J. Linguistics **37** (2001), 527–550. © 2001 Cambridge University Press
DOI: 10.1017/S0022226701001050 Printed in the United Kingdom

A mathematical model for the analysis of variation in discourse[1]

JESÚS ROMERO TRILLO

Universidad Autónoma de Madrid

(Received 12 March 1997; revised 2 May 2001)

The present article discusses the notion of variation in discourse as an essential characteristic of language and the linguistic parameters that can be used for its study. The article describes the different traditions in the study of variation and places special emphasis on the role of prosodic analysis in the study of spoken discourse. The study explores the use of discourse markers in the London-Lund Corpus and describes their linguistic variation by introducing the notion of APPROPRIATENESS. This notion, which combines the quantitative and qualitative presence of elements in discourse, is based on a mathematical index that can describe discourse variation with a sound systematic criterion.

1. Introduction

One of the most puzzling features of language is the way in which certain words collocate in specific positions. In some cases, these collocations are 'fixed', as in the case of proverbs or other standardized expressions, but in other cases, collocations respond to internal features of language or to contextual preferences of the speaker. In the latter case, linguists speak about style, genre or register to account for the alternative lexical or grammatical realizations that express essentially similar meanings in different contexts. It can be postulated that linguistic items, therefore, move between two axes: meaning *per se* (semantic domain), and meaning in context (pragmatic domain).

The aim of this article is to describe the pragmatic preferences that lie behind the variation of discourse markers in spoken English and proposes a way to model their variation from an empirical perspective. I believe that there exist several contextual and prosodic constraints that determine what I call DISCOURSE COLLOCATIONS, which explain the pragmatic appropriateness of a discourse marker in a specific position.

[1] I am grateful to anonymous *JL* referees and to T. Espigares and R. Whittaker for their helpful comments in the revision of this article. I am also indebted to Professor Robert Borsley for his insightful observations and comments on the final draft.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

2. Discourse coherence and cognition in discourse: the case of discourse markers

According to Sanders et al. (1992), any theory of discourse structure should meet two requirements: descriptive adequacy and psychological plausibility. These two features seem to me especially relevant to the analysis of spoken discourse because of its elusive nature. Spoken language, due to its spontaneity, is probably the mode of communication where the description of social and cognitive relationships needs more elaboration and research. In ordinary interaction, speakers and hearers have little time at their disposal to process speech and to construct text structure, i.e., the coherent mental representation of the language that is being exchanged. Discourse markers, in this respect, are essential pointers in the mental map of linguistic relationships.

This mental representation is intimately linked to the notion of coherence, i.e.,

> The relations that hold between its subparts or spans [of a text] … [These] relations can apply between spans as small as clauses, or between larger spans such as paragraphs or groups of paragraphs. (Knott & Dale 1994: 36–37)

In fact, as these same authors state with regard to discourse markers ('cue phrases' in their terminology), 'cue phrases are useful for readers when they construct a mental representation of text' (Knott & Dale 1994: 46). In a more recent paper, Knott & Sanders (1998) apply this concept to a study of coherence in Dutch and English texts.

Knott & Dale (1994: 45) describe discourse markers as a 'reasonably homogeneous group' that tends to be realized by '*simple* linguistic expressions … that have become simplified because they correspond to constructs that are in continual use when we process text' [their italics – JRT]. In their description of the elements that can function as 'cue phrases', they enumerate coordinators, subordinators, conjunct adverbs and phrases that take sentential complements (e.g. 'it follows that', 'it may seem that').

In the case of spoken language, however, the cognitive relations that are established between speaker and hearer derive from linguistic chunks (called phonemic clauses, see Boomer 1965) that need not have a uniform grammatical structure but that are based on pragmatic and prosodic coherence, as Romero Trillo (1994) shows. Therefore, it can be stated that prosodic and pragmatic coherence are the basis of cognitive relationships in speech.

In an excellent account of the different approaches to the nature of discourse markers, Fraser (1999) explores several characteristics of these elements, summarized as follows:

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

> [discourse markers] impose a relationship between some aspect of the discourse segment they are a part of, call it S2, and some aspect of a prior discourse segment, call it S1. (Fraser 1999: 938)

In other words, this author identifies discourse markers with the elements that signal a two-place relationship between adjacent discourse segments, e.g. *therefore*, *however*, *and*, etc., and discards what he calls 'commentary pragmatic markers' (e.g. *frankly*), 'focus particles' (e.g. *even*), 'pause markers' (e.g. *hum*, *well*) and interjections (e.g. *oh*). In my view, commentary pragmatic markers, and especially the last two categories – pause markers and interjections – are discourse markers because they are responsible for the scaffolding of interactional meaning in conversation and need careful investigation.

With regard to the meaning of these particles in real usage, the difference between my approach and Fraser's is that he takes what I call a grammatico-semantic point of view, as shown in what he considers the three basic features of discourse markers: they 'do not constitute a separate syntactic category', their meaning is 'procedural not conceptual' and 'every individual discourse marker has a specific, core meaning' (Fraser 1999: 943–945).

My approach, on the other hand, is geared to the pragmatic and prosodic analysis of the markers. In this respect, I will not comment on the first two characteristics identified by Fraser. Regarding the notion of a 'core meaning', the study of discourse markers shows that each element may appear in such a multiplicity of functional contexts, and with such a varied array of meanings, that it is very difficult to assign a core meaning, especially in the case of pause markers and interjections. The problem at stake here is this: can we say that there are different DISCOURSE ENTRIES for an element which derives from a core meaning, given that such a core meaning is not even clear in many cases? In my opinion, this is not possible because, as will be shown below, the elements may vary in meaning depending on variables dealing with context, position in the turn and intonation.

Due to this multiplicity of conditions that model the choice of an element and give a specific meaning to it, the challenge is to find a principled method of selection and analysis of this phenomenon. In other words, how can we decide about the appropriateness of the meaning of an element in a specific discourse slot? What are the contextual interactions between the marker and the discourse slot in which it occurs?

3. Variation in language, variation in discourse: the notion of appropriateness

The question of choice outlined above leads us to the problem of linguistic variation. Schiffrin (1994: 282) describes the VARIATIONIST APPROACH, which stems from linguistic studies on language change, as an attempt

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

to discover patterns in the distribution of alternative ways of saying the same thing, i.e. the social and linguistic factors that are responsible for variation in ways of speaking.

Labov (1972b), the founder of this approach, tried to apply the notions which had been traditionally applied to semantically equivalent variants to the search for text structure and the analysis of text-level variants.

