Berkeley Linguistics Society. 2013. 108-122. http://dx.doi.org/10.3765/bls.v39i1.3873
Published by the Linguistic Society of America

Discourse Marker Sequencing and Grammaticalization*

CHRISTIAN KOOPS AND ARNE LOHMANN
University of New Mexico and University of Vienna

1 Introduction

This paper deals with the grammatical properties of discourse markers (DMs), specifically their ordering preferences relative to one another. While the data presented here are synchronic, we approach the topic of DM sequencing from the perspective of grammaticalization. From this perspective, DMs can be understood as the result of a process in which elements serving other functions, for example 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 defined as "sequentially dependent elements which bracket units of talk." Here we are concerned with the final outcome of this process. We ask: to what degree do fully formed DMs retain or lose the grammatical properties associated with their previous role, specifically their syntactic co-occurrence constraints? In other words, what degree of syntactic decategorialization (in the sense of Hopper 1991) do DMs display?

This raises the question of how DMs grammaticalize. As they constitute a broad and diverse class of elements with different developmental trajectories, we draw on Auer's (1996) taxonomy of relevant grammaticalization processes, which covers a wide range of diverse types. Auer's analysis deals specifically with grammaticalization in the syntactic position known as the "pre-front field" (*Vor-Vorfeld*) of spoken German. In drawing on his model, we assume that this

* We thank the audiences at BLS 39 in Berkeley and HLDS 10 in Albuquerque for their comments and suggestions on our analysis. All remaining errors and inaccuracies are our own.

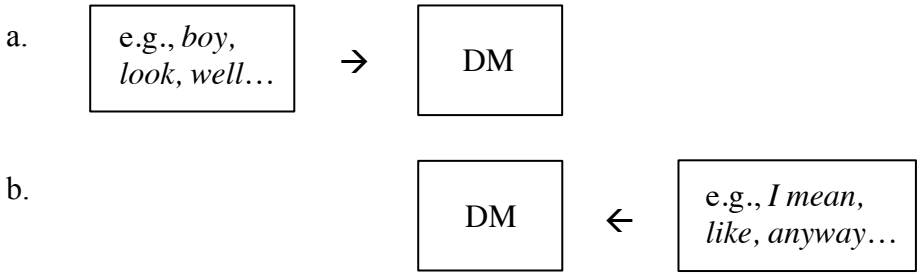
position is broadly comparable to the extra-sentential, utterance-initial position in which many English discourse markers are found (e.g. Schiffrin 2001).

Auer identifies two grammaticalization paths, or clines, along which elements evolve to occupy this position. On the first cline, which we will refer to as the (a)-path, "a dialogical, sequential structure is condensed and 'compacted' into a grammatical one." (313) Elements on this cline include "vocatives and other … constituents which may be used as summons in conversation." (ibd.) There are obvious English equivalents to the types of structures identified by Auer, such as address terms (*boy, man*), imperatives (*listen, look*), interjections (*oh, wow*), as well as forms of assessment and agreeing responses (*well, sure, right*).[1]

On Auer's second cline, which we will call the (b)-path, "a constituent moves out of the grammatical centre of the sentence into its periphery." (ibd.) 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 discussion makes it clear that discourse markers that are identical in form to conjunctions (*and, because*) also fall on this cline.

Figure (1) summarizes these two grammaticalization paths, with DMs on the (a)-path coming to occupy the utterance-initial DM slot from the left, as it were, and DMs on the (b)-path moving into this position from the right.

(1) Two grammaticalization paths for DMs



The (b)-path of DM grammaticalization has been investigated in some detail, for example in classic case studies of English discourse markers such as *like* (Romaine and Lange 1991) and *I think* (Thompson and Mulac 1991), among others (see also Traugott 1997). In these studies, decategorialization phenomena are often cited as evidence for the fact that a particular structure has attained DM status. For example, Thompson & Mulac (1991) show that the disproportionately high rate of omission of the complementizer *that* following *I think* and similar "matrix clauses" shows that these structures are not subject to the rules of

