# Distributed Morphology and Psycholinguistics

Christopher Hammerly and Roland Pfau
(University of Minnesota / University of Amsterdam)

## 1. INTRODUCTION

Generative approaches to language have, for the most part, been concerned with linguistic competence – that is, the knowledge of language for an ideal speaker/hearer. In line with this research agenda, Chomsky (1965: 3) excludes from consideration "such grammatically irrelevant conditions as memory limitations, distractions […], and errors". In contrast, linguistic performance, which is seen as a direct reflection of competence, is concerned with the actual use of language. Indeed, one year earlier, Katz (1964: 126) contended that "purely linguistic theories cannot succeed in predicting and explaining the fact of linguistic performance without making reference to the mental events, capacities, and processes of speakers".

In this chapter, we adopt Katz' view and build the connection between the theoretical model of competence and real-time behavior. We argue that performance data from both production and comprehension can contribute in valuable ways to our understanding of the theoretical constructs assumed in Distributed Morphology (henceforth DM; Halle & Marantz, 1993). On the one hand, we show that various performance phenomena including speech errors and speed of lexical access receive a straightforward explanation within DM; on the other hand, DM also allows for interesting predictions regarding possible and impossible output forms via underspecification and the Subset Principle. Given that the theoretical constructs of DM can be shown to have an effect on linguistic behavior, we conclude that DM makes for a *psychologically real* theory of grammar.

Before moving forward, a note on the scope of this article. First, we restrict focus to evaluating the connections between DM and psycholinguistic theories, rather than comparing how different theories of morphology fare in connecting to what is known within psycholinguistics. This restriction is not intended to imply that other theories of morphology are irrelevant or unable to capture the data that are discussed. Still, we think that there are certain traits of DM that allow for particularly elegant explanations, and we will highlight these. Second, we do not discuss neural data, which is covered in depth by Linnaea Stockall in this volume.

In Section 2, we start with some general remarks on the relationship between DM and processing theories. In Sections 3 and 4, we then discuss studies related to production and comprehension, respectively. In Section 3, we address the theoretical impact of spontaneous speech errors and of bilingual productions. In Section 4, we offer a DM perspective on lexical access and on the processing of polysynthesis. Section 5 concludes the chapter.

## 2. DISTRIBUTED MORPHOLOGY AND PROCESSING THEORIES

In this section, we begin by providing a general overview of how DM relates to theories of languages processing. We organize the section by the two major principles of the theory. Section 2.1 focuses on the principle of Syntax All The Way Down, which most immediately relates to the discussion of language comprehension in Section 4. Section 2.2 focuses on information flow and, more specifically, on Late Insertion, which is further elaborated in the discussion of language production in Section 3.

### 2.1 Syntax All the Way Down and Language Processing

As referenced in the introductory remarks, building a theory of language processing requires a link between our theories of the processing devices (the *generator* and *parser*) and theories of the *grammar*. DM provides a particular view of what representations are consistent with the grammar. In advancing the principles of Syntax All The Way Down – the notion that *phrase structure representations* organize both words and morphemes into constituents – and Late Insertion – the post-syntactic realization of phonological Vocabulary Items at PF and semantic realization of Encyclopedia Entries at LF (Marantz, 2013) – DM allows for a mapping between language processing devices and the grammar that appeals to a single *morphosyntactic* representation across languages with distinct *morphophonological* and *morphosemantic* mappings.

A recent and particularly clear elucidation of this viewpoint in terms of processing has been proposed by Oseki (2018: 2) under the pair of hypotheses in (1).

(1) a. **The Single Engine Hypothesis**
There is only one generative engine to build complex words and sentences.

b. **Hierarchical Morphological Processing Hypothesis**
Morphological processing tracks [the] hierarchical syntactic structure of words.

The anti-lexicalist spirit of DM, summarized under (1a), plus the assumption that there is a relatively straightforward mapping between the grammar and the processing devices (i.e., a maximally transparent mapping between *competence* and *performance*; Chomsky, 1965), allows us to arrive at the hypothesis in (1b): that word-level structures are built with the same representation as sentence-level structure, both in production and comprehension. In short, Oseki (2018) proposes that there is *representational identity* (Momma & Phillips, 2018) between word- and sentence-level processing.[1]

[1] It is worth being abundantly clear about what is meant by "word-level" and "sentence-level" for the purposes of this discussion. We take word-level processing to be processing that occurs *within* a string of dependent pieces of morphology. We take sentence-level processing to be processing that occurs within a sentence at the granularity of the (phonological) word, where words are (perhaps loosely) defined as potentially phonologically independent strings.

However, we may take this hypothesis further by allowing for the possibility that there is no principled distinction between the mechanisms that create hierarchical representations word-by-word and morpheme-by-morpheme (i.e., *mechanistic identity*; Momma & Phillips, 2018). Or, we may extend identity to the final logical conclusion by proposing that the *ordering* of these mechanisms in the creation of hierarchical phrase structure representations is the same in both word- and sentence-level processing (i.e., *algorithmic identity*; Momma & Phillips, 2018). The possible levels of identities are summarized in (2).

(2) **Levels of identity between word-level and sentence-level processing**

a. <u>Representational:</u> Word-level and sentence-level processing occurs over identical hierarchical morphosyntactic representations.

b. <u>Mechanistic:</u> Hierarchical morphosyntactic representations are built with the same mechanisms at both the word-level and sentence-level.

c. <u>Algorithmic:</u> The ordering of the application of the mechanisms that build hierarchical morphosyntactic representations is the same at both the word-level and sentence-level.

In this taxonomy of identity, each subsequent level of identity implies the previous level(s). As a result, much of the focus of the present chapter will be on whether representational identity holds, as mechanistic and algorithmic identity become irrelevant if this is not the case. Indeed, importing the assumptions of DM to theories of language production and parsing at a minimum implies representational identity, but not necessarily the higher levels: we can imagine a world in which distinct word-level and sentence-level mechanisms create identical morphosyntactic representations, and such a world would remain consistent with the core tenets of DM.

However, even under the relatively modest assumption of representational identity, certain burdens of explanation begin to shift. For example, the need to explain how identical morphosyntactic representations can lead to distinct morphophonological structures across languages falls to our aptitude for generating and parsing phonological strings. Importantly, this is a burden that this system must carry independently of whether a word/morpheme distinction is made in the morphosyntactic representation. Whether the parts that make up a given string are phonological dependents (i.e., a string of morphemes) or independents (a string of words), the production and comprehension mechanisms must map between discrete morphosyntactic categories and a continuous phonological externalization (production), or the other way around (comprehension). No additional mechanism that isn't required to parse words is required to capture the processing of morphology from this viewpoint – the processing of words already requires a mapping of a continuous phonological representation to the discrete notion of word. Making the move to mapping between the phonology and morpheme-level categories is simply changing the level of granularity (from word-by-word to morpheme-by-morpheme) and the nature of the output (from a set of words to a set of morphemes). Therefore, parsimony is clearly on the side of maintaining *at least* representational identity between word-level and sentence-level processing.

## 2.2 Information Flow

An important issue regarding both the model of grammar and the processing model concerns the flow of information: how are different types of information packaged and at what point are different sources of information accessed and manipulated?

Evidence from language production is broadly consistent with a theory where processing proceeds in ordered stages, with grammatical encoding preceding phonological encoding; at all stages, specific errors may occur (e.g., Garrett, 1980a; Levelt, 1989). Also, it is generally assumed that the lexicon is split into a semantic (lemma) lexicon, which is accessed at the point of grammatical encoding, and a phonological (form) lexicon, which is accessed at the point of phonological encoding. It is, however, a matter of debate in how far the different stages operate independently of each other, that is, in how far the derivation of an utterance proceeds strictly top-down (Dell & Reich, 1981; Dell, 1986), and in how far the two components of the lexicon are truly separate from each other (Dell & O'Seaghdha, 1992). It is possible that the architecture is strictly serial, where each stage strictly precedes and is informationally encapsulated from all other stages. On the other end, processing may be fully interactive, with no boundaries between processes or types of information, as is characteristic of spreading-activation or connectionist models (e.g., Schade, 1992; Dell et al., 1999; Rohde & Plaut, 2003). In the middle of these two is a cascading architecture, where modules remain distinct, but can be executed in a partially parallel fashion, with information flowing in both directions during these periods of overlap.

The above possibilities can be couched in terms of the assumptions of DM. DM is most immediately consistent with either a serial or cascading architecture, where there are discrete stages of processing. These two possibilities are schematized in Figure X.1. First, given the principle of Late Insertion, grammatical structure building is separate from and precedes phonological spell-out. Crucially, in both models, syntax only manipulates abstract roots and bundles of morphosyntactic features. Secondly, in DM, there is not a single lexicon, but rather specialized lists, with List 1 (the "narrow" lexicon) overlapping to some extent with the lemma lexicon discussed above, and List 2 (the Vocabulary) corresponding to the form lexicon. List 3 (the Encyclopedia) contains the remaining information that would otherwise be stored in the lemma lexicon. Note that in Figure X.1, we adopt the idea that lists can include entries from multiple languages, as indicated by the subscripts (Lillo-Martin et al. 2016; see Section 3.2 for further discussion). With respect to other assumptions regarding representation and derivation, however, the grammar model diverges from the received model in psycholinguistics. In particular, according to DM, roots do not bear category labels, that is, they are acategorial. Also, the DM model includes a level of Morphological Structure (MS), which intervenes between syntax and phonology. At this level, operations like fusion and morpheme insertion may change the composition and number of terminal nodes.