In their research, Labov and his colleagues tried to study the structure of the narrative and its components in relation to the social origins of the speakers. Labov's studies, however, are embedded within a rich linguistic tradition, which can be summarized as follows:

> It is common for a language to have many alternate ways of saying 'the same' thing. Some words like *car* and *automobile* seem to have the same referents; others have two pronunciations, like *working* and *workin'*. There are syntactic options such as *Who is he talking to?* vs. *To whom is he talking?* or *It's easy for him to talk* vs. *For him to talk is easy*. (Labov 1972a: 188)

What are the internal constraints and resources inside the language system and its users that allow for such rich variation? Questions of register, genre, age, social class, cohesion or processing speed, among others, may play an important role in the definition of patterns of use. The problem is, however, that some of these forms alternate and are not in free variation. In other words, many forms have 'sisters' for which they can be substituted in some – hopefully – well-defined circumstances without drastic changes of meaning. Traditionally, the variants that have been studied refer to the fields that Labov describes in the quotation above: phonology, lexical semantics and syntax. However, the level of discourse remains a difficult one to tackle since it is, as Schiffrin (1994: 288) suggests,

> difficult to define the theoretical status of discourse variants. They are not alternative realizations of a single underlying form or representation (as one could argue for phonological, morphological, or syntactic variants); nor do they occupy a 'slot' (a phonological segment, a syntactic constituent) in a grammar or set of grammatical rules.

The difficulty in the analysis of variation in discourse contrasts with its variationist nature in the sense that interactions between speakers are unique. Therefore, it can be postulated that the higher the degree of lexical diversity in a language, the more precise and informational it will be; for example, the availability of a large set of terms for colours will make the descriptions of a painting much more vivid. Likewise, discourse diversity enables speakers to have a wider and more detailed range of possible meanings and functions at their disposal. This can be seen in the availability of choice between several elements to realize the feedback function in a conversation in English: *m*,

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

mhm, *aha*, *yeah*, *yep*, etc. This range of options implies that there is a different degree of appropriateness in terms of context and informational intention.

Therefore, diversity can be seen as a source of richness in language that enables the speaker to formulate his/her ideas with more accuracy, not only from the ideational but also from the interpersonal and the textual points of view.[2]

How can we account for the specific theoretical status of discourse variants? I propose the postulation of the notion of APPROPRIATENESS. I define appropriateness as 'the possibility to choose the most adequate element in the realization of a certain function in a specific context'. By using the notion of appropriateness I highlight the fact that the analysis of a form in discourse is not dependent on any kind of grammatical assessment – as a matter of fact, many forms are considered acategorial, as I will mention below – but on the frequency with which it appears in a significant corpus-based language sample.

Appropriateness is based on the assumption that there are not pre-established rules that can inform us about the correct or incorrect use of a given element in discourse. Speakers rely on the regularities of certain forms that are typically perceived as adequate to a given linguistic situation. These regularities need to be not only linguistic but also extra-linguistic, such as social class, context of the situation, age of the speakers, etc.

The analysis of the data from the London-Lund Corpus[3] in this article will show the distribution of the linguistic behaviour of certain elements at a discourse level, and will model the main patterns of use of these elements in spoken English. In the analysis, all the markers that realize the functions under study will be taken into consideration. The data will show the scalar nature of the use of the markers, i.e., there is a cline in the preference of use of the markers. Therefore, I believe that it will only be through quantitative analyses that the markers' appropriateness for a specific context will be ascertained.

4. THE CONCEPT OF DISCOURSE GRAMMATICALIZATION

The inclusion in the category of discourse markers of elements that have no apparent meaning or grammatical category but play a role in the cognitive structure of interactants led me to develop the notion of discourse grammaticalization. As I have shown in previous studies (Romero Trillo 2000a, b), discourse markers are elements that have undergone a process of discourse grammaticalization and have included in their semantic/

[2] Following the macro-functions described in Halliday (1994).

[3] See Aijmer & Altenberg (1991) for an excellent account on the research possibilities of corpus studies and, specifically, on the London-Lund Corpus.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

grammatical meaning a pragmatic dimension that serves interactional purposes (see also Blakemore 1987). In other words, a grammaticalized marker becomes a homonym in a particular synchronic system which – when appearing with a pragmatic function – constrains the relevance of the proposition it introduces, as Hopper & Traugott (1993) have suggested.

Drawing on the work on grammaticalization by Heine et al. (1991) and Hopper & Traugott (1993), Romero Trillo (2000a) offers the following classification of discourse markers:

- acategorial items such as: *yeah*, *yep*, *m*, *mhm*, etc.;
- lexical items such as: *listen*, *well*, *good*, *fine*, etc.;
- lexical composites such as: *I mean*, *you know*, *the thing is*, *oh my God*, etc.

This grammaticalization approach to discourse markers addresses the question of the versatility of meanings and the non-transparent nature of these elements, which may not have a unique pragmatic meaning. In this area, I disagree with Schiffrin's (1987: 314) claim that not all markers have meaning. She specifically mentions the markers *oh* and *well* which, incidentally, are very prolific in my data. In fact, I think it is linguistically counter-intuitive to believe that the abundance of a discourse marker does not reflect a diversity of pragmatic (not necessarily core) meanings.

This topic leads us to another problem, also mentioned in Knott & Dale (1994: 51), that of 'substitutability': 'if two phrases are intersubstitutable in a passage of discourse, then they should be classified in the same category'. This question, in my opinion, raises the key linguistic question in language variation: can elements belonging to the same class be intersubstituted, and – if so – to what extent? According to these authors, mainly interested in the relationship between function and discourse relations, it is even possible to substitute elements taking into account their position in the clause, e.g. *but* and *nevertheless* in many contexts.

In my view, intersubstitutability of discourse elements depends not only on the realization of the same function, or on the equivalence of a CORE MEANING, following Fraser, but also on FREQUENCY and DISTRIBUTION. In other words, some elements may realize the same function but their frequency or position may be different in a particular context, and so, a speaker would consider the use of one of the elements inappropriate from the point of view of register, social class, etc.

5. Research procedure

In this section, I will address the question of discourse variation and show the mathematical procedure that can be used in order to account for this phenomenon. The analysis will concentrate on the function of discourse markers as elements that have a privileged variationist nature due to their lack of core meaning and their different levels of grammaticalization.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

5.1 *The variables*

The aim of the study is to look into the co-occurring patterns of discourse markers from a pragmatic perspective, i.e. to examine the different uses of discourse markers according to the pragmatic information that they convey when they appear in different contexts.

