Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux

Mike Kestemont
Institute for the Study of Literature in the Low Countries & CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Sara Moens and Jeroen Deploige
History Department, Ghent University, Belgium

Abstract

Hildegard of Bingen (1098–1179) is one of the most influential female authors of the Middle Ages. From the point of view of computational stylistics, the oeuvre attributed to Hildegard is fascinating. Hildegard dictated her texts to secretaries in Latin, a language of which she did not master all grammatical subtleties. She therefore allowed her scribes to correct her spelling and grammar. Especially Hildegard's last collaborator, Guibert of Gembloux, seems to have considerably reworked her works during his secretaryship. Whereas her other scribes were only allowed to make superficial linguistic changes, Hildegard would have permitted Guibert to render her language stylistically more elegant. In this article, we focus on two shorter texts: the *Visio ad Guibertum missa* and *Visio de Sancto Martino*, both of which Hildegard allegedly authored during Guibert's secretaryship. We analyze a corpus containing the letter collections of Hildegard, Guibert, and Bernard of Clairvaux using a number of common stylometric techniques. We discuss our results in the light of the Synergy Hypothesis, suggesting that texts resulting from collaboration can display a style markedly different from that of the collaborating authors. Finally, we demonstrate that Guibert must have reworked the disputed visionary texts allegedly authored by Hildegard to such an extent that style-oriented computational procedures attribute the texts to Guibert.

Correspondence:
Mike Kestemont, Institute for the Study of Literature in the Low Countries & CLiPS Computational Linguistics Group, University of Antwerp, Belgium.
Email:
mike.kestemont@gmail.com

1 Introduction

Since the end of the 1960s, literary studies have seen a clear shift of focus from the analysis of authorial intentions to reader-oriented criticism. The repudiation of the modern idea of autonomous authorship has perhaps gone furthest in medieval studies, with the rise, since the late 1980s, of Material Philology (Nichols, 1997). Medievalists have become increasingly aware of the importance of manuscript culture in their understanding of texts: medieval texts should not primarily be studied, it is argued, as abstract entities resulting from authorial ambitions, but rather as tangible objects, materialized in

Digital Scholarship in the Humanities, Vol. 30, No. 2, 2015.

doi:10.1093/llc/fqt063 Advance Access published on 26 October 2013

specific manuscript contexts. Every material manifestation of a text is unique, because the acts of copying and compiling nearly always resulted in textual changes—from minor changes in orthography to complete rewritings. Our modern post-romantic conception of authorship therefore seems profoundly anachronistic with respect to the Middle Ages (Cerquiglini, 1999, p. 8–10). Yet, even if medieval culture did not share our present-day view on the significance of original authorship, the Middle Ages have known many respected and authoritative individuals who were recognized by their contemporaries and posterior readers as producers of very specific literary works. Some kind of correlation even existed between the degree to which texts were susceptible to alterations and the religious and intellectual authority of their authors (Deploige, 2005).

This did not mean, however, that such recognized authors were necessarily acting individually in the process of conceiving their treatises or narratives—quite the contrary. Writing in the Middle Ages meant entering into a dialogue with a long line of predecessors, whether through citations, paraphrasing, or allusions. In the actual process of literary composition too, medieval authors only seldom worked alone. A 'new' text could be the result of drafts on wax tablets copied by professional scribes, of processes of dictation and subsequent correction, etc. A twelfth-century authority like the Cistercian abbot Bernard of Clairvaux (1090–1153), one of the most prolific and influential medieval authors, is known to have been surrounded by a team of secretaries. For his sermons and letters in particular, he was assisted by a number of collaborators to whom he could dictate his messages or who were asked to produce texts in accordance with his own views. Some of his collaborators were even trained in imitating his writing style, thus facilitating Bernard's work of final editing or correcting (Leclercq, 1962; 1987, pp. 147–52). In the case of the remarkably few medieval female authors known to us, the role of secretaries and collaborators is even more intricate. Women writers like the German nuns Hildegard of Bingen (1098–1179) or Elizabeth of Schönau (1129–1165) were considered unlearned and incapable of independently writing down their visionary experiences, even if these were 'divinely inspired'. These women therefore had to be assisted by male collaborators, often also serving as their spiritual directors. The precise nature and implications of such cross-gender collaborations remain a topic of scholarly debate.

