A quantitative approach to the grammaticalization of discourse markers
Evidence from their sequencing behavior*

Christian Koops and Arne Lohmann
University of New Mexico / University of Vienna

This article takes a quantitative approach to the grammar of English two-part discourse marker sequences like oh well, you know I mean, etc. We investigate the internal ordering preferences of such sequences in spoken American English corpus data from the perspective of grammaticalization. From this perspective, the development of many discourse markers can be understood as involving a process of increasing syntactic de-categorialization (Hopper 1991) as the grammaticalizing element loses its original grammatical constraints and comes to function as a marker at the level of discourse. We test the hypothesis that discourse marker grammaticalization results in largely unconstrained ordering possibilities. Our analysis shows that, on the contrary, discourse marker sequencing is highly constrained. We interpret these constraints in terms of Auer’s (1996) model of discourse marker grammaticalization. Discourse marker sequencing is characterized by strong persistence of a marker’s original syntactic category and reflects its specific grammaticalization trajectory.

Keywords: discourse markers, grammaticalization, discourse marker sequences, English, spoken language corpora

1. Discourse markers and grammaticalization

This article deals with the grammatical properties of discourse markers (DMs), specifically their ordering preferences relative to one another. While the data

* We thank the audiences at HDLS 10 in Albuquerque and BLS 39 in Berkeley as well as the anonymous IJCL reviewers for their comments on earlier versions of this paper. All remaining errors are our own. The analysis presented here is a revised and expanded version of Koops & Lohmann (in press).
presented here are synchronic in nature, we approach the topic from the perspective of grammaticalization. From this perspective, DMs can be understood as the result of a process in which elements serving other functions, including grammatical functions at the level of sentential syntax, come to be conventionally used as markers of discourse-level relations, or what Schiffrin (1987: 31) operationally defines as “sequentially dependent elements which bracket units of talk.” We are concerned with the final outcome of this process. Do fully formed DMs still have grammar? Specifically, to what degree do DMs still retain aspects of their former syntactic ordering restrictions?

Our approach requires us to clarify how DMs grammaticalize in general. To begin, DMs constitute a broad and diverse class of elements with different developmental trajectories. In order to capture this diversity, we draw on Auer’s (1996) taxonomy of relevant grammaticalization processes, which covers a wide range of types. Auer’s (1996) model was developed specifically to describe grammaticalization in the syntactic position of spoken German known as the ‘pre-front field’ (Vor-Vorfeld). We assume that this position is broadly comparable in discourse-functional terms to the utterance-initial, extra-sentential position in which many English discourse markers appear (e.g. Schiffrin 2001).

Auer (1996) identifies two grammaticalization paths, or ‘clines’, along which elements evolve to occupy this position. The first involves elements that already start out as syntactically independent structures. Examples include “vocatives and other […] constituents which may be used as summons in conversation” (Auer 1996: 314). Later on, this “dialogical, sequential structure is condensed and ‘compacted’ into a grammatical one in which the structure is no longer open to interactional negotiation, but weakly (and later on possibly strongly) adjoined grammatically” (Auer 1996: 313). There are obvious English equivalents to the types identified by Auer (1996), including vocatives (boy, man), imperatives (listen, look), interjections (oh, wow), and forms of assessment and agreeing responses (well, sure, right). We will refer to this grammaticalization cline as the ‘(a)-path’.

Auer’s (1996) second cline involves elements originally governed by sentential syntax. As these grammaticalize into DMs, “a constituent moves out of the grammatical centre of the sentence into its periphery” (Auer 1996: 313). The types of elements found on this cline also have well-known counterparts in English, for example adverbials (like, anyway) and matrix clauses (I mean, I guess). Auer’s (1996) discussion makes it clear that discourse markers identical in form to conjunctions (and, because) also evolve on this cline. We will refer to it as the ‘(b)-path’.

Figure 1 summarizes the two grammaticalization paths. Note that DMs on the (a)-path come to occupy the utterance-initial DM slot from the left, as it were, while DMs on the (b)-path move into this position from the right.
The (b)-path of DM grammaticalization has been investigated in detail, for example in early case studies of English discourse markers such as *like* (Romaine & Lange 1991) and *I think* (Thompson & Mulac 1991). In order to show that a structure has attained DM status, many studies point to what Hopper (1991) has called ‘de-categorization’, i.e. the loss of the morphosyntactic properties generally associated with the relevant grammatical category. For example, Thompson & Mulac (1991) argue that the disproportionately high rate of omission of the complementizer *that* after *I think* and similar combinations of main clause subjects and verbs shows that these structures are not subject to the rules of sentential syntax in the same way that genuine matrix clauses are.

Note that in this common view of DM grammaticalization the perspective taken is that of the source construction. In effect, DM status is defined negatively. It is inferred from the lack of some otherwise expected grammatical behavior. Taking this perspective is an important method, especially in order to diagnose early stages of grammaticalization. But how does one best approach the grammar of DMs at more advanced stages of grammaticalization, after they have developed discourse functions that no longer resemble the meaning of their sources in an obvious way? The de-categorialization perspective, in which an increased usage potential as DM is implicitly equated with a decrease in grammatical constraints, suggests that at some point DMs may become so dissociated from their sources that they are essentially devoid of grammar. Their placement then becomes entirely a question of discourse function, rather than grammatical rules. The first aim of this paper is to test this strong hypothesis. Does DM grammaticalization in fact imply a massive loss of grammatical restrictions? Or, alternatively, do even mature DMs still show Hopper’s (1991) ‘persistence’, i.e. the retention of the constraining influence of their diachronic sources?

The specific grammatical behavior we investigate is linear ordering. We use the ordering of DMs relative to other DMs as a window on the question of de-categorialization versus persistence. Consider the English DMs *but* and *I mean*. Both are conventionalized DMs, each with its own range of discourse functions (Schiffrin 1987). Crucial for us, speakers sometimes produce both together in a sequence, in one of two orders: *but I mean* or *I mean but*. Only the former is consistent with the grammar of the source elements, a coordinating conjunction.

![Figure 1. Two grammaticalization paths for discourse markers](image)

a. e.g. *boy, look, well, etc.*  →  DM

b. DM  ←  e.g. *I mean, like, anyway, etc.*
A quantitative approach to the grammaticalization of discourse markers

(but) and a matrix clause (I mean). To the degree that DMs display persistence, this order should be used exclusively or at least preferred. On the other hand, the hypothesis of massive or total de-categorialization predicts that there should be no such constraint or preference.

The question of syntactic persistence applies more readily to DMs on the (b)-path. As their source syntax is well understood, it is easy to anticipate what form syntactic persistence should take (as in the example of but and I mean). This is less clear in the case of (a)-type DMs, such as DMs originating in interjections or vocatives (e.g. oh or boy). As these derive from independent utterances, they do not bring with them any obvious external syntax. Still, the (a)-path makes one general prediction, which applies to this class of DMs as a whole. We should expect (a)-type DMs, which come to occupy the DM slot “from the left”, to continue to precede (b)-type DMs, which enter this position “from the right”. For example, in the case of oh, we should expect that speakers prefer oh I mean over I mean oh. Placement of (a)-type DMs in the initial position would indicate persistence in that they continue to act as relatively self-contained elements, as if they still made their own, independent contribution in addition to prefacing a forthcoming utterance.

