Correspondence Analysis of the Synoptic Gospels

A. J. M. LINMANS
University of Leiden, The Netherlands

Abstract

Taking a recent article by D. L. Mealand (*Literary and Linguistic Computing*, 10: 171–82, 1995) as the starting-point, this study investigates what correspondence analysis (CA) could contribute towards the stylometric study of the synoptic gospels (Mark, Matthew, and Luke). Results from the author's dissertation are presented, based mainly on syntax, in particular the subordination of clauses. Mealand's source investigations of the Gospel of Luke, building on CA and other multivariate techniques, are shown to be flawed because they fail to take into account the discourse-type constitution of the gospels. Stressing the importance of discourse type, the present investigation, by contrast, seeks inspiration from modern sociolinguistic research on register. Because discourse type acts as a third variable—alongside corpus and language—a generalized form of CA is needed, using loglinear analysis to produce restricted models. Generalized CA proves to be a flexible tool to display discourse-type and authorial preferences in an unalloyed form. Finally, it is argued—and exemplified—that multivariate analysis is particularly valuable when used in combination with advanced discourse-functional linguistics.

1. Introduction

Multivariate data analysis has lately been gaining ground in Biblical stylometrics. Some time ago in this journal, David L. Mealand (1995) applied correspondence analysis (CA) to the Gospel according to Luke to shed new light on the intricate problem of the source relationships between the synoptic gospels (Mark, Matthew, and Luke)—the so-called synoptic problem. Mealand basically sees his stylometric results as being compatible with the standard theory, a two-source hypothesis that may be rendered as shown in Fig. 1.

In the figure, Q (from German *Quelle*, i.e. 'source') is the hypothetical Sayings Source which, in addition to Mark (~70 CE), is assumed to be at the origins of both Matthew and Luke (between 80 and 90 CE). Other source theories, however, are not excluded by Mealand at this stage of enquiry, especially the possibility of Luke being dependent on Mark and Matthew.

CA is a multivariate technique that maps both the row and column entries of a frequency matrix as points in a multidimensional space. Within this space, the distances between the row points visualize the extent to which they resemble each other, and so *vice versa* do the distances between the column points. To produce a representation that is easily interpretable on paper, the multidimensional image is commonly reduced to an optimal two-dimensional display. Unless the underlying data are very complex and chaotic, such reduced pictures often account for more than 60–70% of the information content of the complete dataset (also called inertia). Scrutiny of higher dimensions is sometimes needed to achieve a fuller picture.

The interpretation of the cross-relationships between row points and column points is less straightforward than the interpretation of the relationships within each category separately. One way to explain the cross-relationships is in terms of the direction and length of the vectors involved; i.e. the more clearly row and column vectors point in the same direction (starting from the origin), the more they are mutually associated, and the longer a row or column vector, the more the entry involved deviates from the average row or column profile, and the more interesting it is. For a detailed exposition of correspondence analysis, the reader is referred to Greenacre (1984).

In the same year as Mealand published his article on Luke, the present author independently published his doctoral dissertation (in Dutch), applying CA to the synoptics (Linmans, 1995; in English *Subordination in the Synoptic Gospels: Syntax, Discourse-Functions and Stylometry*). The resemblance between both studies notwithstanding, there are important differences as well, some of them far-reaching. Four points may be mentioned in this regard. First, my scope in using statistical data not only from Luke, but from all three synoptics, was broader than Mealand's. This enabled me to address a wider range of topics.

Second, the two studies do not select the same linguistic feautures as a basis for comparison. Mealand uses a mixture of parts of speech, word length measurements, lexemes, word forms, and phonetic features; twenty-five categories in all, most of them inspired by Neumann (1990). In addition to investigating the distribution of the twenty-three commonest lexemes of the synoptics, I extended the field of research to syntax, in particular subordinate clauses, of which I distinguished twenty types.

Third, contrary to Mealand, I chose not to focus exclusively on source-critical issues, but also investigated the styles of the authors/redactors of the gospels in their own right. Moreover, I attempted to deepen the quantitative analysis through state-of-the-art linguistic analysis, mainly from a discourse-functional viewpoint.

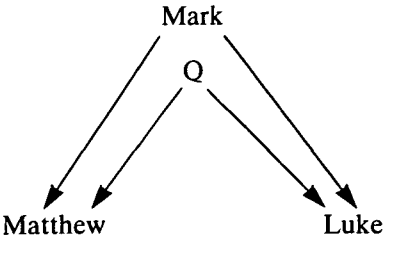

Fig. 1 The standard two-score hypothesis.

Correspondence: A. J. M. Linmans, University Library, Witte Singel 27, PO Box 9501, 2300 RA Leiden, The Netherlands. E-mail: Linmans@Rulub.LeidenUniv.nl

Fourth, and methodically perhaps most importantly, I, in contrast to Mealand, systematically made an effort to take into account the influence of register on the distribution of linguistic features. In this work, inspiration was drawn from modern sociolinguistic studies of register using multivariate analysis to elucidate variation across speech and writing. By contrast, in Mealand, register plays a marginal role only (see the comments about genre (Mealand, 1995, pp. 173 and 177), never becoming the subject of the multivariate analysis itself. Allowance should be made for Mealand's being aware of the likely importance of genre, on which he has lately been working extensively (personal communication).

One of the main goals of the present contribution will be to prove that Mealand's neglect of the register factor seriously impairs his source-historical conclusions. As we proceed, I shall find occasion to present some of the findings of my dissertation. After my critical confrontation with Mealand's views, I shall adduce other potentialities of multivariate analysis. In particular, it will be argued that CA can help us give a description of the personal style (idiolect) of the gospel authors. Multivariate explorations will prove to be especially beneficial when taken in conjunction with advanced linguistic approaches. The critical text used here is the standard Nestle–Aland, 27th edition; the investigations are carried out with an electronic version of it, available at the University of Nijmegen.

2. Corpus Variable and Language Variable

To begin with, in this section and the following, I shall make some preliminary remarks about the selection and interplay of variables in stylometric CA, partly reflecting on Mealand (1995). In applying multivariate stylometrics, at the outset the researcher has to decide which variables he is going to use. Commonly, first of all there are two types of variables to be considered. The first one, naturally, is constituted by the corpus under consideration—this supplies the independent variable. As a second variable, we normally choose linguistic features which serve to distinguish the texts in the corpus—this is the dependent variable. Depending on the purpose one has in mind, the categories of the former variable may be different authors (e.g. in literary attribution studies), different works of one author (e.g. in chronology studies), hypothesized or alleged sources (in source detection studies), genres and registers (in sociolinguistic variation studies), and so forth. The latter variable requires the choice of appropriate linguistic units. These can be features of lexicon, phonology, grammar, writing system, even rhetoric, provided that they are unambiguously definable.

This is the place to clarify a terminological point, as, with regard to the linguistic side, Mealand uses *variable* in a different sense from that adopted here. For example Mealand calls the Greek words *kai* ('and'), *de* ('but'), *gar* ('for'), and so forth, variables, whereas we shall follow the practice, common in qualitative data theory, of calling them categories (or, alternatively, values, attributes, types), to be subsumed together into one variable *linguistic items*. Likewise, on the corpus side, we have, for example, the variable *synoptic gospels*, covering the categories Mark, Matthew, and Luke. Categories are the basis for categorizing or pigeonholing items occurring in the texts, consequently they can be assigned frequencies; there is a type–token relationship between the categories and the countable occurrences. At this point we shall leave terminology.

Turning to the corpus variable, the designation of text categories may cause problems, especially in source detection studies. Our case is an example of this, because the source boundaries within the synoptics are not marked, either intentionally or unintentionally, sources being assimilated into the embedding text and adapted to its style. Therefore, clear fingerprints—here and in other instances—are wanting. Moreover, though sometimes alleged sources may still be extant, for example Mark, more often they have been lost, and are, therefore, purely hypothetical; for example Q. Various strategies to circumvent the problem of constructing a fitting variable can be contemplated and I shall mention two of them.

The first strategy is feasible if the texts under consideration partly overlap in content. In this case, one may establish parallel (or overlap) schemes and partition each text accordingly. Applying simple mathematics produces seven overlap combinations for the synoptics: one triad (Mk/Mt/Lk), three doubles (Mk/Mt, Mk/Lk, Mt/Lk), and three singles (Mt, Mk, Lk). This enables us to divide each gospel into four sections, which we shall call parallel sections. Mark, for example, is divided into four (discontinuous) parallel sections, corresponding respectively to the schemes Mk/Mt/Lk, Mk/Mt, Mk/Lk, and Mk—the other schemes in this case not being applicable. For the synoptics this adds up to twelve (3×4) sections in all. As a result, we have generated, beyond the original corpus variable with the three gospels as categories, an additional corpus variable of twelve categories, especially useful for source studies. Basically this was the strategy which I adopted in my dissertation. It should be noted that the division is carried out fairly mechanically. Accordingly, it is essential to perceive that the parallel sections cannot automatically be claimed to coincide with real source boundaries. Only if further analysis shows that the parallel sections have distinct style profiles, may source relationships be deduced—and the jump be made from mathematics to historical genesis.