[1] Auer (1996) does not consider the final stage of his first grammaticalization cline to be dicourse markers, but merely pre-front field constituents. Our definition of DMs, which follows Schiffrin (1987), is slightly broader and considers many elements at this stage as DMs.

sentential syntax in the same way as genuine, syntactically integrated matrix clauses. In this mode of analysis, then, DM status may be defined negatively, as the lack of some otherwise expected grammatical behavior. But does a DM's dissociation from its syntactic source structure render it devoid of grammar? In other words, is the placement of fully formed DMs wholly determined by discourse-functional constraints, with no persistence of their former grammatical behavior whatsoever? Or, do even fully formed DMs retain properties that are best explained with reference to their former role? These are the question we address in the following.

2 Discourse marker sequencing

It is well known that DMs are often used in direct sequence with other DMs, resulting in two-part sequences like *oh well*, *but then*, etc. It has also been pointed out that such sequences may hold interesting analytical insights. For example, in her discussion of *now*, Aijmer (2002) points to the sequences *so now* and *now therefore* to argue that *now*, unlike *well*, is "oriented toward the upcoming topic." (64) In fact, Aijmer proposes that DM sequences "are perhaps the most important formal indication of what function the discourse particle has." (189) Nevertheless, as noted by Fraser (2011), the phenomenon of DM sequencing has received surprisingly little attention in the literature on discourse markers. The quantitative analyses of DM sequencing we are aware of all come from the field of automatic text generation (Knott 1996, Oates 2000) and have been restricted to DMs in written discourse. The significance of sequencing constraints for theories of DM grammaticalization has not previously been explored.

In two-part DM sequences the question of a DM's relative position becomes relevant. What determines whether a given DM appears in first or second position? Is its placement at least partially determined by its source syntax? From the perspective of decategorialization, one would expect syntactic constraints to loosen, or even disappear, and ordering variability to increase. Indeed, Schiffrin (1987) argues that the use of DMs in syntactically non-canonical combinations, as in (2) and (3), is a formal indicator *that* they are DMs.

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

(3) **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 in the original)

In (2), the apparent co-occurrence violation consists in having a coordinate conjunction preceded by *so*, rather than the other way around. In (3), according to Schiffrin, the illicit co-occurrence of two coordinate conjunctions is made possible because *and* functions as a DM.

This paper can be understood as an empirical investigation of the status of such examples. How regularly do DMs combine in a non-normative order, such as *so and* as opposed to *and so*? In order to answer this question, we will also have to clarify what it means for two DMs to be used 'in sequence.' For instance, is it justified to argue that the DMs *and* and *but* in (3) were uttered as a planned sequence? How can we rule out the possibility that *but* simply replaces *and* in an act of self-repair? The utterance-initial position in which DMs occur is a likely site of repair, as interlocutors start without having fully planned their turn and produce false starts. An empirical analysis of DM sequencing in spoken discourse therefore faces the considerable challenge of distinguishing 'genuine' DM sequences from accidental ones.

3 Hypotheses and predictions

At the most general level, we test the null hypothesis H0 that DMs are in fact devoid of grammar and their sequencing is unconstrained. The prediction of H0 is that the ordering of two DMs should be free and the likelihood of observing one or the other order is indistinguishable from chance. H0 is opposed to H1, according to which DMs do have (some) grammar, which predicts that DM sequencing is not random, but measurably constrained.

To the extent that H1 is borne out, we can further ask whether a DM's sequencing constraints reflect its linguistic origin. One version of H1 posits that DM ordering shows reflexes of the grammaticalization paths shown in Figure 1. DMs that evolved on path (a) should precede DMs that evolved on path (b). We will call this hypothesis H1a. Secondly, coming back to Schiffrin's examples of non-canonical ordering, another version of H1, which is restricted to those DMs on the (b)-path, holds that DM sequencing shows reflexes of a DM's source syntax. We call this hypothesis H1b. The prediction following from H1b is that DMs tend to occur in sequences which don't violate the order predicted by their source syntax, so that, for example, the DM sequence *and so* should be attested more frequently than the DM sequence *so and*.