The immediate incentive for the present article is the preparation of a new critical edition of two lesser known texts attributed to Hildegard of Bingen, supposedly dating from the last years of her life: the *Visio de Sancto Martino*, which is conceived as a letter addressed to the worshippers of Saint Martin, and the *Visio ad Guibertum missa*, containing spiritual advice to an anonymous monk-priest, generally identified as her last secretary, Guibert of Gembloux (1124–1213) (Deploige and Moens, forthcoming). Among the few scholars who paid attention to these texts, there is still no consensus as to the extent to which they should be attributed to either Hildegard herself or to her collaborator Guibert. As neither traditional stylistic analysis nor contextual historical research has so far been able to resolve the problem, we will approach this issue through a stylometric analysis. We will focus on three research questions.

First, does stylometry allow for an authorial differentiation between the writings of twelfth-century Latin authors, belonging to highly similar intellectual circles? To answer this question, we will investigate the letter collections or *epistolaria* of Hildegard of Bingen, her secretary Guibert of Gembloux, and their famous contemporary, Bernard of Clairvaux. Our aim is to assess to what extent we can distinguish stylistic profiles for these authors, despite the marked *variance* within medieval manuscript culture (Cerquiglini, 1999), as well as the fact that these authors, like many of their contemporaries, were often assisted by secretaries. Next, we wish to analyze in more detail to what extent we can discern in Hildegard's epistolary work, the influence of her last secretary, Guibert of Gembloux. Did her style undergo detectable stylistic changes under the editorial assistance of Guibert, or does the same homogeneous authorial voice appear throughout her epistolary work? Finally, we will assess the complex question to which author we should attribute, at least on

stylistic grounds, the *visiones* at stake in this article. In answering these research questions, we do not aim to develop novel stylometric techniques. The originality of this research is to be found in our application of a number of well-established techniques to assess their feasibility when dealing with medieval Latin texts, a textual tradition that until now has only rarely received attention in computational authorship attribution. Before addressing these issues, we will first briefly introduce the state of research with respect to the so-called *Mittarbeiter* problem in the Hildegard scholarship.

2 'Uneducated in the Art of Grammar'

The Benedictine nun Hildegard of Bingen was one of the most productive female authors of the Middle Ages (Newman, 1998). After a youth as anchoress at the abbey of the monks of Disibodenberg in the Rhineland near Mainz, she ended up as abbess of her own convent at the nearby Rupertsberg. Her extensive oeuvre includes genres as diverse as visionary books, letters, hagiographical texts, treatises on monastic life, musical compositions, and some works on physics and medical healing. Considered a true prophetess, receiving revelations and admonitions from God, she enjoyed a special status, even in the highest ecclesiastical milieux. Her extensive circle of correspondents, comprising, among others, popes and the emperor, testifies to her prophetic reputation. She was therefore able to gain an authority unprecedented for a woman, enabling her to even criticize the male clergy of her time. Among the first to approve her visionary gift was Bernard of Clairvaux, in a letter answering her request for support. Her female authorship was built on her recognition as a mouthpiece of God, which caused her to present herself during her entire life as a poor and uneducated woman—uneducated precisely because she was a woman (Deploige, 1998). In one of her *vitae*, her biographer Guibert of Gembloux specifies that she was 'uneducated as to her schooling in the art of grammar' (Derolez, 1988–1989, p. 377). Her status, both as a woman and an allegedly unlearned prophetess who may not have had the same type of schooling as young monks, meant that throughout her life Hildegard had to be assisted by secretaries (Ferrante, 1998).

Her first and principal secretary was Volmar of Disibodenberg, who remained her close associate until his death in 1173. He assisted in the redaction of the majority of her works. As we can learn from a famous miniature in the now lost manuscript (henceforth MS) Wiesbaden, Landesbibliothek, 1, dating from the end of her life, Hildegard dictated and wrote drafts on wax tablets, which were subsequently copied on parchment and linguistically 'polished' in accordance with the rules of grammar (Fig. 1). In addition, several Rupertsberg nuns must have aided their abbess as scribes during this period, given the number of known manuscripts produced in Rupertsberg under Hildegard's supervision (Embach, 2003, p. 76, 128–9, 160, 184–5; Herwegen, 1904, p. 302–8). After Volmar's death, Hildegard had to complete her last major visionary cycle, the *Liber divinorum operum* ('Book of the Divine Works'), with more occasional assistance by a number of different collaborators from her immediate circle of spiritual acquaintances (Herwegen, 1904, p. 308–15). At the very end of her life, however, she was unexpectedly joined by Guibert, a monk from the abbey of Gembloux in Brabant (nowadays Belgium). Himself a fervent letter writer and hagiographer (Moens, 2010), he served as her secretary from 1177 until her death in 1179 (Delehaye, 1889; Ferrante, 1998, p. 122–30).