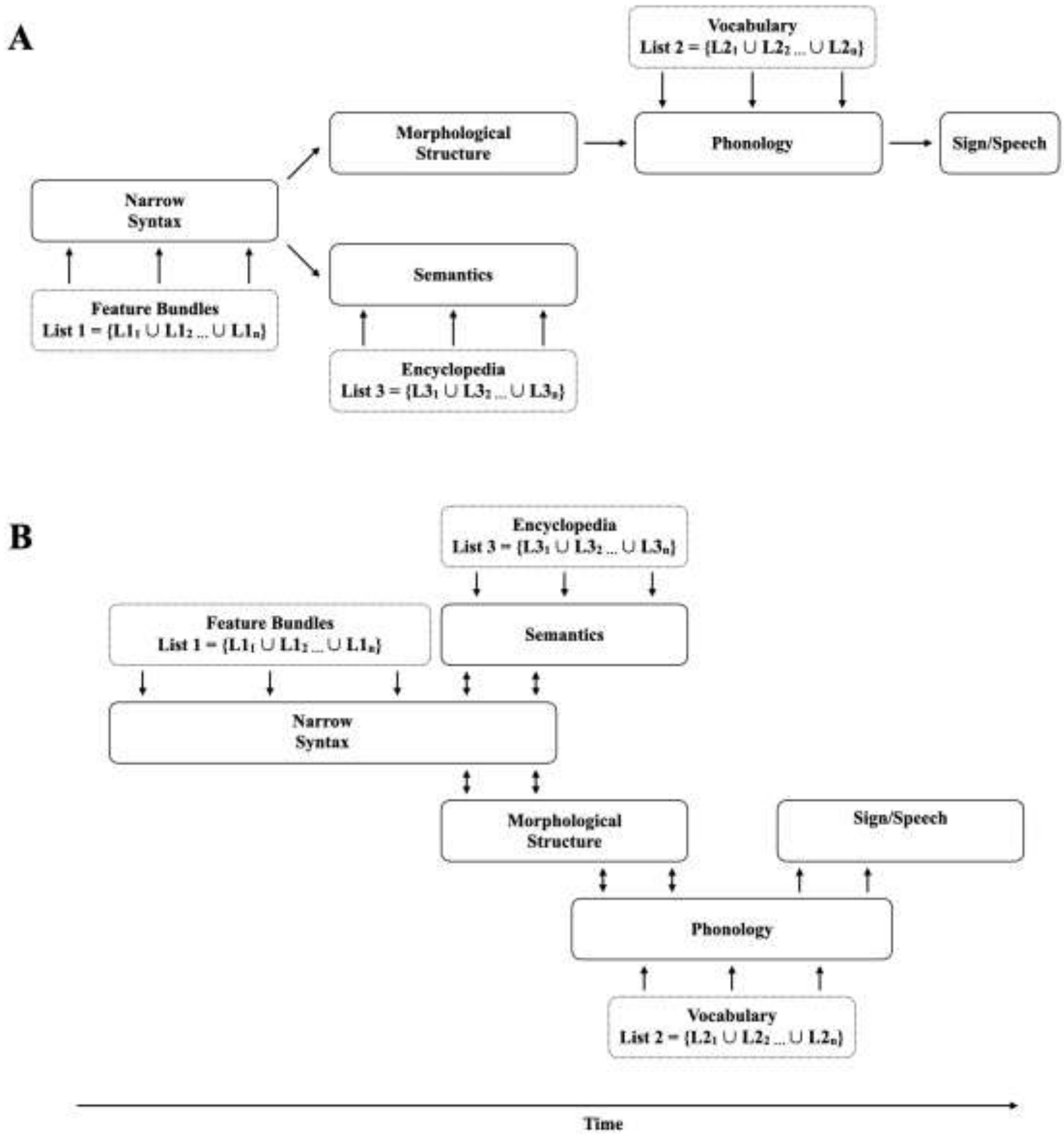


**Figure X.1.** Two options for information flow within the DM model: (A) strictly serial (top-down) flow of information with no overlap between processing levels; (B) cascading flow of information with some overlap between processing levels.

## 3. LANGUAGE PRODUCTION

### 3.1 Spontaneous Speech Errors

Ever since the seminal work of Fromkin (1971) and Garrett (1975), spontaneous speech errors – slips of the tongue – have been considered a rich source of evidence for the processes that mediate between the planning of an utterance and its articulation, as well as for the linguistic units that are manipulated in the course of constructing an utterance. In addition, speech errors have provided evidence that linguistic units such as phonological features, segments, syllable constituents, and morphemes are indeed processed during language production, that is, that these units are *psychologically real* (Matthews, 1991).[2]

In the previous section, we already alluded to the fact that the grammar model and the production model share certain assumptions regarding architecture and information flow. Pfau (2000, 2009) takes these correspondences, as well as the differences, as starting point in his investigation of how (mostly German) spontaneous speech errors can be accounted for within DM. On the one hand, his study demonstrates that certain assumptions of (early) DM have to be reconsidered in light of the speech error data; one such assumption concerns the nature of roots and will be addressed in Section 3.1.1. On the other hand, it also turns out that most error types can be accounted for in a straightforward way within DM. Of particular interest are cases which appear to involve a post-error repair mechanism; these will be discussed in Section 3.1.2. Finally, in Section 3.1.3, we briefly address slips that pose challenges for the DM model.

### 3.1.1 *The Nature of Roots*

In early versions of DM, it has been assumed that only morphosyntactic features and certain compositional semantic features (e.g., [count], [animate]) are present in the syntax. In contrast, roots that enter the computational system have been argued to be void of semantic features, as such features do not play any role within the computational system (Halle & Marantz, 1993; Marantz, 1997; Harley & Noyer, 2003). It was therefore suggested that List 1 does not include different, semantically specified roots, but rather only a single generic √ROOT. As pointed out by Harley and Noyer (2003: 473), "a Root morpheme in an appropriately local relation to a Determiner might be filled by *cat*, *dog*, *house*, *table* or any other Vocabulary item we would normally call a 'noun'".

However, in light of the speech error data, such an "economic" perspective on List 1 cannot be maintained. As discussed in Pfau (2009: 84f), describing a root hosted by a terminal node solely on basis of its syntactic and conceptual semantic features does not suffice to individuate it; that is, if a terminal node contained, for instance, √ROOT, [animate], and [count], insertion of the appropriate Vocabulary item cannot be guaranteed, as is also acknowledged by Marantz (1995: 401) when he writes that "the Vocabulary items for "cat" and "dog" would be equally specified […] and either might be inserted at that node". Beyond this conceptual problem, speech error data, in particular, semantic substitutions, clearly show that the roots manipulated by the syntax must be specific to the particular concept they are linked to, that is, √DOG rather than √ROOT.

A semantic substitution takes place when an erroneous, yet semantically related, element is selected from List 1 and enters the computational system. The existence of such semantically motivated mis-selections strongly suggests that the grammar is sensitive to the difference between, for instance, *cat* and *dog* before PF – contrary to the original proposal which endorses

[2] This issue has been addressed for typologically diverse languages, including languages with non-concatenative morphology (e.g., Arabic: Abd-El-Jawad & Abu-Salim, 1987) and tone languages (e.g., Thai: Gandour, 1977), as well as for sign languages (Newkirk et al., 1980; Hohenberger & Leuninger, 2012).

late (post-syntactic) semantic interpretation (via List 3, the Encyclopedia). An English and a German example are provided in (3).[3]

(3) a. he got hot under the **belt** ← under the collar (Fromkin 1973: 262)

| | | | | | | |
|---|---|---|---|---|---|---|
| b. | hast | du | einen | **Radiergummi** | ← | einen Spitzer |
| | have | you(SG) | an.M | eraser(M) | ← | a.M pencil.sharpener(M) |

'Do you have a pencil sharpener?'

According to Siddiqi (2010: 537), a "large proportion of the literature in DM seems to have assumed Pfau's proposal, using specific roots in derivations, effectively accepting this revision as main stream". In an attempt to avoid the use of language-specific root labels, Harley (2014) suggests to use numerical addresses instead, which serve as the link "between a set of instructions for phonological realization in context and a set of instructions for semantic interpretation in context" (Harley, 2014: 226); that is, e.g. $\sqrt{}_{276}$ rather than √DOG.

