recognition and the learning of rules for derivational morphology. *The Canadian Journal of Linguistics, 21*, 38-66.

Derwing, B. L. (1980b). English pluralization: A testing ground for rule evaluation. In G. D. Prideaux, B. L. Derwing, & W. J. Baker (Eds.), *Experimental linguistics: Integration of theories and applications* (pp. 81-112). Ghent, East-Flanders, Belgium: E. Story-Scievtia.

Derwing, B. L. (1992a). A "pause-break" task for eliciting syllable boundary judgments from literate and illiterate seakers: Preliminary results for five diverse languages. *Language and Speech, 35* (1,2), 219-235.

Derwing, B. L. (1992b). Orthographic aspects of linguistic competence. In P. Downing, S. D. Lima, & M. Noonan (Eds.), *The Linguistics of Literacy: Vol. 21*. (pp. 193-211). Amsterdam & Philadelphia, PA: John Benjamins.

Derwing, B. L. (1997). Testing phonological universals in the laboratory. In P. M. Bertinetto, L. Gaeta, G. Jetchev & D. Michaels (Eds.), *Certamen Phonologicum III (Papers from the Third Cortona Phonology Meeting, April, 1996*; pp. 45-65). Torino, Piedmont, Italy: Rosenberg & Sellier.

Derwing, B. L. & Baker, W. J. (1979). Recent research on the acquisition of English morphology. In P. J. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development* (pp. 209-223). London & New York: Cambridge University Press.

Derwing, B. L., & Baker, W. J. (1980). Rule learning and the English inflections (with special emphasis on the plural). In G. D. Prideaux, B. L. Derwing, & W. J. Baker (Eds.), *Experimental linguistics: Integration of theories and applications* (pp. 247-272). Ghent, East-Flanders, Belgium: E. Story-Scientia.

Derwing, B. L., & Dow, M. L. (1987). Orthography as a variable in psycholinguistic experiments. In P. A. Luelsdorff (Ed.), *Orthography and phonology* (pp. 171-185). Amsterdam: John Benjamins.

Derwing, B. L., Dow, M. L., & Nearey, T. M. (1989). Experimenting with syllable structure. In J. Powers & K. de Jong (Eds.*), Proceedings of the Fifth Eastern States Conference on Linguistics (ESCOL '88;* pp. 83-94). Columbus: The Ohio State University.

Derwing, B. L., & Nearey, T. M. (1986). Experimental phonology at the University of Alberta. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (pp. 187-209). Orlando, FL: Academic Press.

Derwing, B. L., & Nearey, T. M. (1991). The "vowel-stickiness" phenomenon: Three sources of experimental evidence. *Proceedings of the XIIth International Congress of Phonetic Sciences: Vol. 2*. (pp. 210-213). Aix-en-Provence, France: ICPhS.

Derwing, B. L. & Wang, H. S. (1995). Concept formation as a tool for the investigation of phonological units in Taiwanese. *Proceedings of the XIIIth International Congress of Phonetic Sciences: Vol. 3*. (pp. 362-365). Stockholm: KTH & Stockholm University.

Derwing, B. L., & Wiebe, G. (1994). Syllable, mora or segment? Evidence from global sound similarity judgements in Japanese. In P. Koskinen (Ed.), *Proceedings of the 1994 Annual Conference of the Canadian Linguistic Association* (pp. 155-163). Toronto: Toronto Working Papers in Linguistics, Linguistic Graduate Course Union.

Derwing, B. L., Nearey, T. M., & Dow, M. L. (1986). On the phoneme as the unit of the "second articulation." *Phonology Yearbo3*, 45-69.

Derwing, B. L., Smith, M. L., & Wiebe, G. E. (1995). On the role of spelling in morpheme recognition: Experimental studies with children and adults. In L. Feldman (Ed.), *Morphological Aspects of Language Processing: Cross-Linguistic Perspective* (pp. 3-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dominowski, R. L. (1970). Concept attainment. In M. H. Marx (Ed.), *Learning: Interactions* (pp. 152-191). New York: MacMillan.

