Chapter 7

SHARED VOCABULARY AND STRUCTURES

BUILDING ON OBSERVATIONS made in passing by Naetebus (1891, 62, 75), the evidence which Morawski (1939, 1940) considers amongst the most significant in allowing the collective attribution of poems to Jehan is the presence of structures which are shared across the texts of the corpus. The most striking of these is the observation that *Povre chevalier* shares a full stanza with *Abesse*, one of the two texts in Morawski's corpus found outside of fr. 24432.

> Por ce que Nostre Dame scet paier largement,
> Veil ge metre m'entente a moustrer plainement
> Les miracles qu'el fait a tous communement
> Qui la veulent servir de cuer parfaitement
>
> (*Povre chevalier*, vv. 5–8)

> Pour ce que Nostre Dame scet paier largement,
> Weil je metre m'entente a moustrer plainnement
> Les miracles qu'el fait a tous communaument
> Qui la weilent servir de cuer parfaitement
>
> (*Abesse*, vv. 9–12)

ABSTRACT Continuing our examination of repeated words and structures, this chapter considers what can be gleaned about Jehan's written style from an analysis of line-initial words and the structures used to mark the onset of discourse. In the former case, I consider the most frequent line-initial words, demonstrating that while individual texts display variability in the frequency with which these appear, the *J-corpus*, when taken as a whole, is statistically distinct from the *A-*, *F-*, and *R-corpora*. Following this, I propose a new typology for discourse onsets in the *J-corpus*, and perform an analysis which reveals further hallmarks of the *J-corpus* in the way that Jehan introduces direct speech, while also allowing us to observe differences in usage between the "long-text" sub-group of Jehan's *œuvre* and his shorter texts.

DOI: 10.17302/ogkl548612 © Arc Humanities Press

Here we may also repeat the similarly comparable stanzas found in both *Jehan Paulus* and *Chanoinnes*, which we identified above:

> Si prions, touz et toutes, a la Vierge Marie,
> Qui en ces flans porta le disgne fruit de vie,
> Que tel oroyson soit de nous bouches parties,
> Qu'avoir puissions des anges la sainte compaygnie.
>
> (*Jehan Paulus*, vv. 367–70)

> Si prions, touz et toutes, a la Vierge Marie,
> Qui en ces flans porta le tres dous fruit de vie,
> Qu'il nous doint si ouvrer en ceste mortel vie,
> Par quoi aions des angres la sainte compaingnie.
>
> (*Chanoinnes*, vv. 409–12)

Morawski observes that many of the structures shared across Jehan's poems relate to shared elements of narrative,[1] but also that there is a particularly significant selection of shared structures related to the texts' religious contents. God, for example, is described as "le roi de maesté" in a total of three separate texts,[2] while Mary appears as the one "qui le cors Dieu porta" in no fewer than five texts in the corpus.[3]

However, while these observations certainly add to the sense of unity we have already observed across the corpus, Munk Olsen (1963b, 152) appears justified in his criticism of taking religious language as objective indicators of shared authorship since, by their very nature, devotional structures are likely to rely on clichés and commonplaces. As we have already argued in the case of *tous et toutes*, other structures can provide more convincing evidence of shared authorship, but this still relies on the assumption that the works in question were written by an author who had easy access to past output. Otherwise, as Munk Olsen suggests, we risk treating him like "un cerveau électronique dans lequel une impulsion de contenu donné déclenche automatiquement et infailliblement l'impulsion d'expression correspondante,"[4] in

1 For example, "Tost et isnelement entra en la chapelle" (*Povre chevalier*, v. 90; *Chevalier et escuier*, v. 107).

2 *Respon* (vv. 90, 176): "le roy de maesté"; *Pommes* (v. 306): "le roy de maisté"; *Abesse* (v. 131): "le roy de majesté."

3 *Respon* (v. 121), *Romme* (v. 213), *Deux chevaliers* (v. 113), *Chevalier et escuier* (v. 78): "qui le cors Dieu porta"; *Narbonne* (v. 100): "qui le corps Dieu porta." Morawski appears to miss the structure found in *Romme*.

4 "an electronic brain in which an impulse of given content automatically and unfailingly triggers the corresponding impulse of expression."

other words a machine which, in consistent circumstances, will be incapable of varying its output.

As we argued above, even the stanzas shared between *Povre chevalier* and *Abesse*, and between *Jehan Paulus* and *Chanoinnes*, while encouraging as part of a broader case for shared authorship, are not enough on their own to definitively prove the shared-author hypothesis, since it is also plausible that the author of one text or another borrowed these relatively formulaic sections that they had read elsewhere. Indeed, given that the stanzas in both pairs are grammatically and narratively discrete, the inclusion could even have been made by a later scribe familiar with both texts. All of this is not to say that these shared structures are meaningless to debates regarding authorship, only that we cannot see them, as Morawski arguably did, as a silver bullet which can be used to decisively prove shared authorship.

Moving beyond this primary focus on the simple observation of the same structures in more than one text, Kleist (1973, 44–45) builds on a statement made in passing by Morawski (1939, 332), that in cases of shared narrative content, "ce sont toujours les mêmes expressions, formules et rimes qui reviennent en pareille occurrence."[5] For Kleist, this is a particularly significant observation, that not only are there recurring structures within Jehan's texts, but that these recurrences appear within comparable narrative contexts. Examining questions of repetition from this perspective, Kleist argues, allows us to differentiate between conventional repetition and that which may genuinely be an indicator of shared authorship.

To illustrate his point, Kleist considers the case of a shared narrative segment in two texts attributed to Jehan which appear consecutively in fr. 24432, *Povre chevalier* and *Chevalier et escuier*. Both stories share three narrative phases. They tell of a character who, in a moment of desperation, sees a chapel and decides to go inside, in the former case a lady whose husband has promised her to the Devil, and in the latter case a knight who has fallen from grace (phase one). This character turns their horse towards the chapel (phase two), and then enters the chapel and goes towards the image of Mary (phase three). While far from the most original literary motif, it is certainly notable that such a narrative progression is shared between both poems, especially since they are adjacent within fr. 24432. Beyond the identical line already identified by Morawski and given above (p. 116n1), which appears during phase three, Kleist also notes that the stanzas containing the narrative phases two and three

5 "it is always the same expressions, formulas, and rhymes that return in such instances."

share a rhyme in both texts, and argues that, in this case "handelt es sich nicht bloß um zufällige Übereinstimmungen, sondern um ein Stilprinzip, das man als Technik der Versatzstücke bezeichnen könnte" (1973, 46).[6]

Some of Kleist's evidence may be described as tentative, particularly the shared rhymes in phases two and three which contain "torna" and "chapelle," both of which appear especially likely to appear in this narrative context. There is also the argument that narrative similarities such as these may have been the primary reason for the juxtaposition of both texts in fr. 24432, regardless of whether they share an author, rather than their proximity being an indicator of shared authorship. Further, we should note that these motifs are not unique to these two texts, and instead exist within established textual traditions. *Povre chevalier* corresponds to Tubach's (1969) types 5115, *Virgin, Blessed, comes to devil instead of his victim* and 5283, *Wife pledged to devil*, while *Chevalier et escuier* appears close to types 3570, *Pact with Devil V*, and 5133, *Virgin, Blessed, renunciation of refused*. All these text types are well attested, such that the narrative phases and their corresponding shared structures identified by Kleist cannot be said to be unique to these two texts.