2. Discourse marker sequences

It is well-known that DMs are often used in two-part sequences, such as oh well or but then (e.g. Svartvik 1980: 170 on well). It has also been pointed out that such sequences may hold interesting analytical insights. Schourup (1999: 233) suggests that exploring the constraints on DM sequencing “may provide important arguments for semantic analyses of particular DMs.” An example of such an argument is Aijmer’s (2002) analysis of now, in which she points to the sequences so now and now therefore to argue that now, unlike well, is “oriented toward the upcoming topic” (Aijmer 2002: 64). Nevertheless, as noted by Fraser (2011), the phenomenon has received surprisingly little attention in the DM literature. It seems today still true that “the restrictions on such [DM] clustering are not yet well understood” (Schourup 1999: 233). The only quantitative analyses of DM sequencing we are aware of come from the field of automatic text generation (Knott 1996, Oates 2000) and are restricted to DMs in written discourse. The significance of sequencing constraints for theories of DM grammaticalization, which we are dealing with here, remains unexplored.

Some DM sequences, for example oh well or and then, stand out through their particularly high discourse frequency. Such high-frequency sequences are reminiscent of what have been called ‘lexical bundles’ in corpus linguistic research.
Biber et al. (2002: 443) define these as sequences of words that are used at least ten times per million words. One class of lexical bundles that seems to be functionally similar to DM sequences are ‘discourse organizers’ (Biber & Conrad 2004:66). Given this overlap, it is important to clarify the differences between the two notions. One practical difference is that lexical bundles are defined as three- to four-word sequences, while many of our sequences consist of only two words. A more fundamental difference is that our analysis is not primarily concerned with the question whether two DMs form a conventionalized collocation. Rather, we take the general tendency for DMs to be used in multi-part sequences as a window on the constraints on their ordering in such sequences. Put another way, our analysis is less concerned with whether a DM sequence has an unexpectedly high frequency relative to the frequencies of its component parts. Rather, we are interested in whether the frequency of one order is unexpectedly higher than the frequency of the opposite order. That is not to say that DM bundling (or, clustering) and DM sequencing (in the sense of “ordering”) are unrelated. Ultimately, the two phenomena should receive a unified account. However, we believe that they can be investigated independently. Here we deal primarily with the latter.

As noted above, for two-part DM sequences the question of a DM’s relative position can be stated as follows: What determines which DM appears in first and which in second position? For example, is their relative placement determined by their source syntax? Or is it relatively unpredictable because having attained DM status is equivalent to complete syntactic de-categorialization? In line with the latter view, Schiffrin (1987) argues that the combination of two DMs in syntactically non-canonical sequences, as in Examples (1) and (2), indicates that the elements in question are, in fact, DMs.1

(1) They don’t even stop. **So: and** they said that they can’t even accommodate us.

(2) **And** uh … **but** they have that— they’re— they’re so conscious of their um … they’re always sittin’ down and figurin’ out their averages.

(Schiffrin 1987: 39, boldface and transcription conventions as in the original)

In Example (1), the apparent co-occurrence violation consists in having a coordinate conjunction preceded by *so*, rather than the other way around. In Example (2), according to Schiffrin (1987), the illicit co-occurrence of two coordinate conjunctions is made possible because *and* functions as a DM. Our empirical analysis is concerned with the status of examples like these. How regularly do DMs combine in a non-normative order, e.g. *so and* rather than *and so*?

---

1. The colon in Example (1) marks a lengthened syllable (see Schiffrin 1987: x).
Examples like (1) and (2) further raise the question: What exactly does it mean for two DMs to be used ‘in sequence’? In spontaneous spoken language, it does not go without saying that two superficially adjacent DMs were meant to be produced and understood together. One might argue, for instance, that Examples (1) and (2) illustrate self-repair. To decide their status, the prosodic shape of each element needs to be taken into account. Our analytic criteria, described in Section 5, therefore also involve prosody.

In summary, we test the following hypotheses and specific predictions. At the most general level, we test the null hypothesis $H_0$ that DMs are in fact devoid of grammar and that their sequencing is unconstrained. The prediction of $H_0$ is that for a given DM pair the likelihood of observing one or the other order is statistically indistinguishable from chance. $H_0$ is opposed to $H_1$, according to which DMs do have (some) grammar. $H_1$ predicts that DM sequencing is measurably constrained. We should find either a significant distributional bias or a total ban on one order.

To the extent that $H_1$ is supported, we can further test whether the observed constraints reflect the nature of the relevant DMs’ source constructions and their grammaticalization paths. DMs that evolved on the (a)-path should precede DMs that evolved on (b)-path. For example, markers like *oh* should precede markers like *I mean*. We will call this hypothesis $H_{1a}$. Moreover, DMs on the (b)-path should occur in sequences that do not violate the order expected from their source syntax. For example, markers deriving from coordinating conjunctions like *but* should precede markers deriving from matrix clauses like *I mean*. We will call this hypothesis $H_{1b}$.

3. Discourse markers analyzed

To test our predictions we used the set of eleven DMs investigated in Schiffrin’s (1987) foundational study *Discourse Markers*. Drawing on Schiffrin’s (1987) analysis has the advantage of providing us with a set of DMs whose status was independently established on the basis of a unified definition. The eleven DMs are given below, subdivided according to each marker’s assumed path of development:

a. *oh*, *well*
b. *and*, *but*, *or*, *so*, *because*, *now*, *then*, *you know*, *I mean*

Further advantages of Schiffrin’s (1987) set are its size and diversity. It includes DMs of both the (a)-type and the (b)-type, and several subtypes of the latter. This makes it possible to test each of our three hypotheses on the basis of multiple DMs. Also, the broad coverage that Schiffrin’s (1987) set provides allows us to determine
whether positional constraints on specific DM pairs combine into a larger system that governs DMs overall, such as a sequencing hierarchy (Oates 2000).

We assume that *oh* and *well* evolved on the (a)-path of DM grammaticalization, while the other nine DMs evolved on the (b)-path. This classification is based on accounts of their historical development in the literature, which we briefly summarize here. *Oh* is generally understood as belonging to the class of interjections (Schiffrin 1987: 73, Aijmer 2002: 97), which Aijmer (2002), referring to Jespersen (1968: 90), defines as “short elements which have the ability to stand alone as complete utterances” (Aijmer 2002: 97). Thus, *oh* starts out as an independently deployable conversational unit before it comes to be used also as a DM to signal a forthcoming utterance. This describes Auer’s (1996) (a)-path.

The literature includes two competing theories of the diachronic development of *well* (see Lutzky 2012: 74–77 for a summary). Finell (1989) suggests that DM *well* derives from the use of *well* as a predicative adjective in Middle English, as in *that is very well*, via a process of shortening. Jucker (1997), on the other hand, argues that DM *well* developed from Old English *wella* or *wella*, which functioned as an attention-getting device and in Middle English was used to introduce direct speech. Both analyses are compatible with Auer’s (1996) (a)-path because in either case *well* starts out as a syntactically independent element.