The German speech error data further suggest that roots drawn from List 1 are associated with a gender feature. This is evidenced by the fact that an 'identical gender effect' is observed in semantic noun substitutions: in 122 out of 171 semantic substitutions from Pfau's corpus (71.3%), the intended and the intruding noun share the same gender; this also holds for the slip in (3b), where both nouns are masculine.[4] That is, while the intended noun in (3b) may also have activated the semantically related noun *Lineal* ('ruler'), *Radiergummi* ('eraser') makes for a more likely intruder, as *Lineal* is of neuter gender.

We are aware that the issue of gender specification is currently much debated in DM. Kramer (2015) and Hammerly (2019), for instance, refute the idea that roots can be specified for gender. Rather, it is assumed that gender is associated with a functional projection above NP, for instance, *n* ("little n"; see also Kihm, 2005). In such models, the identical gender effect might then be due to selection restrictions of *n*. According to this line of reasoning, inserting a semantically related noun of different gender would imply substituting the root and *n*, while the same is not true for a substitution involving nouns of the same gender. Note that this implies (i) that the system somehow "knows" which variant of *n* will be merged with a given root, as the gender of German roots is almost never variable; (ii) that the substitution takes place at a point in the derivation when *n*P has already been built, as the gender-specific *n* needs to be present in order to account for the identical gender effect. Clearly, the latter is only possible in an architecture that assumes cascading information flow (see Figure X.1B).

[3] In the speech error examples, the actual utterance (with error element(s) in bold) appears on the left side of the arrow, the intended utterance on the right side. When there is no arrow, the slip was self-corrected by the speaker. In the interlinear translation, only the grammatical information that is relevant to the error is provided; the gender of nouns is indicated between brackets. Only the intended utterance is translated. All German speech errors come from Pfau's error corpus (which consists of 829 errors and includes those examples from the Frankfurt speech error corpus that were relevant to his study).

[4] Similarly, Marx (1999) reports that in her corpus of German slips of the tongue, 78% of the semantic noun substitutions obey the identical gender effect.

3.1.2 *On So-called 'Accommodations'*

An error type that has intrigued scholars for a long time are the so-called 'accommodations', that is, slips which appear to involve a post-error repair process. For Garrett (1980b), who assumes a strictly serial model of language production (see Figure 1.XA), accommodations are a particularly clear piece of evidence for the existence of distinct processing levels: at one level, the actual error occurs, while at a subsequent level, the erroneous utterance is brought in line with some grammatical constraint. The repair may affect the error element itself (error accommodation) and/or its environment (context accommodation).

Two examples, both self-corrected by the speaker, are provided in (4). According to an accommodation account, in (4a), the anticipated verb stem *les* ('read') undergoes a stem-internal change due to the [+past] feature at its landing site (which is spelled out by the regular past tense affix *-te* in the intended utterance) – this is an error accommodation. In contrast, in the exchange in (4b), the stem *nahr* ('feed') triggers two changes in the error-induced environment, i.e., context accommodations: insertion of the nominalizing suffix *-ung*, which is not part of the intended utterance, as well as accommodation of the determiner to the feminine gender associated with this suffix.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| (4) | a. | ich | **las** | ihr | fürs, | äh, | dank-te | ihr | fürs | les-en […] |
| | | I | read.PAST | her | for.the, | er, | thank-PAST | her | for.the | read-INF |

'I thanked her for reading (my handout).'

| | | | | | | |
|---|---|---|---|---|---|---|
| b. | **nerv**-e | die | **Nahr**-ung, | nähr-e | den | Nerv |
| | nerve-IMP | the.ACC.F | food-NMLZ(F), | feed-IMP | the.ACC.M | nerve(M) |

'Feed the nerve!'

The fact that the processor manages to deliver a grammatical – albeit at times semantically awkward – utterance is without doubt impressive. However, Pfau (2007, 2009) argues that once one adopts DM mechanisms, it is no longer necessary to assume costly post-error repair strategies. In (4a), the relevant mechanism is phonological readjustment: at the point of Vocabulary insertion (VI), the anticipated abstract root √LES will be spelled out as /le:z/, and subsequently, the phonological readjustment rule in (5a) will apply in a [+past] context.[5] In (4b), the error element √NAHR surfaces in a nominal context, that is, in a position where it is licensed by a determiner (Harley & Noyer, 1998), and in this context, the morpheme insertion rule in (5b) is triggered (Pfau, 2009: 247).[6] The rule applies at MS, and it must precede feature copy, as it is only the nominalizing suffix that contributes the gender feature, which is then copied onto D (⇐ = 'licensed by').

---

[5] See Siddiqi (2009) for an account that does away with phonological readjustment rules, suggesting instead that after fusion at the level of MS, a terminal node may contain a root and morphosyntactic features (e.g. √LES and [+past] in (4a)), and that this node will be spelled out by a specific Vocabulary item (e.g. /la:z/).

[6] Root-specific morpheme insertion also plays a role in English speech errors, e.g. *it's **care**-ful to measure with **reason*** ← *it's reasonable to measure with care*, where an adjectival suffix is adjusted (Fromkin, 1973: 31).

(5) a. /e:/ $\rightarrow$ /a:/ / √X + [+PAST]
(where √X = *les* ('read'), *geb* ('give'), *seh* ('see'), ...)
b. Insert [-ung(F)]$_{\mu}$ / √X $\Leftarrow$ D
(where √X = √NAHR ('feed'), √ERZÄHL ('tell'), √WARN ('warn'), …)

Crucially, all of the relevant mechanisms apply in the course of the derivation anyway; they do not impose any extra effort on the processor. Hence, Pfau (2009: 274) concludes "that the concept accommodation is unnecessary and should therefore be abandoned". In contrast, lexicalist theories, which claim that narrow syntax manipulates words rather than roots, would have a harder time in accounting for such errors in a comparably elegant fashion, i.e., without involving actual repair processes.

Based on the order of operations as assumed in DM, we can make another interesting prediction with respect to the grammaticality of an erroneous utterance. In Section 3.1.1, we mentioned that an 'identical gender effect' is observed in semantic substitutions. Actually, the same has been shown to hold for phonological substitutions, which result from the erroneous selection of a phonologically related Vocabulary item from List 2 at the point of Vocabulary insertion. Remember, however, that the identical gender effect is just a tendency; that is, we occasionally find substitutions in which the intended and the intruding noun are of different gender. Given that DM assumes that morphosyntactic features are copied before spell-out, we predict that in such cases, the determiner will surface with the correct gender specification only in semantic substitutions. This prediction is borne out in the German speech error data: only 1/22 semantic substitutions leads to a feature mismatch between noun and determiner, while such a mismatch is observed in 12/14 phonological substitutions (see (6a) for one of the two counterexamples).

3.1.3 *Challenges*

Pfau (2009) discusses the above and various other errors types that involve the manipulation of semantic or morphosyntactic features and argues that these errors receive a straightforward explanation within the DM architecture. Moreover, the model allows us to explain why certain error types are not attested. He thus concludes that the grammar model and the psycholinguistic production model can be mapped onto each other, that is, that sequences of grammatical operations, as assumed in DM, in a real sense constrain language processing.

Still, it has to be acknowledged that some potentially problematic cases remain. One challenging error type has already been mentioned in the previous section: phonological substitutions in which intended and intruding noun are of different gender, and yet, the determiner is spelled out with the correct gender feature. This is what we observe in (6a), where the feature [masculine] of the intruding noun *Kalender* ('calender') has been copied onto the determiner. Clearly, the intended noun *Geländer* ('railing') and the intruder are not semantically related, and we must therefore assume that the error takes place at the point where Vocabulary items are selected from List 2. However, at this point in the derivation (i.e., at PF), it is too late for feature copy to take place – at least in a model that assumes strictly serial processing (Figure X.1A). In contrast, in a cascading model, where processing at the levels of MS and PF may overlap, the processor may, after the error has occurred, feed information back from PF to MS, where the feature is copied. Two things should be emphasized. First, we reiterate that cascading does not mean interactive, but only implies partial overlap between adjacent processing levels. Second, in production, interaction between processing levels must still be the exception, as otherwise gender accommodation following a phonological substitution would be expected to be more common.

(6) a. wo sie über den **Kalender** guckt ← über das Geländer
where she over the.M calendar(M) looks ← over the.N railing(N)
'[…] where she looks over the railing'

b. man muss die **Wurzel** an der **Übel** pack-en
one must the.F.ACC root(F) at the.F.DAT evil(N) grab-INF
← das Übel an der Wurzel pack-en
← the.N.ACC evil(N) at the.F.DAT root(F) grab-INF
'One has to tackle the root of the problem.'

The noun exchange in (6b) is intriguing, as it involves 'partial repair'; that is, in the first error position, the gender feature [feminine] is copied onto D, while in the second error position, feature copy fails, and we observe a feature mismatch within DP (**der Übel*, while the expected outcome would be *dem Übel*). It thus appears that the gender feature associated with √WURZEL is copied twice, in both positions, while the feature associated with √ÜBEL is not copied at all – clearly a challenging scenario for the theory.