Mealand combines this first strategy with a second, sampling. Within the confines of the parallel sections of Luke, he takes samples of 500 consecutive words. The assumption is that if it can be shown by CA that in fact the samples cluster along parallel section lines, these parallel sections somehow point to different origins. In such cases, sources may be postulated that are co-extensive with parallel sections. For Mealand, working within Luke, sampling is a necessary means to reach conclusions about sources. By contrast, intertextual comparison (i.e. between the gospels) can dispense with sampling, since cross-gospel clustering of—unsampled—parallel sections (e.g. Q–Matthew and Q–Luke, or the triple material of Matthew, Mark, and

Luke) indicates of its own accord source relatedness, which will be elaborated below in Section 10.

3. Side Effects and the Need for Control Variables

The variables mentioned above, pertaining respectively to corpus and language, are relatively clear-cut. However, we have to reckon with other, possibly interfering, influences. Language behaviour may be shaped to a large degree by factors that cut through the text delimitations of the corpus, since many quantitative characteristics of language use are not due to the individual, stylistic usages of the authors in the corpus. In multivariate analysis, in order to control these possible side effects, and to grasp the extent of their impact, we need additional variables.

Potential side effects, structurally apt to give rise to variation in speech and writing, are known from sociolinguistics, for instance dialects (based on geography) and sociolects (based on class, gender, age, education, etc.). The kind of variation which appears most relevant to us is called variously register, text type, or genre; i.e. variation correlated not to the group to which a person continuously belongs (as in dialects and sociolects), nor to his idiolect, but to the type of speech or writing situation—in other words, to how one normally expresses oneself given the circumstances of use. (On the different meanings attributed by scholars to register, text type, genre, style, etc., see Biber (1994, pp. 51–3).)

Register investigation through multivariate analysis has proved to be especially worthwhile when based on a large corpus of different sorts of situationally distinct texts (e.g. Biber, 1988, who has concentrated on English material; as far as I know, so far his method has not been adopted in either biblical or classical studies). However, when, as in our case, we want to examine variation *within* texts, without there being an immediately obvious falling apart of the texts into registers, we face a much more complex task (cf. Biber and Finegan (1994), on intra-textual variation within medical research articles). We might try to circumvent the problem by first taking arbitrary samples, subsequently subjecting these to multivariate analysis, and only then trying to explain the emerging similarities and differences within a sociolinguistic framework. Varieties would then appear in a fairly late, interpretative stage of the investigation. By so doing, we forgo the benefit of having sociolinguistic varieties directly as input variables. Therefore, provided the material permits us to do so, from the outset it appears preferable to conceive a providentially fertile sociolinguistic stratification of the text, and to scrutinize this as a statistical variable.

When attempting to do this for the synoptics, we meet several problems. There appear to be at least three constraints to be borne in mind. Firstly, there is the huge handicap of the non-linguistic correlates (region, class, type of situation, etc.) of ancient texts/writers no longer being accessible for direct observation, as they would be in modern language research. Therefore, we miss the mutual clarification of linguistic and extra-linguistic factors. Secondly, as far as register is concerned, written texts are often less easy to pinpoint than are the spoken utterances of everyday life. Narrative in particular can pose major problems because it may report speech and thought from many different types of situation, as well as describe the surrounding events and setting. The gospels are a case in point. Thirdly, there is a general methodological requirement we have to face, i.e. that to enable counting, the register categorization has to be clear-cut and unambivalent. Diffuseness of categories, which would not necessarily hamper free argued discourse, ought to be avoided in quantitative research.

There may not be any unique and definitive register categorization for all purposes. One obvious solution, which we have decided not to follow entirely here, might be to partition the gospels into small rounded episodes of varying nature, such as miracle stories, controversy stories, parables, aphorisms, and so forth. This type of breaking up of the text is standard practice in traditional form-critical studies. Here we will adopt a slightly different way of taking the text apart, which has proved to be capable of casting the stylistic properties of sources and end-authors into relief remarkably well. The four categories to be distinguished below (cf. Morgenthaler, 1971) will be called discourse types, because this term is better suited than text type or genre to the fragmented character of the pieces. As a predecessor to the approach followed here, it might be possible to cite Radday and Shore (1985), who divide the text of Genesis into Narrator's, Human, and Divine speech.

4. Discourse Type as Control Variable

In my dissertation, I have proposed and elaborated the distinction between the following four discourse types:

1. Narrative framework	NAR
2. Dialogue	DIA
3. Aphorisms	APH
4. Parables	PAR

The narrative framework (Morgenthaler's 1971 'Begleittext') comprises all those parts which are not directly reported speech. Here the author/narrator speaks without quoting others. It seems superfluous to underscore that the distinction proposed here is of a more mechanical kind than the distinction between real author, implied author, and narrator, which has proved to be valuable in certain types of narrative analysis. The three remaining discourse types—dialogue, aphorisms, and parables—are located on a different level. Taken together, they cover the cited speech of participants in the story (among whom, for that matter, can be the author/narrator himself, especially in first-person narrative). They are embedded in the narrative framework, which fulfils different functions: narrating the sequential actions and events, describing their setting, inserting speech, and commenting. In the gospels, the last function turns up only incidentally, which contributes to the impression of impersonal narration.

The border transitions into and out of the framework involve some typical shifts in deixis, e.g. pertain-

ing to the use of first and second person (within the narrative framework referring to narrator and reader, elsewhere to persons in the story, speaking and being addressed), deictic adverbs ('here', 'there', 'now', and so forth), and verbal tense. Even the distribution of verbal mood and aspect often tends to change dramatically. Other conspicuous transitions could be cited, but for the moment this point may suffice for our purposes.

Among the three types of cited speech, dialogue is typically speech that cannot stand on its own feet and relies on the context to be really comprehensible. One example is Pilate's question 'Are you the king of the Jews?', and Jesus's answer 'The words are yours' (Mark 15:2). Dialogue is mostly interactive (with questions, replies, and repartees), and often consists of short bouts of speech. By contrast, aphorisms and parables—both universally acknowledged forms in the gospels—are self-contained literary forms, which, lifted out of their context, remain complete utterances (which is not to say that the context does not enrich their meaning). For aphorisms, see, for example, the many sayings in the Sermon on the Mount and, for parables, the parable of the Sower and the Seed. Of the two, aphorisms are non-narrative, argumentative/discursive sayings, in contrast to the parables, which are short narratives (destined to make a point about man in his relationship with God, its message occasionally made explicit at the end by a moral). Alternative labels for aphorisms are sayings, logia (from Greek), or meshalim (from Hebrew). Parables often incorporate some form of dialogue; we have, somewhat arbitrarily, assigned this parable-internal dialogue to the discourse-type parable, rather than to dialogue.

In the synoptic gospels as a whole, 39.4% of all words belongs to the narrative framework, 19.7% to dialogue, 29.5% to aphorisms, and 11.4% to parables.

5. Lukan Samples m and Q in Mealand (1995)

Having introduced discourse type as a control variable, we are able to review Mealand's case afresh. We select Mealand's figure 5 (reproduced here as Fig. 2) because it encapsulates his method and thesis in a nutshell. The criticism which will be voiced about it might, *mutatis mutandis*, be extended to other parts of Mealand's study.

Examination of Fig. 2 demonstrates that CA in the first two dimensions separates m-samples (Lukan fragments co-occurring at least in Mark) quite well from Q-samples (fragments co-occurring only in Matthew), based on five linguistic features: *kai* ('and'), *de* ('but'), *gar* ('for'), *autos* ('he/she/it'), and adjective. From this separation, Mealand infers that the two series of samples derive from two real-life sources, i.e. Mark and Q (which is, as we saw, the standard theory).

The separation, however, is not perfect, the most distinct exception being the sample mu (Luke 21:1–34, Jesus's apocalyptic speech), which penetrates far into the Q-area. This observation induces Mealand to express doubts about the Markan origin of Luke's Chapter 21.

I believe that these conclusions do not stand up to careful scrutiny. As will be shown, instead the separation reflects discourse-type differentiation; it does not allow us to make assessments on sources. Also, Mealand's exceptions will prove to have a natural explanation independent of source theories as soon as discourse type is accounted for.