There are thus reasons to be cautious about such evidence, yet the combination of these observations with other similarities, not to mention the texts' codicological context, certainly reinforces the possibility that there was an association between these texts that was acknowledged by contemporary readers, whether or not that means they were written by the same individual. Through the digital and statistical analysis presented in this chapter, we will examine similar ways in which the shared-author hypothesis can be strengthened, or not, by the presence of shared structures across the *J-corpus*. In so doing, we will identify further evidence of structural coherence displayed by its texts beyond Morawski's shared structures and Kleist's points of narrative correspondence, allowing us to further progress the shared-author hypothesis.

Line-Initial Words

Munk Olsen (1963b, 155) argues, counter to the suggestions of Morawski (1939, 1940) and Kleist (1973), that there is a danger to solely considering repetitions which contain specific, narrative-relevant meaning, since these are more likely to be required by the subject matter than because of authorial preference. Instead, Munk Olsen argues, we should consider the use of

6 "it is not just a matter of coincidental similarities, but of a stylistic principle that could be called the technique of set pieces."

more semantically "empty" words, such as *et* and *quand*, particularly those at the beginnings of lines of verse, to discover whether texts contain a consistent stylistic profile since, compared to content words which may be determined by narrative requirements, in the case of function words "il est plus probable qu'ils représentent des 'tics' chez l'auteur."[7]

In doing so, Munk Olsen is aligned with developments in the methodologies which have come to be known as "stylometry," pioneered by Mosteller and Wallace (1963, 275). Published in the same year that Munk Olsen released his first works on Jehan's *œuvre*, and considering a dataset far removed from the present corpora, Mosteller and Wallace's opening arguments are strikingly similar to Munk Olsen's:

> The filler words of the language such as *an*, *of*, and *upon*, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates [in the work of a single author], whereas more meaningful words like *war*, *executive*, and *legislature* do not.

Mosteller and Wallace go on to show through statistical methods that "the function words of the language appear to be a fertile source of discriminators" (1963, 306). That is, as Munk Olsen suggested, it is those semantically "empty" words which best enable discrimination between potential authors for a given text.

Munk Olsen applies his method to some effect in his comparison of Jehan's poems to the alexandrine *Alexis* included in our *A-corpus* (1963b, 159), using comparative counts of line-initial words to further his claim that the text cannot be considered part of Jehan's *œuvre*. However, the totals are not drawn from the full 1978 corpus of texts he assembled later, and which we have taken as the *J-corpus*, only the earlier corpus given by Morawski (1939). Further, perhaps for want of a means to automate what would otherwise be a drawn-out process of counting the frequency with which each word appears at the beginning of a line, Munk Olsen also only offers totals for a few words, and his conclusions for *Alexis* are therefore not driven by the whole data, and make no attempt to determine statistical significance. Munk Olsen apparently also largely limits his examination to function words which are in some ways more contentful,[8] thus going against his otherwise valid suggestion that it is the semantically "empty" words which are of greatest interest for an analysis of authorial "tics."

---

7 "it is more likely that they represent authorial 'tics.'"

8 *mais*, *car*, *lors*, *or*, *puis*, *adont*, *atant*, and *adoncques*, alongside the semantically "emptier" *et* and *quand*.

There is space, therefore, for a more data-driven approach to the same questions. And yet, as noted in our discussion on pp. 17–18, stylometry in its purest form falls outside of the scope of this study, the aim of which is to make use of explicitly bespoke methods, designed for a particular corpus of texts based on a humanistic understanding of those texts. With that in mind, an updated version of Munk Olsen's approach, drawing from stylometry but using our understanding of the structure of alexandrine verse to look only at function words which come at the beginning of a line, promises to be more fruitful.

The focus here, as in our work on line-final forms in Chapter 6, is on machine-tagged lemmata, rather than observed word forms, since this approach allows us to bypass any scribal or inflectional variation. In the case of these data, where function words such as *le* and *la* are common, it also enables us to distinguish between uses of these same word forms which can serve as both determiners and pronouns. Grammatical gender, number, and case distinctions between function words are all assumed here to be examples of inflectional rather than lexical variation, such that the determiners *li*, *le*, *la*, *les*, etc. are subsumed under the single lemma *le*, while the pronominal forms *il*, *elle*, *le*, *li*, etc. are all assigned to the single lemma *il*.

Looking at the six most common lemmata in each of the texts of the *J-corpus*, presented in Table 34, the most striking feature is their variability.[9] Several lemmata that appear, such as *le*, *que*, and *et*, are consistently high in frequency, but there is no established order across all of the texts. Most of these words are, unsurprisingly, function words, but there are a few appearances of more contentful words, most notably of *mais* as the third most common lemma at the beginnings of lines in *Florence*, and *seignor* in sixth place in *Abesse*.

We should of course not expect the frequencies of line-initial lemmata to be completely consistent across each text of the *J-corpus*, since while an author may have preferences, as we have already acknowledged, they cannot be treated as a machine who will always give an identical output based on a given set of inputs. That said, and on the understanding that an author's *œuvre* will never be truly homogeneous, it is possible to make some sense out of these apparently noisy data. By examining the corpus of texts as a whole, we extract the lemmata which appear most commonly at the beginnings of lines across the entire *J-corpus*, and we can then compare the rates of use of each of these words for individual texts, which will help to identify

9 The decision to look specifically at the six most frequent lemmata is largely arbitrary, but examination of the most frequent lemmata is likely to yield the most reliable data.

Table 34. The six most common lemmata at the beginning of a line for each of the texts in the *J-corpus*.