While the challenge posed by slips such as those in (6) has to be acknowledged, it is also clear that such errors are very rare.[7] We thus conclude that spontaneous speech errors provide intriguing evidence for the psychological reality of derivational mechanisms, as assumed in DM.

### 3.2 Code-switching and Code-blending

In the preceding section, we have been concerned with erroneous utterances of monolingual speakers, and we have argued that oftentimes, DM mechanisms guarantee a grammatical outcome. We now turn to utterances produced by bilinguals in which elements from two languages are combined, as such productions have also been analyzed from the perspective of DM. In Section 3.2.1, we discuss intra-sentential code-switching, which involves the mixing of two spoken languages. Another type of language mixing is observed in bimodal bilinguals, that is, subjects who are native users of a spoken and a sign language. This type of mixing is referred to as code-blending (Emmorey, Borinstein & Thompson, 2005; Emmorey et al., 2008) and will be addressed in Section 3.2.2.

[7] See Albright (2007) for discussion of further examples that possibly contradict the assumption of an entirely repair-free derivation.

### 3.2.1 *Code-switching in Bilinguals*

Studies on bilingual code-switching, involving various different language pairs, have demonstrated that intra-sentential code-switching – the type that we will be concerned with in the following – does not proceed at random; rather, the switch point as well as the grammatical shape of the mixed utterance are subject to certain constraints (see, e.g., Myers-Scotton, 1993; Muysken, 2000). MacSwan (1999, 2000) suggested that the grammatical restrictions imposed on code-switching are no different from those at work in monolingual speech, and that this can be modeled within a Minimalist framework (Chomsky, 1995). The bilingual speaker can select lexical items from two different lexicons, but feature checking applies in the same way as in monolingual speech, and uninterpretable features have to be deleted in the course of the derivation in order to yield a well-formed outcome. This approach is attractive, as it does away with previously proposed constraints that are specific to bilingual code-switching (as, e.g., the Functional Head Constraint proposed by Belazi, Rubin & Toribio (1994)).

Subsequent work on code-switching, however, departed from the Minimalist idea that fully specified lexical items enter the syntactic derivation. Liceras and colleagues (Liceras, Spradlin & Fernández Fuertes, 2005; Liceras et al., 2008) and Pierantozzi (2012), among others, discuss code-switching data that, according to them, receive an elegant explanation once basic tenets of DM are adopted. It is argued that the idea of late insertion in particular allows for a straightforward explanation for a range of data. A case in point is the phenomenon of 'mixed agreement' in code-switched noun phrases, as illustrated in (7a), where a gender-marked Spanish determiner combines with an English noun (Muysken, 2000: 23). Clearly, this gender feature cannot come from the English root; it must come from the Spanish competitor *casa*. The argument thus goes as follows: √CASA is drawn from the Spanish List 1; this root is associated with a gender feature (cf. the discussion in Section 3.1.1). Prior to spell-out, the gender feature is copied onto the determiner, but at the point of Vocabulary insertion, the corresponding Vocabulary item is selected from the English List 2. This Vocabulary item is compatible with the gender-specified root due to the Subset Principle (Halle, 1997): given that it is not specified for gender, no feature conflict arises.[8]

[8] Various studies have demonstrated that the opposite pattern (i.e., *the casas*) is also attested, although it appears that code-switched DPs which contain the Spanish determiner are the preferred option (e.g., Jake, Myers-Scotton & Gross, 2002; Liceras et al., 2008; Pierantozzi, 2012). In addition, Jake et al. (2002) and Liceras et al. (2008) discuss cases like *el chair*, which is different from (7a), as the Spanish translation equivalent *silla* is feminine, but the determiner is masculine. It is argued that such cases may occasionally surface, as it has been proposed that the masculine D is the default form.

(7) a. Veo las house-s [Spanish – English]
see.1.SG the.F.PL house-PL
'I see the houses.'

b. Veo el Sonne [Spanish – German]
see.1.SG the.M.SG sun(F)
'I see the sun.'

As pointed out by Pfau (2016: 804), the above line of reasoning allows for an interesting prediction: in case of code-switching involving two languages that both have a gender system, cases like the hypothetical example in (7b) should be much less likely to occur. Note that the Spanish and German translation equivalents of *sun* are of different gender. Consequently, the German Vocabulary item *Sonne* is not compatible with the gender of the root from which the gender feature has been copied onto the determiner. Indeed, for children who are bilingual Italian–German, Cantone and Müller (2008) report that mismatches, as in (7b), were observed in only 16% of the cases in which the nouns were specified for different genders.

The conclusion from the discussion of DP-internal code-switching thus is that an account based on DM can derive the various patterns observed within mixed DPs, while a Minimalist checking account, which assumes that syntax manipulates fully specified lexical items, will have a harder time explaining attested cases in which the features on D and N do not match. Crucially, appealing to constraints that are specific to bilinguals is uncalled for in the DM model, just as it is in MacSwan's account.[9]

### 3.2.2 *Code-blending in Bimodal Bilinguals*

More recently, the same line of reasoning has been applied to productions of bimodal bilinguals, that is, hearing native signers (sometimes referred to as CODAs: Children of Deaf Adults) that mix a sign language and a spoken language. What is special about their productions is that, given different articulators, the languages can be mixed simultaneously – a phenomenon that is referred to as 'code blending' (Emmorey et al., 2005, 2008). In fact, the available studies suggest that sequential code-switching is rare in the utterances of bimodal bilinguals (e.g., attested in only 6.26% of the utterances produced by American Sign Language (ASL) – English bilinguals analyzed in Emmorey et al. (2008)).

Lillo-Martin, de Quadros, and Chen Pichler (2016) offer an account of code-blended utterances within DM. Their model, which they refer to as "Language Synthesis" model, shares many properties with previous models of bilingual productions: (i) the numeration may contain roots and features from either language (L); (ii) syntactic operations then operate on these elements from $L_a$ and $L_b$; (iii) at the point of Vocabulary insertion, Vocabulary items may be selected from two vocabularies (List 2) and inserted into appropriate terminal nodes. In contrast to previous models, however, the Language Synthesis model allows for two phonological levels

[9] Another phenomenon that has been argued to speak in favor of a late-insertion account is language mixing below the word level (e.g., Bandi-Rao & den Dikken, 2014; Grimstad, Lohndal & Åfarli, 2014); cf. footnote 11.

– one for speech and one for sign – and thus for the simultaneous production of two phonological representations.

Before turning to code-blending, let us briefly address another phenomenon that Lillo-Martin et al. discuss: syntactic synthesis, where a spoken language utterance displays structural properties that are grammatical in the sign language, but are not allowed in the spoken language. The authors consider various phenomena (e.g., word order, argument omission), but here we focus on wh-questions for illustration. Both sign languages they consider, ASL and Brazilian Sign Language (Libras), allow for clause-final wh-signs and doubling of wh-signs, structures which are grammatical in neither English nor Brazilian Portuguese. In spoken utterances produced by 2-year-old bimodal bilingual children, they observe corresponding structures, for instance, a clause-final wh-sign in the English example in (8a) and an instance of wh-doubling in the Portuguese wh-question in (8b) (Lillo-Martin et al., 2016: 734).

| | | | | | | |
|---|---|---|---|---|---|---|
| (8) | a. | Bug go where? | | | | [Tom 2;04] |
| | b. | Que | eu | quere | que? | [Igor 2;01] |
| | | What | I | want | what | |
| | | ‘What do I want?’ | | | | |

According to Lillo-Martin et al. (2016: 730f), these structures are “simply generated by an architecture that allows feature bundles entering into a derivation to come from a different ‘language’ than the one that the vocabulary items are drawn from”. Following previous work on ASL and Libras (Petronio & Lillo-Martin, 1997; Nunes & de Quadros, 2008), they assume that the ASL version of a [+wh] C head permits in situ wh-questions, while a [+emphatic] focus head is responsible for doubling in Libras. That is, the child selects functional elements from the sign language List 1, builds the structure based on these elements, but then inserts Vocabulary items from the spoken language List 2 at the point of spell-out.