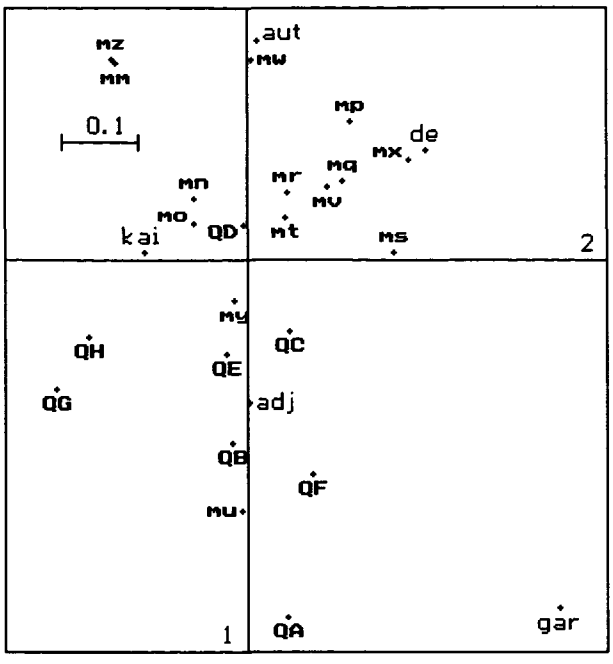

Fig. 2 Correspondence analysis of Lukan samples m and Q (from Mealand, 1995, Figure 5) Note. the first two dimensions represent 76.9% of the total information content. The second letter of the sample symbols is to identify the sample.

Before carrying on, I would like to observe that Mealand's use of discriminant analysis (Stepdisc, Candisc, and Discrim) raises some questions with me. Though nothing seems wrong in principle with using discriminant analysis to reduce the number of variables after an exploratory examination by CA of the full data set, there seems to me a risk connected with it, if the different effects at work in it have not been sorted out fairly satisfactorily. Since discriminant analysis selects uncorrelated variables from prior *known groups* (m, Q, etc.), the basic grouping should be grounded *firmly* and *independently* for the few discriminators to become reliable signposts, otherwise there is the implication of a certain measure of circularity. This objection grows even weightier whenever features with low frequencies are selected, as for instance is true for *gar* ('for'). What is gained in economy through discriminant analysis, seems to me to be lost in unobtrusiveness and scope.

6. Discourse Type Distribution in Luke

At this point, we have to examine whether it is true that discourse type is a predominant factor in the occurrence of linguistic items. Table 1 presents the distribution of the five items on which Fig. 2 was based.[1]

The dependencies, already showing through in the row percentages of Table 1, become fully visible when the data are standardized as in Table 2. The measures of association used are the adjusted standardized χ^2

Table 1 Discourse types and linguistic features in the Lukan samples m and Q (with row percentages)

	NAR	PAR	APH	DIA
kai	388 (46.8)	57 (6 9)	270 (32.6)	114 (13.8)
de	246 (73.4)	13 (3.9)	59 (17.6)	17 (5.1)
gar	9 (14.3)	1 (1.6)	39 (61.9)	14 (22 2)
autos	433 (71.7)	23 (3.8)	102 (16.9)	46 (7.6)
adjective	335 (30 8)	79 (7.3)	439 (40 4)	234 (21.5)

Table 2 Standardized residuals of Table 1

	NAR	PAR	APH	DIA
kai	-1.1	1.4	1 0	-.8
de	9.8	-1.7	-5.7	-5.2
gar	-5.5	-1.5	5.3	1 7
autos	12 9	-2.5	-8.5	-5.4
adjective	-14 6	2.4	8.3	8 2

Table 3 Discourse type distribution in Mealand's samples m and Q (standardized residuals)

	NAR	PAR	APH	DIA
mm	8.7	-3.1	-8.4	.8
mn	1.5	-3 1	-1 8	2.2
mo	-2	13.7	-3 8	-3.9
mp	10 1	-2 9	-7 8	-2.0
mq	4.4	-3.2	-3.7	2.9
mr	3 5	-3 1	-3 2	1 3
ms	2.0	-3.0	-2.7	2 6
mt	-2.6	13.4	-7.3	4.3
mu	-7.2	-2.8	9.2	1
mv	2.1	-2.5	-1.6	.8
mw	6.0	-3.0	-8.0	4.0
mx	5.3	-3 2	-6.5	3.1
my	7 9	-3 0	-8 2	1.6
mz	3.7	-3.3	-8.8	8.5
QA	-9 6	-2 4	15 2	-4.7
QB	-1.8	3.2	-1.2	1.9
QC	-3.8	-2.8	8.4	-3.7
QD	-5.1	-2.6	9.7	-3.8
QE	-5.8	-2 7	10 6	-3.9
QF	-9.8	-2 7	14.7	-3.7
QG	-4 7	3.7	6.8	-4.7
QH	-9.3	15 7	5.5	-4 5

residuals (formula $f_{ij} - e_{ij} \div \sqrt{e_{ij}}$, where f stands for observed frequency and e for expected frequency).

Tables 1 and 2 show unmistakably that the distribution of the five linguistic items over the four discourse types is anything but homogeneous. *Autos* and *de* are typical of the narrative framework, compared with *gar* and the adjective, and to a lesser extent *kai*, which appear more frequently in aphorisms and dialogue. These tendencies are also demonstrated when mapped through CA; see Fig. 3.

Given these correlations, in the event of discourse types being unevenly spread over Mealand's samples, these samples will be inherently biased towards specific linguistic features. Now Table 3 demonstrates that the samples m and Q do indeed present different mixtures of discourse types.

Table 3 demonstrates that, on the whole, m and Q have distinct discourse-type profiles. Especially important are NAR and APH, because they are the largest. In general, m-samples have high proportions of both NAR and DIA, but low proportions of APH. Conversely, Q-samples show low proportions of NAR and DIA, and high proportions of APH. Turning to PAR, the picture is mixed: both in m and Q we generally meet low proportions, but individual high peaks occur in mo, mt, and QH, and smaller peaks in QB and QG. Adherents of the standard theory will not be surprised. These acknowledge that Q—which they regard as a genuine source—is mainly a collection of Jesus's sayings, without much narrative framework, whereas, in the Gospel according to Mark, which is at the origin of m, events are being narrated and therefore the narrative framework is substantive.

Having thus established a close association between, first, discourse types and linguistic features, and, second, discourse types and m-/Q-samples, we are able to predict that the samples m and Q favour different linguistic features. It is the latter association which now has to be recognized as the real explanation for the separation of m-samples and Q-samples in Fig. 2. Likewise, the anomalies of the figure are explained. The Q-like linguistic behaviour of mu (Luke 21:1–34) is caused by it being the only m-sample consisting chiefly of aphorisms (s.r. +9.2) and only parsimoniously exhibiting narrative framework (s.r. -7.2). Interestingly, its parallel in Mark, Chapter 13, happens to contain Jesus's longest uninterrupted speech in this gospel (the 'apocalyptic speech'), and thereby stands out from the chain of narrated events. The conclusion seems inescapable that Mealand's putting forward the suggestion (following others) that 'either Luke has

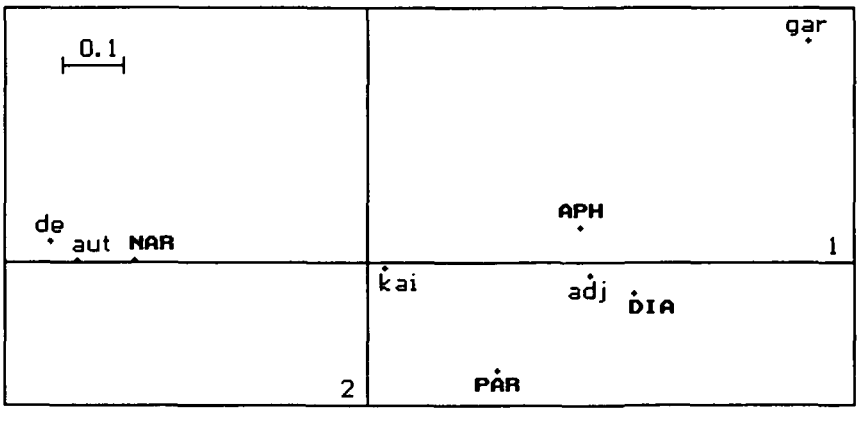

Fig. 3 Correspondence analysis of *kai*, *de*, *gar*, *autos*, and adjective in discourse types in Luke. Note: the first two dimensions represent 99.2% of the total information content.

edited his source rather freely here, or he is using another source, or perhaps both explanations are true', in combination with Mealand's own guess that 'there could be a genre factor here which could be investigated on another occasion' (Mealand, 1995, p. 177), does not help us much further. Actually, idiolect, source, and genre are all cited here as possible explanations—while, if what is said above is true, genre suffices to evaluate the pecularities of Luke 21.