While even the authenticity of her female authorship had not always gone uncontested, until the seminal work by Schrader and Fürhrkötter (1956), a lot of scholarly efforts have been concerned with the precise role of Hildegard's secretaries. Just as for other female writers working under the direction of father confessors (Coakley, 2006), the question has been raised to what extent Hildegard's secretaries interfered with the final versions of her works, possibly generating male, clerical interpretations rather than original female viewpoints. Following the pioneering research by Herwegen (1904), most specialists now agree that the role of Hildegard's collaborators was restricted to minor grammatical and stylistic alterations. Generally speaking, they had to copy her words *verbatim* unless they received Hildegard's

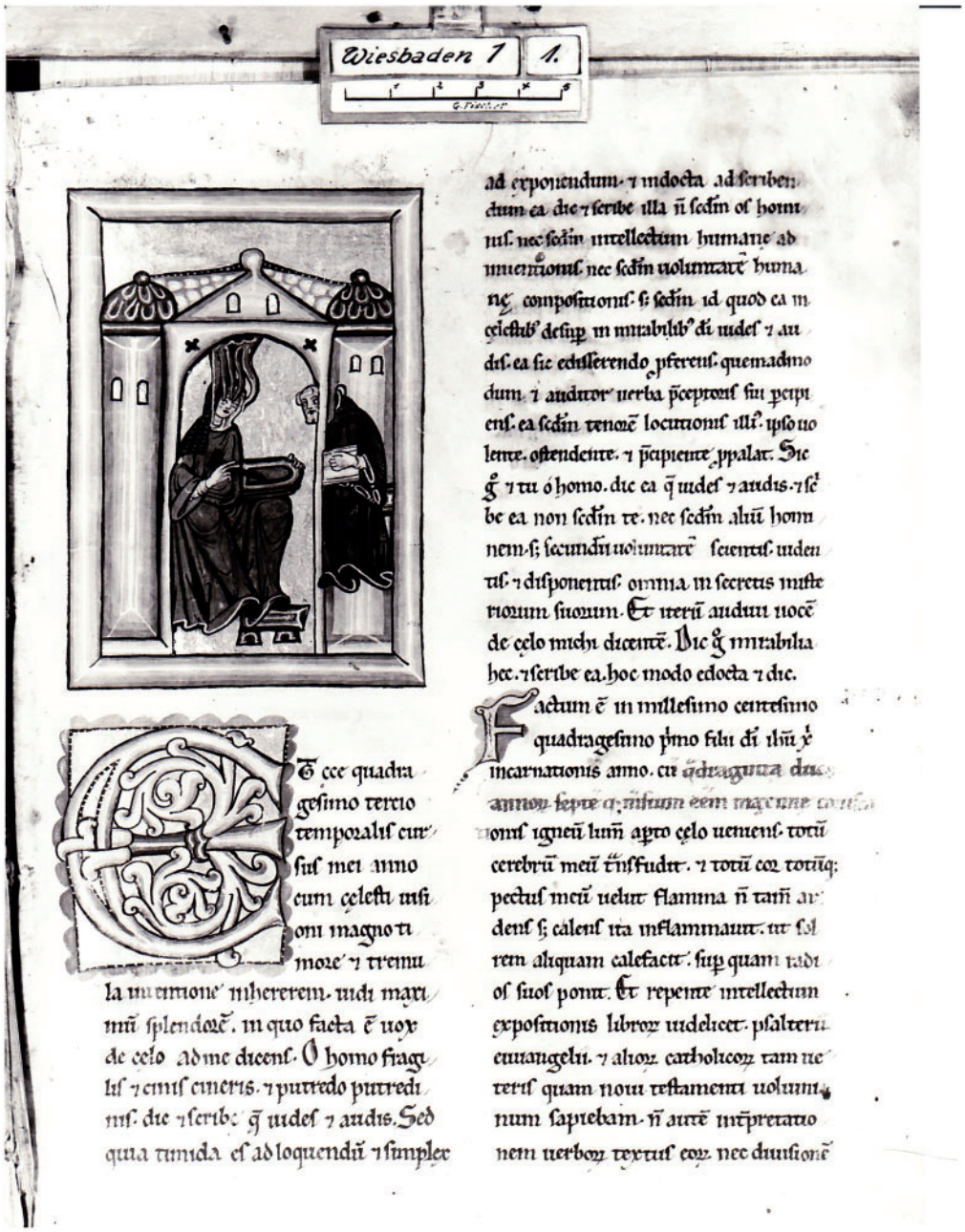

Fig. 1 MS Wiesbaden, Landesbibliothek, 1, fol. 1r. (lost since 1945). Photo: Rheinisches Bildarchiv Köln 13321

explicit authorization for corrections (Schrader and Führkötter, 1956, p. 182–3; Ferrante, 1998, p. 104).