One obvious premise in this kind of analysis is that the study of formal differences in discourse needs much rigour because these variations always reflect subtle differences in pragmatic meaning. Biber & Finegan (1991) suggest using the following variables in this respect:[4]

- the work (function) that a form does in discourse, e.g. the use of *now* to start a topic and the use of *well* to finish;
- the situational (contextual) or processing (cognitive) constraints that a form reflects, e.g. the inappropriateness of the use of hesitations like *ah* in the middle of a proverb or a fixed expression: 'it's an ill *ah* wind that blows … ' (James 1974);
- a situational or social distinction that a form 'arbitrarily' indexes, e.g. the use of certain expressions in some specific situational contexts, like the different use of *look* and *listen* to draw the attention of a hearer to a topic (Romero Trillo 1997).

The present analysis incorporates the first two features mentioned above. The last characteristic, which has a sociolinguistic basis, has been addressed in earlier work and will not be relevant to this paper.

Thus, the pragmatic categories that will be studied operate between two axes: the meaning of a given discourse marker and the contextual and cognitive constraints that model the meaning of that element in a specific discourse slot.

The discourse markers under analysis are the fifty-four most frequent ones present in the London-Lund Corpus of Spoken English (Svartvik & Quirk 1980).[5] These elements have been classified according to eleven independently motivated functions that were identified for this study. By the notion of INDEPENDENT MOTIVATION I refer to the complete dichotomy between the core meaning, if any,[6] of the marker, and the meaning of the function(s) it may realize. In other words, I do not presuppose any specific one-to-one relationship between form and function. I consider this one of the advantages of the present study: the description of the multifunctionality of discourse elements in conversation stemming from the detachment from their original core meaning. This perspective suggests that it is possible to incorporate new

[4] Author's examples.

[5] The fifty-four discourse markers are the most frequent ones in the whole corpus. The present results were obtained from a detailed analysis of the first ten conversations (over 50,000 words) with a total number of 4,296 elements.

[6] As stated above, elements do not have any core meaning at all in the classical sense, but acquire a functional meaning in the pragmatic organization of a conversation.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

discourse markers into a language, regardless of their core meaning, provided that they fulfil the functional criteria described here, and that they are used in the appropriate slot (Fraser 1999).[7]

In the analysis, each occurrence of the selected fifty-four markers was assigned to one of the functions presented below. In order to validate the analysis, I repeated the classification on two different occasions, and also presented the data with the classification criteria to a group of doctoral students. The coincidence of the analyses showed the reliability of the classification and analysis that I present below.

5.2 *The prosodic component*

As mentioned above, one of the components that model the function of discourse markers is intonation. Few studies contemplate intonation as an essential factor in communication and try to account for prosodic variation in discourse. However, a close look at discourse markers gives evidence of the importance of intonation in describing pragmatic meanings, especially in the case of acategorial elements, and in explaining the multiple meanings that an item may acquire in conversation.

In order to classify the elements, I followed Heritage & Atkinson's (1984) differentiation between neutral and emphatic modes of communication. I applied this classification to discourse markers because one of the most interesting facts in the data was precisely the appearance of the same marker with a range of different intonational contours. In this way, I could differentiate between the neutral and the emphatic functional realizations of the markers by analysing the prosodic contours of the elements according to the guidelines suggested in Halliday (1967, 1970): tones 1 and 2 (falling and rising, respectively) were considered neutral in statements/*wh*-questions and polar questions, and tone 3 (level-slight rise) was also regarded as neutral. Tone 2 with *wh*-questions, and tones 4 (rise-fall-rise) and 5 (fall-rise-fall) were considered emphatic in all cases. In other words, the intonation contours are the key to the further classification of the functions into neutral and emphatic in the cases where this kind of prosodic variation is found.

Obviously, the study of prosody was only considered in the cases when the tonic (prosodic stress) fell on the marker itself. When this happens, the element carries important information for the message, i.e., the NEW information, according to Halliday (1967, 1994). The fact that such cases are found corroborates the importance of these elements in the transmission of information and their centrality in the communicative process (Romero Trillo 1994). There were other, comparatively few cases in which the discourse markers did not realize the tonic but did fulfil a certain function in

[7] This is probably one of the reasons why discourse markers evolve with some freedom and are so varied.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

the turn. For these instances I invented the concept of 'tone 0', or absence of tone and tonicity in the marker.

Other examples led to interesting observations, as in the case of the appearance of two elements together. In these cases, when both prosodic realizations were neutral, I classified the second element as emphatic because of the value of the repetition for the overall meaning of the markers.

In the examples below I indicate the prosodic contours that are realized by the elements in the corpus. These three analytic axes, namely, intonation, core meaning and cognitive and contextual information, led to a large number of possible combinations for each of the fifty-four elements under study. In this way, the appearance of an element, with a certain intonation contour, with a particular position inside a turn and in a specific context, were the variables that constituted the functional assignment parameters.

6. Description of the functions[8]

6.1 *Start of the turn*

An element is used by the speaker to indicate the beginning of the turn.

(1) (a) *WELL: TONES 0, 1, 3*

9 5 430 1 1 A 21 ^finishing 'up with *a :sort* /
9 5 440 1 1 B 11 * ^where've you* !b\een#- /
9 5 450 1 2 A 11 **well** I ^had a sort of ''!dinner 'given for me by /
9 5 450 1 1 A 11 some '':Gr\eeks# /
9 5 460 1 1 A 11 which was. ^very !t\/errible#. /

(b) *YOU KNOW: TONE* 1

9 14 1310 1 1 A 11 and then ^in the /end *of course# /
9 14 1320 1 1 A 11 ^people 'always !d\o#* /
9 14 1330 1 1 B 11 ***you** ^**kn\ow**#;.*; /
9 14 1340 1 1 B 11 [@:m] - it ^has its 'reper:c\/ussions# /
9 14 1350 1 1 B 11 you ^you can 'have !too m/\any 'people# /

6.2 *End of the turn*

An element is used by the speaker to indicate the end of the turn or a summary of his/her information.

(2) (a) *ANYWAY: TONES 0, 1*

7 32 2770 1 1 A 12 !s\ugar# /
7 32 2780 1 1 A 21 ^y\/ou know it's. /
7 33 2790 1 1 a 20 I suppose so . well cheers **anyway** . /
7 33 2800 1 1 A 11 ^ch\eers# . /
7 33 2810 1 1 A 11 I ^had a {b/eer the 'other} :d/ay#- /

[8] In the illustrations of the examples I only show four lines of co-text. The original analysis was done with five lines of co-text before and after the element

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

(b) *THAT'S THAT: TONE 1*

5 18 2680 1 1 B 11 in ^L/ondon# /
5 18 2690 1 1 A * ^n\o#* /
5 18 2700 1 1 (B 11 and ^**that's !th\at**# **-** /
5 18 2710 1 1 B 11 it's ^d\ifferent {in a ^sm\all_place#}# *-* /
5 18 2720 2 1 B 21 - - but in ^London /

6.3 *Sympathetic circularity function*

The current speaker tries to keep the attention and understanding focused on his/her message by showing an intellectual collaboration with the audience.