While the phenomenon illustrated in (8) is also observed in spoken language bilinguals (cf. Tieu (2010) for wh-in-situ questions produced by English–Cantonese bilingual children), code-blending is clearly a modality-specific phenomenon, as it involves the simultaneous use of different articulators: the vocal apparatus and the hands. We illustrate the phenomenon in (9) by means of two examples involving different language pairs. In (9a), the matrix language is clearly Sign Language of the Netherlands (NGT), as the utterance is produced fully in sign, following also the word order of NGT. Only the clause-final adjective is accompanied by the corresponding Dutch word (adapted from Baker & van den Bogaerde, 2008: 8). Lillo-Martin et al. refer to this as ‘co-insertion’. Note that in co-insertion, the two elements are not always perfectly aligned temporally, and they may also be semantically non-equivalent (e.g., Baker & van den Bogaerde (2008: 8) report the combination of NGT SHOOT with Dutch *doodmaken* (‘kill’)). Example (9b) illustrates that co-insertion may occur multiple times in a single utterance. As ASL and English use the same word order in this particular context, it is only the missing indefinite determiner that suggests that ASL is the matrix language (Lillo-Martin et al., 2016: 746).[10]

(9) a.

| | | | | |
|---|---|---|---|---|
| Dutch: | | | blauw | [Mother of Jonas (3;00)] |
| NGT: | INDEX$_3$ | COAT | BLUE | |
| | 'He has a blue coat.' | | | |

b.

| | | | | | | |
|---|---|---|---|---|---|---|
| Eng.: | I | thought | I | saw | cat | [Adult – Adult] |
| ASL: | IX$_{(self)}$ | THINK | IX$_{(self)}$ | SEE | CAT | |
| | 'I thought I saw a cat.' | | | | | |

Just as in (8), the syntactic derivation employs abstract roots and features from $L_a$, where the latter ensure that the syntactic structure is built according to language-specific feature values. However, at the point of Vocabulary insertion, the Vocabularies of both $L_a$ and $L_b$ are accessed, and Vocabulary items from the two languages are inserted into terminal nodes. A compromise is uncalled for, as the use of different articulators affords simultaneous articulation.[11]

Based on (9b), a further interesting observation can be made. Note that the co-inserted elements are semantically equivalent but that the English string contains tense inflection that is not present in the ASL string; in other words: ASL only expresses a subset of the features expressed in English. However, given that tense is not marked overtly on ASL verbs, no feature clash arises. Following the Subset Principle (Halle, 1997), insertion of, for example, the sign THINK without tense marking does not result in a conflict with √THINK [+past]. Just as in a monolingual English utterance, the Vocabulary item that spells out √THINK will undergo phonological readjustment in a [+past] context. Only in cases, where features are in conflict, the derivation will crash. Plural, for instance, can be marked on some ASL nouns, and hence the combination of ASL BOOK++ (where '++' signals reduplication) with English *book* might cause such a crash.

Taken together, the DM-inspired Language Synthesis model can account for the code-blending cases discussed above, and it also allows us to make predictions regarding simultaneous combinations that should not occur. It has to be acknowledged, however, that more complex instances of code-blending are attested which do not receive a straightforward explanation within DM, in particular cases where full clauses are generated in both modalities (as in (9b)), but with different word orders. While Lillo-Martin et al. (2016) mention that such

[10] Notational conventions: signs are glossed in SMALL CAPS; INDEX/IX represents a pointing sign towards the signer's body (self) or a location in the signing space (subscript 3), indicating first or third person reference; alignment of glosses and spoken words indicates simultaneous production. In example (9a), we translated the glosses into English for ease of representation; the spoken word, however, is Dutch *blauw* ('blue').

[11] As pointed out by Pfau (2016), node sharing is also attested in unimodal productions. In spontaneous speech errors, node sharing is observed in so-called word blends, such as German *Plitz* resulting from a competition of the semantically related nouns *Platz* ('place') and *Sitz* ('seat'). Grosjean (1982: 184) reports similar examples produced by English–French bilingual children, such as *shot* resulting from a combination of French *chaud* [ʃo] and English *hot*. In crucial contrast to (9), however, a compromise in the form of a fusion/blend is required in both cases.

examples with incongruent syntax are very rare in their data set, Donati and Branchini (2013: 109) report such cases for Italian – Italian Sign Language (LIS) CODAs. A point in case is the simultaneous combination of a LIS interrogative with clause-final wh-subject (CALL WHO) with the corresponding Italian string with initial wh-subject. The LIS verb is articulated simultaneously with the Italian wh-word (*chi*), and the LIS wh-sign with the Italian verb (the participle *chiamato*). Clearly, a co-insertion account where two Vocabulary items compete for the same terminal node cannot explain these cases, and further research is necessary to uncover if and how such complex blendings can be accounted for in DM (see also Branchini & Donati, 2016).

## 4. LANGUAGE COMPREHENSION

When we encounter linguistic input, we do not wait for a pause, the end of a sentence, or even the end of a clause to begin parsing what has been said. A central tenet of all viable theories of language comprehension is that processing occurs *incrementally*. That is, the processing of linguistic input takes place *in real time* as it is being heard, rather than in a separate stage after a chunk has been encoded. What is far less clear, however, is just how radical to be about incrementality, particularly when it comes to the processing of morphosyntax. How do we build morphosyntactic representations in real-time? Are morphologically complex words processed as a unit, or does something akin to word-by-word incrementality extend to the level of the morpheme? And if incremental morpheme-by-morpheme parsing does occur, what is the nature of the representations and processes that underlie processing at this level? These questions – and the possible answers that become available under Distributed Morphology – are the focus of this section. This amounts to seeking an answer to the core question raised at the outset of the chapter: does representational (and mechanistic and algorithmic) identity hold between morphological and syntactic parsing?

Much of the work in comprehension has focused on *lexical access* – the processing of single (orthographic) words. The first section focuses on reviewing results from the two major methodological areas within this domain: lexical decision tasks and eye tracking while reading. We also review evidence for *underspecification*, a core tenet of DM, and its role in morphosyntactic processing. In Section 4.2, we then turn to an area relevant to DM that falls outside the direct purview of lexical access: the study of languages with significant morphological structure (e.g., polysynthetic languages).

### 4.1 Lexical Access

What is the status of a "word" as a psychological primitive? In the study of lexical access, this question is central. By taking seriously the possibility of parity between morphological and syntactic processing, "words" have the potential to become entirely irrelevant as a level of representation. Indeed, whether or not we can dispose of the notion of "word" in models of morphological processing gets to the core of a distinction between *storage-based* versus *decomposition-based* models. Both models provide a view of how lexical access proceeds. To take an example, consider the word *devalued*, which all speakers of English would recognize as a word. Storage-based models contend that no morphological decomposition occurs when accessing words from the lexicon – morphologically complex words such as *decommissioned* are stored, retrieved, and recognized in unanalyzed forms. In this model, words are elevated as the fundamental building blocks of (morpho)syntactic processing. In decomposition models, words are consistently broken into component parts (i.e., *de-value-d*), and storage and retrieval occurs with respect to these morphemes. This class of models essentially dispenses with the notion of word for the purposes of morphosyntactic processing.

*Single-route models* of lexical access form the endpoints of the continuum between storage and decomposition-based theories, where morphosyntactic processing is claimed to be either fully storage-based (Butterworth, 1983; Rumelhart & McClelland, 1986; Plaut et al., 1996; Seidenberg & McClelland, 1999; McClelland & Patterson, 2002) or fully decomposition-based (Taft, 1979, 2004; Taft & Forster, 1975, 1976). In the middle of the continuum are dual-route models, where storage and decomposition coexist. *Dual-route models* can take one of two basic shapes: in serial dual-route models (Pinker, 1999; Alegre & Gordon, 1999), storage and decomposition routes are fully dichotomized; in parallel models (Baayen, Dijkstra & Schreuder, 1997; Kuperman, Bertram & Baayen, 2008, 2010; Kuperman et al., 2009), storage and decomposition routes occur at once.

In this section, we focus on the evidence in favor of models with a decompositional component. Only models that include some degree of decomposition are consistent with the principles of DM. The review is by no means exhaustive, but is meant to provide a starting point by introducing the basic areas of research. For a full review of the lexical processing literature, see Baayen (2014), Marantz (2013), Taft (2014), and Zwitserlood (2018).

### 4.1.1 *Lexical Decision*

The vast majority of the literature on lexical access has employed some version of the *lexical decision task*, where participants classify stimuli as "words" or "non-words" by pressing one of two response keys (note, there are also "go"/"no-go" versions, where participants only make a response when they see a real word; see Gordon, 1983). Researchers commonly report two measures: (i) accuracy of response, and (ii) response latency, usually measured as the time between the onset of the visual stimulus and the execution of a button press. Among the first researchers to employ this methodology were Taft and Forster (1975, 1976), who found early evidence in favor of lexical decomposition with both affixed and compounded words. For example, Taft and Forster (1975) showed that non-words that are stems of prefixed words (e.g., *juvenate*, which is a stem of the word *rejuvenate*) take longer to classify as non-words than those which are not stems (e.g., *pertoire*, which is formed from the word *repertoire*), suggesting that prefixed words are broken into their constituent parts prior to lexical access. Similar effects have been shown for stems that constitute a real word (e.g., *darkness–dark*; Bertram, Schreuder, & Baayen, 2000; Niswander, Pollatsek, & Rayner, 2000) and pseudo-derived stems (e.g., *corner–corn, brother–broth*; Rastle, Davis, & New, 2004).