It is generally assumed, however, that Hildegard must have granted a somewhat greater liberty to Guibert, who only entered into her life when she was already at the very advanced age of 79. Although their involvement was short, Guibert nevertheless had a significant impact on Hildegard's literary

legacy. For example, he may have assisted her as one of the correctors in the final redaction of the *Liber divinorum operum*, of which MS Ghent, University Library, 241 (Fig. 9), can be considered the autograph copy most true to Hildegard's own words (Derolez and Dronke, 1996, pp. XCI–XCIV). He also aided her in both the writing and compilation of portions of her *epistolarium*. On the basis of manuscript evidence, content, and dating, we can distinguish in Hildegard's letter collection a part that must have been written and compiled with the help of Volmar and another group of letters that must have been written or transmitted under Guibert's supervision.[1] Last but not least, Guibert is also thought to have directed the compilation of the so-called *Riesenkodex* (MS Wiesbaden, Landesbibliothek, 2), the manuscript in which, by the end of her life, Hildegard had collected all the authorized versions of her works (Van Acker, 1989, pp. 129–34).

3 Two Suspect Visions

The *Visio de sancto Martino* ('Vision of Saint Martin') and *Visio ad Guibertum missa* ('Vision sent to Guibert'), which are at stake in this article, cannot be found in the *Riesenkodex*. They are only preserved in three manuscripts that can be linked to the abbey of Gembloux and Guibert's own oeuvre.[2] Therefore, both texts are traditionally not included in the core of Hildegard's canon (Schrader and Führkötter, 1956, p. 182; Embach, 2003, p. 469). Whereas the titles in the manuscripts (Fig. 2), as well as Guibert's accompanying letters, firmly attribute these *visiones* to Hildegard, there are good reasons to suspect that Guibert must have been extensively involved in their final redaction. The figure of Saint Martin for instance—the main topic of the *Visio de sancto Martino*—is entirely absent from Hildegard's oeuvre. Guibert, on the other hand, developed a lifelong fascination for this saint and devoted nearly half of his life to spreading his cult. The *Visio ad Guibertum missa* discusses the role of the priest as well as the topic of literary collaboration, both issues of direct relevance to Guibert. Moreover, the end of the latter text contains a passage of particular interest in which Hildegard grants Guibert the exceptional right to revise her texts more fundamentally than simply at the level of style and grammar:

> When you correct [the *Visio de sancto Martino*] and the other works, in the emending of which your love kindly supports my deficiency, you should keep to this rule: that adding, subtracting, and changing nothing, you apply your skill only to make corrections where the order or the rules of correct Latin are violated. Or if you prefer—and this is something I have conceded in this letter beyond my normal practice—you need not hesitate to clothe the whole sequence of the vision in a more becoming garment of speech, preserving the true sense in every part. For even as foods nourishing in themselves do not appeal to the appetite unless they are seasoned somehow, so writings, although full of salutary advice, displease ears accustomed to an urbane style if they are not recommended by some color of eloquence (translated by Newman, 1987, p. 23)

With this statement, Hildegard allegedly granted Guibert editorial privileges that she had not allowed any other previous collaborator. The passage also prompted scholars to have a closer look at the authorship, style, and content of these visionary texts. Already in his 1882 edition, Pitra voiced doubts with respect to Hildegard's alleged authorship. He stated that Guibert, if not their original author altogether, must at least have reworked the texts profoundly. Pitra based his verdict on a number of syntactical features, on metaphors which he considered typical of Guibert, and on the extensive insertion of Biblical quotations (Pitra, 1882, p. 370–1, 375). Herwegen remained more cautious: although he accepted that Guibert had refined the texts stylistically, he still discerned Hildegard's authorial voice shimmering through Guibert's multiple corrections. He recognized Hildegard's genius in the overall structure of the visions and in some typically Hildegardian vocabulary. He also rejected Pitra's assertion that the numerous Biblical quotations could only have been inserted by Guibert (Herwegen, 1904, p. 394–6).