4 Methodology

In order to test these hypotheses, we used the set of eleven DMs investigated in Schiffin's (1987) foundational study of discourse markers. Drawing on Schiffrin's analysis has the advantage of providing us with a relatively large and diverse set of DMs whose status as DMs has been independently established on the basis of a

unified definition. The set is given in (4), subdivided according to each marker's historical route of development within the taxonomy of DM grammaticalization paths discussed above. For this analysis, we assume that *oh* and *well* evolved on the (a)-path, while the other nine DMs evolved on the (b)-path.

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

We quantified the ordering preferences of these eleven DMs relative to one another by examining the rate of occurrence of all 110 theoretically possible pairwise combinations of them in the Fisher corpus (Cieri et al. 2004, 2005), a telephone speech corpus of North American English. Our first step in the analysis was to obtain exhaustive concordances of each sequence on the basis of the corpus transcripts. This resulted in over 150,000 hits. In the next step, we examined each of the 110 concordances more closely to obtain an estimate of how many of the matches of a given orthographic sequence represent 'genuine' DM sequences. Our selection criteria are discussed below. Because the total number of hits was too large for us to manually edit all concordances, our procedure was to inspect in detail a random sub-sample of 100 hits in each concordance (or all hits, in cases of concordances with 100 or fewer hits) and then to extrapolate the ratio of spurious to 'genuine' sequences to the whole concordance.

4.1 Data selection criteria

Our method of determining whether the elements contained in a superficial match, for example a sequence of the words *so* and *and*, both function as DMs in the context in which they were uttered was closely based on Schiffrin's (1987, 2001) definition of discourse markers, specifically her criteria for distinguishing the DM use of particular structures from their use in other functions.

4.1.1 Lack of obligatoriness

Our first criterion was syntactic obligatoriness. Non-obligatoriness is a key operational criterion in distinguishing DMs from their formally identical non-DM counterparts (Schiffrin 1987:64, 2001:57). If omitting one or both elements in a given sequence resulted an incomplete syntactic structure, or where doing so significantly changed the semantics of the utterance, the item was not analyzed as a DM sequence. To illustrate, the phrases in (5a-c) and (6a-c) contain the superficial sequences *and so* and *so and,* but none of them qualify as DM sequences because in each case the word *so* is obligatorily present. It is part of a larger syntactic construction from which it cannot be omitted.

(5) a. *and so* did everyone else
b. *and so* on *and so* forth
c. *and so* many of them…

(6) a. I would say *so and*…
b. we gave it to *so and* so
c. once a year or *so and*...

We did not analyze *and, but,* and *or* as DMs when they were followed by a constituent smaller than a complete clause (Schiffrin 1987:128). The greatest analytical challenges were posed by the DMs *now*, *then*, and *because.* For the former two, we were able to rely on Schiffrin's semantic and formal criteria (Schiffrin 1987:230-232, 246-248), for example by generally excluding cases in which *now* and *then* function semantically as temporal modifiers of an event. We had to slightly supplement Schiffrin's criteria for *because* (Schiffrin 1987:191-217). We did not analyze *because* as a DM when the *because*-clause preceded the "main clause", i.e. the clause or clauses containing the proposition(s) that the speaker is giving a reason for. We also only included cases in which the *because*-clause had the form of a fragment, i.e. separated from the "main clause" by another syntactic construction or by some discontinuity, or without a clear antecedent in the prior discourse.

4.1.2 Prosodic integration

Having reduced the data to cases in which both sequence elements qualify as DMs, we coded the remaining data for whether the DMs constitute 'genuine' or accidental sequences (see above). For this decision, we used the parameter of prosodic integration. The more integrated two DMs are prosodically, the more certain we can be that they were planned to be uttered and understood together, and the less likely we are to deal with a case of self-repair. Nevertheless, given that DMs are frequently followed by a minor prosodic boundary, the lack of full prosodic integration does not in itself disqualify particular cases. This meant that we also had to distinguish between prosodic boundaries of different strength.

As the Fisher transcripts include no prosodic mark-up, our analysis involved listening to all random sub-samples of our 110 concordances. While time-consuming, the auditory analysis was also an opportunity to verify the accuracy of the transcripts and to discard cases in which the words in question were mistranscribed, untranscribed words intervened between the DMs, or one of the two DMs was not fully produced. This analysis was primarily auditory. In difficult cases, pitch tracks were also inspected.