(3) (a) *YOU KNOW: TONES 1, 2, 3, 4*

9 62 5920 1 1 a 20 yeah /
9 62 5930 1 1 A 11 ''^this 'is a v\ery 'bad th/ing# /
9 62 5940 1 1 A 12 **you** ^**kn\ow** I . but ^n/\obody# /
9 62 5950 1 1 A 11 could do ^/\anything {a^b/\out it#}# /
9 62 5960 1 1 A 11 be^cause a :uni!versity com!m\ittee was 'formed# /

(b) *YOU SEE: TONES 0, 1, 2, 3* (*always at the end of Tone Unit*)

5 35 5300 1 1 B 11 ^y\es# /
5 35 5310 1 1 B 11 well ^qu\ite# /
5 35 5320 1 1 B 11 they ^do 'that sort of :th\/ing **you see**# /
5 35 5330 1 1 B 11 and then they ^see what they've prod\uced# /
5 35 5340 1 1 B 12 and ^then they *'sort of* they ^score them \up# /

6.4 *Feedback*

The addressee uses some linguistic elements to show that he/she is following the ideas expressed in the message.

(4) (a) *YEAH: TONES 1, 2, 3*

7 70 6340 1 1 B 11 *^tw\/o of them# /
7 70 6350 2 1 B 22 ^had to 'keep* a :hundred and _eighty
7 70 6360 1 1 A 11 ^**y\eah**# . /
7 70 6350 1 1 (B 12 :children :quiet a''^m\used for the /after'noon#/
7 70 6370 1 1 B 11 ^so they 'thought the !b\/est thing to 'do#- /

(b) *M: TONES 1, 2, 3, 4, 5*

10 30 2290 1 2 b 20 again and again and again except I just did escort /
10 30 2290 1 1 b 20 yesterday /
10 30 2300 1 1 A 11 ^**[\m]**# /
10 30 2310 1 1 b 20 ((do you)) see what I mean /
10 30 2320 1 1 A 11 ^[\m]# /

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

6.5 *Agreement with the message*

The addressee shows his/her agreement.

(5) (a) *MHM: TONES 1, 2*

1 18 2840 1 1 A 11 it ^makes it pretty!\awkward# /
1 18 2850 1 1 A 11 ^d\oesn't it#- /
1 19 2860 1 1 B 11 ^**[mh\m]**# /
1 19 2870 1 2 A 11 do you ^happen to know whether . [@:m]: Sparrow /
1 19 2870 1 1 A 11 has got an :image . of the man he /

(b) *RIGHT: TONES 0, 1, 2, 3*

9 55 5240 1 1 A 11 * ^that would be l\ovely# /
9 55 5250 1 1 A 11 ^y\es# . /
9 55 5260 1 1 A 11 ^**r\ight**# . /
9 55 5270 1 1 A 21 ^that would be !very* ((nice)) /
9 55 5280 1 1 C 11 *^y\es# . /

6.6 *Disagreement with the message*

The addressee shows his/her disagreement.

(6) (a) *NO: TONES 1, 3, 4, 5*

9 24 2400 1 1 A 11 ^what a br/illiant i:d\ea# /
9 24 2410 1 1 A 11 ^did they d\/o it# /
9 24 2420 1 1 C 11 [@m] . ^**n\o**# *.* /
9 24 2430 1 1 C 21 it was ^cheaper to 'pay . [@] /
9 24 2440 1 1 A 11 *^\oh#* /

(b) *WELL: TONES 0, 1, 2, 3, 4, 5*

9 9 810 1 1 B 11 ^are you in t\ouch with the St {B/ee's} cr/owd#/
9 9 820 1 1 B 20 or /
9 9 830 1 1 A 11 ^**w\ell**# /
9 9 840 1 1 A 11 you ^kn/ow# . /
9 9 850 1 1 A 11 to a ^certain ext/ent# /

6.7 *Agreement acknowledgement*

The current speaker replies to the agreement expressed by the addressee.

(7) (a) *YEAH: TONE 1*

7 12110850 1 1 B 12 ^oh it was a ^started off ''!v\iciously# /
7 12110860 1 1 A 21 *(- . laughs) ^really ((vicious)) (- laughs)* /
7 12110870 1 1 (B11 *^**y\eah**#- /

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

7 12110880 1 1 B 11 ^real 'high-powered 'sort of* . th\ing# /
7 12110890 1 1 B 11 but ^now it's !cooled d\own **((_quite a _ bit))# /

(b) *M: TONE 1*
9 17 1680 1 1 A 11 ^[\m]#* /
9 17 1690 1 1 a 20 *oh God yes it's appalling* . yes /
9 17 1700 1 1 A 11 ^**[\m]**# - - /
9 17 1710 1 1 B 11 have ^you 'got a [?] - !l\eave n/ow# - /
9 18 1720 1 1 B 11 *sab^b\/atical#* /

6.8 *Attention-getting*

The speaker uses some linguistic elements to draw the attention of the addressee to the message.

(8) (a) *LOOK: TONES 1, 3*
2 2 270 2 1 B 21 1 *((was when))* /
2 2 280 1 1 A 11 1 *^y/es#* /
2 2 270 1 1 (B 11 1 ^I said ((it was)) :**l\ook**# /
2 2 290 1 1 B 13 1 ^I ^I'm . we're ^n\ot prep/ared# /
2 2 300 1 1 B 11 1 to ^go on being p/art# /

(b) *LISTEN: TONE 1*
7 13412060 1 2 a 20 oh well that's very good if you can fool him just /
7 13412060 1 1 a 20 for an instant - - *((2 to 3 sylls))* /
7 13412070 1 1 B 11 *^**l\isten**# /
7 13412080 1 2 B 11 ^((if* you)) !feel like a f\ilm to'morrow 'night /
7 13412080 1 1 B 11 M/ike#- /

6.9 *Self-initiated self-correction*

The speaker corrects him/herself.