The conclusion will be that, although nothing of what we have said so far about Fig. 2 effectively contradicts the standard source theory, it may be adduced as neither direct evidence in favour of this theory nor, for that matter, of any other source theory.

7. Widening the Scope: Discourse Type Distribution in the Synoptics

We proceed to show that the association phenomenon we have met is not restricted to Luke alone. In my book, I have presented evidence demonstrating the prevailing influence of discourse type in each of the three synoptics at various linguistic levels. The first level at which we look involves the twenty-three commonest words of the synoptics (Fig. 4).[2] A comparison is made of the twelve categories of the compound variable gospel (3) × discourse type (4).

The behaviour of the words across the twelve sections, as displayed by Fig. 4, is extremely interesting. The words principally cluster not according to gospel, but to discourse type. Hence we may state that discourse type has a more powerful effect on word use than authorial preferences. At least this is true for frequently used words (the twenty-three word types, by the way, being responsible for no less than 48% of all word-tokens in the synoptics). Less frequent words, i.e. generally those endowed with more information content, cannot of necessity be presumed to behave in the same way. These might be connected more with subject matter rather than with discourse type.

As, for the moment, we are especially interested in discourse types, I have sought to present Fig. 4 in a simplified form, by eliminating the effects attributable to authorial preferences. This is done in Fig. 5.

In the first and most important dimension (x-axis) of Fig. 5, NAR stands opposite APH and DIA, while PAR, not unexpectedly, is positioned in between. It is only along the second dimension (y-axis) that the difference between APH and DIA emerges clearly, the personal pronoun *humeis* (plural 'you') being characteristic for APH, whereas *egô* ('I') and *su* (singular 'you') are frequent in DIA.

When the investigation is extended to other linguistic levels, the very same pattern re-emerges over and over again, which ensures its basic stability. This can be verified when we adopt some major grammatical categories as our linguistic variable (Fig. 6), or, turning to an area of syntax, subordinate clauses (Fig. 7). The former is based on a verb-oriented division of all extant words of the synoptics into five categories: independent verbs, dependent finite verbs, participles, infinitives, and words not being a verb.

Figure 7 is based on a subdivision of dependent verbs into twenty types of subordination, including both finite (1–10) and non-finite (11–20) clauses. It is presented in Table 4. The division is in accordance with traditional grammar (e.g. Blass–Debrunner), and does not involve newer discourse-functional rationales, which will occupy us further on. This fairly traditional approach at this stage seems justified to me, as it secures a widely understood and approved basis. The encoding of individual places (verb forms) was performed manually by the author himself. In such an undertaking, there are always going to be some cases which are open to dispute. In my dissertation, I have expounded my choices extensively; on an accompanying disk all assignments are listed.[3]

On the basis of the evidence presented in the current section, it is possible to draw the conclusion that, as

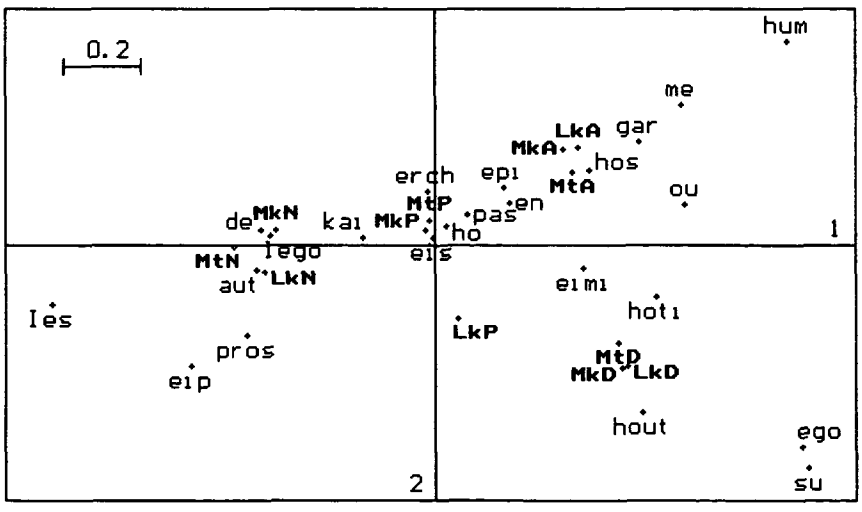

Fig. 4 Correspondence analysis of the twenty-three most common words in the synoptics split according to discourse type. Note the first two dimensions represent 81.9% of the total information content. Key to symbols: N[AR], P[AR], A[PH], D[IA]. *aut[os]* ('he/she/it'), *de* ('but'), *egô* ('I'), *eimi* ('be'), *eip[on]* ('say' aorist), *eis* ('into', 'toward'), *en* ('in'), *epi* ('on'), *erch[omai]* ('come'), *gar* ('for'), *ho* ('the'), *hos* ('who' rel. pron.), *hoti* ('that' conj.), *hout[os]* ('this' demonstr. pron.), *hum[eis]* ('you' plural), *Iês[ous]* ('Jesus'), *kai* ('and'), *legô* ('say' present), *mê* ('not'), *ou* ('not'), *pas* ('every'), *pros* ('towards'), *su* ('you' singular)

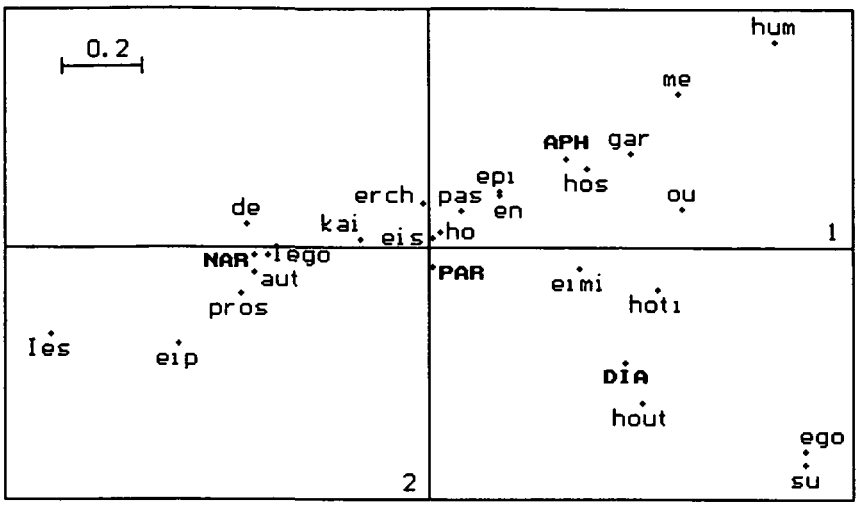

Fig. 5 Correspondence analysis of the twenty-three most common words and discourse types in the synoptics, gospel effects eliminated. Note: the first two dimensions represent 97.3% of the total information content.

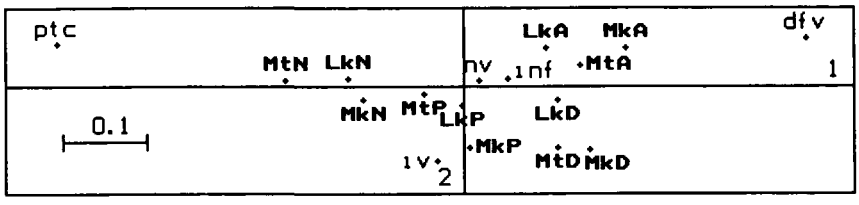

Fig. 6 Correspondence analysis of five major grammatical categories in the synoptics split according to discourse type. Note. the first two dimensions represent 94.7% of the total information content Key to symbols iv = independent verbs, dfv = dependent finite verbs, ptc = participles, inf = infinitives, nv = words not being a verb.