Text	1	2	3	4	5	6
Respon	le	il	de	que	a	et
Chanoinnes	le	et	a	que	quand	de
Pommes	le	que	quand	et	a	se
Romme	le	de	et	que	a	il
Deux chevaliers	le	que	de	et	a	lors
Enfant rosti	le	et	a	lors	son	de
Povre chevalier	que	le	de	a	car	en
Chevalier et escuier	et	le	je	que	quand	il
Narbonne	le	a	et	lors	quand	que
Chevalier hermite	le	et	que	quand	de	a
Cordouanier	le	et	de	a	en	or
Juitel	le	quand	et	que	a	car
Enfant sauva mere	le	et	que	a	de	quand
Eaue et vergier	le	que	et	par	de	en
Pain	le	que	et	quand	de	por
Mescreant	et	le	que	il	quand	si
Pecherresse	que	de	le	en	cel	quand
Merlin	le	que	et	quand	de	a
Florence	que	le	mais	a	et	quand
Anelés	que	le	et	mais	de	il
Buef	le	et	que	de	a	mais
Beguine	le	et	quand	que	a	de
Abesse	le	que	et	quand	il	seignor
Sauveur	le	quand	a	de	et	que

whether there is indeed a unity to the frequency with which Jehan uses particular words at the beginning of each line.

Table 35 contains the total counts of the six words which appear most frequently within the *J-corpus*. These combined data appear to consolidate the trends observed in Table 34, namely that *le* is by far the most common lemma line-initially across the corpus, with *que* and *et* appearing next, before *de*, *quand*, and *a* form a group with reasonably similar rates.

Table 35. The six most common lemmata at the beginning of a line across the entire *J-corpus*.

le	*que*	*et*	*de*	*quand*	*a*
825	505	482	292	283	279

By comparing the rates of the use of each of the above lemmata within each text with the rate at which the lemma is used in the rest of the *J-corpus*, we are able to calculate a p-value to determine whether there is a statistically significant difference between an individual text's use of a particular lemma in line-initial position, and the use observed across the rest of the *J-corpus*, given here in Table 36. Lemmata with a p-value under the threshold of 0.05 display a use of a particular line-initial lemma that can be said with some certainty not to match that of the rest of the *J-corpus*, and are marked as achieving a PASS result.

The first observation of note from these tests is that the variability that we observed anecdotally in Table 34 appears to be borne out by the data. Only seven texts do not display any statistically significant differences from the rest of the *J-corpus* for any of these lemmata, with nine differing for a single lemma, five texts for two lemmata, and two texts (*Chevalier et escuier* and *Sauveur*) for three lemmata. *Florence* differs from the usage observed for the rest of the *J-corpus* in no fewer than four instances.

It is also the case that on occasion, even when a text makes use of a given lemma at a rate which locates it in an expected position in the order of the most frequent line-initial lemmata, its rate of use of that particular lemma still differs in a statistically significant way from that observed for the rest of the *J-corpus*. This is the case for *le*, for example, in *Pommes*, *Romme*, *Enfant rosti*, *Cordouanier*, *Buef*, and *Abesse*. In all except for *Buef*, where the rate is lower, this results from the fact that *le* is used even more frequently in these texts than we might expect, based on the rates observed across the rest of the *J-corpus*.

It would be tempting, at least initially, to assume that this variability provides evidence against the shared-author hypothesis. However, while we have already observed areas of significant variation within the *J-corpus*, we have also seen that this internal variation is often still less than that observed for our other reference corpora. Indeed, such variation also does not necessarily detract from the fact that the texts of the *J-corpus*, however internally variable, still form a cohesive unit that is collectively distinct from other corpora.

Table 36. Results of p-value tests comparing use of the six most frequent lemmata at the beginning of a line across the entire *J-corpus*.

Text	***le***	***que***	***et***	***de***	***quand***	***a***
Respon						
Chanoinnes						PASS
Pommes	PASS					
Romme	PASS			PASS		
Deux chevaliers						
Enfant rosti	PASS	PASS				
Povre chevalier			PASS			
Chevalier et escuier	PASS		PASS	PASS		
Narbonne						
Chevalier hermite						
Cordouanier	PASS	PASS				
Juitel					PASS	
Enfant sauva mere						
Eaue et vergier						
Pain		PASS				
Mescreant			PASS			
Pecherresse	PASS					
Merlin						
Florence	PASS	PASS	PASS	PASS		
Anelés	PASS	PASS				
Buef	PASS					
Beguine			PASS		PASS	
Abesse	PASS					
Sauveur		PASS	PASS		PASS	

By taking a more global approach to the *J-corpus*, we can build this sense of cohesion, by identifying ways in which Jehan's texts differ collectively from trends observed elsewhere. Here we compare the corpus to our three textually coherent reference corpora, namely the *A-corpus* of narrative poems in alexandrine monorhyme quatrains, the *F-corpus* of *fabliaux*, and the *R-corpus* of Rutebeuf's works. By comparing the *J-corpus* to these corpora, we thus compare it to poems of a comparable type to the *J-corpus*,

works of a consistent genre, and works by a single author. In doing so, we observe that, despite their surface-level variation, the texts of the *J-corpus* do, in fact, form a coherent corpus in how, collectively, they differ from these other corpora of text.

Table 37. The six most common lemmata at the beginning of a line across the *J-*, *A-*, *F-*, and *R-corpora*.

Corpus	**1**	**2**	**3**	**4**	**5**	**6**
J	le	que	et	de	quand	a
A	que	le	et	de	a	mais
F	que	et	le	a	si	de
R	que	et	le	ne	de	si

Table 37 displays the six most frequent lemmata in line-initial position for the *J-*, *A-*, *F-*, and *R-corpora*. By contrasting our reference corpora with the *J-corpus*, we notice in a superficial way that there is variation in which lemmata are the most frequent. While *quand* is the fifth most frequent line-initial lemma in the *J-corpus*, it drops from the list entirely for the other corpora, appearing in positions seven, twenty-three, and eleven for the *A-*, *F-* and *R-corpora* respectively. The use of *le*, meanwhile, is especially significant as it is the most frequent line-initial lemma in the *J-corpus* (by some margin, as we saw in Table 35), while it is only the second most frequent in the *A-corpus*, and third in the *F-*, and *R- corpora*.

We note that both these examples, while reinforcing the cohesion of the *J-corpus*, also align it with the *A-corpus* as the closest reference corpus, reinforcing the suggestion that at least some of this order may be determined by generic and versificatory requirements, even if authorial preference does play an additional role. The appearance of *mais* in sixth position for the *A-corpus* also reinforces this association, given that it also appears in seventh position in the *J-corpus*.