The DMs *and*, *or*, *but*, *so* and *because* have clear counterparts in coordinating and subordinating conjunctions (Schiffrin 1987). The process whereby conjunctions lose their syntactic dependency on sentence structure and come to be used as optional, utterance-initial markers of discourse relations is captured by Auer’s (1996) (b)-path. The case of the DMs *now* and *then* is similar (see Schiffrin 1987: 228–266). They originate as temporal adverbials used sentence-internally and then develop into DMs occurring at the periphery, including utterance-initially.

For *I mean*, Brinton (2008: 124–127) discusses three possible origins. First, it may have developed out of a matrix clause *I mean (that) + Subject*, similar to the development of *I guess* and *I think* proposed by Thompson & Mulac (1991). A second possibility is an adverbial or relative structure of the form (*as/so/which*) *I mean* (Brinton 1996). The third possibility, advocated by Brinton (2008), is that DM *I mean* originates in contexts where it is used to govern a phrasal element, e.g. *I mean + Noun Phrase*. Through grammaticalization, the bonds between *I mean* and the NP are loosened and *I mean* comes to function as an independent parenthetical with a larger scope (Brinton 2008: 127). All three analyses have in common that *I mean* was first syntactically integrated into a sentence before relaxing its syntactic restrictions. This development is captured by Auer’s (1996) (b)-path.

A similar scenario as for *I mean* has been put forward for *you know*. According to Fitzmaurice (2004: 431), *you know* first functioned as “a matrix verb phrase
which governs a complement clause”. As a DM, it was then used “as an unanalyzable whole with considerable freedom of occurrence” (Fitzmaurice 2004: 437). Another possibility would be a development from a relative structure as you know, as suggested by Brinton (2008: 141–161) for you see. Whatever the case may be, you know parallels I mean in that during its development you know apparently severed its syntactic ties to the following clause. We therefore understand it as having evolved on the (b)-path as well.

4. Database and methodology

Our database is the Fisher corpus (Cieri et al. 2004a, b, 2005a, b), a collection of 16,000 10-minute telephone conversations between speakers of American English. It contains about 20 million words. In employing this database, our study is in the spirit of other recent work that has demonstrated the importance of corpora for the study of discourse markers (e.g. Andersen 2001; Aijmer 2002, 2013). Moreover, our study forms part of the growing trend in the field of corpus linguistics to work closely with spoken corpus data. This includes, for example, analyzing prosodic information on the basis of the associated audio files (see Adolphs & Carter 2013).

We determined the corpus frequencies of all 110 theoretically possible pairwise combinations of the eleven DMs. For several reasons, determining these frequencies was not a straightforward process. To begin, the Fisher transcripts contain no lexical, syntactic, or prosodic mark-up. This makes it impossible to automatically identify all instances in which a particular form, say the word so, functions as a DM rather than, for example, as an adverb (e.g. so good). Our approach to this problem was to start with exhaustive searches of each orthographic sequence in the corpus transcripts, for example the contiguous word sequence and so. The search results for each sequence were stored in separate concordance files. The number of hits in these 110 original concordance files ranged from as few as 14 hits for because or to as many as 32,298 hits for and then (median number of hits: 487).

We then analyzed the hits contained in each concordance file more closely to determine how many of the superficial matches are DM sequences. Given the large number of hits for many of the orthographic sequences, we opted for an approach that provided us with a reasonably accurate estimate for the high-frequency sequences. In the case of concordance files containing 120 hits or less (27 of the 110 files), we inspected all hits. In the case of concordance files containing more hits than that, we inspected a random sample of 120 and estimated the number of genuine DM sequences in the whole file on the basis of that sample. Our method of frequency estimation was to multiply the total number of hits in the concordance
file by the ratio of genuine-to-spurious hits in the sample of 120. For example, the raw number of hits for and so was 11,961. Applying the selection criteria discussed below to a random sample of 120 revealed 75 genuine DM sequences. The estimated total number of genuine and so sequences in the whole concordance file was then determined by multiplying 11,961 by 0.625 (i.e. 75/120), which comes to 7,476 cases.

4.1 Selection criteria

In the following sections, we explain the four sets of criteria by which we decided whether a superficial match counts as a genuine DM sequence. These are: lack of obligatoriness (Section 4.1.1), prosodic integration (Section 4.1.2), utterance-initial position (Section 4.1.3), and criteria pertaining to sequences of more than two DMs (Sections 4.1.4).

4.1.1 Lack of obligatoriness

Our first criterion in distinguishing DMs from their formally identical non-DM counterparts was the lack of obligatoriness (Schiffrin 1987: 64, 2001: 57). If omitting one or both elements in a given sequence resulted in either an incomplete syntactic structure or a significant change in semantic meaning, the item was discarded. In addition, we applied the following criteria pertaining to individual DMs. We discarded cases in which and, but, and or precede a constituent smaller than a complete clause (Schiffrin 1987: 128). In the case of now and then, we discarded those cases in which they function semantically as temporal modifiers of an event (Schiffrin 1987: 230–232, see also 246–248). For because we slightly refined Schiffrin’s (1987: 191–217) criteria. We did not analyze because as a DM when the because-clause preceded the main clause, i.e. the clause or set of clauses describing the situation that the speaker is explaining. We also only included cases in which the because-clause had the form of a fragment, i.e. where it was separated from the main clause by another syntactic construction or by some form of discontinuity, or where there was no clear antecedent at all.

4.1.2 Prosodic integration

Having reduced the data to sequences in which both elements qualify as DMs, we went on to analyze their prosodic form. The goal of this analysis was to use prosodic criteria to eliminate spurious DM sequences, i.e. sequences whose phonetic form makes it unlikely that the two DMs were meant to be understood together (recall this possibility from Examples (1) and (2) above). Spurious sequences may arise specifically from what Levelt & Cutler (1983) call ‘instant repair,’ such that the speaker places one DM immediately after the other to indicate that the former
replaces the latter. Instant repair is typically marked prosodically. The associated phonetic features found in our data are consistent with those described by Cutler (1983) and Levelt & Cutler (1983). Either the end of the originally produced structure, the beginning of the corrected structure, or both are produced in a manner that creates a salient prosodic discontinuity at the editing point.

To keep our analysis consistent with published work on American English prosody, we used the analytic categories of Du Bois et al.'s (1993) discourse transcription system. Like other prosody frameworks, their system recognizes that speech is produced in terms of smaller prosodic constituents, which are called ‘intonation units’ (IUs) in this framework but known in other traditions as ‘intonational phrases’ (e.g. Pierrehumbert 1980) or ‘tone groups’ (e.g. Cruttenden 1997). Du Bois et al. (1993: 47) define an IU broadly as “a stretch of speech uttered under a single coherent intonation contour”. The most important phonetic correlate of such a contour is a continuous pitch development. An interrupted pitch contour creates a prosodic boundary. As a secondary cue besides pitch, the end of an IU is also often characterized by a decrease in tempo. A third potential cue are pauses. Following Du Bois et al. (1993), we did not consider a pause per se as an indicator of an IU boundary, but only where there were concomitant discontinuities in pitch or tempo.