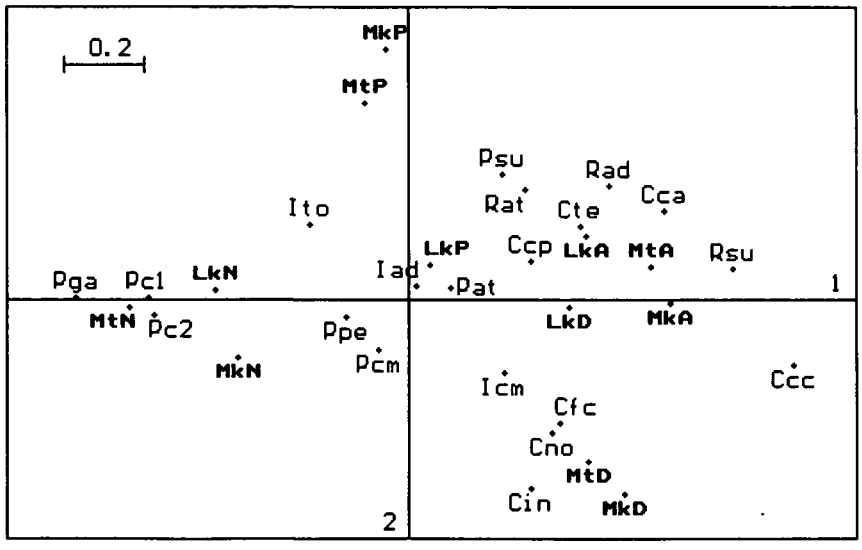

Fig. 7 Correspondence analysis of subordinate clauses in the synoptics split according to discourse type (for key to symbols, see Table 4). Note: the first two dimensions represent 78.1% of the total information content.

far as the influence on language use is concerned, discourse type takes precedence over authorial preferences. Ignoring the discourse-type factor in stylometrics will almost invariably vitiate the outcomes. In addition to this general conclusion, the pictures can, of course, help us provide quick, yet detailed, insights into the discourse-type preference patterns for individual linguistic features. We will make use of this in later sections.

8. Methodological Interlude: How to Eliminate Side-Effects

Introducing, as suggested earlier on, extra variables to cope with side effects raises problems for CA, because this technique accepts no more than two variables for input, i.e. the rows and columns of a two-way contingency table. How, then, can we deal with discourse types, in view of the fact that the text-delimiting variable and the linguistic variable already occupy the

Table 4 Types of subordinate (dependent) clauses

1	Nominal clause (*hoti, hina*, etc.)	Cno
2.	Indirect interrogative clause	Cin
3.	Temporal clause	Cte
4.	Conditional/concessive clause	Ccc
5.	Causal clause	Cca
6.	Final/consecutive clause	Cfc
7.	Comparative clause	Ccp
8.	Attributive relative clause	Rat
9.	Substantival relative clause	Rsu
10.	Adverbial relative clause	Rad
11	Attributive participle	Pat
12.	Substantival participle	Psu
13.	Periphrastic participle	Ppe
14.	Complementary participle	Pcm
15.	Preposed conjunctive participle	Pc1
16	Postposed conjunctive participle	Pc2
17.	Absolute participle (genitive absolute)	Pga
18.	Complementary infinitive	Icm
19.	Adverbial infinitive	Iad
20.	Infinitive with definite article *to*	Ito

rows and columns? Essentially, we want to use the information contained in a three- or higher dimensional dataset (i.e. a multiway contingency table with observed frequencies) in its entirety, and yet in the end somehow compress this full information into a flat, two-way contingency table, usable for CA. It should be clear from the outset that this target cannot be achieved simply by adding up the category cells of surplus variables, for this would in fact amount to using a two-dimensional dataset from the start, the third variable and onwards being erased.

Van der Heijden and De Leeuw (1985; see also Van der Heijden *et al.*, 1989), inspired by Escofier, have shown that remodelling three-way or higher-way tables by means of loglinear analysis (LLA) provides the solution to our problem. LLA was developed by Bishop *et al.* (1975), and is now a standard technique in statistical software packages. Lack of space compels us to keep the explanation brief, the reader seeking more detail should consult the references.

Given a contingency table, whatever its dimension, LLA lays bare the interaction effects between the variables represented in the table. Interaction between variables emerges because their categories are correlated—or associated—to a greater or lesser degree. So, for instance, LLA confirms the conclusion we had reached earlier along different lines, namely that the association between discourse types and linguistic features is stronger than that between texts and linguistic features. To do this, LLA decomposes observed frequencies into a set of interaction effects. Since each effect is quantified, the contribution of every variable interaction to every single observed frequency is determined exactly.

Once the decomposition has been made, we can deliberately remodel the original table by eliminating one or more of the interaction effects. In this way several new, restricted models can be deduced from the original table, each new model giving, as it were, a different X-ray of the data. The new model for the moment retains the original number of dimensions, but the observed frequencies are replaced by expected frequencies, conditioned by our choice of the interactions to be retained. It reveals what the data would have been like had the influence of one or more of the interaction effects been zero.

The next step is to cut back the full-dimensional model into a two-way table by collapsing (deleting one or more variables by adding up over them), or by compounding two variables into one. With regard to the resulting map, we may state that, though it excludes one or more interactions, it still provides information to which the entire dataset has contributed, not merely two variables. Successively filtering out different interaction effects enables us to extract various partial representations of the dataset, each in its own way shedding light on the data and lending itself to inspection by CA. By generalizing CA in this way, we have solved our problem.

Without noticing it, we have already applied this procedure in Fig. 5, where we had an underlying three-dimensional dataset of gospels (3) × discourse types (4) × words (23), i.e. 276 frequency cells in all. Figure 5 mapped the variables discourse types and words, after the interaction effects between gospels and words, and between gospels and discourse types had been discarded (i.e. the gospel effects had been eliminated). The result was a focus on the correlation between discourse types and words. In the next sections, we make further use of this combination of LLA and CA.[4]

9. Once Again, Lukan Samples m and Q

Figure 8 epitomizes our discussion so far in that it recaptures Fig. 2 (Mealand's CA of samples m and Q) in a new fashion, namely, removing the discourse-type effects with the help of LLA.

The map convincingly reveals that, after the linguistic preferences of discourse type have been put at zero, the m-cloud and Q-cloud, which were clearly separated

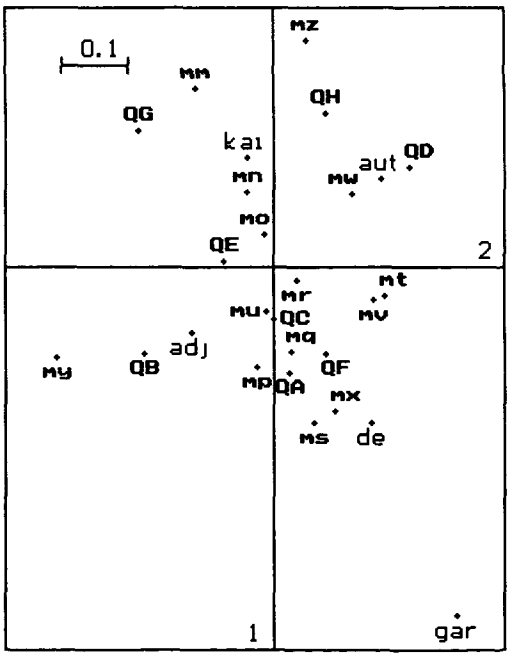

Fig. 8 Correspondence analysis of Lukan samples m and Q (from Mealand, 1995, Figure 5), discourse-type effects eliminated (cf. Fig. 2). Note. the first two dimensions represent 65 9% of the total information content

in Fig. 2, are no longer distinct. Both the samples m and Q are represented in each of the four quadrants. The plot confirms that Mealand's source-critical conclusions are no longer warranted and that the separation of the clouds in Fig. 2 is due to discourse type above all.

Accordingly, with regard to the single points that were still deviant in Fig. 2, new facts are now brought to light that must alter our view. Mealand, as we have seen, has argued for a non-Markan origin of the sample mu by virtue of its resemblance to Q-samples. In Fig. 8, however, mu is no longer accorded a special position. Its affinity with Q has vanished and it now invokes a fairly 'average' image, as it stays close to the centre of the map.

10. An Alternative Way of Tracing Synoptic Sources with CA

As hitherto we have been unable either to prove or to disprove the standard source theory, or any other viable source theory, in this section we shall proceed to see whether using data from all three synoptic gospels could advance the case.