Considering the data from the other direction, we find several lemmata which appear with a high frequency in our reference corpora, but which do not appear as frequently in the *J-corpus*. Again, the association between the *J-* and *A-corpora* is strengthened by their comparable use of *et*, given that it is the third most frequent lemma in line-initial position in both, and yet it appears in second position in the *F-* and *R-corpora*. Similarly, *si* appears at a relatively high position for the *F-* and *R-corpora*, in positions five and six respectively, but only appears in positions fourteen and ten for the *J-* and *A-corpora* respectively. This could go some way towards suggesting a degree

of similarity between the *F-* and *R-corpora*, which is unsurprising given the frequently similarities of content and form of many of the texts which they contain, but it also once again has the effect of separating the texts of the *J-corpus* from these reference corpora.

Simply ordering the most frequent line-initial lemmata in this way provides a useful yardstick by which to gain a sense of how our corpora may be distinct, but it ultimately does not allow us to reach any statistically assured conclusions. It is therefore also worthwhile comparing the *J-corpus* to our reference corpora by means of the specific rates at which these lemmata are used at the beginning of a line.

Table 38 contains the results of Welch Two Sample t-tests comparing the *J-corpus* to our reference corpora for each of the six most common line-initial lemmata within the *J-corpus*. Overall, and specifically for *le* and *quand*, the results reveal that many line-initial lemmata are used significantly differently in the *J-corpus*, when compared to the rates observed in our reference corpora.

Table 38. Average rates of the use of the six most common lemmata at the start of lines in the *J-corpus*, compared to the *A-*, *F-*, and *R-corpora*.

	J-corpus	***A-corpus***		***F-corpus***		***R-corpus***	
Lemma	**Rate**	**Rate**	**P-value**	**Rate**	**P-value**	**Rate**	**P-value**
le	12.08%	8.64%	0.00001	6.01%	0.00000	5.07%	0.00000
que	7.39%	9.36%	0.08833	12.15%	0.00000	13.85%	0.00000
et	7.06%	7.79%	0.1778	10.01%	0.00040	9.28%	0.00397
de	4.28%	3.94%	0.7611	3.02%	0.00672	3.93%	0.05749
quand	4.14%	3.16%	0.00912	1.40%	0.00000	2.09%	0.00000
a	4.09%	3.54%	0.6083	3.07%	0.0124	2.73%	0.00003

In addition, we note that the *F-* and *R-corpora* are also distinct from the *J-corpus* for more lemmata (six and five respectively) than the *A-corpus*, which is only statistically distinct on two occasions. These data therefore reinforce the impression gained from the simple ordering of lemmata given in Table 34, that the *A-corpus* is most closely aligned with the *J-corpus*. This of course highlights the texts' narrative and versificatory similarities, but the presence of clear distinctions for *le* and *quand* also reinforces the fact that the *A-corpus* still differs from the *J-corpus* in a way that strengthens the shared-author hypothesis.

We are also able to draw broader linguistic conclusions from these data, namely that while uses of *le*, *que*, *et*, *quand*, and *a* vary significantly across all the corpora we have examined, the rate of use of *de* appears to be more

static, suggesting that the rates in this case may be determined by linguistic factors, rather than by the norms of a given genre, or the preferences of an individual author.

These comparative data therefore allow us to reach a valuable conclusion for the *J-corpus*: as we have seen, Jehan's use of line-initial words varies from text to text, with some poems displaying significantly different rates of use compared to the rest of the corpus. When taken together, however, the usage observed in the *J-corpus* does reveal itself to be significantly distinct from that found in texts by other authors, across several metrics. Thus, Munk Olsen (1963b, 152) is correct, of course, to argue that an author is not straightforwardly "un cerveau électronique" with predictable outputs,[10] but as modern stylometry has proven in less bespoke ways, there are nevertheless macro-level trends which can unite a given corpus of texts through distinctiveness relative to some other corpus, in turn contributing to a sense that the texts share the same author.

Having established the validity of this method as a means of providing a sense of unity to the *J-corpus*, we may therefore also usefully examine the extent to which *Guillaume*, *Robert*, *Richard*, and *Jehan Paulus*, the texts of the *A-corpus* to which we have extended the shared-author hypothesis, line up with the trends observed for line-initial lemmata in the *J-corpus*. As we performed above for each of the texts of the *J-corpus*, Table 39 presents the results of p-value tests comparing these three poems to the *J-corpus*.

Table 39. Results of p-value tests comparing use of the six most frequent lemmata at the beginning of a line across the entire *J-corpus* to *Guillaume*, *Robert*, *Richard*, and *Jehan Paulus*.

Text	*le*	*que*	*et*	*de*	*quand*	*a*
Guillaume	PASS	PASS	PASS			
Robert		PASS	PASS	PASS	PASS	
Richard	PASS	PASS	PASS		PASS	
Jehan Paulus			PASS		PASS	

From these tests, we once again get the sense that for *Jehan Paulus* in particular, the text is sufficiently similar to the broader *J-corpus* to support collective attribution, given that it differs in a statistically significant way from the *A-corpus* on only two occasions, in the cases of *et* and *quand*, and is thus well within the range of variation observed across the *J-corpus*. This is of

10 "an electronic brain."

course especially notable, given our suggestion that the text has at some point been *remanié*, and indicates that in this text, the later modifying poet made minimal changes to structures at the beginning of a line.

Guillaume, *Robert*, and *Richard*, however, appear somewhat more problematic. *Guillaume* displays statistically significant differences from the *J-corpus* for three lemmata (*le*, *que*, and *et*), and *Robert* (*que*, *et*, *de*, and *quand*) and *Richard* (*le*, *que*, *et*, and *quand*) do so for four lemmata each. If we compare these tests to those performed on the individual texts of the *J-corpus*, we therefore find that *Robert* and *Richard* sit alongside *Florence*, which is also distinct from the rest of the *J-corpus* in the case of four lemmata, and indeed overlaps on three of these in each case. One might therefore contemplate the exclusion of *Florence*, *Guillaume*, *Robert*, and *Richard* from Jehan's *œuvre*, but it is noteworthy that, as we have seen and unlike *Jehan Paulus* which sits more comfortably within the usual range, these four texts are amongst the longest which have been attributed to Jehan. Indeed, they constitute a group apart from the rest of his *œuvre*, alongside *Anelés* and *Buef*, forming what we have termed the "long-text" sub-group.