Our prosodic analysis was based on the notion of an *intonation unit* (IU), aka. *intonational phrase* or *tone unit*. IUs are fundamental to Du Bois et al.'s (1993)

discourse transcription system, which served as our practical framework. We first determined whether both DMs fell within the same IU, i.e. within a "stretch of speech uttered under a single coherent intonation contour" (47). In the following, we will refer to cases that meet this criterion as *strongly integrated* sequences.

Where a prosodic boundary separated the DMs, we further distinguished two types. The first, which we will call *non-integrated* sequences, and which we discarded, includes a variety of prosodic phenomena which can all be interpreted as signals that the second IU was not intended to be understood as a continuation of the larger prosodic structure that includes the prior IU. Very clear instances of this, although not the majority of the cases, are those in which the first DM ends in either Du Bois et al.'s 'final' intonation or in their 'appeal' intonation, i.e. in a fall to a very low pitch or a very high rise (transcribed "." and "?", respectively). More often, non-integration was evident from the onset of the second DM, specifically where the onset of the second DM was much higher in pitch than the offset of the first DM (or, less frequently, where the onset had a much lower pitch), resulting in a salient prosodic discontinuity. Such sudden, dramatic pitch increases were typically accompanied by equally sudden increases in amplitude and tempo. Any one of these three parameters was considered sufficient to identify a sequence as non-integrated. We also considered as non-integrated cases in which the first DM "trails off", i.e. where it was produced with a drawn out, low-pitched quality that, though not sufficiently low to qualify as 'final', clearly indicates that the speaker is opening the floor. Such cases were almost always followed by pauses, sometimes extended ones. However, we did not consider a pause in itself as an indicator of non-integration.

Our last prosodic category, which we will call *weakly integrated* sequences, were cases in which the DMs are separated by a prosodic boundary, but one that doesn't meet the criteria for a non-integrated sequence, as defined above. In these cases, the end of the first IU and the beginning of the second IU were similar in pitch, amplitude, and tempo, resulting in a relatively soft prosodic boundary. [2]

4.1.3 Utterance-initial position

Our third criterion 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. Although intuitively obvious, the distinction between utterance-initial or utterance-final occurrence is often difficult to draw in practice. Our method was to first exclude all cases in which the second DM was not followed by any talk by the same speaker. In addition, we applied the same

[2] In terms of Du Bois et al.'s transcription system, our 'weakly integrated' sequences are all cases in which the first DM ends in 'continuing' intonation (transcribed ","). However, there is no one-to-one relationship between Du Bois et al.'s 'continuing' intonation and our 'weakly integrated' category, because our 'non-integrated' category also includes cases of 'continuing' intonation.

prosodic criteria that we used 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 above) of the second DM and the following utterance.

4.1.4 Sequences of more than two DMs

Our final selection criterion addresses sequences of more than two DMs. To see what the problem is with these, consider the 3-DM sequence *so and then*. In this sequence, the part *and then* is a highly conventionalized sequence or 'chunk.' It would therefore be problematic to treat the sequence *so and* as part of *so and then* the same as *so and* occurring by itself as a 2-DM sequence. Doing so runs the risk of artificially inflating the number of *so and* cases because in some instances *and* in *so and* may be licensed only by the larger structure *and then*.

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 such a 'chunk.' In a first step, we coded separately all cases in which one or more additional DMs precede or follow a 2-DM sequence. In doing this, we considered as DMs not only those 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 many more. Having identified all sequences of three or more DMs (about 1000 instances), we excluded those cases in which the sequence included a pair of DMs occurring together more than five times in this subset. For example, all instances of *and but then* were excluded because *but then* constitutes a chunk according to this heuristic, so that the sequence *and but* could be an artifact.