The figure to be considered, Fig. 9, is based on a four-dimensional dataset, made up from gospels (3) × parallel schemes (5) × discourse types (4) × subordinate-clause types (20). From this dataset, we have extracted a restricted LLA model by dropping both discourse-type effects and gospel effects, thereby concentrating on the association between parallel sections and language. In this way, the possibility is excluded

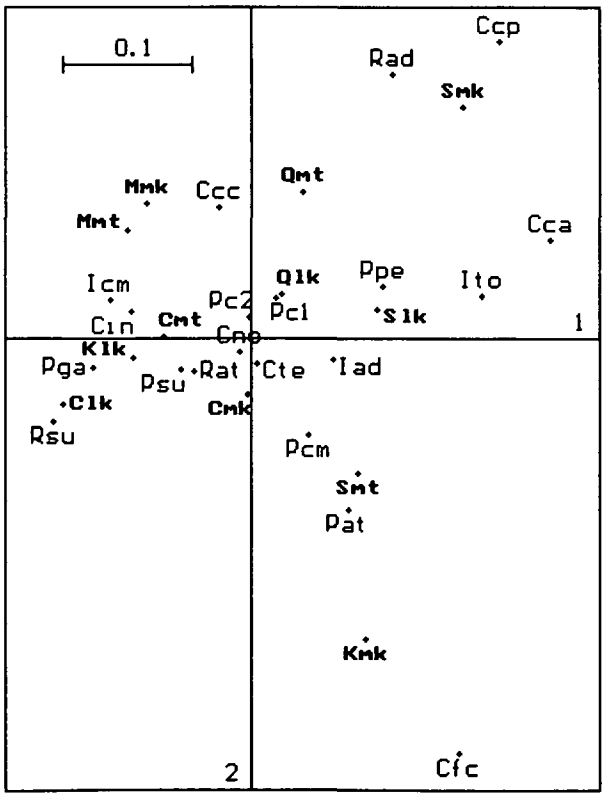

Fig. 9 Correspondence analysis of subordinate clauses in parallel sections of the synoptics; discourse-type effects and gospel effects eliminated (for key to symbols, see Tables 4 and 5) Note: the first two dimensions represent 45.8% of the total information content.

that sources could enter obliquely as a corollary of discourse type, i.e. because of a coincidence of parallel-section boundaries and discourse-type boundaries. After all, we want to avoid the risk of unjustifiably introducing source claims while in fact discourse type can take on the job of explaining on its own.

The CA analysis of the restricted model is based on two variables: the twelve parallel sections (compound variable of gospels×parallel schemes, see Table 5) and subordination.

We call parallel sections having the same parallel scheme (e.g. Cmk, Cmt, and Clk) counterparts. Together, counterparts form a parallel group (e.g. M is a parallel group, consisting of counterparts Mmk and Mmt). Note that, in spite of the common symbol S, Smk, Smt, and Slk are not counterparts; therefore too, S—in contrast to C, Q, M, and K—does not constitute a parallel group.

In Figure 9, the parallel groups C, Q, and M emerge as clearly identifiable. Kmk and Klk, however, do not lie in close proximity on the first two dimensions. However, closer scrutiny reveals that on the third dimension, not given here, in their case even the clustering is established unequivocally. This is one of those instances where, with regard to a specific point, essential information is held by higher dimensions of the CA space; such may be the case especially when, in the first two dimensions, only a relatively small percentage of the total variance (information content) is accounted for (in Fig. 9 this is 45.8%).

Making the step from parallel groups to genuine sources, however, proves to be fairly difficult. Three reasons may be adduced. First, the overriding importance of discourse-type and authorial preferences tends to blur the outlines of sources. Second, it remains an intriguing fact that C, M, K, and S do not really intermingle, nor as a group do they distance themselves neatly from Q (though, admittedly, on the first axis C, M, and K on the one hand, and Q on the other are on different sides of the origin); on the basis of the standard theory, a more articulate intermingling and distinctiveness was to be expected.[5] Third, and maybe most importantly, though CA may have helped underpin the parallel groups linguistically, it does not give away any clues about the direction of the historical development. For example, the association between Qmt and Qlk could indicate that (i) Qmt is a source of Qlk, (ii) Qlk is a source of Qmt, or (iii) both derive

Table 5 Parallel schemes and sections in the synoptics

	Parallel scheme	Parallel section	Mealand (1995)
Mark	Mk/Mt/Lk	Cmk	
	Mk/Mt	Mmk	
	Mk/Lk	Kmk	
	Mk	Smk	
Matthew	Mk/Mt/Lk	Cmt	
	Mk/Mt	Mmt	
	Mt/Lk	Qmt	
	Mt	Smt	
Luke	Mk/Mt/Lk	Clk	m
	Mk/Lk	Klk	m
	Mt/Lk	Qlk	Q
	Lk	Slk	I,L

from a lost source. As long as these choices remain open, no real progress has been made.

Therefore, the conclusion is that the whole case seems to hang in the balance. The evidence supplied does not contradict the standard hypothesis, but this is true for quite a few other theories as well. It remains, therefore, doubtful whether in the final analysis CA mapping is very helpful in the settling of source arguments. It is far more likely that we should rely on more direct methods, like, for example, arguments of order (cf. Honoré, 1968; Morgenthaler, 1971).

11. Gospel Preferences (Idiolects)

Turning away from source-historical questions, we shall now use generalized CA to focus on the styles of the gospels as they are. That is, we make CA describe the personal idiolects of the evangelists, looking beyond supra-personal discourse types, and leaving sources aside. The validity of this undertaking cannot be called into question, since LLA shows the interaction between gospels and language to be statistically significant, though to a lesser degree than the interaction between discourse types and language (Linmans, 1995, p. 97). This attempt to establish the linguistic profiles of the evangelists should in no way be seen as a dismissal of literary-historical approaches; on the contrary, we believe that literary history might benefit considerably from stylometrics.

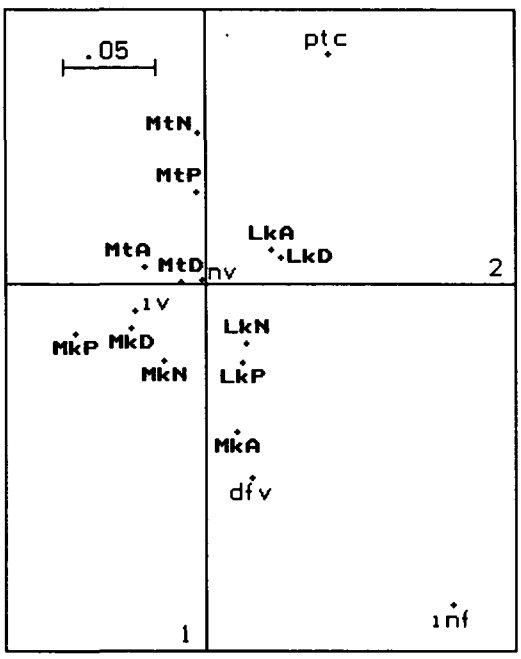

Fig. 10 Correspondence analysis of five major grammatical categories in the synoptics; discourse-type effects eliminated (for key to symbols, see Fig. 6). Note: the first two dimensions represent 82.1% of the total information content.

Within the New Testament, the gospels form a remarkably homogeneous group, due to agreements in style, discourse type, source relationship, subject matter, and possibly more (cf. Linmans, 1995, p. 77). Nevertheless, their having different identities and interests is unmistakable, both stylistically and in their way of viewing the subject matter. As the gospels betray scarcely a hint of the biography of their writers, and even independent attestation is sparse and contested, every scrap of information about their individuality on the basis of multivariate analysis should be welcomed. The study of their style, alongside that of their theology and world view, seems to be the main avenue to discovering the individuality of the evangelists.

At different levels, Figs 10, 11, and 12 offer us stylistic profiles of the synoptics. Naturally, discourse-type effects have been removed. Not without some hesitation, I decided not to remove parallel-section effects as well. Had I done otherwise, I would have expunged, among other characteristics, the linguistic preferences of the single sections Smk, Smt, and Slk as clues to gospel styles, which seemed to be a retrograde step.

I want to draw attention to the fact that in Figs 10 and 12 the clustering of the parallel sections is predominantly gospel-oriented. For example, in Fig. 10, Mark is concentrated mainly in the lower left quadrant, Matthew in the upper left quadrant, and Luke in the two right quadrants. This indicates that, as far as the lexemes (Fig. 12) and the major grammatical categories (Fig. 10) are concerned, each of the gospels shows coherent stylistic preferences across the discourse types. Turning to the subordination types (Fig. 11), the picture is different. Here the gospels are much more

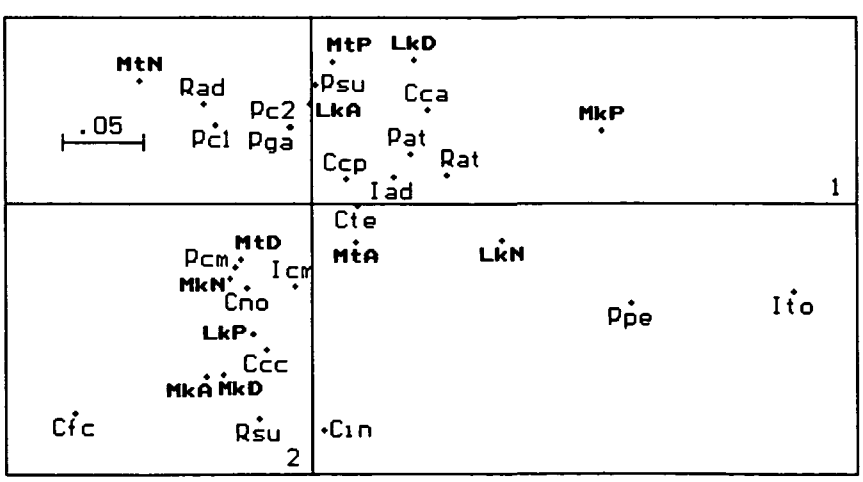

Fig. 11 Correspondence analysis of subordinate clauses in the synoptics, discourse-type effects eliminated (for key to symbols, see Table 4) Note: the first two dimensions represent 55.2% of the total information content.