As is common practice in discourse transcription, our analysis was largely auditory. To confirm the interpretation of ambiguous cases, we also occasionally inspected pitch tracks using the acoustic analysis software Praat (Boersma & Weenink 2014). While listening to the data, we further identified and discarded any remaining cases in which the Fisher corpus transcribers failed to record intervening words or other vocalizations, and we discarded a small number of instances in which, upon auditory inspection, the recording did not match the transcript, or where a DM was incompletely produced.

We first distinguished cases in which the two DMs are separated by a prosodic boundary, regardless of its type, from those in which they were produced as part of one intonation unit. In the following, we call DM sequences that display a continuous prosodic contour ‘strongly integrated’. An example is given in (3). Note that in Du Bois et al.’s (1993) transcription system, which we use for all examples cited from the corpus, each line represents one IU. The nature of the IU-final prosodic

2. Du Bois et al.’s (1993) system is an established framework for the transcription of conversational discourse. It was used, for example, to transcribe the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000). While its inventory of distinctions is smaller than that of other prosody frameworks such as ToBI (Beckman et al. 2005), it includes several particularly useful distinctions pertaining to our phenomenon of interest, prosodic boundary types.
shape is marked at the end of each line using one of several symbols (discussed below). The relevant DM sequence is marked by boldface.\(^3\)

\(3\) Caller A: And that still wasn’t enough. Considering it was only part-time, you know.

Caller B: Yeah.

Caller A: But I mean I, I did work for a little bit,

(Fisher corpus file 3513, 396.4–401.5 sec)

The pitch and amplitude contours associated with the sequence but I mean in Example (3) are shown in Figure 2. Note that the transition between the DMs shows a generally continuous pitch development. For example, there is no drop to a relatively low pitch at the end of but.

![Figure 2. Pitch contour (middle) and amplitude contour (top) associated with the strongly integrated DM sequence but I mean in Example (3)](image)

Strongly integrated sequences constitute the most definitive kind of DM sequencing data. The lack of an intervening prosodic boundary of any kind almost guarantees that the speaker intended for the two DMs to be understood together. Therefore, in order to avoid artifacts, we could have restricted the analysis to this type. However, there are reasons for considering cases in which DMs are not fully integrated prosodically as well. First, most prosodic boundaries are unrelated to repair. Therefore, excluding all cases that fall short of full prosodic integration would mean to ignore potentially relevant data. Second, the DM literature suggests

\(^3\) To maintain easy readability, our transcriptions utilize only a subset of Du Bois et al.’s (1993) distinctions. In addition to those explained in the text, we use the equal sign for lengthening, two dots for pauses between 0.3 sec and 0.5 sec, three dots for pauses longer than 0.5 sec, square brackets for overlapped speech, and capitalization for IU-initial words following any prosodic boundary other than a ‘continuing’ transition.
that the lack of complete prosodic integration into the following utterance is, in fact, an expected feature of some DMs (e.g. Hirschberg & Litman 1993, cf. Dehé & Wichmann 2010). It would therefore not be surprising to find some DMs also prosodically set off from a following DM.

To allow for a certain degree of prosodic separation of two DMs in a genuine DM sequence, we further distinguished two categories of boundaries and, accordingly, two further sequence types. The first boundary type is what Du Bois et al. (1993) call a ‘continuing’ transition, which in their system is marked by a comma. This transition type indicates, broadly speaking, that the speaker’s utterance is still unfolding, i.e. more talk is forthcoming. Phonetically, it is “often realized in English as a slight rise in pitch at the end of an intonation unit (beginning from a low or mid level), but it may have other realizations as well […] [including] a terminal pitch that remains level […] [or] a terminal pitch that falls slightly, but not low enough to be considered final” (Du Bois et al. 1993: 54, see below on ‘final’ contours). Perceptually, these patterns share in common the impression of a relatively inconspicuous, soft boundary. We will refer to cases in which the two markers show this boundary type as ‘weakly integrated’. An example is the sequence *I mean but* in Example (4).

(4) Caller B: O=h, that’d be great.
Caller A: Yeah.
   I mean,
   but I just,
   I don’t talk a lot,
   (Fisher corpus file 7381, 518.1–522 sec)

The pitch and amplitude contours of *I mean but* in Example (4) are shown in Figure 3. The pitch pattern seen here is the one Du Bois et al. (1993: 54) describe
as “a terminal pitch that falls slightly, but not low enough to be considered final”. Note the pitch reset back to the speaker’s middle range at the onset of *but*.

Our last category of prosodic boundaries was designed specifically to identify spurious DM sequences. It includes four fairly diverse phonetic types that share in common the auditory impression of a sharp prosodic boundary, i.e. a salient discontinuity. We will call the corresponding DM sequences ‘non-integrated’. One of the four subtypes occurs far more frequently than the other three. We therefore discuss the three less frequent types only briefly.

The first of the three rare patterns is a high rising final pitch, which Du Bois et al. (1993) call ‘appeal’ intonation. It is the contour associated with English *yes-no* questions, and indicated by a question mark. It is found with some regularity on the DM *you know*. The second type generally corresponds to the intonation contour Du Bois et al. (1993) call ‘final’, which is defined as a fall to a low pitch. It is indicated by a period. The third of the less frequent types is a prosodic pattern associated with truncated words. We use it to capture cases where, even though all segments are identifiable, the word offset coincides with an abrupt cessation of the projected pitch contour and sudden, prominent glottalization. In Du Bois et al’s (1993) system this type is marked by a double hyphen.

The fourth and more frequent type has the features Cutler (1983) describes as typical of self-repair. The most common feature is a pitch reset to a considerably higher level than that at which the first DM was produced. Often the sudden and conspicuous rise in pitch is accompanied by a simultaneous, prominent increase in amplitude, tempo, or both. We took any of these three features as sufficient to identify this type. Note that unlike the prosodic phenomena discussed thus far, in this case it is the onset of the second DM that is modified, not the end of the first DM. For example, in the sequence *but and* in (5), repair is not marked on *but*, which ends in a ‘continuing’ intonation. Rather, the discontinuity is created by modifying *and*. This can be seen in Figure 4, which shows the associated pitch and amplitude contours. Note the much higher pitch at the onset of *and*. Also, note that *and* exceeds *but* in amplitude.

(5) Caller B:  you can really get,
    ...a lot of stuff for cheap.
Caller A:  Right.
    ..But,
            *and* we used to eat out,
    ...actually m- --
            much more often,
(Fisher corpus file 10517, 532.9–540.3 sec)
Figure 4. Pitch contour (middle) and amplitude contour (top) associated with the non-integrated DM sequence but and in Example (5)

As Levelt & Cutler (1983) point out, the effect is not strictly that of the high pitch on the repaired structure, but rather that of the increased prosodic contrast with the preceding word. Our data show that the default direction is in the direction of a pitch increase, but we also found a few cases in which the repaired structure had a noticeably lower pitch.

In summary, we classified all DM sequences in our data into three categories depending on their degrees of prosodic integration: (i) strongly integrated, (ii) weakly integrated, or (iii) non-integrated. The likelihood of a sequence being a genuine, planned sequence decreases from (i) to (iii), while the likelihood of it being an artifact increases. Because of the ambiguous nature of the weakly integrated category, which may or may not indicate self-repair, we performed two parallel analyses in which this category was treated differently. Both analyses excluded cases of non-integration. In the first analysis, we excluded weakly integrated cases as well, at the risk of missing sequences that are actually genuine. In the second analysis, we included the weakly integrated cases at the risk of inviting a certain amount of noise in the form of spurious sequences.