5 Results

The estimated frequencies of all 110 theoretically possible DM sequences in the corpus are given in Table 7. This table is the result of applying the various selection criteria discussed in Section 4 to our sub-samples of the raw, unedited concordances (see above) and extrapolating from the resulting number of 'genuine' cases to the corpus frequencies. Rows represent DMs in initial position, and columns represent DMs in second position. The first value in each cell is the estimated frequency of prosodically strongly integrated sequences. The second value, given in parentheses, is the estimated frequency of strongly and weakly integrated sequences added together.

As can be seen by inspecting the cells associated with opposite orders of the same DM pair, e.g. the frequencies of *oh well* and *well oh*, there are many cases in which two DMs are used much more frequently in one order than in the reverse order. *Oh well* is an extreme example, occurring 1,558 times as a strongly integrated sequence, compared to only a single case of a strongly integrated *well oh*. This asymmetry can be expressed in the form of a ratio of 0.9994 for *oh well*

(1,558/1,559) and a ratio of 0.0006 for *well oh* (1/1,559). In the following discussion, we refer to these as *ordering ratios*. The ordering ratios of all 110 combinations are shown in Table 8. As in Table 7, the two values per cell reflect the frequencies of the different prosodic types. Again, the first value refers to strongly integrated sequences, while the second value, given in parentheses, refers to both strongly and weakly integrated sequences combined.

To determine how many of the asymmetrical pairwise distributions seen in Table 7 deviate significantly from chance, we performed a series of binomial tests over the estimated token frequencies. For the strongly integrated sequences, we find that 82 of the 106 combinations attested in at least one order (77.4%) show a significant asymmetry (*or* and *oh* as well as *or* and *so* are not attested as strongly integrated sequences in either order). For the prosodically weakly integrated sequences, 86 out of 110 combinations (78.2%) show a significant asymmetry. Thus, for most DM pairings one order is significantly preferred over the other.

As can be seen in Table 8, some DMs exhibit consistent ordering preferences. For instance, *oh* occurs in first position with all other DMs, as reflected in the values above 0.5 in the row labeled "oh." The opposite is the case for *I mean*, which is preferred in second position with all other DMs, as reflected in consistent values below 0.5 in the column labeled "I mean." One way to aggregate and summarize these general preferences is in the form of a sequencing hierarchy that ranks all eleven DMs according to their preference for one or the other position. Such a hierarchy predicts one preferred order for each theoretically possible 2-DM sequence. Different hierarchies are possible and can be compared on the basis of their predictive power. For example, a hierarchy which ranks *oh* before *I mean* will make better predictions than one which ranks *I mean* before *oh*. The ideal hierarchy is the one that accounts for the greatest amount of attested orderings. This provides us with a measure of how well individual preferences are accounted for, as well as how strictly constrained DM ordering is in general.

We calculated two such hierarchies: one for the ordering of the strongly integrated sequences only, and one for the ordering of the strongly and weakly integrated sequences combined. The predictive accuracy of different hierarchies was determined on the basis of the cumulative explained ratios, rather than on the basis of the cumulative explained token numbers, to avoid skewing of the results due to some DMs being much more frequent than others. For the mathematical calculation we used a script written in the R programming language (R Development Core Team 2012). The script generates all ~40 million possible permutations of our 11 DMs and for each permutation calculates the total amount of explained ordering ratios.[3] The resulting ideal rank orders are given in (9) and (10). DMs further to the left are predicted to occur in initial position, while DMs further to the right are predicted to occur in second position.

[3] The permutations were generated using the `permn()` function of the `combinat` package.