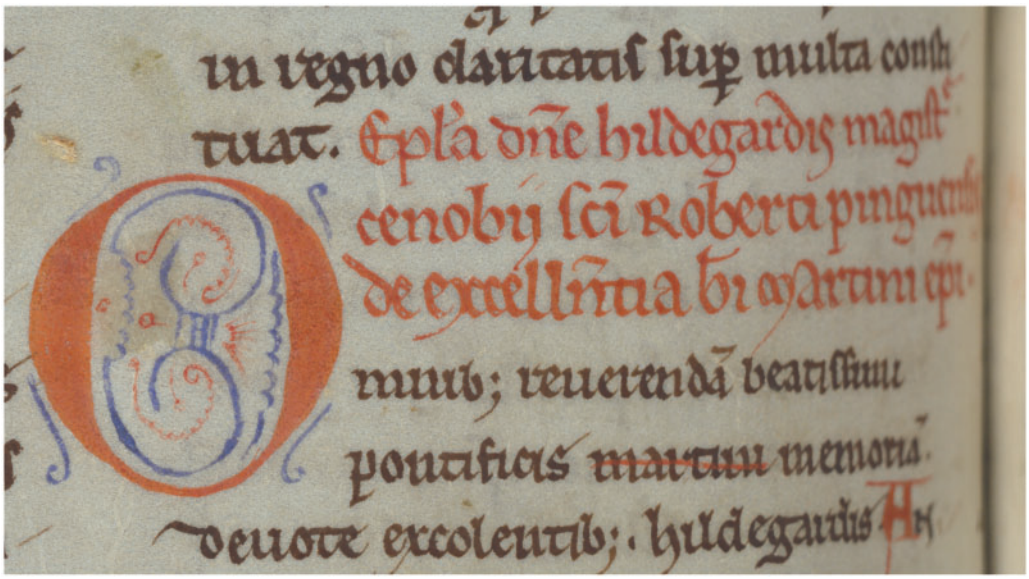

Fig. 2 MS Brussels, Royal Library, 5527–5534, fol. 141v. *Epistula domine Hildegardis magistre cenobii sancti Roberti Pinguensis de excellentia beati Martini episcopi* – 'Letter of lady Hildegard, *magistra* of the monastery of saint Rupert in Bingen, on the excellence of the blessed bishop Martin'

Newman recently stated that the *Visio ad Guibertum missa* was 'written by Guibert in Hildegard's persona' (Newman, 1987, p. 24), although Van Acker (1989, p. 130) and Coakley (2006, p. 61) continued to consider Hildegard as the text's author and Guibert as a mere stylistic reviser.

These assertions concerning the authorship of the *visiones* seem to have been predominantly based on subjective appreciations of style and content and the arguments used in this debate remain, at best, intuitive. The appearance of a new critical edition of the *visiones* once more put the question of their authorship at the forefront: should the texts be regarded as Hildegardian or pseudo-Hildegardian? Stylometric methods may provide a more objective basis for disentangling the issue and to re-assess the nature of Guibert's secretaryship.

4 Corpus Preparation

For the present study, Brepols Publishers generously provided a digital corpus containing the nearly complete works of Hildegard, Guibert, and Bernard of Clairvaux. We obtained these texts in raw format, corresponding to the way they are included in the Brepols electronic *Library of Latin Texts*, on the basis of modern critical editions.[3] Fortunately, these editions are all based on manuscripts that were compiled under the supervision of the original authors or at least in their close vicinity, so that we do not have to worry about major scribal interventions. The fact that all three authors in our corpus have been productive letter writers rendered their *epistolaria* an attractive point of departure. Moreover, the two short visionary texts of dubious origin that are at issue in this article are mostly comparable with Hildegard's letters with respect to length, topics, and manuscript tradition. Obviously, we restricted our authors' letter collections to the letters they wrote themselves, leaving aside the letters that were merely addressed to them and that were usually contained in the same manuscripts (Constable, 1976). For Bernard, this resulted in a sub-corpus of 166,063 words and for Guibert of 124,580 words.[4] Hildegard's letter collection contained 109,633 words, 82,154 of which are contained in the part compiled with the help of her first secretary

Volmar, while the remaining 27,479 words constitute the letters that, as discussed earlier, have most probably been edited in some way by Guibert.[5]

Medieval Latin is characterized by unstable orthography. As even a single scribe often used different spellings for the same word, modern editors already tend to silently normalize minor orthographic variants. We have normalized the orthography in our corpus even further via lemmatization, a useful procedure in stylometry for medieval texts (Kestemont *et al.*, 2010). The texts were first tokenized using the *Natural Language Toolkit* (Bird *et al.*, 2009). The coordinating conjunction *–que* ('and') was not realized as a separate word in medieval Latin, but it was appended to the preceding word (e.g. *terra aquaque*, 'land and water'). To automatically isolate the clitic, we have stripped the suffix ('x*que*') from every word that did not occur in a list of words proposed by Schinke *et al.* (1996, p. 180–1).[6] We have also split up the medieval contraction of the reflexive pronoun *se* and the idiomatic reinforcement *ipsum* in *seipsum* (or *teipsum*, *teipsam*, etc.).