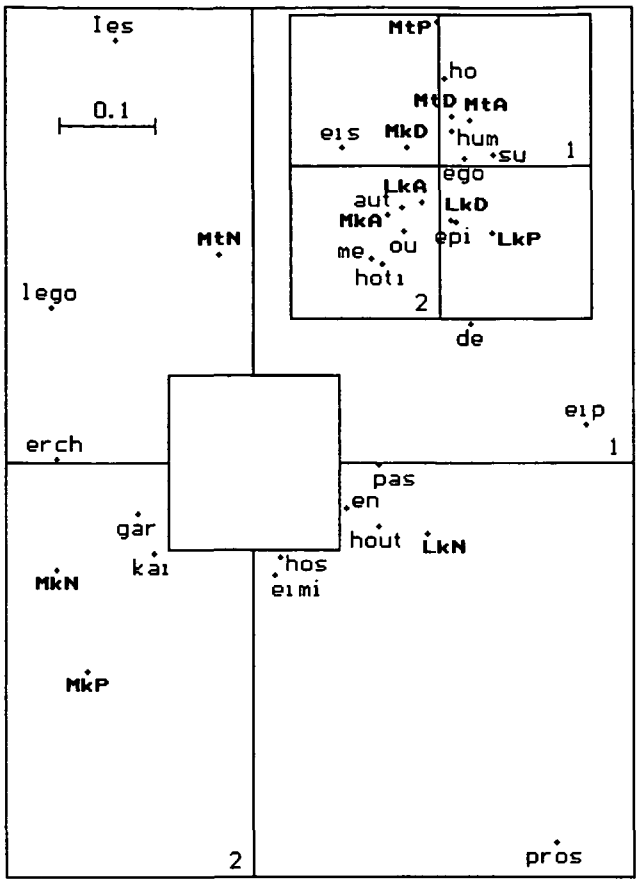

Fig. 12 Correspondence analysis of the twenty-three commonest words in the synoptics; discourse-type effects eliminated (for key to symbols, see Fig. 4) Note: the first two dimensions represent 72.6% of the total information content.

scattered in an unsystematic way, demonstrating that in this particular linguistic area the preferences of each evangelist tend to diverge depending on discourse type. Borrowing a term from the natural sciences, one could say that the entropy is greater here. A further symptom of this is the relatively low information content (55.2%) of Fig. 11.

12. Multivariate Analysis and Linguistic Analysis

Thus far, language in its own right has not been much of our concern. Linguistic features have above all been considered to be helpful in characterizing sources, discourse types, and authors, and no special efforts have been made to analyse the characteristics more throughly in linguistic terms. Yet it is beyond doubt that linguistic and multivariate analysis can enrich each other considerably. Since it operates in a multi-dimensional space of sociolinguistic variation, language performance is naturally suitable to multivariate analysis—not merely through literary stylometrics and register studies, but also, for instance, through typological (cross-language) studies. Alternatively, a theoretical linguistic framework is indispensable in order to establish proper linguistic variables for quantification as well as to interpret the output data in a sensible way.

Gradually I have become convinced that pre-eminence has to be claimed for recent discourse-functional approaches that seek to find linguistic explanations above the level of the sentence (in the co-text) and that try to discover how the speech situation (con-text) affects the organization of the text. They devote much attention to how, with every step of the communicative process, the language producer makes assumptions about what is going on in the mind of the language receiver, and how the former moulds the discourse to direct the latter's comprehensions. Among the scholars who should be mentioned in this connection are Wallace Chafe, Talmy Givón, Knud Lambrecht, Sandrah Thompson (all USA), M. A. K. Halliday (UK), Simon C. Dik (The Netherlands), and Jan Firbas (Czech Republic). Lack of space prevents us from doing more than scratching the surface. Therefore, we shall only touch on a few linguistic facts of this kind brought to light by the CA maps.

To begin with, one interesting field is that of the discourse aspects of clause combining and subordination. It has been shown, e.g. by Sandrah Thompson, that subordination is not fully understood when viewed from an innersentential standpoint alone. Subordination also grammaticalizes the information flow over larger stretches of text, for instance by creating an alternation of background (less important) versus foreground (more important). In this, adverbial clauses especially—both finite and non-finite—often differ as to their discourse function depending on whether they stand at the beginning or the end of a sentence. Quantitatively, examination of our CA figures for discourse-type preferences shows that subordinate finite clauses with conjunction are favoured in aphorisms and dialogue (see Fig. 6, dfv), of the two types aphorisms having a liking for several kinds of adverbial clauses (Fig. 7, Cte, Cca, Ccp), as opposed to dialogue which has more complementary clauses (Fig. 7, Cno, Cin; cf. non-finite Icm). Both are rich in conditional/concessive clauses (Fig. 7, Ccc). In contrast, the narrative framework abounds in adverbial participles (Fig. 7, Pc1, Pc2, Pga). When these precede the predicate, they are typically aorist forms with the same subject as the predicate, not seldom—though semantically quite empty ('having gone', 'having seen', 'answering', etc.)—endowed with important paragraph-marking functions (e.g. recapitulating, setting frames, giving temporal or spatial orientation). Conversely, when following the predicate, adverbial participles are mainly present forms, often used to introduce direct speech ('saying'). As far as relative clauses—as well as substantive participles—are concerned, they are common in aphorisms, but also in parables (Fig. 6, Rat, Rsu, Psu; Figs 4 and 5, *hos*). Regarding the latter discourse type, one can point to parable introductions like 'The kingdom of Heaven is like a mustard-seed, which a man took and sowed in his field' (Matthew 13:31). Finally, in dialogue and parables, a precedence of independent over dependent verbs, i.e. less reliance on subordination, can be observed (Fig. 6, iv).

Taking a look in turn at gospel preferences with regard to subordination, it can be seen that Matthew NAR is even fonder of preposed adverbial participles than his counterparts, particularly Luke NAR (Fig. 11, Pc1). He appears to exploit the text-organizing capabilities of this construction to the utmost, albeit in a

somewhat stereotypical fashion. Inspired by the Greek Septuaginta translation of the Hebrew Bible, Luke NAR, for his part, displays a liking for (mainly adverbial) infinitives with the definite article, especially temporal *en tôi* + inf. (Fig. 11, Ito; Fig. 12, *en*). Mark has a high proportion of independent verbs and shuns participles (Fig. 10, iv, ptc), thereby betraying himself perhaps as the least proficient of the three in Greek.

Another important field of discourse analysis is sentence-linking devices used to establish cohesion in narrative, e.g. *kai* ('and'), *de* ('but', 'however'), *tote* ('then'), *gar* ('for'), and *oun* ('so, therefore'). In the synoptics, *kai* and *de* are abundant in the narrative framework, and *gar* in aphorisms (Fig. 5). Recently, Bakker (1993; see also Levinsohn, 1987) has propounded a discourse analysis of the particle *de* in ancient Greek writers which can be transferred fairly easily to the gospel narratives. The specific capabilities of *de* are grounded on its being linked up with the first constituent of the sentence (mainly not the predicate). One type of use is more local, i.e. when *de* is used to signal subject-shift (by some scholars called topic-shift), usually implying some degree of contrastiveness. Sometimes *de* marks local, temporally unsequenced background information, in the same way as many narrative *gar* clauses do. We meet a different type when, over a more extended range, *de* has paragraph-marking force. This emerges most clearly when *de* is joined to preposed adverbial clauses or participles, to install a demarcation. *De*, thus used, is a 'higher' narrative boundary than *kai*; the latter, under the scope of *de*, primarily indicates narrative continuity ('and then').

Without doubt, it would be worthwhile to investigate the individual connecting strategies of the synoptic writers in their use of *de* and *kai*, as well as other connectives in depth. By so doing, the high proportion of *kai* compared with *de* in Mark NAR and PAR, in comparison with Matthew and Luke (Fig. 12), could find new interpretations. Hopefully, the traditionally sweeping verdicts of Semitic interference, or even bad style in the latter's case, will be superceded by such careful analyses as those of Bakker and Levinsohn.