(7) Estimated token frequencies

	and	*because*	*but*	*I mean*	*now*	*oh*	*or*	*so*	*then*	*well*	*you know*
and		53 (79)	70 (140)	782 (1,564)	206 (274)	82 (151)	15 (38)	6,418 (8,898)	21,990 (21,990)	60 (131)	3,596 (6,934)
because	2 (14)		0 (2)	133 (247)	2 (4)	8 (11)	1 (2)	1 (2)	166 (171)	6 (31)	246 (982)
but	17 (78)	25 (35)		1,806 (2,796)	118 (135)	24 (63)	0 (7)	33 (120)	4,163 (4,702)	5 (46)	1,049 (4,546)
I mean	410 (614)	61 (104)	92 (234)		14 (19)	1 (6)	0 (3)	5 (101)	24 (41)	66 (126)	537 (1,201)
now	0 (27)	0 (5)	0 (0)	18 (32)		0 (2)	0 (0)	0 (11)	2 (2)	3 (4)	9 (62)
oh	228 (294)	16 (29)	39 (120)	33 (74)	43 (69)		0 (3)	330 (860)	22 (37)	1,558 (2,147)	147 (401)
or	9 (29)	0 (4)	1 (12)	27 (84)	1 (2)	0 (1)		0 (0)	15 (19)	7 (27)	304 (1,822)
so	197 (666)	2 (7)	0 (370)	691 (1,788)	81 (99)	0 (13)	0 (9)		704 (794)	19 (80)	511 (2,360)
then	14 (29)	0 (2)	0 (18)	27 (49)	0 (6)	2 (7)	1 (1)	0 (6)		5 (10)	528 (1,201)
well	460 (884)	39 (132)	19 (133)	903 (1,478)	137 (195)	1 (15)	8 (24)	50 (138)	505 (665)		1,237 (2,959)
you know	5,217 (7,192)	561 (1,031)	64 (1,096)	1,700 (2,703)	48 (63)	50 (119)	247 (393)	520 (2,211)	156 (344)	14 (202)	

(8) Ordering ratios calculated from the estimated token frequencies

	and	*because*	*but*	*I mean*	*now*	*oh*	*or*	*so*	*then*	*well*	*you know*
and		0.964 (0.849)	0.805 (0.642)	0.656 (0.718)	1 (0.91)	0.265 (0.339)	0.625 (0.567)	0.97 (0.93)	0.999 (0.999)	0.115 (0.129)	0.408 (0.491)
because	0.036 (0.151)		0 (0.054)	0.686 (0.704)	1 (0.444)	0.333 (0.275)	1 (0.333)	0.333 (0.222)	1 (0.988)	0.133 (0.19)	0.305 (0.488)
but	0.195 (0.358)	1 (0.946)		0.952 (0.923)	1 (1)	0.381 (0.344)	0 (0.368)	1 (0.245)	1 (0.996)	0.208 (0.257)	0.942 (0.806)
I mean	0.344 (0.282)	0.314 (0.296)	0.048 (0.077)		0.438 (0.373)	0.029 (0.075)	0 (0.034)	0.007 (0.053)	0.471 (0.456)	0.068 (0.079)	0.24 (0.308)
now	0 (0.09)	0 (0.556)	0 (0)	0.563 (0.627)		0 (0.028)	0 (0)	0 (0.1)	1 (0.25)	0.021 (0.02)	0.158 (0.496)
oh	0.735 (0.661)	0.667 (0.725)	0.619 (0.656)	0.971 (0.925)	1 (0.972)		0 (0.75)	1 (0.985)	0.917 (0.841)	0.999 (0.993)	0.746 (0.771)
or	0.375 (0.433)	0 (0.667)	1 (0.632)	1 (0.966)	1 (1)	0 (0.25)		0 (0)	0.938 (0.95)	0.467 (0.529)	0.552 (0.823)
so	0.03 (0.07)	0.667 (0.778)	0 (0.755)	0.993 (0.947)	1 (0.9)	0 (0.015)	0 (1)		1 (0.993)	0.275 (0.367)	0.496 (0.516)
then	0.001 (0.001)	0 (0.012)	0 (0.004)	0.529 (0.544)	0 (0.75)	0.083 (0.159)	0.063 (0.05)	0 (0.008)		0.01 (0.015)	0.772 (0.777)
well	0.885 (0.871)	0.867 (0.81)	0.792 (0.743)	0.932 (0.921)	0.979 (0.98)	0.001 (0.007)	0.533 (0.471)	0.725 (0.633)	0.99 (0.985)		0.989 (0.936)
you know	0.592 (0.509)	0.695 (0.512)	0.058 (0.194)	0.76 (0.692)	0.842 (0.504)	0.254 (0.229)	0.448 (0.177)	0.504 (0.484)	0.228 (0.223)	0.011 (0.064)	

(9) Ideal rank order for strongly integrated DM sequences
oh > well > and > or > but > you know > so > because > now > then > I mean

(10) Ideal rank order for strongly and weakly integrated sequences combined
oh > well > and > so > or > but > because > then > now > you know > I mean

The percentage of explained ordering ratios is 82.3% for the hierarchy in (9) and 79.7% for the hierarchy in (10).