In part, this outcome is likely therefore a feature of the statistical methods used, since the t-test becomes more certain in its result the greater the sample size, and thus a longer text is more likely to trigger the threshold for statistical significance. However, there is more to this variation than simple variations in sample size. One final example will reinforce the cohesion of this grouping, allowing us to conclude that, just as Jehan's versificatory complexity increased when writing longer poems, resulting in lesser use of so-called "easy" rhymes, so too did the complexity of his line-initial structures, hence his longer texts are more likely to deviate from expected patterns seen elsewhere in the corpus. This, in turn, grants us a surer footing from which to attribute *Guillaume*, *Robert*, and an earlier version of *Richard* to Jehan.

The evidence for this claim comes from the lemma *mais*, which only appears in the top six lemmata in the three longest texts of the *J-corpus*: *Florence*, *Anelés*, and *Buef*, appearing at positions three, four, and six respectively. It also appears at a comparable position in *Guillaume*, *Robert* and *Richard* (two, three, and two respectively). That *mais* is used a great deal less in the shorter texts of the *J-corpus* therefore suggests that Jehan's word choices may have varied depending on features of the text he was writing, such as its intended length and thus its narrative complexity. When writing longer texts, Jehan appears to have preferred to use more contentful words at the beginning of lines, such as *mais*, perhaps to better engage an audience or simply to avoid significant repetition of standard structures. That this is true of all of the texts of the "long-text" group suggests that what matters more

than the texts' points of difference from the *J-corpus* as a whole is therefore the fact that they share notable similarities with other texts from the "long-text" group, to which we now add *Guillaume*, *Robert*, and a putative earlier version of *Richard*, the latter of which again does not appear to have been especially heavily *remanié* for this feature. We shall observe further evidence for the validity of this subdivision in the next section of this chapter.

The method pursued in this section has once more allowed us to affirm the shared-author hypothesis. It has done so while continuing to treat the texts in question in a humanistic way, acknowledging the undeniable truth that an author cannot be simply treated as a machine who, given a certain input, will always produce the same output. It is therefore worthwhile expanding this approach to consider other types of shared structures across the *J-corpus*. In doing so, we will gain a clearer idea of both the diversity and the unity of Jehan's texts.

Onset of Speech

A further object of analysis proposed by Munk Olsen (1963b, 150) for the examination of common structures, but once again not widely applied by him across the *J-corpus*, is the moment of the onset of discourse in the line.[11] He proposes a typology for the ways in which direct speech is introduced within Jehan's poems, identifying three types, as below:

Type A: Discourse begins at the start of a new line, with the speaker indicated in the previous line.

Type B: Discourse begins at the start of a new line, with the speaker indicated later in the same line, before discourse resumes.

Type C: The speaker is indicated at the beginning of a new line, followed by the onset of discourse.

11 Here I talk of the "onset of discourse" or "discourse onsets" to refer to that moment at which direct speech begins within a text, as marked in modern editions with an opening set of quotation marks. A great deal of useful work has been done on different kinds of discourse, the analysis of which falls outside our scope here (see, for example, Ducrot [1984, 1989], Cerquiglini [1981, 1989] and Marnette [1998, 2005]). Avenues for further examination within the *J-corpus* include the frequency of the use of direct discourse compared to other kinds of discourse, details of the identity of the speaker, and where in a text discourse appears.

Examples of each of these types abound:

Type A: Le chevalier l'apelle et si l'araisonna:
"Amis, celi t'en gart qui le monde forma!" (*Chevalier et escuier*, vv. 25–26)

Type B: "Sire," dit la pucelle, "je ne vous desdi mie." (*Respon*, v. 102)

Type C: Le chevalier respont: "Voulentiers le ferai." (*Chanoinnes*, v. 136)

There is, however, little concrete evidence provided by Munk Olsen to support his claims as to the frequency of use of these types, and it is equally unclear to what extent the corpus is unified in the texts' relative use of them. A corpus-wide approach, aided by computational and statistical methods, can therefore shed further light on the matter.

The first observation from an examination of all onsets of discourse across the *J-corpus* is that Munk Olsen's types are too restrictive, and do not encapsulate all instances. For Type A, for instance, we notice that when direct discourse begins at the start of a new line, the speaker is not necessarily identified in the preceding line, and may in fact have been identified earlier, as in this passage from *Pain*, in which the direct discourse appears after an intervening line of indirect speech:

Li ennemis parole a l'ange d'une part
Por quoi Diex veult avoir en cestë ame part:
"Il m'est avis qu'il s'est aperceüs trop tart,
[…]" (*Pain*, vv. 113–15)

For Type B, in cases where the direct discourse given by a character is particularly short, direct discourse may indeed begin at the start of a new line, with the speaker's identity given afterwards, but discourse does not then resume:

"Non ferai je," dist cil, qui bien la sot entendre. (*Beguine*, v. 46)

It is also possible, despite Munk Olsen's arguments to the contrary, for direct discourse to resume in a following line, rather than within the same line as the initial onset:

"Par ma loy," dit li peres, qui fu fel et engrez,
"Qu'il a passé .vij. ans, nous le suivrons de prez.
[…]" (*Juitel*, vv. 37–38)

As for Type C, it is not always the case that when the introduction of a speaker occurs before but in the same line as their direct discourse, this introduction is the first element of the line, since there are examples of elements of narration appearing as the first element of such lines:

L'ostel vit d'un hermite, si a dit: "Diex aïde!" (*Sauveur*, v. 180)

Indeed, on occasion, the first element of a line may be the end of a previous unit of discourse:

"[…]
Qu'il m'ont fait." La voiz dist: "Tu ne fais pas folie." (*Florence*, v. 724)

Beyond the types proposed by Munk Olsen, we also observe a further means of introduction of direct discourse, namely the often stichomythic juxtaposition of two units of speech with no intervening element, with the switch in speaker usually identifiable in the manuscript only through the content of the discourse.[12] These onsets of discourse may take place between one line and the next, or within the same line:

"Dame, vous estiez morte. Qui vous resuscita?"
"Eu non Dieu, dous amis, cele qui Dieu porta." (*Cordouanier*, vv. 131–32)

"Sire, quel non avez?" "On m'appelle Merlin." (*Merlin*, v. 54)

Finally, examples can be found in which more than one type of discourse onset intersects:[13]

[…] Et celle respondi:
"Certes, sire," dist elle, "ne vous mentiré mie.
[…]" (*Beguine*, vv. 68–69)

We may therefore amend Munk Olsen's typology to the following:[14]

Type A: Discourse begins at the start of a new line, with the speaker indicated earlier.