A number of specific character combinations were freely interchangeable in medieval Latin, such as *ph* for *f*, *v* for *u*, *oe* or *ae* for *e* (or for ę, the so-called '*e caudata*') (Rigg, 1996). We have therefore lifted the difference between *v* and *u*, as well as between *ae*, *oe*, and *e*, by substituting all *v*s for *u*s and all *ae*s and *oe*s for *e*s. For the substitution of *ae* and *oe* by *e*, this actually meant that we were sometimes forced to erase the distinction between grammatically important morphemes (e.g. between the male vocative singular *domine* and the female nominative plural *dominae*). Yet, this was unavoidable, as a good deal of the *ae*s and *oe*s in our corpus were already contracted to *e*s, making it nearly impossible to automatically normalize them the other way round. Subsequently, we checked whether the surface tokens in our corpus were present in a large and representative word list from the *Perseus Project* (Tufts University). When a token was not, we used a permutation algorithm to generate plausible spelling variants for it. If one of these newly generated forms was contained in the word list, the original form was replaced by its newly generated counterpart. To generate these variants, we constructed an array with all possible variations for the consecutive character groups. Next, we combined these options through the Cartesian product in the matrix by means of a permutation algorithm (Kestemont *et al.*, 2010). Table 1 lists the series of common alternative character combinations we have considered, loosely based on Riggs (1996).[7] An example matrix for a word like *chirographum* would be: {[c], [h | Ø], [i | y], [r], [o], [g], [r], [a], [ph | f], [u], [m]}. All unique, alternative word spellings that can be generated on the basis of the matrix are: *chirographum*, *cirographum*, *chyrographum*, *cyrographum*, *chirografum*, *cirografum*, *chyrografum*, and *cyrografum*.

Table 1 Interchangeable medieval Latin character combinations allowed in our permutation algorithm

ci vs. *ti*
ch vs. *h*
ph vs. *f*
h vs. Ø
w vs. *uu* vs. *vv* vs. *uv* vs. *vu*
i vs. *j* vs. *y*
k vs. *c* vs. *ch*
g vs. *gu*

Finally, we automatically annotated the tokens with lemmas using the medieval *Index Thomisticus Treebank* (*IT-TB*: Passarotti and Dell'Orletta, 2010) as training material (ca. 170,000 tokens; ca. 9,000 sentences).[8] For the lemmatization of our corpus we have used *Morfette* (Chrupała *et al.*, 2008). Unlike other popular lemmatization tools, such as *TreeTagger* (Schmid, 1994), *Morfette* also lemmatizes input tokens that the tagger did not already encounter verbatim in the training data. Morfette considers pairs of input tokens and lemmas in the training material. From these pairs it learns 'shortest edit scripts' or ways to transform tokens into their lemmas using character insertions, deletions, and replacements. An annotated sample from the *Visio ad Guibertum missa* is listed as an example (Table 2), illustrating how this procedure did not manage to identify all lemmas correctly. Especially content words that are not typical of Thomas Aquinas's scholastic vocabulary were not always recognized. For the function words used in our analyses (see below), this problem was fortunately hardly an issue.

Table 2 Example of lemmatization based on *Morfette*

Original	Lemma	Translation
in	in	'in'
uisionem	uisio	'vision'
anime	anima	'soul'
mee	meus	'my'
,	/	/
uidi	uideo	'I see'
ingentem	ingentem	*not recognized* [ingens = 'gigantic']
rutilantis	rutilo	'glow'
ignis	ignis	'fire'
nubem	nubem	*not recognized* [nubes = 'cloud']

Translation: 'In a vision of my soul, I saw a gigantic cloud of glowing fire.'

5 Feature Selection

Today's stylometry has become an umbrella term for a still growing number of techniques for authorship analysis. Each of these has been the subject of both criticism and praise, making it hard to discern a consensus on best practice in this field. For this research too, we had to balance the pros and cons of a number of tried and tested methodologies. Recent studies still tend to agree on the undeniable methodological advantages of using function words in authorship attribution (Binongo, 2003, p. 11). An author's use of function words is said, for instance, to be relatively unaffected by a text's topic or genre. (Dis-)similarities between texts regarding function words are therefore to a certain extent content-independent and can be more easily associated with authorship than e.g. content words or other topic-specific stylistics (Juola, 2006, p. 264–5). Numerous empirical studies have effectively demonstrated that analyses of the high-frequency strata of function words yield reliable indications about a text's authorship (Koppel *et al.*, 2009, p. 11–12; Stamatatos, 2009, p. 540–1). In this research, we have therefore restricted our analyses to function words, using a number of approved methods—many of them implemented in the publicly available script suite 'Stylometry with R' (Eder *et al.*, 2013).