(9) (a) *I MEAN: TONES 0, 1, 3*
4 15 2270 1 1 A 11 I^h\ope 'so# /
4 16 2280 1 1 B 11 ^I'm sure we {c/an} r/eally# /
4 16 2290 1 1 B 21 **I** ^**mean** it's [@] . /
4 16 2300 2 1 A 21 ^so much /
4 16 2310 1 1 20 (telephone starts ringing) /

(b) *WELL: TONES 0, 1, 3*
9 88 8500 1 1 A 11 to ^d\eal with n/ow# - /
9 88 8510 1 1 A 11 ((^th/en)) the dr\ainage# /
9 88 8520 1 1 A 11 ^**w = ell**# - - /
9 88 8530 1 1 A 11 ^/I don't kn\ow# /
9 88 8540 1 1 A 11 ^how long :that'll 'take m/ore# /

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

6.10 *Word-search*

The speaker uses a discourse marker to avoid silence while searching for the appropriate word or expression.

(10) (a) *MHM: TONE 0*

10 40 3140 1 2 A 14 ^they ^they ^they ^they say !{v\est'meaning}/
10 40 3140 1 1 A 14 !\/under'garment# /
10 41 3150 1 1 b 20 **[mhm]** - - - we're fairly strong ((today on)) /
10 41 3160 1 1 A 11 they ^have their !\/own 'words of 'course# /
10 41 3170 1 3 A 14 ^there's an ^there's a !nice 'word for the [@m] /

(b) *M: TONES 0 (middle of the turn), 3 (beginning of the turn)*

9 82 7990 1 1 A 11 but ^all the !s\ame# . /
9 82 8000 1 1 A 20 [@:m] /
9 82 8010 1 2 a 20 **[m . m . m]** - *[t] it's a convenience for you
9 82 8010 1 1 a 20 sylls))* /
9 83 8020 1 1 A 11 they ^ought to 'be *. sort of t/ied \up#* /

6.11 *Phatic*

A participant in the conversation makes use of linguistic elements to avoid being silent while there is no exchange of information.

(11) (a) *M: TONES 1, 4*

9 18 1750 1 1 A 11 ^\ever# /
9 18 1760 1 1 A 20 [@m] /
9 18 1770 1 1 B 11 ^**[\/m]**#- /
9 18 1780 1 1 A 11 and I ^started /off# /
9 18 1790 1 1 A 11 you ^kn/ow# /

(b) *YEAH: TONE 1*

7 13912540 1 1 B 11 *^y\eah# . /
7 13912550 1 1 B 11 ^[\m]#* /
7 13912560 1 1 A 11 ^**y\eah**# /
7 14012570 1 1 B 11 [j] you ^just cut 'out the !br\eaks# /
7 14012580 1 1 B 11 and you ^get con_tinuous **!f\ilm#** /

7. Discussion of the analysis

After the classification of the markers, the aim was to determine the appropriateness of a specific marker in the realisation of a particular function.

In general terms, the first result of the analysis showed that there exists a positive correlation between the number of functions that a particular element realizes, and its absolute frequency of appearance ($r = 0.7068$, $p < 0.001$). That is, there is a direct link between the number of functions

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

realized and the frequency of a particular element in language. For example, we can compare the element *yes*, which represents 19.45% of all the markers that appear in our corpus and realizes ten different functions, with *I agree*, which appears only twice and realizes only one function. According to the data, the high frequency of *yes* enables it to realize various functions, and be used in many different contexts with a rich diversity of intonation contours.

This turned out to be a trend in the data. In my opinion, though it seems logical that the more frequent an element, the more functions it realizes, this does not seem to be shared by other aspects of language. For example, in lexical semantics, the increase in the meanings of an element may not depend on its raw frequency, but on other reasons such as etymology.

In the case of discourse markers, the relation between frequency and functionality increases because of the process of grammaticalization and the absence of a core meaning in discourse. In fact, these items may take a number of functions, by appearing in different prosodic and discoursal contexts that give them a wide range of meanings. On the other hand, the few elements that reflect their core lexical meanings in discourse, like *listen* or *look*, do not easily lend themselves to being used in many functional contexts.

In my view, here lies one of the crucial differences between grammatical(ized) items and lexical items: the adaptability to new meanings and/or functions in direct correlation with their frequency. In other words, the frequency of occurrence of grammatical(ized) items enables the elements to fill a large diversity of functional slots, whereas lexical items, on the other hand, show a negative correlational tendency that is related to their multiplicity of meanings and their occurrence in language. This is simply because there is usually a clear cline in the frequency of use in the different meanings of polysemous items. In more grammatical terms, Ellis (1993) explains the formal differences between grammatical and lexical items by stating that the former tend to acquire external markers simply because they are more frequent in the language. He, in fact, compares the elements *rabbit*, colour terms and plural markers as forming a cline from lexicality towards grammaticality.[9]

In a different perspective, I also believe that the main difference between meaning and function in discourse is that the concept of meaning is inherently linked to a linguistic item, whereas the idea of function is perceived as an open category realized by a certain number of linguistic items. A function, although also a carrier of meaning, is not autonomous because it depends on elements that change their original meanings when they enter a fuzzy set of grammaticalizing relations subjected to the external factors described above: meaning, context, prosody, etc. In this way, a

[9] I thank M. Romano for this reference.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

function is equivalent to a sort of slot filled by elements originally belonging to various word classes.

In the present case, this implies that the idea of meaning is blurred when we speak about discourse because we can no longer describe meanings but functions. If we apply this notion to our three groups of discourse markers, we find that in the case of acategorial items or lexical composites this does not make any difference, simply because the former do not have core meanings, and the latter cannot be directly studied from a lexical perspective, but from a pragmatic one. However, the group of lexical items poses some difficulties because the analyst is forced to decide whether the appearance of an element in a piece of discourse is an instance of a lexical or of a pragmatic item. In order to differentiate between the two uses of the same item, I had to resort to the three parameters explained above, meaning, context and prosody. Only this analysis, based on discourse criteria, will help us to determine the semantic or pragmatic use of an element (Romero Trillo 1997).

Keeping these problems in mind, I decided to design an analytical model that could explain the relationship between frequency and function in order to learn more about the appropriateness of discourse elements. It seemed that many of the markers showed a certain SPECIFICITY in their realization of a given function, i.e., they 'specialized' in their appearance in language; and conversely, some functions were only realized by one or two elements.

In order to systematize this, I devised the concept of SPECIFICITY, which will be later developed in mathematical terms. Specificity can be defined as the degree of uniqueness that a function shows towards an element, and an element towards a function, in discourse. Specificity is, therefore, the criterion which shows the regularities of discourse by considering the functions and the appearance of the elements which realize those functions. Specificity determines the degree of appropriateness of an element in a certain context. Discourse appropriateness, therefore, is based on a cline that ranges from more to less specific elements in a particular context.