4.1.3 Utterance-initial position
Our third major criterion to identify genuine DM sequences was designed to ensure that both DMs are in utterance-initial position, in keeping with Schiffrin’s (2001: 57) definition and Auer’s (1996) grammaticalization model. We first discarded all cases in which the speaker did not produce an utterance after the second DM. This still left instances in which the discourse context makes it clear that the DMs were not designed to introduce the following utterance. Such turn-final DM sequences are often prosodically identifiable, just as single turn-final DMs are, because the second of the two DMs typically has either a falling intonation or some other form of prosodic non-integration with the following utterance. Therefore, we applied the same criteria as for the between-DM boundary to the transition
from the second DM to the following utterance and excluded all cases of prosodic non-integration, as defined in Section 4.1.2.

4.1.4 Sequences of more than two DMs
Sequences of three or more DMs can be analyzed as several two-DM sequences that overlap. For example, one might count so and then as one instance of so and and one instance of and then. However, this analysis is sometimes problematic. So and then is a case in point. The substring and then is a highly frequent sequence in our data. This suggests that it has attained a certain degree of formal unity. On the other hand, the frequency and, by implication, degree of unity of the substring so and is low. The internal syntactic structure of so and then could therefore be represented as [so [and then]]. In this structure so and is not a constituent element. This makes it problematic to treat so and here the same as so and occurring by itself as a two-DM sequence. Doing so runs the risk of artificially inflating the frequency of so and.

Our solution to this problem was to exclude all cases in which there was quantitative evidence that two markers contained in a longer sequence formed a conventionalized sequence in the sense just described. In a first step, we coded separately all cases in which one or more additional DMs precede or follow a two-DM sequence. In doing this, we considered as DMs not only the items in our set of eleven DMs, but also any other structure that might conceivably qualify as a DM, e.g. I guess, anyway, gosh (as part of oh gosh) and others. Having identified all sequences of three or more DMs (about 1,200 instances), we excluded those that include a DM pair occurring together more than five times in the subset. For example, all instances of and but then were excluded because but then constitutes a conventionalized sequence according to this heuristic.

4.2 Calculation of ordering constraints
Using the criteria discussed above we obtained frequency values for all 110 theoretically possible pairwise combinations of our eleven DMs. Given that we performed two alternative prosodic analyses, the result were two sets of 110 frequency values. We performed several statistical tests over each set to evaluate our hypotheses about DM sequencing.

To obtain a general sense of how common it is for DM pairs to show a quantitative bias toward one order, we first analyzed each pairing of two DMs individually. As an example, consider the hypothetical case of two DMs, A and B. If their ordering is biased, we might find only one order, say, only AB but not BA. Alternatively, we might find a preference for one order, say, 800 cases of AB but only 200 cases of BA. In either case, we performed a binomial test to test any existing asymmetries for statistical significance. The larger the share of significant effects is overall, the
more support there is for $H_1$, the hypothesis that DM sequencing is measurably constrained.

Next, we assumed a more global perspective from which the biases observed for individual pairings are viewed together as a whole. The guiding idea behind this analysis was that of a sequencing hierarchy, as discussed by Oates (2000). To demonstrate the concept, consider the case of three DMs, A, B, and C. Let us say the data show that A generally precedes B and C, and that B generally precedes C. We can summarize this pattern in terms of a hierarchy $A < B < C$. Such a hierarchy makes a prediction for all two-part sequences of A, B, and C. A given DM is predicted to precede all elements to its right on the hierarchy and to follow all elements to its left.

In a first step we determined which hierarchy that can be constructed from our eleven DMs is the optimal one. We did this by comparing the amount of ordering variability accounted for by different hierarchies. Consider again our hypothetical example of three DMs A, B, and C. Let us now say that A categorically precedes B and C, but that B precedes C in only 800 of 1,000 cases. We can express this variability in terms of ordering ratios. The ordering ratios of AB and BA are 1 and 0, respectively, as only AB is attested but there are no cases of BA. The same holds for AC relative to CA. For B and C, however, the ordering ratio is 0.8 for BC and 0.2 for CB. The strength of a given hierarchy can then be determined by adding up the ordering ratio values it is able to account for. For instance, $A < B < C$ explains a greater amount of variability than $A < C < B$. The former explains 2.8 out of 3 combined ratio values ($1 + 1 + 0.8$), i.e. 93% of all existing variability. The latter accounts for only 2.2 out of 3 ($1 + 1 + 0.2$), i.e. 73%.

We automatized the process of picking the optimal hierarchy through a script we wrote in the programming language R (R Core Team 2014) as there are about 40 million different ways to linearly arrange eleven elements. Our script successively generates every possible permutation of the eleven DMs using the function “permn()” of the package “combinat”. For each permutation, the total amount of ordering variability accounted for is calculated as explained above. The script then selects as the optimal hierarchy the one that explains the most variability. Note that because we have two sets of frequency values, corresponding to our two alternative prosodic analyses, we calculated two hierarchies.

The last step was to test statistically whether the optimal hierarchy for a given dataset significantly predicts the data. We performed a linear regression analysis over each dataset in which the 110 ordering ratio values were the dependent variable. The independent variable was binary. It was the prediction made for each sequence by the optimal hierarchy, i.e. whether it should or should not occur. The regression analysis thus tests whether the systematicity that is assumed by the hierarchy fits the data to a degree that has explanatory value.
In this way, we evaluated H₁, the hypothesis that DM ordering is measurably constrained. The same test served to evaluate the secondary hypotheses H₁a and H₁b. For H₁a, we performed a regression analysis over the subset of sequences in which an (a)-type DM is combined with a (b)-type DM. The dependent variable was again the attested ordering ratio, and the independent variable was the prediction made by H₁a, i.e. that only orders in which an (a)-type DM precedes (b)-type DMs should occur. To test H₁b, we included the subset of sequences in which both DMs are of the (b)-type. Again, the dependent variable were the attested ordering ratios. The independent variable was the prediction made by the two source elements’ canonical syntactic behavior. We decided for every sequence of two (b)-type DMs whether the corresponding sequence of source elements is well-formed according to sentential syntax. For example, we judged the sequence and you know to be well-formed because it can introduce a clause, while you know and is not well-formed as it cannot occur within one clause, but only when there is a clause-boundary in between you know and and.

5. Results

Table 1 shows the token frequencies of all 110 theoretically possible two-part sequences of our eleven DMs (recall that about three quarters of the values are estimated; see beginning of Section 4). In the table, rows represent the initial sequence position and columns represent the second position. The table combines the results of both prosodic analyses. Each cell shows the frequency of prosodically strongly integrated cases first, followed in parentheses by the combined frequency of strongly and weakly integrated cases (see Section 4.1.2 on these categories). Table 2 gives the corresponding ordering ratios, as explained in Section 4.2.