Another way to test the validity of the two hierarchies is through a linear regression analysis. For this analysis, we created a binary independent variable that indicates whether a certain sequence is predicted or not predicted. For example, *but so* is predicted by the first hierarchy but not by the second one. This variable was used to predict the ordering biases given in Table 8. The regression analyses yield highly significant results for both hierarchies ($p<0.001$). The model fit is reasonably good with R-squared values of 0.68 and 0.75, respectively.

The position of the DMs on the hierarchies in (9) and (10) is quite similar overall. In both cases, *oh* and *well* are most strongly associated with the initial position. In the first hierarchy, they are followed by the group of DMs identical in form to coordinating conjunctions (*and, but, or*), which are followed by DMs identical in form to subordinating conjunctions (*so, because*), which are themselves followed by DMs identical in form to adverbs (*now, then*). As for DMs that look like matrix clauses, while *I mean* is strongly associated with the final position, *you know* appears in the center of the hierarchy, i.e. showing no consistent ordering preference. The two DMs which show the greatest difference between the hierarchies are *so,* which precedes both *but* and *or* in the second hierarchy, and *you know*, which here patterns with *I mean* at the right end. A minor difference is that the position of *now* and *then* is reversed.

We also calculated a measure of how stable the rank order of individual DMs is on each hierarchy. For this measure we examined the 1000 'best' hierarchies, in terms of explanatory accuracy, and calculated the standard deviation of each DM's rank order across the 1000 hierarchies. The results are shown in Table 11.

Table 11 shows that the positional variability of most DMs is fairly low, as reflected in standard deviations of about 1. This suggests that individual ordering preferences are generally captured well by the two hierarchies. Still, two DMs stand out as harder to pin down. First, *or* shows extreme variability in the first hierarchy. This may be an artifact due to the very low token frequencies in strongly integrated sequences, leading to less reliable ordering information (see Table 7). More interesting is the case of *you know*, which is the second most variable DM in the first hierarchy and the most variable DM in the second one. The low predictability of *you know* within each hierarchy dovetails with its variability across the two hierarchies.

(11) Positional variability

	strongly integrated sequences		strongly and weakly integrated sequences	
	rank order	variability	rank order	variability
oh	1	0.59	1	0.43
well	2	0.83	2	1.07
and	3	1.16	3	1.13
or	4	2.45	5	1.19
but	5	1.06	6	0.98
you know	6	1.81	10	1.36
so	7	1.06	4	0.92
because	8	1.19	7	1.16
now	9	0.76	9	1.36
then	10	0.81	8	1.07
I mean	11	0.99	11	1.06

Finally, note that the rank orders, especially those in the first hierarchy, strongly suggest that the grammatical categories from which these DMs derive (excepting *oh* and *well*) influence the DMs' ordering preferences. Speaking in terms of traditional grammatical categories, coordinators precede subordinators, which precede adverbs, which precede matrix clauses. To quantify the extent to which canonical syntactic ordering constraints predict the attested DM orderings, we summed the ordering ratios explained by traditional syntactic constraints and compared them to those associated with orders that violate them. The percentage of ordering ratios explained by traditional syntactic constraints is 71.5% for the first hierarchy and 66.7% for the second hierarchy.

6 Discussion

The results show that DM sequencing is clearly not random. This allows us to reject H0 and further pursue H1. The ordering effects captured in our two hierarchies show that a DM's sequencing behavior does indeed reflect its grammaticalization history. First, as predicted by H1a, DMs that derive from independent sequential moves (the a-path of Figure 1) precede those that derive from sentence-level structures (the b-path in Figure 1). This can be seen in the fact that *oh* and *well* remain strongly associated with initial position. Second, as predicted by H1b, the sequencing of DMs that develop on the (b)-path remains to a large extent constrained by the syntax of their source structures.