Four different models of specificity can thus be envisaged in this respect:

1. Specificity of function by an element; i.e. an element realizes only one function. For example, the element *Christ* appears once in the corpus and realizes the function of Emphatic Feedback.
2. Specificity of element in a function; i.e., a function is only realized by one element. This case does not occur in the data.
3. Multiple specificity of functions by an element; i.e., an element realizes several functions. This is very common in the corpus: for example, *anyway* realizes five different functions (and it only appears 24 times!): Neutral and Emphatic Sympathetic Circularity, End of Turn, Start of Turn, and Neutral Disagreement.
4. Multiple specificity of elements in a function; i.e., a function can be realized by several elements. This is also very frequent in our data: for

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

example, the Neutral Feedback Function appears 1,307 times and is realized by 27 different elements.

The cases can be seen in the diagrams in figure 1.

The overall distribution of elements and functions in the corpus is presented in figures 2 and 3.

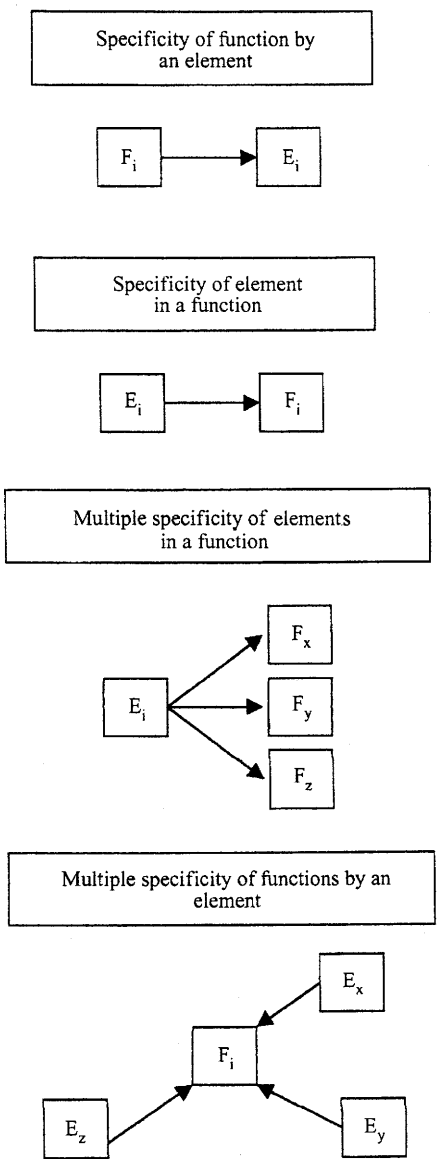


Figure 1
Types of specificity between elements and functions

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

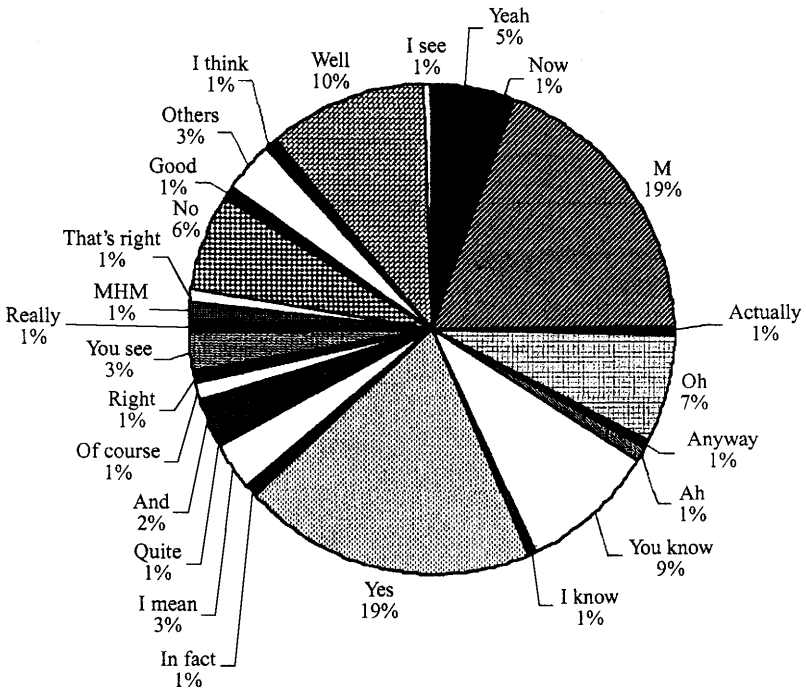


Figure 2
Distribution of elements in the corpus

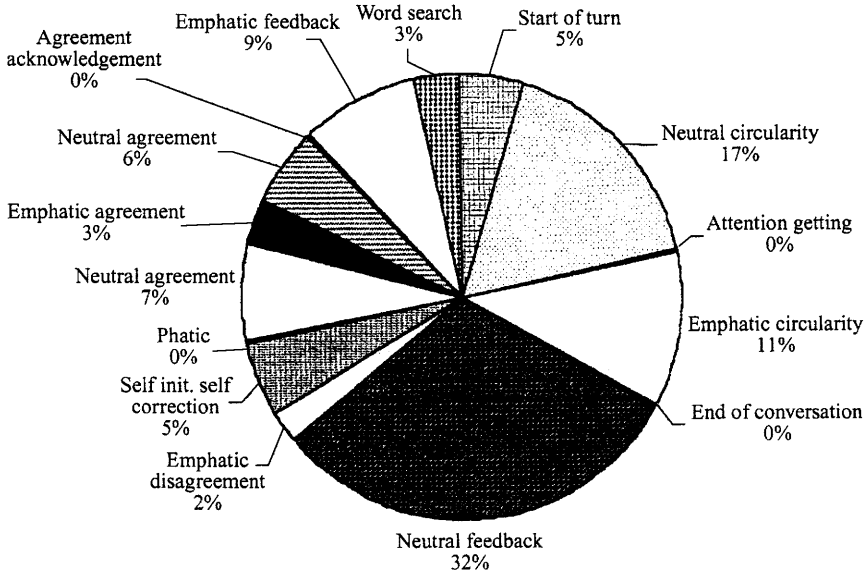


Figure 3
Distribution of functions in the corpus

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

In order to study the specificity of elements and functions as a means to ascertain discourse appropriateness, I designed what I shall call SPECIFICITY INDEXES that will enable us to measure the relationship between the functions and the elements. The indexes range from 0 (the lowest value) to 1 (the highest value).