Starting with these early studies, differences in whole-word versus root frequency have played a central role in distinguishing storage versus decompositional accounts. The reasoning behind the use of frequency is as follows. In general, frequency stands in an inverse relationship to reaction time in lexical decision tasks: the more frequent a word, the faster the response (see Jacobs et al., 2016 for a review). However, in a morphologically complex word, multiple different measures of frequency are potentially relevant. Taking an example from Taft (2004), both of the words *seeming* and *mending* have the same *whole-word frequency*. However, the stem *seem* has a much greater *base frequency* than the stem *mend* due to the fact that words like *seemed*, *seems*, and *seem* have a greater cumulative frequency than words like *mended*, *mends*, and *mend*. Many studies, beginning with Taft (1979), have shown that a higher base frequency of the stem leads to faster word recognition even when controlling for whole-word frequency, suggesting that words are decomposed and recognized via their stem, rather than accessed as a whole.

We would be remiss not to note that despite strong evidence for decomposition, there is evidence that frequency at both the whole-word and even multi-word level affects language processing. For example, Arnon and Snider (2010) argue that frequency information for multi-word phrases such as *I don't know why* is stored and used during language comprehension based on the fact that higher frequency strings are processed faster than lower frequency strings in a way that does not fully reduce to the constituent frequencies. At first glance, such findings appear problematic for the radical decompositionality of morphology in DM. However, adopting the tenets of DM does not require full morphosyntactic decomposition to occur in all cases. For example, DM provides pathways for understanding idioms (e.g., Marantz (1997); see Harley (2019) for a recent review), where multiple morphemes and words are treated as a unit for the purposes of semantic interpretation or phonological spell-out. It is possible that these pathways are generalized not only to idioms, but also to strings of varying frequency and length, and that these representations are tied to frequency.

#### 4.1.2 *Eye Tracking*

Eye tracking while reading arguably provides a more naturalistic measure of processing than the lexical decision task. In this method, participants are generally shown a complete sentence, which they are asked to read as naturally as possible for comprehension. In most cases, participants answer a comprehension question after reading a sentence. While the subject is reading, the experimenter measures both the location and duration of each fixation (maintaining gaze at a single location) using an eye tracker. This detailed record can then be used to derive a wide variety of measures.

Measures are calculated with respect to different *regions* of a sentence, where each region generally consists of at least one orthographic word. The most common measures are first-fixation (the duration of the first fixation in a region), first-pass (the sum of all fixations from when a region is initially fixated, to when a fixation is made outside of the region, given that the region is fixated at least once; often called gaze duration if the region is a single word), go-past or regression path (the sum of all fixations from when a region is first fixated to when a fixation is made to the *right* of the region), regressions-out (the probability of fixating to the left of a region following the first-pass fixations), second-pass (the sum of all fixations in a region following first-pass fixations, including zero if the region is not re-fixated), and total reading time (the sum of all fixations in a region given that the region was fixated). A figure summarizing these measures is shown in Figure X.2.

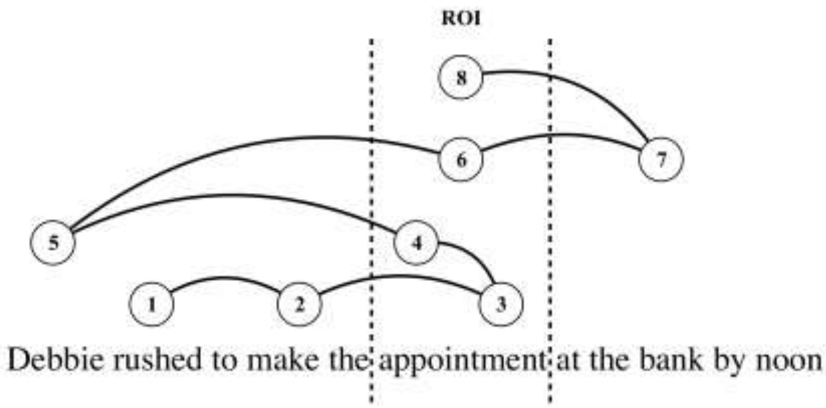


**Figure X.2.** Hypothetical record of fixations for a sentence taken from Niswander et al. (2000). Each circle represents a single fixation. The region of interest (ROI) is a single word *appointment* (it is conventional to include the space before the word in the ROI, but not the space following the word). The common measures for this ROI would be calculated by adding up the duration of the following fixations: first-fixation = 3; first-pass/gaze duration = 3 + 4; go-past/regression path = 3 + 4 + 5 + 6; second pass = 6 + 8; total time = 3 + 4 + 6 + 8. There is a regression out as a result of the leftward saccade out of the ROI between fixations 4 and 5.

A small number of studies have used eye tracking to examine prefixed verbs. One early study from Beauvillian (1996) using a non-naturalistic reading task failed to find effects of root frequency in prefixed verbs in French (e.g., *interchangé* = *inter* + *changé*) with first-fixation reading time, but did find whole-word frequency effects, initially supporting a non-decompositional account of lexical processing in reading, where whole-word representations are accessed. However, later work by Niswander et al. (2000) on normal reading in English found the negative correlation between root frequency and gaze duration expected under a decompositional account for prefixed verbs such as *remove* (= *re* + *move*). Finally, Pollatsek, Slattery, and Juhasz (2008) showed that root frequency affected gaze duration (but not first fixation time) for both novel (e.g., *overmelt*) and lexicalized (e.g., *overload*) prefixed verbs. The lack of difference in root frequency effects for novel words, which have no way of being stored in whole-word form, and lexicalized words, which could in principle be stored and accessed as whole words, suggests that some degree of decomposition feeds lexical access.

In addition to the literature on prefixed verbs, a significant number of studies have examined the processing of compounds to examine whether complex words are decomposed during lexical access. Some of the earliest studies in this domain were on the reading of Finnish compounds (Hyönä & Pollatsek, 1998; Pollatsek, Hyönä, & Bertram, 2000), where the first word of the compound was shown to consistently affect first-fixation times, while whole-word frequency had little effect. Similar results have been obtained for English compounds (Andrews, Miller, & Rayner, 2004; Juhasz et al., 2003). In contrast, a number of studies have shown whole-word frequency effects in early measures such as first fixation in addition to the effects of the first word (Juhasz, 2008; Kuperman et al., 2008, 2009). Recent work by Kush et al. (2019) on compounds in Norwegian has clarified this picture, arguing that compounds are initially decomposed, leading to root frequency effects. However, these constituent parts are associated with whole-word representations in the lexicon, leading to concurrent effects of whole-word frequency. In short, they find that there is little evidence for a model where lexical access proceeds solely over whole-word representations, favoring a decompositional account broadly consistent with the principles of DM.

#### 4.1.3 *Underspecification*

As previously discussed in Section 3.2, one of the fundamental tenets of DM is underspecification: the notion that the externalization of syntax need not express all of the underlying features, and that in these cases, a form that expresses the maximal subset of features is expressed (e.g., Harley & Noyer, 1999). One question that can be asked is to what extent underspecification of this sort is reified in the mental lexicon.

The most radical version of underspecification holds that none of the morphosyntactic categories (e.g., noun, verb, adjective) are directly represented in the mental lexicon. Such a position is taken by Barner and Bale (2002, 2005) on the basis of linguistic productivity, neurological aphasia, and vocabulary learning in acquisition (see Panagiotidis (2005) for a critical reply). On such a theory, the lexicon consists of category-less roots, which are syntactically formed into nouns, verbs, and the like (see Section 3.1.1 above for a review of the nature of roots in DM and theories of processing).

Even if one does not accept the notion that lexical items may be category-less, underspecification can hold for individual features such as number and gender. For example, work by Opitz and Pechmann (2013) and Opitz et al. (2016) showed that masculine adjectives and nouns in German have an increased processing load as compared to feminine nouns and adjectives, which they take as evidence that masculine nouns are specified for a gender feature while feminine nouns are underspecified. From this they suggest the following principle: if a feature is not used morphosyntactically, then it is not represented mentally. This suggests that underspecification effects such as those observed for gender in German should be widely observed. Given the significant role of underspecification in linguistic theory, much work remains to see if this prediction will be borne out in psycholinguistic research.

### 4.2 Beyond Lexical Access: Processing Polysynthesis

The vast majority of studies on language processing, and most of the studies reviewed so far, have focused on languages that are more analytic than synthetic. However, languages on the synthetic end of the continuum provide an important testing ground for the relationship between morphological and syntactic processing. DM maintains that one language’s word is another languages morpheme, but both can share the same underlying morphosyntactic representation (i.e., representational identity holds).

Indeed, many polysynthetic languages have single words containing an impressive amount of morphology. While fully storage-based theories have at least an *a priori* tenability with languages on the analytic side of the spectrum, with synthetic languages, where a single word can contain a root, voice morphology, multiple exponents or agreement, tense, and aspect – indeed conveying what in languages like English takes a multi-word string – storage theories miss the mark. In particular, they miss capturing the fundamental fact that languages are generative at all levels of representation, including the morphological level. Indeed, a strongly storage-based theory is forced to contend that synthetic languages fail to do much of the "syntactic" processing that more analytic languages do, while having a far more detailed lexicon where complex syntactic structures are stored and accessed in non-decomposed forms. In short, a fully storage-based theory must maintain that mechanistic and algorithmic identity does not hold across languages that differ in morphological complexity.