Type B: Discourse begins at the start of a new line, followed by an indication of the speaker later in the same line. The discourse may then resume, either later in the same line, or in a following line.

12 We could take this type of discourse onset as evidence for the texts' possible orality, since it would be far easier to identify a change in speaker through the actions and tone of voice of a storyteller than it is on the manuscript page, with fr. 24432 only rarely containing any punctuation to indicate the beginning of a unit of discourse. On these rare occasions where such a feature is marked, it is usually with a *punctus*. For more on the use of punctuation to introduce direct discourse, see Marnette (2006a).

13 When this occurs, both types of discourse onset are included in our counts.

14 Types A and C combine to form a category of discourse onset which corresponds with what Cerquiglini (1981) and Marnette (2006a) have called *prolepse*, that is, the introduction of a speaker to direct discourse before it begins, while Type B is *analepse*, being the introduction of a speaker after direct discourse has already begun.

Type C: The introduction of the speaker is followed in the same line by the onset of discourse.

Type D: Discourse begins immediately following the end of a previous unit of discourse by another speaker, as a continuation of the same line or at the beginning of a new line.

With our amended typology, we can thus examine the extent to which the use of each type is consistent across the *J-corpus*. From the data in Table 40,

Table 40. Counts of each type of discourse onset for each text in the *J-corpus*.

Text	Type A	Type B	Type C	Type D	Total
Respon	6	11	10	6	**33**
Chanoinnes	11	8	12	1	**32**
Pommes	11	19	9	7	**46**
Romme	6	10	1	0	**17**
Deux chevaliers	5	16	7	1	**29**
Enfant rosti	3	11	4	0	**18**
Povre chevalier	4	1	7	12	**24**
Chevalier et escuier	4	19	11	2	**36**
Narbonne	5	16	3	0	**24**
Chevalier hermite	6	9	9	1	**25**
Cordouanier	3	10	5	2	**20**
Juitel	9	2	5	4	**20**
Enfant sauva mere	4	9	13	0	**26**
Eaue et vergier	3	4	2	0	**9**
Pain	4	5	7	0	**16**
Mescreant	0	11	12	0	**23**
Pecherresse	2	2	6	0	**10**
Merlin	3	3	12	20	**38**
Florence	13	9	50	0	**72**
Anelés	11	3	44	1	**59**
Buef	13	23	17	1	**54**
Beguine	5	11	5	0	**21**
Abesse	3	23	4	3	**33**
Sauveur	15	32	5	2	**54**
Total	**149**	**267**	**260**	**63**	**739**

we observe that of a total of 739 onsets of discourse, Types B and C count for roughly a third each, while the totals for Types A and D are much lower. Munk Olsen's claim that Type B is "de beaucoup le plus fréquent" (1963b, 160),[15] in a simply numerical sense at least, therefore appears to be overblown, since while onsets of Type B are indeed the most frequent kind, they cannot be said to be far and away the preferred method, given how frequently we also find onsets of Type C.

By considering the relative rates of use of each type per text, as represented in Figure 15, it becomes clear that there is a great deal of variation in the types used by individual texts, especially for Type D. Except for *Mescreant*, which has no examples of Type A, all texts contain examples of Types A, B, and C. For Type D, ten of the twenty-four texts contain no examples of this type at all. In *Povre chevalier* (50%) and *Merlin* (52.6%), by contrast, at least half of the total onsets of discourse can be classified under Type D. There does not appear to be a narrative explanation for this distinction either, since while *Povre chevalier* and *Merlin* are indeed texts which contain a lot of the fast-paced dialogue that might lend itself to unindicated switches in speaker, so too are many other texts of the corpus, such as *Florence*, which contains no such examples.

Of further note is that *Florence* and *Anelés* appear to favour Type C onsets, almost to the exclusion of all other types, resulting in a much lower than expected use of Type B. This impression of a negative correlation between Type B and Type C onsets is borne out by a Pearson's product-moment correlation test, which reveals a strong, negative correlation between the two onset types (-0.65328, p = 0.00054), such that in the *J-corpus* Types B and C can be seen as variants in alternation, with a high rate of use of one type leading to fewer possibilities for the other. This is a surprise, given the narrative difference between the proleptic Type C and the analeptic Type B,[16] but it can perhaps be explained by the fact that, particularly in the alexandrine lines of the *J-corpus*, discourse onsets of Types B and C are usually those which take up roughly half an individual hemistich. They are thus roughly interchangeable from a versificatory perspective, and so were an author to decide to make frequent use of one type for a given text, then the opportunities for the other would be significantly diminished.

A similarly strong, negative correlation is observed between Types B and D (-0.50540, p = 0.01176), such that a higher rate of one is associated

15 "by far the most frequent."