Preliminary analyses showed that the upper tail of the frequency spectrum in our corpus still contained a good deal of content-rich lemmas. Among the ca. 200 most frequent lemmas in our entire corpus, listed in Table 3, we came across multiple topic-specific nouns like *deus*, *dominus*, *sanctus*, . . . and verbs like *facio*, *uideo*, *uiuo*, . . . The inclusion of such lemmas obviously reflects the corpus's fairly specific, religious semantics. It is also related, however, to the simple fact that a highly inflected language like Latin with its many declensions makes less use of function words than weakly inflected languages like English. A third explanatory factor might be the fact that we worked with the frequencies of lemmas instead of surface forms. It thus seemed advisable to remove these content words from our data tables.

The content-rich words we chose to remove are marked by a hashtag (#) in Table 3.[9] The words followed by an asterisk (*) in the same Table 3 are non-reflexive personal pronouns, which are also often culled in stylometry to avoid the intrusion of genre-related or topic-specific features. Naturally, a collection of letters will contain more instances of the second-person pronouns *tu*/*vos* ('you') or *tuus*/*vester* ('your') than a saint's life. In our analyses, we have deleted this kind of pronoun. Just as in Table 2, one can still distinguish a certain number of wrongly lemmatized tokens in Table 3. The surface form *sui*, for example, often seems to have remained unchanged, whereas it should have been transformed into *suus*. This particular error, however, is neutralized by our elimination of non-reflexive personal pronouns.[10] In sum, our culling of the lemmas in Table 3 resulted in 65 function words with which to form the basis for the actual analyses.

It should be noted, however, that character *n*-grams might have been an attractive additional feature type for our research, as these have often been shown to be excellent features in authorship attribution (Koppel *et al.*, 2009, p. 12–13; Stamatatos, 2009, p. 541–2). This method, which does not require any kind of normalization or lemmatization, segments texts into consecutive, partially overlapping groups of *n* characters—the word 'bigram' for instance contains the bigrams '_b', 'bi', 'ig', 'gr', 'ra', 'am', 'm_'. Contrary to a word-level approach, character *n*-grams are also sensitive to stylistic information below the word level, like case endings or other grammatical morphemes that are

Table 3 Most frequent lemmas in the corpus (# = content words; * = non-reflexive pronouns)

et	e	quoniam	#caritas	#consilium	contra
qui	uel	#uerbum	#uenio	#rex	#pono
in	#possum	aut	quasi	dum	#amicus
#sum	pro	idem	scilicet	#talis	#honor
non	quam	super	#causa	#ceterus	#nomen
#tu*	#uester*	#terra	#manus	#caro	uelut
#is*	autem	#uolo	#iustitia	#fides	ante
#ego*	#multus	nunc	#modus	#res	#ta
#deus	#habeo	iam	#primus	#paruus	#iudicium
ad	ne	#uita	semper	apud	usque
hic	#sanctus	ac	#audio	#pax	quantum
sed	enim	#cor	#mundus	#salus	#lex
ut	etiam	#nam	#debeo	siue	#fidelis
de	#noster*	#do	#uiuo	#eternus	#sol
#suus*	#uerus	#solus	#cado	#inuenio	#celestis
#ille*	#uideo	unde	inter	#frater	#potior
a	sicut	quidem	#o	#uir	uidelicet
cum	#alius	tam	#diligo	magis	tunc
quod	ita	propter	#uoluntas	#fors	#angelus
ipse	tamen	#quidam	#gloria	#us	#diuinus
#tuus*	#filius	#bonus	quoque	#certus	#summus
#omnis	#spiritus	ergo	atque	#loquor	#ideo
si	#christus	#tempus	#aliqui	#uox	#prior
#sui*	#bonum	sine	#malum	#iustus	#populus
per	#ecclesia	nisi	#mens	post	#episcopus
#facio	#opus	#unus	#oculus	#misericordia	#similis
#homo	xque	#dies	#nihil	#celum	#os
#dico	sic	#nullus	#secundum	adhuc	#nouus
quia	#magnus	ubi	#pars	#domus	#tantum
#dominus	#iste*	#corpus	#mors	#uis	#uia
#meus*	#anima	#locus	#peccatum	#beatus	licet
nec	#pater	#uirtus	#scio	#quomodo	#predico
#quis	#gratia	#totus	#hildegars	#ueritas	#fratres
#duo	#quero				

not realized as separate words (Rybicki and Eder, 2011, p. 320). Latin, for instance, is a heavily inflected language that makes use of affixes to mark the grammatical functions of words—'*by* iron, not *by* sword' being for example '*ferr*o *non gladi*o' (Sapir, 1921, ch. VI). Therefore, it would have made sense to additionally study the character *n*-grams in the corpus.