Coming back to Schiffrin's (1987:39) argument that non-canonical ordering is an expected feature of DMs, we have found that on the whole such combinations are not typical, at least not in the sense that they are more likely to be observed

than canonical combinations. Nevertheless, Schiffrin's observation is supported in that for some DMs non-canonical ordering is very well attested. A case in point is the DM *so*, for which our second hierarchy actually predicts non-canonical sequencing relative to *but* and *or*. Another clear case is *you know*, for which non-canonical ordering is even predicted by our first hierarchy, i.e. for prosodically fully integrated sequences, which are arguably the more conventionalized ones. In fact, Schiffrin's examples (cf. [1] and [2]) turn out to be quite representative. Among the most frequent non-canonical combinations in our data are: *so* preceding coordinators (*so but, so and*), *you know* and *I mean* preceding coordinators or subordinator *because* (*you know and*, *I mean but, you know because*), as well as combined coordinators (*and but*, *but and, and or*).

Future research will address DM sequences like these, whose order regularly violates traditional syntactic constraints, now that their significance has been empirically established. It is an interesting question what motivates such cases. We suspect that as DMs grammaticalize, their 'pragmatic scope' expands, which allows them to precede a greater number of other DMs. Those DMs in our data for which this is best attested can be understood as having reached a relatively higher degree of syntactic decategorialization.

7 Conclusion

Although decategorialization is often taken as a defining criterion in identifying DMs, decategorialization in terms of sequencing constraints appears to be rather weak. Even in grammaticalized DMs, persistence of source constraints appears to be the norm. Thus, there is no contradiction between functioning as a DM and retaining clear ordering preferences.

References

Aijmer, Karin. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: Benjamins.

Auer, Peter. 1996. The Pre-front Field in Spoken German and its Relevance as a Grammaticalization Position. *Pragmatics* 6:295-322.

Cieri, Christopher, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004. *Fisher English Training Speech Part 1, Transcripts*. Philadelphia, PA: Linguistic Data Consortium.

Cieri, Christopher, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2005. *Fisher English Training Speech Part 2, Transcripts*. Philadelphia, PA: Linguistic Data Consortium.

Du Bois, John, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of Discourse Transcription. In J. Edwards and M. Lampert, eds., *Talking Data. Transcription and Coding in Discourse Research*, 45-89. Hillsdale, NJ: Erlbaum.

Fraser, Bruce. 2011. The Sequencing of Contrastive Discourse Markers in English. *Baltic Journal of English Language, Literature, and Culture* 1:29-35.

Hopper, Paul. 1991. On Some Principles of Grammaticalization. In E. Traugott and B. Heine, eds., *Approaches to Grammaticalization*, 17-35. Amsterdam: Benjamins.

Knott, Alistair. 1996. A Data-Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh.

Oates, Sarah Louise. 2000. Multiple Discourse Marker Occurrence: Creating Hierarchies for Natural Language. *Proceedings of the 3rd CLUK Colloquium, Brighton*, 41-45.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Romaine, Suzanne and Deborah Lange. 1991. The Use of *Like* as a Marker of Reported Speech and Thought: A Case of Grammaticalization in Progress. *American Speech* 66(3):227-279.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schiffrin, Deborah. 2001. Discourse Markers: Language, Meaning, and Context. In D. Schiffrin, D. Tannen and H. E. Hamilton, eds., *The Handbook of Discourse Analysis*, 54-75. Malden, MA: Blackwell.

Thompson, Sandra and Anthony Mulac. 1991. The Discourse Conditions for the Use of the Complementizer *that* in Conversational English. *Journal of Pragmatics* 15:237-251.

Traugott, Elizabeth. 1997. The Role of the Development of Discourse Markers in a Theory of Grammaticalization. Paper presented at the ICHL XII, Manchester, 1995.

Christian Koops
Department of Linguistics
University of New Mexico
MSC03 2130
1 University of New Mexico
Albuquerque, NM 87131

ckoops@unm.edu

Arne Lohmann
Department of English
Universität Wien
Spitalgasse 2-4
1090 Wien
Austria

arne.lohmann@univie.ac.at