16 See p. 130n14.

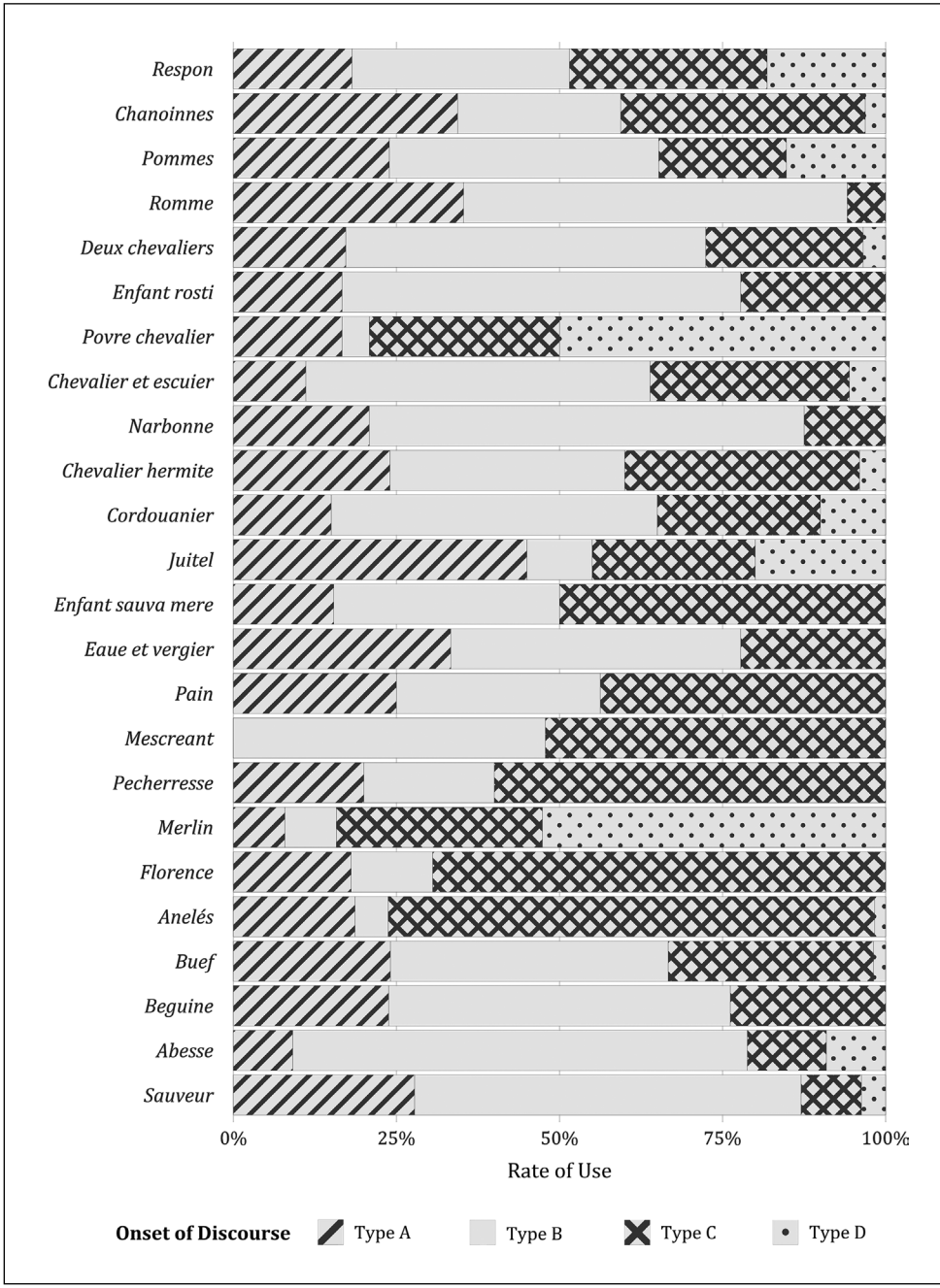

Figure 15. Rates of use of each type of discourse onset in the *J-corpus*.

with a lower rate of the other. This is again surprising, but can be explained through the fact that Types B and D in some sense seek to achieve the same goals: by placing the discourse as early as possible, it is granted a sense of immediacy. Then, once the discourse has begun, indication of the speaker can be given if it would be unclear in context (Type B), or the shift in speaker can be marked in more subtle ways (Type D). Looking again at our Type D example from *Cordouanier*, for instance, we see that "dous amis" functions as a marker of turn-taking, and indicates that the addressee is male (and therefore in this dialogue between a husband and wife, that the speaker must be female). In a more ambiguous case, the phrase could just as easily have been replaced with the trisyllabic "ce dist el" or similar, shifting the Type D onset to one of Type B.

> "Dame, vous estiez morte. Qui vous resuscita?"
> "Eu non Dieu, dous amis, cele qui Dieu porta."
>
> (*Cordouanier*, vv. 131–32)

A final correlation of significance is that of length. Given that *Florence* and *Anelés* are two of the longest in the *J-corpus*, and contain the greatest number of discourse onsets, their greater use of Type C onsets means that we are able to reinforce our hypothesis above that the longer texts within the *J-corpus* form a sub-group distinct from the rest of Jehan's *œuvre*.[17] A correlation test reveals that, across the *J-corpus*, there is a weak, positive correlation between text length and the rate of use of Type C onsets (0.48798, $p = 0.01556$) such that, as a general rule, the longer a text is in the *J-corpus*, the higher its expected rate of Type C discourse onsets.[18] Notably, however, *Buef*, the other text within the *J-corpus* belonging to the "long-text" sub-group, while not completely out of alignment with these texts, nevertheless displays rates of use of Types B and C more in keeping with the rest of the corpus. The uses observed in *Florence* and *Anelés* are therefore explicable, and in line with a trend that exists across the whole corpus, but high use of Type C onsets is not a universal feature of the whole of this sub-group.

---

17 Indeed, if one excludes both *Florence* and *Anelés* from the data, then Munk Olsen's suggestion that Type B is "de beaucoup le plus fréquent" (1963b, 160; "by far the most frequent") appears far more plausible, since we observe 255 instances of Type B, but only 166 instances of Type C.

18 A correspondingly inverse correlation between length and Type B onsets appears plausible, but does not reach our threshold for statistical significance ($p = 0.1492$). Neither Type A nor Type D onsets display a correlation with length at anything close to statistical significance.

Given these points of similarity between *Florence* and *Anelés*, we are reminded of Morawski's claim that "nous croyons que [*Florence* et *Anelés*] sont l'œuvre d'un même auteur, mais nous n'avons pas acquis la conviction qu'ils soient de Jean" (1939, 351–52).[19] The broader correlation between length and Type C onsets gives us reason to disagree with Morawski's claim, especially given the unifying features we have seen elsewhere. Our analysis here has nevertheless enabled us to encounter further stylistic similarities between these two texts, suggesting that if indeed they are the work of the same author, then they may have originated from a similar context of composition.

Comparing these data to those observed for the *A-corpus* in Figure 16,[20] we once again gain the sense that we have often seen for these corpora that many of the poems we are attributing to Jehan are broadly aligned with those of a similar poetic form and narrative style. In this case, this is true both in the distribution of different quote types, and in the very fact of their variability. As in the *J-corpus*, Types A, B, and C are prevalent in the *A-corpus*, with only *Déz* lacking instances of Type A, and *Confesseur* examples of Type B. All contain examples of Type C, while Type D discourse onsets are again much more variable, being absent in ten of the twenty-three texts of the *A-corpus* that contain direct discourse. *Guillaume*, *Robert*, *Richard*, and *Jehan Paulus*, meanwhile, again align with the rates we observed in the *J-corpus*, suggesting that even the apparent *remaniement* of the latter two texts had minimal impact on their use of discourse onsets. *Guillaume* and *Robert* display a comparable use of discourse onsets to *Anelés* and *Florence*, two texts of the "long-text" grouping alongside which *Robert* appears in fr. 24432, further strengthening the attribution of these texts.