The relatively small number of zero values in Table 1 shows that only very few sequences do not occur at all. Unattested sequences are more common in the case of strong prosodic integration, where they make up 16.4% (18 out of 110). For the combined cases of strong and weak prosodic integration, on the other hand, unattested sequences make up only 1.8% (2 out of 110). This suggests that the conditions that allow two DMs to occur within the same intonation unit are somewhat more restrictive. However, the fact that the vast majority of sequences are attested in both strongly and weakly integrated form indicates that complete prosodic integration, though less common, is not unusual.

Instead of categorical effects, we see evidence for a large number of probabilistic effects. When comparing the cell values associated with opposite orders of two DMs, it becomes clear that often one order is clearly preferred over the other. A case in point is the asymmetry of but I mean and I mean but. The former
Table 1. Token frequencies of the 110 theoretically possible DM sequences.

<table>
<thead>
<tr>
<th></th>
<th>and</th>
<th>because</th>
<th>but</th>
<th>I mean</th>
<th>now</th>
<th>oh</th>
<th>or</th>
<th>so</th>
<th>then</th>
<th>well</th>
<th>you know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(82)</td>
<td>(114)</td>
<td>(1,405)</td>
<td>(226)</td>
<td>(119)</td>
<td>(37)</td>
<td>(7,476)</td>
<td>(20,994)</td>
<td>(136)</td>
<td>(4,584)</td>
<td>(59,598)</td>
<td>(29,799)</td>
</tr>
<tr>
<td>and</td>
<td>52</td>
<td>41</td>
<td>758</td>
<td>170</td>
<td>71</td>
<td>12</td>
<td>5,482</td>
<td>20,994</td>
<td>62</td>
<td>2,157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>because</td>
<td>2</td>
<td>0</td>
<td>131</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>147</td>
<td>6</td>
<td>194</td>
<td>490</td>
<td></td>
</tr>
<tr>
<td>but</td>
<td>15</td>
<td>22</td>
<td>1,185</td>
<td>139</td>
<td>17</td>
<td>0</td>
<td>24</td>
<td>3,837</td>
<td>3</td>
<td>909</td>
<td>6,151</td>
<td></td>
</tr>
<tr>
<td>I mean</td>
<td>309</td>
<td>62</td>
<td>58</td>
<td>11</td>
<td>1</td>
<td>17</td>
<td>405</td>
<td>927</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>now</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>oh</td>
<td>224</td>
<td>15</td>
<td>51</td>
<td>39</td>
<td>0</td>
<td>310</td>
<td>1,151</td>
<td>139</td>
<td>1,979</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>or</td>
<td>4</td>
<td>0</td>
<td>32</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>6</td>
<td>267</td>
<td>324</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>206</td>
<td>2</td>
<td>13</td>
<td>542</td>
<td>90</td>
<td>0</td>
<td>500</td>
<td>427</td>
<td>1,798</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>then</td>
<td>7</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>455</td>
<td>491</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>well</td>
<td>380</td>
<td>42</td>
<td>20</td>
<td>682</td>
<td>140</td>
<td>0</td>
<td>445</td>
<td>45</td>
<td>1,133</td>
<td>2,894</td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know</td>
<td>4,481</td>
<td>502</td>
<td>58</td>
<td>1,453</td>
<td>49</td>
<td>48</td>
<td>194</td>
<td>460</td>
<td>171</td>
<td>24</td>
<td>7,440</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5,628</td>
<td>697</td>
<td>242</td>
<td>4,849</td>
<td>642</td>
<td>145</td>
<td>215</td>
<td>6341</td>
<td>1,323</td>
<td>6,093</td>
<td>(10,399)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Ordering ratios calculated from the token frequencies of the 110 theoretically possible DM sequences.

<table>
<thead>
<tr>
<th></th>
<th>and</th>
<th>because</th>
<th>but</th>
<th>I mean</th>
<th>now</th>
<th>oh</th>
<th>or</th>
<th>so</th>
<th>then</th>
<th>well</th>
<th>you know</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>and</td>
<td>0.963 (0.863)</td>
<td>0.734 (0.67)</td>
<td>0.711 (0.911)</td>
<td>0.239 (0.286)</td>
<td>0.756 (0.579)</td>
<td>0.964 (0.927)</td>
<td>1 (0.999)</td>
<td>0.139 (0.152)</td>
<td>0.325 (0.446)</td>
<td>0.683 (0.659)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>because</td>
<td>0.037 (1.37)</td>
<td>0 (0.595)</td>
<td>0.678 (0.435)</td>
<td>0.286 (0.25)</td>
<td>0.333 (0.333)</td>
<td>1 (0.182)</td>
<td>0.125 (0.162)</td>
<td>0.279 (0.426)</td>
<td>0.474 (0.37)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>but</td>
<td>0.734 (0.911)</td>
<td>0 (0.938)</td>
<td>0.953 (0.217)</td>
<td>0.246 (0.357)</td>
<td>0 (0.284)</td>
<td>0.649 (0.997)</td>
<td>1 (0.254)</td>
<td>0.138 (0.162)</td>
<td>0.94 (0.62)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I mean</td>
<td>0.29 (0.248)</td>
<td>0.32 (0.276)</td>
<td>0.047 (0.088)</td>
<td>0.397 (0.056)</td>
<td>0.032 (0.036)</td>
<td>0 (0.074)</td>
<td>0.472 (0.074)</td>
<td>0.208 (0.0197)</td>
<td>0.187 (0.226)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>now</td>
<td>0 (0.09)</td>
<td>0 (0.578)</td>
<td>0 (0.064)</td>
<td>0.609 (0.663)</td>
<td>0 (0.034)</td>
<td>0 (0.082)</td>
<td>0.667 (0.015)</td>
<td>0.14 (0.142)</td>
<td>0.14 (0.142)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>oh</td>
<td>0.759 (0.714)</td>
<td>0.714 (0.784)</td>
<td>0.749 (0.944)</td>
<td>0.968 (0.966)</td>
<td>1 (0.75)</td>
<td>0 (0.985)</td>
<td>0.1 (0.85)</td>
<td>0.744 (0.781)</td>
<td>0.744 (0.852)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>or</td>
<td>0.228 (0.418)</td>
<td>0 (0.667)</td>
<td>0 (0.643)</td>
<td>0.964 (0.964)</td>
<td>1 (0.25)</td>
<td>0 (0.941)</td>
<td>0.1 (0.435)</td>
<td>0.744 (0.831)</td>
<td>0.744 (0.625)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>0.036 (0.073)</td>
<td>0.637 (0.815)</td>
<td>0.359 (0.716)</td>
<td>0.97 (0.925)</td>
<td>1 (0.921)</td>
<td>0 (0.105)</td>
<td>0 (0.991)</td>
<td>0.96 (0.351)</td>
<td>0.481 (0.476)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>then</td>
<td>0 (0.001)</td>
<td>0 (0.007)</td>
<td>0 (0.003)</td>
<td>0.53 (0.539)</td>
<td>0.363 (0.695)</td>
<td>0.91 (0.015)</td>
<td>0.078 (0.009)</td>
<td>0.96 (0.014)</td>
<td>0.727 (0.225)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>well</td>
<td>0.86 (0.848)</td>
<td>0.876 (0.838)</td>
<td>0.867 (0.744)</td>
<td>0.936 (0.926)</td>
<td>0.986 (0.985)</td>
<td>0 (0.006)</td>
<td>0.538 (0.465)</td>
<td>0.716 (0.648)</td>
<td>0.99 (0.014)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know</td>
<td>0.675 (0.554)</td>
<td>0.721 (0.574)</td>
<td>0.06 (0.145)</td>
<td>0.782 (0.681)</td>
<td>0.875 (0.557)</td>
<td>0.258 (0.169)</td>
<td>0.519 (0.057)</td>
<td>0.273 (0.231)</td>
<td>0.021 (0.059)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.316 (0.341)</td>
<td>0.523 (0.631)</td>
<td>0.382 (0.392)</td>
<td>0.814 (0.803)</td>
<td>0.862 (0.775)</td>
<td>0.115 (0.148)</td>
<td>0.279 (0.375)</td>
<td>0.421 (0.773)</td>
<td>0.824 (0.261)</td>
<td>0.54 (0.631)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
is estimated to occur 1,185 times as a strongly integrated sequence, but the latter only 58 times. This translates into ordering ratios of 0.953 and 0.047, respectively. In the second prosodic analysis the imbalance is also strong, with 1,794 cases of *but I mean* compared to 174 cases of *I mean but*, i.e. ordering ratios of 0.912 and 0.088, respectively.