Deixis ('pointing') and the narrative representation of referents are other promising fields of research. Not surprisingly, CA (Fig. 5) indicates several deictic devices to be characteristic of dialogue, i.e. *egô* 'I', *su* 'you' singular, *humeis* 'you' plural, *houtos* 'this', while *humeis* is also prolific in aphorisms, where Jesus mostly addresses audiences of more than one person. Figure 5, like Fig. 3, also demonstrates the NAR-relatedness of *autos* 'he/she/it'. Against the background of the prolificness of *houtos* in dialogue, it is interesting to note Luke's personal preference for it in NAR (Fig. 12), where it serves co-textual anaphora. On the whole, the lack of deictic devices in NAR—apart from anaphora—helps convey a sense of objectivity. In a full study of deixis of course, the person suffixes of verbs should also be considered.

As far as third-person referents are concerned, it is important to note that, especially in narrative, the choice between full noun phrases, pronouns (*autos*, *houtos*, *ho*, *de*), and verbal suffixes is tied up with the speaker/writer's assumptions about the activation state (new/given) of the referent in the mind of the listener/reader (cf. Chafe, 1994; Lambrecht, 1994). To judge by his fairly frequent use of nouns, Matthew appears to have a tendency to explicitness and redundancy (cf. Fig. 12, *Iêsous*, definite article *ho*; see also Kenny, 1986, p. 59). A special topic is the first-time introduction of participants, when for the language receiver they have the status 'non-identifiable'; this is signalled by the absence of the definite article. The grammatical and lexical devices to introduce new referents and anchor them to known items is an interesting field of study; e.g. *tis* 'somebody', *heis* 'one', (*kai*) *idou* '(and) see', presentative *eimi* 'be', *hôi onoma* 'whose name'.

The areas just mentioned are only a few examples. I shall conclude this section by naming four other topics suggested by the CA maps. The first of these is tense/aspect. Matthew has significantly more present-stem forms than the others (exemplified in Fig. 12 by present-stem *legô* versus aorist *eipon* 'say'; more generally, cf. Kenny, 1986, p. 70). Part of Luke's imitation of the Greek Septuaginta is his rather un-Greek use in NAR of the periphrastic conjugation (*eimi* 'be' + participle, especially of the present) (see Fig. 11, Ppe, and Fig. 12, *eimi*, cf. recently Verboomen (1992)). Recent work on aspect in the New Testament by S. E. Porter and B. M. Fanning should be resumed from a consistently discourse-functional viewpoint. The second point is modality. This is represented in our CA figures by grammatical negation only, which is characteristic of aphorisms and dialogue (Fig. 5, *ou*, *mê*; on negation as modality, see Givón, 1984, pp. 321–51). In the third place, there is *eimi* 'be' and the adjective. They are shown to be characteristic of aphorisms and dialogue (Figs 3 and 5)—not unexpectedly, because nominal clauses are most common in them. The fourth point concerns the introduction of direct discourse. The formulaic character of these introductions is the main reason that *legô*, *eipon*, and *pros* (a preposition often used in combination with *eipon*, especially by Luke) are typical of the narrative framework, and that *hoti*, owing to the frequency of *hoti* immediately preceding direct speech, is frequent in aphorisms and dialogue (Fig. 5).

These suggestions may be enough to give an idea of how quantitative and linguistic analysis might be united. More details are to be found in my dissertation.

13. Conclusion

The overall thrust of the present article does not diverge greatly from Mealand's in that both appreciate the usefulness of CA. My contribution has, hopefully, shown that, in its generalized form especially, CA is a powerful instrument for visualizing the linguistic usages and preferences of both discourse types and authors. In one form or another, discourse types reflecting sociolinguistic variation appear pivotal. Their neglect by Mealand, as I have pointed out, has seriously flawed his conclusions. As has been argued, even if discourse types are accounted for, CA seems to be rather less effective in matters of literary contact and source relatedness.

It would be useful to amplify multivariate analysis of the synoptics by adding other contemporaneous writings, from both within and outside the New Testament (cf. Linmans, 1995, pp. 75–9 and 94–6). The isolated treatment of the synoptic gospels seems only fully justified when we deal with source problems narrowly. Moreover, future multivariate research might be given a considerable boost if scholars had at their disposal a standard electronic text of the New Testament with each word already coded for lemma, part of speech, categories of grammar (including syntax), discourse type, and—in the event of the synoptics—parallel schemes, each record also being easily extendable by codings of the researcher's own devising, the whole enhanced by counting procedures. Should this not be so, the creation of datasets will remain cumbersome, especially in syntax. Extant tools (grammatical analyses of the New Testament, such as GRAMCORD and Friberg, and the Thesaurus Linguae Graecae), though immensely valuable, do not yet fulfil all exigencies mentioned.

Sociolinguistic variation is certainly an extremely engaging and worthwhile field of research, for New Testament scholars no less than for others. Alternative discourse-type stratifications, differing from that presented here, should be tried out. Real progress will very much depend on bold linguistic reasoning, especially in view of discourse-functional aspects. Quantitative, particularly multivariate analysis can play its role as a guide to tell us what phenomena it might be worthwhile concentrating on, and as an expert help, to confer a firm grip on the quantitative properties of language performance in all of its dimensions, on a personal as well as a supra-personal level.

Notes

1. There are some minor counting differences between the present article and Mealand (1995), but their impact is negligible.
2. Here and hereafter a normalization option is used that treats rows and columns symmetrically, unlike Linmans (1995).
3. If requested, a copy of the disk will be sent (e-mail: Linmans@Rulub.LeidenUniv.nl).
4. For six of the CA maps (hierarchical) LLA was used. For those interested, I give the generating LLA formulas (cf. Van der Heijden and De Leeuw, 1985):

Fig. 5	[TD][TL]
Fig. 8, 10, 11, 12	[TD][DL]
Fig. 9	[TDP][TDL].

   Key to symbols. T = gospels/Mealand samples, D = discourse types, L = linguistic features, P = parallel schemes.
5. Further research may show whether in addition removing the effect [TPL]—which would require non-hierarchical LLA—will reveal new aspects. In accordance with Fig. 9, cluster analysis connects C, M, and K only on higher nodes of the dendrogram.

References

Bakker, E. J. (1993). Boundaries, topics, and the structure of discourse: an investigation of the Ancient Greek particle *dé*. *Studies in Language*, 17: 275–311.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

—— (1994). An analytical framework for register studies. In Biber, D. and Finegan, E. (eds), *Sociolinguistic Perspectives on Register*. Oxford University Press, Oxford, pp. 31–56.

—— and Finegan, E. (1994). Intra-textual variation within medical research articles. In Oostdijk, N. and De Haan, P. (eds), *Corpus-Based Research into Language. In Honour of Jan Aarts*. Rodopi, Amsterdam, pp. 201–21.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Chafe, W. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago University Press, Chicago, IL.

Givón, T. (1984). *Syntax: A Functional-Typological Introduction*. I. Benjamins, Amsterdam/Philadelphia.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Honoré, A. M. (1968). A statistical study of the synoptic problem. *Novum Testamentum*, 10: 95–147.

Kenny, A. (1986). *A Stylometric Study of the New Testament*. Clarendon, Oxford.

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge.

Levinsohn, S. H. (1987). *Textual Connections in Acts*. Scholars Press, SBL Mon. 31, Atlanta, GA.

Linmans, A. J. M. (1995). Onderschikking in de synoptische evangeliën: syntaxis, discourse-functies en stilometrie. Ph.D. thesis, University of Nijmegen, The Netherlands.

Mealand, D. L. (1995). Correspondence analysis of Luke. *Literary and Linguistic Computing*, 10: 171–82.

Morgenthaler, R. (1971). *Statistische Synopse*. Gotthelf, Zürich/Stuttgart.

Neumann, K. J. (1990). *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis*. Scholars Press, SBL. Diss 120, Atlanta, GA.

Radday, Y. T., Shore, H., *et al.* (1985). *Genesis: An Authorship Study in Computer-Assisted Statistical Linguistics*. Bibl. Instit. Press, Rome.

Van der Heijden, P. G. M. and De Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, 50: 429–47.

——, De Falguerolles, A., and De Leeuw, J. (1989). A combined approach to contingency table analysis and log-linear analysis. *Applied Statistics*, 38: 249–92.

Verboomen, A. (1992). L'imparfait périphrastique dans l'Évangile de Luc et dans la Septante. *Académie Royale de Belgique, Lettres X*, Peeters, Louvain.