Dow, M. L., & Derwing, B. L. (1989). Experimental evidence for syllable-internal structure. In R. Corrigan, F. Eckman, & M. Noonan (Eds.), *Linguistic categorization* (pp. 81-92). Amsterdam: John Benjamins.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology*. New York: Dover.

Ehri, L. C., & Wilce, L. S. (1980). The influence of orthography on reader's conceptualization of the phonemic structure of words. *Applied Psycholinguistics, 1*, 371-385.

Esper, E. A. (1925). A technique for the experiment investigation of associative interference in artificial linguistic material. *Language Monograph I.*

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Fodor, J. A., Garrett, M. F., Walker, E. C. T., & Parkes, C. H. (1980). Against definitions. *Cognition, 8*, 263-367.

Foss, D. J. (1968). An analysis of learning in a miniature linguistic system. *Journal of Experimental Psychology, 76*(3), 450-459.

Gergely, G., & Bever, T. G. (1986). Related intuitions and the mental representation of causative verbs in adults and children. *Cognition, 23*(3), 211-277.

Grace, H. K. & Rosanoff, A. J. (1910). A study of association in insanity. *American Journal of Insanity, 67*, 37-96, 317-390.

Gray, V. A., & Cameron, C. A. (1980). Longitudinal development of English morphology in French immersion children. *Applied Psycholinguistics, 1*(2), 171-181.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English, I and II. *Word, 20*, 157-178.

Greenberg, J. H., & Jenkins, J. J. (1966). Studies in the psychological correlates of the sound system of American English, III and IV. *Word, 22*(1-3), 207-242.

Halle, M. (1964). On the basis of phonology. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language: Readings in the philosophy of language* (pp. 324-333). Englewood Cliffs, NJ: Prentice Hall.

Hanson, V. L., & Bellugi, U. (1982). On the role of sign order and morphological structure in memory for American Sign Language sentences. *Journal of Verbal Learning and Verbal Behavior, 21*(5), 621-633.

Heringer, J. (1970). Research on quantifier-negative idiolects. Papers from the Sixth Proceedings from the Regional Meeting of the *Chicago Linguistic Society*, *6*, 287-296.

Hombert, J. M. (1986). Word games: Some implications for analysis of tone and other phonological constructs. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental Phonology* (pp. 175-186). Orlando, FL: Academic Press.

Hooper, J. B. (1972). The syllable in phonological theory. *Language, 48*, 525-540.

Hooper, J. B. (1976). *An introduction to natural generative phonology*. New York: Academic Press.

Hörmann, H. (1971). *Psycholinguistics: An introduction to research and theory* (H. H. Stern, Trans.). New York: Springer-Verlag.

Hymes, D., & Fought, J. (1975). American structuralism. In T. A. Sebe(Ed.), *Current trends in linguistics:Vol. XIII. Historiography of linguistics* (pp. 903-1176). The Hague & Paris: Mouton.

Innes, S. J. (1974). *Developmental aspects of plural formation in English*. Unpublished master's thesis, University of Alberta, Edmonton, Canada.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge: MIT Press.

Jaeger, J. J. (1980a). *Categorization in phonology: An experimental approach* . Unpublished doctoral dissertation, University of California, Berkeley.

Jaeger, J. J. (1980b). Testing the psychological reality of phonemes. *Language and Speech, 23*(3), 233-253.

Jaeger, J. J. (1986). Concept formation as a tool for lexical research. In J. J. Ohala & J. J. Jaeger (Eds.), *Eperimental Phonology* (pp. 211-237). Orlando, FL: Academic Press.

Jaeger, J. J., & Ohala, J. J. (1984). On the structure of phonetic categories. *Proceedings of the 10th Annual Meeting of the Berkeley Linguistics Society* (pp. 15-26). Berkeley, CA: Berkeley Linguistics Society.

Jenkins, J. J. (1985). Nonsense syllables: Comprehending the "almost incomprehensible variation." *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(3), 455-460.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Kac, M. B. (1980). In defense of autonomous linguistics. *Lingua, 50*, 243-245.

Kahn, D. (1976). *Syllable-based generalizations in English phonology*. Bloomington: Indiana University Linguistics Club.