The SPECIFICITY INDEX OF ELEMENTS (SI_e), which deals with the relationship between an element and all the functions it realizes, is calculated with the following formula:

$$1 - \frac{\text{No. of functions}}{\text{Overall presence of the element}}$$

The values of this index will range from 0 to 1. The lower values, i.e., close to 0, indicate the presence of a high number of functions realized by an element, or a low presence of the element in comparison with the functions it realizes. This implies that when the numerical result of the index is very low (close to 0), there is a low specificity of the element, and, consequently, it tends to realize a not very specific set of functions. On the other hand, when the index is close to 1, the appearance of the element is very specific in discourse and, therefore, one can be quite sure about the function it realizes.

I calculated the Specificity Index of all the elements in the corpus with the formula and classified them according to three levels of specificity as shown in table 1.

- When the index is from 0.9 to 1: high specificity.
- When the index is from 0.5 to 0.89: medium specificity.
- When the index is lower than 0.49: low specificity.

After the calculation of the index for each element, I wanted to check the relationship between the frequency and its Specificity Index in order to see whether the frequency influenced the functional specialization of an element. The results showed a positive and statistically significant correlation ($r = 0.42$, $p = 0.001$). This means that the more frequent an element, the higher its Specificity Index. In other words, when an element is very frequent in language, its functional meaning tends to be very specific and we can take its most frequently realized function as the prototype.

This means that the elements with a high Specificity Index can be used with high reliability in the functions that they realize, while the elements with a Specificity Index close to 0 can be used with certain guarantees in those functions, although they are not statistically significant or, in other words, they are not prototypical in any particular function.

In terms of language description, it is also interesting to mention the cline between the most grammatical(ized) elements (the most specific) towards the

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

High Specifity		Medium Specificity		Low Specificity	
you know	1.00	really	0.89	that's it	0.46
mhm	0.99	say	0.88	oh yeah	0.40
you see	0.99	of course	0.88	God	0.33
well	0.99	right	0.86	yep	0.25
yes	0.99	I see	0.86	for God's sake	0
m	0.99	I know	0.84	listen	0
and	0.98	and then	0.83	good God	0
I mean	0.98	that's right	0.83	aha	0
yeah	0.97	anyway	0.79	hey	0
oh	0.97	hm	0.78	God knows	0
in fact	0.94	ah	0.78	bloody	0
actually	0.94	quite	0.77	ugh	0
look	0.91	goodness	0.75	Christ	0
I think	0.91	oh my God	0.75		
good	0.90	my goodness	0.75		
now	0.90	heavens	0.72		
		fine	0.70		
		good Lord	0.60		
		gosh	0.58		
		I agree	0.50		
		OK	0.50		
		indeed	0.50		

Table 1
Levels of specificity of the elements

most lexical ones (the least specific), many of the latter with religious connotations.

Nevertheless, when we speak about appropriateness in discourse we are evidently dealing with a question of degree. This is why in many cases speakers are not able to give exact and definite explanations about the use of an element in discourse and rely on their 'intuitions'. Some uses of the elements may be considered marginal, although appropriate, due to their low specificity. This is the reason why the notion of appropriateness is also based on the CENTRALITY (prototypicality) of a given element for a certain function. The present analysis could enable speakers to decide on their discourse style by using the most central or the most peripheral elements, while maintaining appropriateness.

With regard to the study of the functions, a parallel analysis was carried

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

out with the design of a SPECIFICITY INDEX OF FUNCTIONS (SI_f) which is calculated in the following way for each function:

$$1 - \frac{\text{No. of elements}}{\text{Overall presence of function}}$$

As in the case of the elements, the values for the specificity of the functions also range from 0 to 1. The high values indicate a small quantity of elements and/or a high presence of the function. They indicate, then, a high specificity of the function in its realization by a given set of elements. For example, the most specific function in our data, Self-Initiated Self-Correction, appears 221 times and is only realized by two elements: *I mean* and *well*.

On the other hand, low values express a high frequency of the elements and/or a low frequency of the function under analysis. Therefore, they reflect a low specificity of the function in its realization by these elements. For instance, the least specific function, Agreement Acknowledgement, appears seven times and is realized by four different elements: *yeah*, *yes*, *m* and *oh*.

The degree of specificity for each function is given in table 2. The values of specificity are the same as for the elements:

- when the index is from 0.9 to 1: high specificity;
- when the index is from 0.5 to 0.89: medium specificity;
- when the index is lower than 0.49: low specificity.

High Specificity		Medium Specificity		Low Specificity	
self-initiat. self correction	0.991	emph. agreem.	0.899	agreement acknowl.	0.429
neutral feedback	0.979	end conver.	0.818		
neutral disagreement	0.975	attent. get.	0.769		
searching	0.957	phatic	0.722		
start of the turn	0.596				
neutral agreement	0.946				
emphatic disagreement	0.945				
emphatic sympath. circularity	0.941				
emphatic feedback	0.93				

Table 2
Levels of specificity of the functions

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

In this case, I also calculated the correlation between the frequency of appearance of each function and the number of elements that realize it. The correlation results were statistically positive and significant ($r = 0.78$, $p = 0.006$). This indicates that the larger the number of elements that realize a function, the more frequent this function is.

The results also show that one central element prototypically realizes the functions which are more specific. In fact, the most specific functions are typically realized by one element with a relatively high percentage of use.

- Self-Initiated Self-Correction: *I mean*, 58.4%.
- Neutral Feedback: *m*, 45.0%.
- Neutral Disagreement: *well*, 62.2%.

Therefore, it could be argued that in the case of discourse markers, and probably in other aspects of language, the speakers' intuition comes from a subconscious awareness of these measures when they identify the central, less central and the peripheral elements which realize a given function.

8. Conclusions

The present article has discussed the notion of variation in discourse from a mathematical perspective. For this purpose, I have exemplified the basic approach to the issue by using discourse markers and their relationship with the discourse functions that they realize. Unlike the traditional analysis of these elements, which relies on their original semantic meaning, my approach focuses on the function of these elements when they appear in discourse. The analysis takes into consideration three axes that contribute holistically to their meaning in discourse: core meaning, prosody and context. These three aspects together make up the new discourse 'face' of the elements.

It is very interesting to note that the newly acquired discourse function of the elements is achieved by going through a process of grammaticalization. I have shown that these elements do behave in a principled way in their discourse functions, and that their use by the speakers is based on what I have called appropriateness, i.e., the suitability of use of a discourse form in a certain context. This notion is deemed parallel to the idea of grammaticality at the sentence level.