We may therefore ask to what extent these two corpora do really differ from each other. A series of Welch Two Sample t-tests, which calculate the probability that the rates for each discourse onset in the *J-* and *A-corpora* are statistically distinct, here produce results (A: $p = 0.226$; B: $p = 0.1072$; C: $p = 0.1228$; D: $p = 0.2873$) that are often close to but all ultimately above our threshold for statistical significance. This remains true even if we exclude *Guillaume*, *Robert*, *Richard*, and *Jehan Paulus* from the *A-corpus*, or include them alongside the *J-corpus*. As a result, we can suggest that, overall,

19 "we believe that [*Florence* and *Anelés*] are the work of the same author, but we have not reached the conviction that they are by Jean."

20 Here we exclude *Thomas*, *Yves*, and *Dieu d'Amours*, which contain no examples of direct speech. In the case of the latter two texts, this is perhaps only because they are fragmentary.

Figure 16. Rates of use of each type of discourse onset in the texts of the *A-corpus* that contain instances of direct discourse.

Jehan's use of discourse onsets aligns with the expected patterns observed across other narrative poems from the same period, that were written in alexandrine monorhyme quatrains.

That is not to say, however, that discourse onsets are a feature totally removed from the question of authorial preference. As in the *J-corpus*, there is a strong, negative correlation in the *A-corpus* between the use of Type B and C onsets (-0.71602, p = 0.00012), once again suggesting that this is a feature determined by generic or versificatory expectations, but the same is not true of the correlation between Types B and D (p = 0.5824). Again, this remains true even if we exclude *Guillaume*, *Robert*, *Richard*, and *Jehan Paulus* or move them to the *J-corpus*. The latter correlation therefore appears to be a feature only of the *J-corpus*, and thus while this is ultimately an indicator of variation, it is nevertheless evidence of an authorial preference, adding to our sense of cohesion for the *J-corpus*, and the sense of a distinction between the *J-* and *A-corpora* for this feature.

As a final direction of analysis, it is worthwhile examining discourse onsets in the *F-corpus* of *fabliaux*, to observe what this corpus may tell us about the extent to which the features we have observed thus far are limited to texts of this type, or shared with texts which are generically and metrically aligned, and yet collectively distinct from the *J-* and *A-corpora*. Whereas Welch two sample t-tests showed that there is no statistically significant distinction between the rates of any onset type when comparing the *J-* and *A-corpora*, when compared to the *F-corpus*, both display rates of use that are distinct in a statistically significant way for Types C (*J*: p = 0.00564; *A*: p = 0.00013) and D (*J*: p = 0.04029; *A*: p = 0.00002). We can therefore say with some certainty that the use of these types of discourse onset is determined by generic or versificatory features. Most likely, the contributing factor is the use of octosyllabic lines in the *fabliaux*, compared to the twelve-syllable alexandrine lines of the *J-* and *A-corpora*, which appears to result in a lower use of Type C onsets, which require greater space in the line, and an increase in Type D onsets, which do not. A statistically significant distinction also appears between the *A-* and *F-corpora* for Type B onsets (p = 0.01014), which again aligns with the hypothesis that octosyllabic lines favour discourse onsets, like Types A and D, which allow for minimal material to appear in the same line as the discourse itself.[21]

21 Indeed, even when Type B discourse onsets do appear in the *F-corpus*, the discourse often begins on one line, before the speaker is introduced in the following line. This does not occur in the *J-corpus*, and only very rarely in the *A-corpus*. Again, this is unlikely to have been the result of authorial choice, rather a function of the

Turning more broadly to the similarities in structure which we have observed for the *J-corpus* and beyond in this chapter, it is worth once more revisiting Munk Olsen's observation (1963b, 152) that an author ought not solely to be thought of as "un cerveau électronique dans lequel une impulsion de contenu donnée déclenche automatiquement et infailliblement l'impulsion d'expression correspondante."[22] As we have seen for both line-initial words and the onset of discourse, if any conclusions are to be drawn at all from such data about whether a group of texts share a single author, then this must be done alongside an acknowledgement that the mind of an author is never fixed; no text is ever written with the same combination of inspiration and creativity as any other, and so it is far from surprising that variation should be inevitable, even in the works of a single author.

Overall, however, we have observed that there are many similarities of vocabulary and structure across the texts of the *J-corpus*, both those determined by Morawski, Kleist, and Munk Olsen, and those uncovered using statistical and digital methods which were unavailable to them. While each of these similarities are not in themselves sufficient to prove shared authorship, they do once again contribute to the unity of the *J-corpus*, alongside *Guillaume*, *Robert*, *Richard*, and *Jehan Paulus*, allowing us to be more assured in the shared-author hypothesis.

octosyllabic metre of the *fabliaux*, since there is less space than in the alexandrines of the *J-corpus* in which to fit both the direct discourse and its analeptic discourse marker, which is therefore pushed onto the following line.

22 "an electronic brain in which an impulse of given content automatically and unfailingly triggers the corresponding impulse of expression."

Building the Case for the Shared-Author Hypothesis

The statistical examination of line-initial lemmata and discourse onsets carried out in this chapter reinforces the internal coherence of the *J-corpus* by identifying the authorial "tics" initially discussed by Munk Olsen (1963b) which distinguish Jehan's writing from that found in our other reference corpora. The "long-text" sub-group, including the newly attributed poems *Guillaume*, *Robert*, and *Richard*, is further unified by their shared approach to line-initial lemmata, and their comparable use of discourse onsets, in a way that differentiates them from Jehan's shorter poems. Meanwhile, the fact that both *Richard* and *Jehan Paulus* remain statistically aligned with the *J-corpus* in both the use of line-initial words and discourse onsets, despite their likely *remaniement*, supports the status of both texts as echoes of earlier poems by Jehan, and suggests that the later *remanieur(s)* carried over units of direct discourse much as they appeared in Jehan's work.

Methodological Contribution

The approach presented in the first part of this chapter aligns closely with the core assumption of stylometric approaches to attribution, namely that focusing on semantically "empty" function words can have significant value when seeking to attribute authorship. However, by examining lemmata specifically at the beginnings of lines, as well as by examining the corpus-specific case of discourse onsets, I have moved beyond a purely stylometric approach to one which builds first and foremost from a humanistic understanding of the texts under consideration. My revised typology of discourse onsets, meanwhile, designed for the unique features of poems written in alexandrine monorhyme quatrains but applicable in other contexts too, offers an example of how one can quantify and compare an author's use of specific narrative structures between and within corpora. In addition, this section also highlights the importance of accounting for corpus-internal variables when drawing conclusions on a given feature. For example, the differences in the use of discourse onsets observed between Jehan's longer and shorter poems could easily be mistaken for evidence of multiple authors, were a unifying explanation for this variation not to be found in the correlation between text length and the use of discourse onsets.