5.1 Individual DM pairings

As explained earlier, we carried out a series of binomial tests over the token frequency values of all opposite orders. For prosodically strongly integrated sequences, we find that 84 of the 106 attested combinations (79.2%) show a statistically significant asymmetry (*or* and *oh* as well as *or* and *so* are not attested as strongly integrated sequences in either order). In the set of combined weakly and strongly integrated sequences, 86 out of the 110 combinations (78.2%) show a significant asymmetry. Thus, for almost four in five DM pairings, one order is significantly preferred over the other.

5.2 Sequencing hierarchies

Another look at Table 1 and Table 2 shows that some DMs exhibit consistent ordering preferences relative to the others. For instance, *oh* is generally preferred in first position, as reflected in consistent ordering ratio values larger than 0.5 in the *oh* row (with the exception of fully integrated *oh or*, which is unattested). *I mean* shows the opposite tendency, as reflected in values below 0.5 throughout its row. Thus, there is evidence of more general sequencing preferences. As explained in Section 4.2, to capture the aggregate effect of all individual asymmetries, we calculated a sequencing hierarchy for each dataset. The optimal hierarchies are shown below:

i. Optimal sequencing hierarchy for strongly integrated DM sequences only:

\[oh > well > and > or > but > you know > so > because > now > then > I mean\]

ii. Optimal sequencing hierarchy for strongly or weakly integrated DM sequences:

\[oh > well > and > so > or > but > because > then > you know > now > I mean\]

The hierarchies predict the order of every possible DM combination. A given DM should precede the DMs to its right and follow those to its left. We discuss the position of specific DMs on the hierarchies and the differences between (i) and (ii) below. The total share of the cumulative ordering ratio values accounted for by the hierarchy in (i) is 81.8%. For the hierarchy in (ii) it is 80.1%. These relatively high values support the idea that the individual ordering preferences add up to a more general pattern.
5.3 Testing $H_1$, $H_{1a}$ and $H_{1b}$

To test hypothesis $H_1$, according to which DM sequencing is measurably constrained in general, we performed a regression analysis that evaluates the fit of the hierarchies in (i) and (ii) to the datasets on which they are based, as explained in Section 4.2. Both hierarchies significantly predict the respective orderings ($p < 0.001$). The model fit is reasonably good, with adjusted R-squared values of 0.68 and 0.76, respectively.

We proceeded to test hypotheses $H_{1a}$ and $H_{1b}$, also using regression as explained in Section 4.2. Note that the hierarchies both fully support $H_{1a}$. The (b)-type DMs in our set (oh and well) indeed tend to precede the others, which are of the (a)-type. Our regression analysis shows that this effect is statistically significant ($p < 0.001$). The adjusted R-squared values of 0.88 and 0.85 are higher than for the global analysis. This shows that the ordering restrictions between the (a)- and (b)-type DMs are more rigid than DM ordering is overall.

The two sequencing hierarchies also generally support $H_{1b}$, although not as clearly. In this case, they do not fully agree with each other, and some DMs do not behave as predicted. To begin, there is a general pattern, variably reflected in (i) and (ii), such that the grammatical categories from which the (b)-type DMs derive are indeed echoed in their ordering preferences, as predicted by $H_{1b}$. Speaking in terms of grammatical categories, coordinators (and, but, or) tend to precede subordinators (so, because), which precede adverbs (now, then), which precede matrix clauses (I mean, you know). Our regression analysis shows that, indeed, these grammatical constraints significantly predict the (b)-type DMs’ ordering ($p < 0.001$). The adjusted R-squared values of 0.31 and 0.35 are lower than in the previous regression analyses, suggesting less rigid ordering here.

Reasons for the lower predictive success of $H_{1b}$ can be seen in the unexpected position of some (b)-type DMs. Two in particular, you know and so, stand out through their non-canonical behavior. On both hierarchies you know appears further to the left than expected. Unlike I mean, it is not strongly associated with the second sequence position. The other salient exception is so, which unexpectedly precedes both but and or on the second hierarchy, although it appears in its canonical position on the first hierarchy. We discuss the theoretical relevance of these effects in the next section.\(^4\)

\(^4\) We also examined the degree to which each DM’s position shifts across the 1,000 “best” hierarchies (see Section 4.2). The rank order of most DMs varies only by about one slot to the left or right on average. Only or shows high variability in the first hierarchy. This is likely due to its low token frequency and the relatively large number of zero observations for strongly integrated sequences with or, which lead to extreme ordering biases. You know is the second most variable DM in the first hierarchy and the most variable DM in the second one. The loose connection of
6. Discussion

We can now revisit our hypotheses. To begin, the results demonstrate that DM sequencing is clearly not random, thus falsifying H₀. DMs are not characterized by massive or total syntactic de-categorialization. Given this finding, we pursued H₁ and examined the nature of the syntactic effects. Our results show that DM sequencing can indeed be explained to a large extent with reference to a DM’s grammaticalization history. First, as predicted by H₁ᵃ, DMs that evolve on the (a)-path (“from the left”) continue to precede those that develop on the (b)-path (“from the right”) (cf. Figure 1). In our data, this is seen in the fact that oh and well remain associated with the initial sequence position, in fact very strongly so. Second, as predicted by H₁ᵇ, DMs that develop on the (b)-path remain syntactically constrained in the way that the syntax of their source structures predicts. This finding stands in spite of the fact that for (b)-type DMs the degree of syntactic persistence is lower overall.

Coming back to Schiffrin’s (1987: 39) suggestion that non-canonical ordering is an expected feature of DMs, discussed at the beginning of this paper, we can now say that this is not strictly the case. Sequences like so and or and but (cf. Examples (1) and (2)) are not typical, at least not in the sense that they are more likely or as likely to be observed than canonical combinations. We found no evidence showing that non-canonical sequencing behavior is a necessary and predictable feature of DMs. Nevertheless, our data confirm that DM status and syntactic de-categorialization are correlated in general. We can see this in the fact that for a few of the DMs we investigated, specifically you know and so, non-canonical ordering is quite common. Such DMs are of particular interest from a grammaticalization perspective and therefore merit further discussion.