This criticism raises an empirical question that is in principle testable: does the sort of "word-by-word" incrementality clearly present in analytical languages extend to the morphological level in polysynthetic languages? To our knowledge, there are only a small number of studies that have examined the consequences of polysynthesis in language processing (Rood, 2002; Rice, Libben, & During, 2002; Baker & Bundgaard-Nielsen, 2016). While the sample remains small, the findings of these studies are consistent with the idea that words in polysynthetic languages are decomposed into their constituent parts rather than processed via a whole-word representation. For example, Baker & Bundgaard-Nielsen (2016) show that word-internal pauses in Wubuy (a language indigenous to Australia) are only acceptable when pauses occur at morphemic breaks, and that word-internal pauses are in some cases more acceptable than pauses that occur between word boundaries. These results demonstrate that speakers of morphologically complex languages access the internal structure of complex words.

## 5. CONCLUSION

The goal of this chapter was to build a link between grammatical and processing theories, with a particular focus towards showing the value of the representational assumptions of DM for psycholinguistics. DM provides a path for explaining the distribution of speech errors and the nature of code-switching/blending via the abstract and acategorial nature of roots, is consistent with the significant evidence in favor of a decompositional view of lexical access, and allows for parity between languages with varying degrees of morphological complexity. In short, the evidence supports a view where at least representational identity holds between the word-level and sentence-level, allowing incremental processing to be radically defined as a function of morphosyntactic representations that are independent of orthographic and phonological externalizations.

## References


Abd-El-Jawad, H. & Abu-Salim, I. (1987). Slips of the tongue in Arabic and their theoretical implications. *Language Sciences*, 9(2), 145–71.

Albright, A. 2007. Cheap but not free: Commentary on Pfau 'Cheap repairs'. In C. T. Schütze and V. S. Ferreira (eds.), *The State of the Art in Speech Error Research: Proceedings of the LSA Institute Workshop* (*MIT Working Papers in Linguistics*, 53). Cambridge, MA: MITWPL, 35–51.

Alegre, M. & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40, 41–61.

Andrews, S., Miller, B. & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, 16(1-2), 285–311.

Arnon, I. & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67–82.

Baayen, R. H. (2014). Experimental and psycholinguistic approaches to studying derivation. In R. Lieber and P. Stekauer (eds.), *Handbook of Derivational Morphology*. Oxford: Oxford University Press, 95–117.

Baayen, R. H., Dijkstra, T. & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language,* 37, 94–117.

Baker, A. & van den Bogaerde, B. (2008). Codemixing in signs and words in input to and output from children. In C. Plaza Pust and E. Morales Lopez (eds.), *Sign Bilingualism: Language Development, Interaction, and Maintenance in Sign Language Contact Situations*. Amsterdam: John Benjamins, 1–27.

Baker, B. & Bundgaard-Nielsen, R. (2016). Pause acceptability is predicted by morphological transparency in Wubuy. In *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology*, 197–200.

Bandi-Rao, S. & den Dikken, M. (2014). Light switches: On v as a pivot in codeswitching, and the nature of the ban on word-internal switches. In J. MacSwan (ed.), *Grammatical Theory and Bilingual Codeswitching*. Cambridge, MA: MIT Press, 161–83.

Barner, D. & Bale, A. (2002). No nouns, no verbs: Psycholinguistic arguments in favor of lexical underspecification. *Lingua*, 112, 771–91.

Barner, D. & Bale, A. (2005). No nouns, no verbs? A rejoinder to Panagiotidis. *Lingua*, 115, 1169–79.

Beauvillain, C. (1996). The integration of morphological and whole-word form information during eye fixations on prefixed and suffixed words. *Journal of Memory and Language*, 35(6), 801–20.

Belazi, H. M., Rubin, E. & Toribio, A. (1994). Code switching and X-bar theory: The Functional Head Constraint. *Linguistic Inquiry*, 25, 221–37.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(2), 489–511.

Branchini, C. & Donati, C. (2016). Assessing lexicalism through bimodal eyes. *Glossa: a Journal of General Linguistics*, 1(1), 48. 1–30.

Butterworth, B. (1983). Lexical representation. In B. Butterworth (ed.), *Language Production, Vol. 2: Development, Writing, and Other Language Processes*. London: Academic Press, 257–94.

Cantone, K. F. & Müller, N. (2008). *Un nase* or *una nase*? What gender marking within switched DPs reveals about the architecture of the bilingual language faculty. *Lingua*, 118, 810–26.

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

Chomsky, N. (1995). *The Minimalist Program.* Cambridge, MA: MIT Press.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3), 283–321.

Dell, G. S., Chang, F. & Griffin, Z. M. (1999). Connectionist models of language production: lexical access and grammatical encoding. *Cognitive Science*, 23(4), 517–42.

Dell, G. S. & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42, 287–314.

Dell, G. S. & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior,* 20, 611–29.

Donati, C. & Branchini, C. (2012). Challenging linearization: simultaneous mixing in the production of bimodal bilinguals. In T. Biberauer and I. Roberts (eds.), *Challenges to Linearization*. Berlin: Mouton De Gruyter, 93–128.

Emmorey, K., Borinstein, H. B. & Thompson, R. (2005). Bimodal bilingualism: Code-blending between spoken English and American Sign Language. In J. Cohen, K. T. McAlister, K. Rolstad, and J. MacSwan (eds.), *ISB 4: Proceedings of the 4th International Symposium on Bilingualism*. Somerville, MA: Cascadilla Press, 663–73.

Emmorey, K., Borinstein, H. B., Thompson, R. & Gollan, T. H. (2008). Bimodal bilingualism. *Bilingualism: Language and Cognition*, 11(1), 43–61.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47(1), 27–52.

Fromkin, V. A. (1973). Introduction. In V. A. Fromkin (ed.), *Speech Errors as Linguistic Evidence*. The Hague: Mouton, 11–45.

Gandour, J. (1977). Counterfeit tones in the speech of Southern Thai bidialectals. *Lingua*, 41, 125–43.

Garrett, M. F. (1975). The analysis of sentence production. In G. Bower (ed.), *Psychology of Learning and Motivation, vol. 9*. New York: Academic Press, 133–77.

Garrett, M. F. (1980a). Levels of processing in sentence production. In B. Butterworth (ed.), *Language Production, Vol.1: Speech and Talk*. London: Academic Press, 177–220.

Garrett, M. F. (1980). The limits of accommodation: Arguments for independent processing levels in sentence production. In V. A. Fromkin (ed.), *Errors in Linguistic Performance. Slips of the Tongue, Ear, Pen, and Hand*. New York: Academic, 263–71.

Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior*, 22(1), 24–44.

Grimstad, M. B., Lohndal, T. & Åfarli, T. A. (2014). Language mixing and exoskeletal theory: A case study of word-internal mixing in American Norwegian. *Nordlyd*, 41, 213–47.

Grosjean, F. (1982). *Life with Two Languages: An Introduction to Bilingualism*. Cambridge, MA: Harvard University Press.

Halle, M. (1997). Distributed Morphology: Impoverishment and fission. In B. Bruening, Y. Kang, and M. McGinnis (eds.), *PF: Papers at the Interface* (*MIT Working Papers in Linguistics* 30), 425–49.

Halle, M. & A. Marantz. (1993). Distributed morphology and the pieces of inflection. In K. Hale and S. J. Keyser (eds.), *The View from Building 20. Essays in Linguistics in Honour of Sylvain Bromberger*. Cambridge, MA: MIT Press, 111–76.

Hammerly, C. (2019). Limiting gender. In É. Mathieu, M. Dali, and G. Zareikar. (eds.), *Gender and Noun Classification*. Oxford: Oxford University Press, 93–118.

Harley, H. (2014). On the identity of roots. *Theoretical Linguistics*, 40(3/4), 225–76.

Harley, H. (2019). Semantics in Distributed Morphology. In C. Maienborn, K. Heusinger, and P. Portner (eds.), *Semantics – Interfaces*. Berlin: De Gruyter Mouton, 143–186.

Harley, H. & Noyer, R. (1998). Licensing the non-lexicalist lexicon: Nominalizations, Vocabulary Items and the Encyclopaedia. *MIT Working Papers in Linguistics*, 32, 119–37.

Harley, H. & Noyer, R. (2003). Distributed Morphology. In L. Cheng and R. Sybesma (eds.), *The Second Glot International State-of-the-article Book. The Latest in Linguistics*. Berlin: Mouton de Gruyter, 463–96.

Hohenberger, A. & Leuninger, H. (2012). Production. In R. Pfau, M. Steinbach, and B. Woll (eds.), *Sign Language. An International Handbook* (*HSK – Handbooks of Linguistics and Communication Science*). Berlin: De Gruyter Mouton, 711–38.

Hyönä, J. & Pollatsek, A. (1998). Reading Finnish compound words: Eye fixations are affected by component morphemes. *Journal of Experimental Psychology: Human Perception and Performance*, 24(6), 1612–27.