Kintsch, W. (1974). *The Representation of Meaning in Memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kintsch, W. (1985). Reflections on Ebbinghaus. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 461-463.

Kiparsky, P., & Menn, L. (1977). On the acquisition of phonology. In J. Macnamara (Ed.), *Language learning and thought* (pp. 47-78). New York: Academic Press.

Kling, J. W., & Riggs, L. A. (1971). *Woodworth & Schlosberg's experimental psychology* (3rd ed.). New York: Holt, Rinehart & Winston.

Lakoff, G. (1982). Categories: An essay in cognitive linguistics. In The Linguistic Society of Korea (Ed.), *Linguistics in the morning calm*, (pp. 139-193). Seoul, Korea: Hanshin Publishing.

Laycock, D. (1972). Towards a typology of play-languages, or ludlings. *Linguistic Communications, 6*, 61-113.

Levelt, W. J. M. (1970). A scaling approach to the study of syntactic relations. In G. Flores-d'Arcais & W. J. M. Levelt (Eds.), *Advances in Psycholinguistics* (pp. 109-121). Amsterdam: North-Holland.

Libben G., & Libben, M. (in press). Ecological validity in experimental psycholinguistic research. *Thirtieth LACUS Forum*.

Liberman, I. Y., Schankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18*, 201-212.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* p. 55.

Marques, J. F. (2000). The "living things" impairment and the nature of semantic memory organisation: An experimental study using PI-release and semantic cues. *Cognitive Neuropsychology, 17*(8), 683-707.

Marques, J. F. (2002). An attribute is worth more than a category: Testing different semantic memory organisation hypotheses in relation to the living/nonliving things dissociation. *Cognitive Neuropsychology, 19*(5), 463-478.

McCawley, J. D. (1979). Some ideas not to live by. In J. D. McCawley, *Adverbs, vowels, and other objects of wonder* (pp. 234-246). Chicago & London: University of Chicago Press.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81-97.

Mobayyen, F. & de Almeida, R. G. (in press). The influence of semantic and morphological complexity of verbs on sentence recall: Implications for the nature of conceptual representation and category-specific deficits. *Brain and Cognition*.

Neisser, U. (1976). *Cognition and reality*. San Francisco: W.H. Freeman.

Nelson, D. L., & Nelson, L. D. (1970). Rated acoustic (articulatory) similarity for word pairs varying in number and ordinal position of common letters. *Psychonomic Science, 19*, 81-82.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/

Ohala, J. J. (1986). Consumer's guide to evidence in phonology. *Phonology Yearbook, 3*, 3-26.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49*, 197-237.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*(3), 193-198.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*(1-2), 73-193.

Prideaux, G. D. (1980). In rejection of autonomous linguistics. *Lingua, 50*, 245-247.

Rappaport Hovav, M. & Levin, B. (1998). Building verb meanings. In M. Butt & W. Geuder (Eds.), *The Projection of Arguments: Lexical and Compositional Factors* (pp. 97-134). Stanford, CA: CSLI Publications.

Rice, S., Libben, G., & Derwing, B. (2002). Morphological representation in an endangered, polysynthetic language. *Brain and Language, 81*, 473-486.

Rollins, W. C. (1980). *Laboratory vs. "free" testing situations in language acquisition research*. Unpublished undergraduate honors thesis, University of Alberta, Edmonton, Canada.

Rosch, E. H. (1973a). Natural categories. *Cognitive Psychology, 4*, 328-350.

Rosch, E. H. (1973b). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.

Rosch, E. H. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorisation* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ross, J. R. (1979). Where's English? In D. K. Charles, J. Fillmore, & W. S-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 127-163). New York: Academic Press.

Rumelhart, D. E. & McClelland, J. L. (1986). On Learning the Past Tenses of English Verbs. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol 2*. (pp. 216-271). Cambridge, MA: MIT Press.

Sachs, J. S. (1967). Recognition Memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics, 2*(9), 437-442.

Schirmeier, M. K., Derwing, B. L. & Libben, G. L. (in press). Morpheme-based lexical processing of German *ver*-verbs: The complementarity of on-line and off-line evidence. *Brain and Language*.