We interpret the high degree of syntactic de-categorialization in the case of you know and so as a symptom of a particularly advanced stage of grammaticalization. If this is the case, one would expect such markers to also display other signs of highly grammaticalized DMs. For you know, we see evidence of this in Table 3, which shows the ten most frequent non-canonical DM sequences, based on their token frequencies.

The frequencies in Table 3 show that sequences containing initial you know constitute the bulk of the unpredicted cases. This is consistent with our earlier finding that you know stands out in general. It displays exceptional behavior on all
of our measures. We can now relate these findings to you know’s very high token frequency. In fact, looking at all DM sequences in our database, 38.1% include you know. A high discourse frequency is a hallmark of highly grammaticalized forms. Judging by you know, then, it appears that advanced syntactic de-categorialization may not be an expected feature of DMs in general, but one of DMs that have an unusually high discourse frequency.

Another predictable feature of highly grammaticalized DMs is the expansion to more abstract discourse functions. Such expansion is seen in the case of so, the other marker we found to exhibit regular non-canonical sequencing behavior. A preliminary discourse analysis of so in our data reveals an interesting correlation between so’s sequence position and the degree of abstractness of its function, which we briefly illustrate. Consider first the use of so in Example (6), where it appears in the canonical sequence and so.

(6) and everything they do on Law and Order now they say, you know is … ripped from the headlines. …And so they’re not really reality TV, …but they’re …taking stories, …that are, you know, …that were real stories, and writing it into their story lines. (Fisher corpus file 4895, 415–432.4 sec)

### Table 3. The ten most frequent non-canonical DM sequences

<table>
<thead>
<tr>
<th>DM sequence</th>
<th>Token frequencies strongly integrated sequences only</th>
<th>strongly or weakly integrated sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>you know and</td>
<td>4,481</td>
<td>5,683</td>
</tr>
<tr>
<td>you know I mean</td>
<td>1,453</td>
<td>2,230</td>
</tr>
<tr>
<td>you know because</td>
<td>502</td>
<td>916</td>
</tr>
<tr>
<td>you know so</td>
<td>460</td>
<td>1,940</td>
</tr>
<tr>
<td>I mean and</td>
<td>309</td>
<td>464</td>
</tr>
<tr>
<td>so and</td>
<td>206</td>
<td>585</td>
</tr>
<tr>
<td>you know or</td>
<td>194</td>
<td>317</td>
</tr>
<tr>
<td>you know then</td>
<td>171</td>
<td>322</td>
</tr>
<tr>
<td>I mean but</td>
<td>58</td>
<td>174</td>
</tr>
<tr>
<td>and but</td>
<td>41</td>
<td>114</td>
</tr>
</tbody>
</table>
The topic of the conversation is reality TV. The speaker explains why the popular TV crime program *Law and Order* does not fall into this category. His *so* in the sequence *and so* marks the utterance *they’re not really reality TV* as a conclusion based on his immediately preceding description. *So* used in this function can be paraphrased as “therefore” or “hence”. Compare this to *so* in Example (7). This example illustrates the non-canonical sequence *so and* cited by Schiffrin (1987: 39), which is among the ten most frequent unpredicted sequences in our data (see Table 3).

(7) Caller B: now do you have children,  
[Jennifer],  
Caller A: [I do].  
Yes.  
Caller B: Yeah?  
How many children?  
Caller A: ..Three.  
Caller B: Okay.  
[How]--  
Caller A: [So and] they all get home from school after three o’clock too.  
(Fisher corpus file 1272, 190.1–197.1 sec)

To understand caller A’s use of *so*, one has to know that earlier in the conversation the other caller had explained that her own children come home at 3:30pm. The utterance prefaced by *so and* connects back to this point, re-initiating the previous topic “how one’s children determine one’s daily schedule”. Note the thematic discontinuity between this utterance and the other caller’s preceding line of questions. When used in this topic management function, *so* applies at a more global level of discourse than it does in (6), where it links up directly with the preceding discourse. This makes *so* in (7) more abstract, which is also evident from the fact that it cannot be paraphrased as “therefore” here.

The two discourse functions of *so* seen in (6) and (7) have both been described in the pragmatics literature. For example, in Müller’s (2005: 68) taxonomy, the former is called ‘marking of result or consequence’ and the latter a ‘main idea unit marker.’ What is interesting and novel about *so* here is the correlation of these two functions with the sequence position of *so*. We regularly find the more abstract function of *so* when it appears in initial position but rarely when it appears in second position. This suggests that for markers like *so*, which “should” appear in second position when paired with a marker like *and*, occurrence in the initial position is more than a symptom of syntactic de-categorialization. It also indicates functional differentiation such that the initial position is associated with the marker’s more abstract, extended discourse functions. In this way,
7. Conclusion

The aim of this paper was to shed light on the grammatical status of DMs by means of an analysis of their ordering preferences in two-part sequences with other DMs. We found that DM sequencing is remarkably constrained overall, and that the sequencing constraints strongly echo the positional restrictions of the DMs’ historical sources. Even DMs that have developed fully conventionalized discourse functions for the most part exhibit a rather shallow degree of syntactic de-categorialization. The strong persistence that seems to characterize DMs overall shows that there is no contradiction between serving as a DM and maintaining clear syntactic preferences.

This conclusion is not meant to deny the existence of a more general correlation between functional change and syntactic de-categorialization in DM grammaticalization. Our findings show that clear de-categorialization effects do occur, but they appear more likely at highly advanced stages of grammaticalization. A close look at these cases shows that syntactic ‘violations’ correlate with innovative DM use. In the case of two-part DM sequences, alternative orders have functional correlates. The unpredicted occurrence of markers in the initial sequence position appears to be associated with particularly abstract discourse functions. We are intrigued by this correlation and by the possibility of using sequencing behavior as an instrument in the analysis of the functional spectrum covered by particular DMs. It also appears fruitful to pursue the idea that sequencing slots are meaningful themselves. That is, for a marker to be placed in a particular position may imply a particular level of abstractness. We are pursuing this functionally oriented perspective in our continuing research on discourse marker sequencing.

References

Andersen, G. (2001). *Pragmatic Markers and Sociolinguistic Variation*. Amsterdam, Netherlands: Benjamins. DOI: 10.1075/pbns.84
A quantitative approach to the grammaticalization of discourse markers

Auer, P. (1996). The pre-front field in spoken German and its relevance as a grammaticalization position. Pragmatics, 6(3), 295–322. DOI: 10.1075/prag.6.3.03aue


Dehé, N., & Wichmann, A. (2010). Sentence-initial I think (that) and I believe (that): Prosodic evidence for use as main clause, comment clause and discourse marker. Studies in Language, 34(1), 36–74. DOI: 10.1075/sl.34.1.02deh


Authors’ addresses

Christian Koops
Department of Linguistics
University of New Mexico
1 University of New Mexico
Albuquerque, NM 87131
USA
ckoops@unm.edu

Arne Lohmann
Department of English
University of Vienna
Spitalgasse 2-4
1090 Wien
Austria
arne.lohmann@univie.ac.at