Jacobs, C. L., Dell, G. S., Benjamin, A. S. & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, 87, 38–58.

Jake, J., Myers-Scotton, C. & Gross, S. (2002). Making a minimalist approach to codeswitching work: Adding the matrix language. *Bilingualism: Language and Cognition*, 5, 69–91.

Juhasz, B. J. (2008). The processing of compound words in English: Effects of word length on eye movements during reading. *Language and Cognitive Processes*, 23(7-8), 1057–88.

Juhasz, B. J., Starr, M. S., Inhoff, A. W. & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology*, 94(2), 223–44.

Katz, J.J. (1964). Mentalism in linguistics. *Language*, 40(2), 124–37.

Kihm, A. (2005). Noun class, gender, and the lexicon-syntax-morphology interfaces: A comparative study of Niger-Congo and Romance languages. In G. Cinque, and R. S. Kayne (eds.), *The Oxford Handbook of Comparative Syntax*. Oxford: Oxford University Press, 459–512.

Kramer, R. (2015). *The Morphosyntax of Gender*. Oxford: Oxford University Press.

Kuperman, V., Bertram, R. & Baayen, H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23, 1089–132.

Kuperman, V., Bertram, R. & Baayen, H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, 62, 83–97.

Kuperman, V., Schreuder, R., Bertram, R. & Baayen, H. (2009). Reading polymorphemic Dutch compounds: Toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 876–95.

Kush, D., Dillon, B., Eik, R. & Staub, A. (2019). Processing of Norwegian complex verbs: Evidence for early decomposition. *Memory & Cognition*, 47(2), 335–50.

Levelt, W. J. M. (1989). *Speaking. From Intention to Articulation*. Cambridge, MA: MIT Press.

Liceras, J. M., Fernández Fuertes, R., Perales, S., Pérez-Tattam, R. & Spradlin, K. T. (2008). Gender and gender agreement in bilingual native and non-native grammars: A view from child and adult functional-lexical mixings. *Lingua*, 118, 827–51.

Liceras, J. M., Spradlin, K. T. & Fernández Fuertes, R. (2005). Bilingual early functional-lexical mixing and the activation of formal features. *International Journal of Bilingualism*, 9(2), 227–52.

Lillo-Martin, D., de Quadros, R. M. & Chen Pichler, D. (2016). The development of bimodal bilingualism: Implications for linguistic theory. *Linguistic Approaches to Bilingualism*, 6(6), 719–55.

MacSwan, J. (1999). *A Minimalist Approach to Intrasentential Code Switching*. New York, NY: Garland Publishing.

MacSwan, J. (2000). The architecture of the bilingual language faculty: Evidence from intrasentential code switching. *Bilingualism: Language and Cognition*, 3(1), 37–54.

Marantz, A. (1995). A late note on late insertion. In Y.-S. Kim, K.-J. Lee, B.-C. Lee, H.-K. Yang, and J.-Y. (eds.), *Explorations in Generative Grammar: A Festschrift for Dong-Whee Yang*. Seoul: Hankuk Publishing, 396–413.

Marantz, A. (1997). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania Working Papers in Linguistics*, 4(2), 201–25.

Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, 28(7), 905–16.

Marx, E. (1999). Gender processing in speech production: Evidence from German speech errors. *Journal of Psycholinguistic Research*, 28(6), 601–21.

Matthews, R. J. (1991). Psychological reality of grammars. In A. Kasher (ed.), *The Chomskyan Turn*. Oxford: Blackwell, 182–99.

McClelland, J. & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences*, 6, 465–72.

Momma, S. & Phillips, C. (2018). The relationship between parsing and generation. *Annual Review of Linguistics*, 4, 233–54.

Muysken, P. (2000). *Bilingual Speech: A Typology of Codemixing.* Cambridge: Cambridge University Press.

Myers-Scotton, C. (1993). *Duelling Languages: Grammatical Structure in Codeswitching.* Oxford: Clarendon Press.

Newkirk, D., Klima, E. S., Pedersen, C. C. & Bellugi, U. (1980). Linguistic evidence from slips of the hand. In V. A. Fromkin (ed.), *Errors in Linguistic Performance. Slips of the Tongue, Ear, Pen, and Hand*. New York: Academic Press, 165–97.

Niswander, E., Pollatsek, A. & Rayner, K. (2000). The processing of derived and inflected suffixed words during reading. *Language and Cognitive Processes*, 15(4-5), 389–420.

Nunes, J. & de Quadros, R. M. (2008). Phonetically realized traces in American Sign Language and Brazilian Sign Language. In J. Quer (ed.), *Signs of the Time. Selected Papers from TISLR 8*. Hamburg: Signum, 177–90.

Opitz, A. & Pechmann, T. (2016). Gender features in German. *The Mental Lexicon*, 11(2), 216–41.

Opitz, A., Regel, S., Müller, G. & Friederici, A. D. (2013). Neurophysiological evidence for morphological underspecification in German strong adjective inflection. *Language*, 89(2), 231–64.

Oseki, Y. (2018). *Syntactic Structures in Morphological Processing.* New York, NY: New York University PhD dissertation.

Panagiotidis, P. (2005). Against category-less roots in syntax and word learning: objections to Barner and Bale (2002). *Lingua*, 115, 1181–94.

Petronio, K. & Lillo-Martin, D. (1997). WH-movement and the position of Spec-CP: Evidence from American Sign Language. *Language*, 73(1), 18–57.

Pfau, R. (2000). *Features and Categories in Language Production*. Frankfurt: Goethe University Frankfurt PhD dissertation.

Pfau, R. (2007). Cheap repairs: A Distributed Morphology toolkit for sentence construction. In C. T. Schütze and V. S. Ferreira (eds.), *The State of the Art in Speech Error Research: Proceedings of the LSA Institute Workshop* (*MIT Working Papers in Linguistics*, 53). Cambridge, MA: MITWPL, 9–33.

Pfau, R. (2009). *Grammar as Processor: A Distributed Morphology Account of Spontaneous Speech Errors* (Linguistik Aktuell 137). Amsterdam: John Benjamins.

Pfau, R. (2016). Switching, blending … and slipping (commentary on Lillo-Martin, Müller de Quadros & Chen Pichler). *Linguistic Approaches to Bilingualism*, 6(6), 802–7. [doi:10.1075/lab.6.6.11pfa].

Pierantozzi, C. (2012). Agreement within early mixed DP: What mixed agreement can tell us about the bilingual language faculty. In K. Braunmüller and C. Gabriel (eds.), *Multilingual Individuals and Multilingual Societies.* Amsterdam: John Benjamins, 137–52.

Pinker, S. (1999). *Words and Rules*. New York, NY: Basic Books.

Plaut, D., McClelland, J., Seidenberg, M. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115.

Pollatsek, A., Hyönä, J. & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2), 820–33.

Pollatsek, A., Slattery, T. J. & Juhasz, B. (2008). The processing of novel and lexicalised prefixed words in reading. *Language and Cognitive Processes*, 23(7-8), 1133–58.

Rastle, K., Davis, M. H. & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11(6), 1090–8.

Rice, S., Libben, G. & Derwing, B. (2002). Morphological representation in an endangered, polysynthetic language. *Brain and Language*, 81(1-3), 473–86.

Rohde, D. L. T. & Plaut, D. C. (2003). Connectionist models of language processing. *Cognitive Studies*, 10(1), 10–28.

Rood, D. S. (2002). Polysynthetic word formation: Wichita contributions to the morphology/syntax debate. In S. Bendjaballah, W. U, Dressler, O. E. Pfeiffer, and M. D. Voeikova (eds.), *Morphology 2000: Selected Papers from the 9th Morphology Meeting.* Amsterdam: John Benjamins, 293–394.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland and D. E. Rumelhart (eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition, Vol.2: Psychological and Biological Models.* Cambridge, MA: MIT Press, 216–71.

Schade, U. (1992). *Konnektionismus. Zur Modellierung der Sprachproduktion.* Opladen: Westdeutscher Verlag.

Seidenberg, M. & McClelland, J. (1999) A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523–68.

Siddiqi, D. (2009). *Syntax Within the Word: Economy, Allomorphy, and Argument Selection in Distributed Morphology*. Amsterdam: John Benjamins.

Siddiqi, D. (2010). Distributed Morphology. *Language and Linguistics Compass*, 4(7), 524–42.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, 7(4), 263–72.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57, 745–65.

Taft, M. & Forster, K. I. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14, 638–47.

Taft, M. & Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, 15(6), 607–20.

Tieu, L. (2010). Re-examining cross-linguistic influence in the bilingual acquisition of WH-questions. In M. Heijl (ed.), *Proceedings of the Annual Conference of the Canadian Linguistic Association*. Toronto: University of Toronto, 1–15.

Zwitserlood, P. (2018). Processing and representation of morphological complexity in native language comprehension and production. In G. Booij (ed.), *The Construction of Words: Advances in Construction Morphology*. Dordrecht: Springer, 583–602.