In general terms, this article proposes a path for the study of discourse by creating a model of analysis. In the first place, there is a tight bond between the concepts of appropriateness and variation in discourse. Very often, certain elements may sound strange, even though they may be absolutely correct from a lexico-grammatical perspective. Obviously, the language user, especially the non-native, should be given objective criteria to determine the appropriateness of an expression in a given context, and should be shown the limits of variation as well as the borders of register freedom.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

It is an obvious fact that when one studies language in grammatical terms, the speaker has at his/her disposal a series of rules which determine what can and what cannot be said in a language. However, for discourse I advocate the notion of discourse variation, although never from a prescriptive perspective. My approach, which takes into consideration the fuzzy limits of the meaning of discourse items and their mobility, since they depend on the interplay of three different components, is to resort to a quantitative model that could help us describe the language user's intuitions. That is why I propose the idea of creating a Specificity Index for each of the elements and for each of the functions as an objective and clear paradigm to determine the appropriateness of a given expression in a specific context. As shown above, by applying Specificity Indexes it is possible to determine the correlation between the appearance of a form and the number and diversity of elements that realize it. This method of analysis also enables the linguist to characterize variation from a historical and dialectal perspective because the functions of language remain constant and the variation occurs in the appearance of a certain form in a particular context.

This mathematical model also opens some other interesting possibilities of research. The immediate one is the pedagogical application in the teaching of discourse. Very often students of a foreign language have to face 'linguistic' explanations of the kind: 'we say it because it sounds better', or 'it is more frequent', etc. It is obvious that these explanations about the use of certain forms in discourse suffer from the difficulty of not having a clear model for deciding on the appropriateness of a form in a given context. Perhaps one of the reasons why teachers tend to focus on grammatically oriented lessons is the availability of prescriptive grammars and textbooks. The present approach would enable linguists to start creating models of usage in discourse that could offer both teachers and students a clear pattern of use.

The second obvious application of this approach is the possibility of modelling language. The natural sciences have been trying for a decade or so to model natural behaviour in biology, physics, geology, etc. The research method in these fields is based on the accurate observation of reality, which then leads to predictions of future situations. In the specific case of discourse variation, it seems evident to me that by using the model presented in this article, linguists would be able to record data of language in use longitudinally throughout a certain number of years, and then predict the evolution of certain categories, either elements or functions, according to the specificity showed throughout time. This can be clearly observed, for example, in the study of the grammaticalization processes that affect certain linguistic categories, and in the different functional roles that lexical(ized) and grammatical(ized) elements may adopt, as explained above.

In a similar way, the analysis of variation in discourse may help linguists to identify what I call DISCOURSE DENSITY, parallel to Halliday's 'lexical density', i.e., the frequency of use of discourse elements in a particular text.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

It is important to stress the relevance of this notion in discourse studies because in the same way that lexical density captures the level of ideational information present in a text, discourse density describes the interpersonal stance of the speaker towards a piece of discourse. In other words, the amount and degree of variation present in discourse will determine the attitude of the speaker/writer towards the recipients of the message.

To sum up, I believe that the use of mathematical patterns to explain linguistic phenomena may lead to very interesting discoveries about some aspects of language, like discourse variation, which are usually denied to the linguist due to the large amount of data that needs to be analysed.

REFERENCES

Aijmer, K. & Altenberg, B. (eds.) (1991). *English corpus linguistics: studies in honour of Jan Svartvik*. New York: Longman.

Biber, D. & Finegan, E. (1991). On the exploitation of computerised corpora in variation studies. In Aijmer, K. & B. Altenberg (eds.), 204–220.

Blakemore, D. (1987). *Semantic constraints on relevance*. Oxford: Blackwell.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech* **8**. 148–158.

Ellis, J. M. (1993). *Language, thought and logic*. Chicago: Northwestern University Press.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics* **31**. 931–952.

Halliday, M. A. K. (1967). *Intonation and grammar in British English*. The Hague: Mouton.

Halliday, M. A. K. (1970). *A course in spoken English: intonation*. London: Oxford University Press.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd edn.). London: Arnold.

Heine, B., Claudi, U. & Hünnemeyer, F. (1991). *Grammaticalization*. Chicago: University of Chicago Press.

Heritage, J. & Atkinson, J. M. (1984). Introduction. In Atkinson, J. M. & Heritage, J. (eds.), *Structures of social action: studies in conversation analysis*. Cambridge: Cambridge University Press. 1–15.

Hopper, P. J. & Traugott, E. C. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

James, D. (1974). Another look at, say, some grammatical constraints on, oh, interjections and hesitations. *Papers from the 10th Regional Meeting of the Chicago Linguistic Society*. 242–251.

Knott, A. & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes* **18**. 35–62.

Knott, A. & Sanders, T. (1998). The classification of coherence relations and their linguistic markers. An elaboration of two languages. *Journal of Pragmatics* **30**. 135–175.

Labov, W. (1972a). The study of language in its social context. In *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press. 183–259.

Labov, W. (1972b). The transformation of experience in narrative syntax. In *Language in the Inner City*. Philadelphia: University of Pennsylvania Press. 354–396.

Romero Trillo, J. (1994). Ahm, ehm, you call it theme? … A thematic approach to spoken English. *Journal of Pragmatics* **22**. 495–509.

Romero Trillo, J. (1997). Pragmatic mechanisms to obtain the addressee's attention in English and Spanish conversations. *Journal of Pragmatics* **28**. 205–221.

Romero Trillo, J. (2000a). Los procesos de gramaticalización desde una perspectiva discursiva. In Ruiz de Mendoza Ibáñez, F. (coord.) *Panorama actual de la Lingüística Aplicada. Conocimiento, procesamiento y uso del lenguaje* (vol. 1). Logroño: Universidad de la Rioja. 361–374.

Romero Trillo, J. (2000b). Systemic tools to establish appropriacy in discourse: the case of interjections in English and Spanish conversations. *Revista Canaria de Estudios Ingleses* **40** (Special volume in honour of M. A. K. Halliday). 59–76.

Sanders, T., Spooren, W. & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes* **15**. 1–35.

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.
Schiffrin, D. (1994). *Approaches to discourse*. Oxford: Blackwell.
Svartvik, J. & Quirk, R. (1980). *A corpus of English conversation*. Lund: Lund University Press.

Author's address: Departamento de Filología Inglesa, Facultad de Filosofía, Universidad Autónoma de Madrid, 28049-Madrid, Spain.
E-mail: jesus.romero@uam.es

https://doi.org/10.1017/S0022226701001050 Published online by Cambridge University Press