Schlesinger, I. M. (1977). Miniature artificial languages as research tool. In J. Macnamara (Ed.), *Language learning and thought* (pp. 251-260). New York: Academic Press.

Segalowitz, N., & de Almeida, R. G. (2002). Conceptual representation of verbs in bilinguals: Semantic field effects as a second-language performance paradox. *Brain and Language, 81*, 517-531.

Selkirk, E. O. (1984). On the major class features and syllable theory. In M. Aronoff & R.T. Oehrle (Eds.), *Language sound structure: Studies in phonology presented to Morris Halle by his teacher and students*. Cambridge: MIT Press.

Spencer, N. J. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research, 2* (2), 83-98.

Stemberger, J. P. & Lewis, M. (1986). Reduplication in Ewe: Morphological accommodation to phonological errors. *Phonology Yearbook, 3*, 151-160.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)Consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition, 15* (1-3), 49-74.

Treiman, R. (1985). Onsets and rimes as units of spoken syllables: Evidence from children. *Journal of Experimental Child Psychology, 39*, 161-181.

Treiman, R. (1986). The division between onsets and rimes in English syllables. *Journal of Memory and Language, 25*(4), 476-491.

Treiman, R. (1988). Distributional constraints and syllable structure in English. *Journal of Phonetics, 16*, 221-229.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language, 27*(1), 87-104.

Treiman, R., & Kessler, B. (1995). In defense of an onset-rime syllable structure for English. *Language and Speech, 38*(2), 127-142.

Valian, V. (1976). The relationship between competence and performance: A theoretical review. *CUNY Forum, 1*, 64-101.

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany: State University of New York Press.

Vitz, P. C., & Winkler, B. S. (1973). Predicting the judged "similarity of sound" of English words. *Journal of Verbal Learning & Verbal Behavior, 12*(4), 373-388.

Wang, H. S. & Derwing, B. L. (1986). More on English vowel shift: The back vowel question. *Phonology Yearbook, 3*, 99-116.

Wang, H. S., & Derwing, B. L. (1993). Is Taiwanese a "Body" Language. In C. Dyck (Ed.), *Proceedings of the 1993 Annual Conference of the Canadian Linguistic Association* (pp. 679-694). Toronto: Toronto Working Papers in Linguistics, Linguistic Graduate Course Union.

Westbury, C., Buchanan, L. & Brown, N. R. (2002). Remembering the neighbourhood: The effects of phonological overlap on false memory. *The Journal of Memory and Language,* 46, 622-651.

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review, 77*, 1-15.

Wickens, D. D., Dalezman, R. E., & Eggemeier, F. T. (1976). Multiple encoding of word attributes in memory. *Memory & Cognition, 4*(3), 307-310.

Wiebe, G. E. & Derwing, B. L. (1994). A forced-choice word-blending task for testing intra-syllabic break points in English, Korean and Taiwanese. In M.J. Powell (Ed.), *The Twenty-First LACUS Forum 1994* (pp. 142-151). Chapel Hill, NC: LACUS.

Yin, H., Derwing, B. L. & Libben, G. (2004). Branching preferences for large lexical structures in Chinese. Paper presented at the 4th International Conference on the Mental Lexicon, Windsor, Ontario. July.

Yoon, Y. B., & Derwing, B. L. (1994). Sound similarity judgements for Korean words by Korean and English speakers. In P. Koskinen (Ed.), *Proceedings of the 1994 Annual Conference of the Canadian Linguistic Association* (pp. 657-655). Toronto: Toronto Working Papers in Linguistics, Linguistic Graduate Course Union.

Yoon, Y. B., & Derwing, B. L. (2001). A language without a rhyme: Syllable structure experiments in Korean. *The Canadian Journal of Linguistics*, 46(3/4), 187-237.

Yoon, Y., & Derwing, B. L. (1995). Syllable saliency in the perception of Korean words. *Proceedings of the XIIIth International Congress of Phonetic Sciences: Vol 2*. (pp. 602-605). Stockholm, Sweden: KTH & Stockholm University.