However, one runs into the aforementioned problem that historical languages are characterized by unstable orthography (Piotrowski, 2012). Although Latin spelling variation seems to have been less pronounced than in vernacular medieval languages, it does constitute a serious issue. When comparing two texts written by the same author, surviving in manuscripts with a strongly divergent orthography, stylometric methods may detect artificially large differences. Conversely, and likewise due to scribal interference, texts of non-identical authorial provenance may show artificial similarities when they survive in manuscripts with a similar orthographical profile. In medieval manuscripts, we might even find inconsistent word spellings for the same words throughout the same text (Rigg, 1996). This ultimately implies that an approach based on character *n*-grams is unadvisable for medieval Latin (cf. Kestemont and Van Dalen Oskam, 2008). Unfortunately, this means that our approach based on lemmatization cannot take into account stylistic subtleties below the word level (e.g.

indicative versus subjunctive mood, as expressed in case endings). However, we will demonstrate that our method is still able to harvest sufficient stylistic information from the texts. Indirectly, our results will therefore even serve to emphasize how much grammatical information is in fact still expressed by isolated function words in medieval Latin.

6 Testing Principal Components Analysis

The first stylometric technique we adopt is principal components analysis (PCA), a procedure derived from multivariate statistics and commonly used to reduce the dimensionality of a data set (Binongo, 2003). By combining the original variables of a data table into new, uncorrelated compound variables or 'principal components', PCA is able to summarize large and complex data sets into insightful lower-dimensional scatterplots. When applied to the frequencies of high-frequency items in texts, this technique often successfully reveals the authorial structure in a data set. PCA's good performance in authorship attribution is due to the fact that it explicitly tries to model correlations between word frequencies. Especially the frequencies of function words show complex correlations that are related to stylistic, arguably authorial choices between small sets of alternative options. A mere visual inspection of the samples' positions in PCA scatterplots often shows that samples written by the same author will cluster, whereas groups of samples written by distinct authors lie further apart.

Because of the considerable size of the *epistolaria* in the corpus, we could start with a large sample size of 10,000 lemmatized words per sample. Recent research has demonstrated that the accuracy of most authorship attribution techniques is likely to increase when larger samples are taken (Eder, 2010; Luyckx and Daelemans, 2011). Our selection of the *epistolaria* of exactly three authors—Hildegard of Bingen, Guibert of Gembloux and Bernard of Clairvaux—respects the fact that it is theoretically unadvisable to include more than three authors in a PCA, especially when the discussion of the results is restricted to the two first Principal Components (PCS) (Binongo and Smith 1999, p. 464). As is customary since Burrows (1987), our PCA is based on the correlation matrix, appropriately scaling the original word frequencies.

Fig. 3 shows the scatterplot that results from our first experiment. Each author's samples are visualized as black letter combinations: the first letter of the author's name is followed by a digit, indicating the sample's indexed position in the respective *epistolaria*. G_EP-4, for instance, is the fourth sample of 10,000 lemmatized words taken from Guibert's *epistolarium*.[11] At this stage, we are restricting Hildegard's *epistolarium* to the letters that are not associated in any way with Guibert's secretaryship. Fig. 3 displays a remarkably clear authorial separation of the samples. Guibert's samples (G_EP) are concentrated in the upper-right quadrant, whereas the samples from Hildegard's *epistolarium* (H_EPNG) are invariably positioned to the left. Finally, Bernard's samples (B_EP) form a tight cluster of samples in the lower-right half of the plot. The density of this last cluster thus points at a clear stylistic unity, despite the fact that, as noted earlier, Bernard must have been assisted in his epistolary work by a true personal chancellery consisting of at least five different collaborators (Leclercq, 1987, p. 147–52).

Additionally, the plot in Fig. 3 contains a series of high-frequency items in light grey, the 'component loadings', visualizing how strongly the 65 lemmas have contributed to the creation of the PCS. If a word can, for instance, be found to the far left of the scatterplot, this demonstrates that it is relatively more frequent in samples with a similar position in the plot. Our first scatterplot thus shows that the use of *et* ('and') and *a* ('from') is surprisingly typical of Guibert's writings, whereas the use of the preposition *in* ('in') is very characteristic of the Hildegard samples. In comparison, the use of the lemmas *non* or *si* seems to be relatively more typical of Bernard's writing. The scatterplot does not reveal any anomalies and it is safe to assume that the high-frequency grammatical lemmas argue in favor of a clear stylistic differentiation between our authors.

The remarkable stylistic differences with respect to a number of specific lemmas used by our authors can be highlighted in another way. The boxplots in Fig. 4 visualize information about the absolute

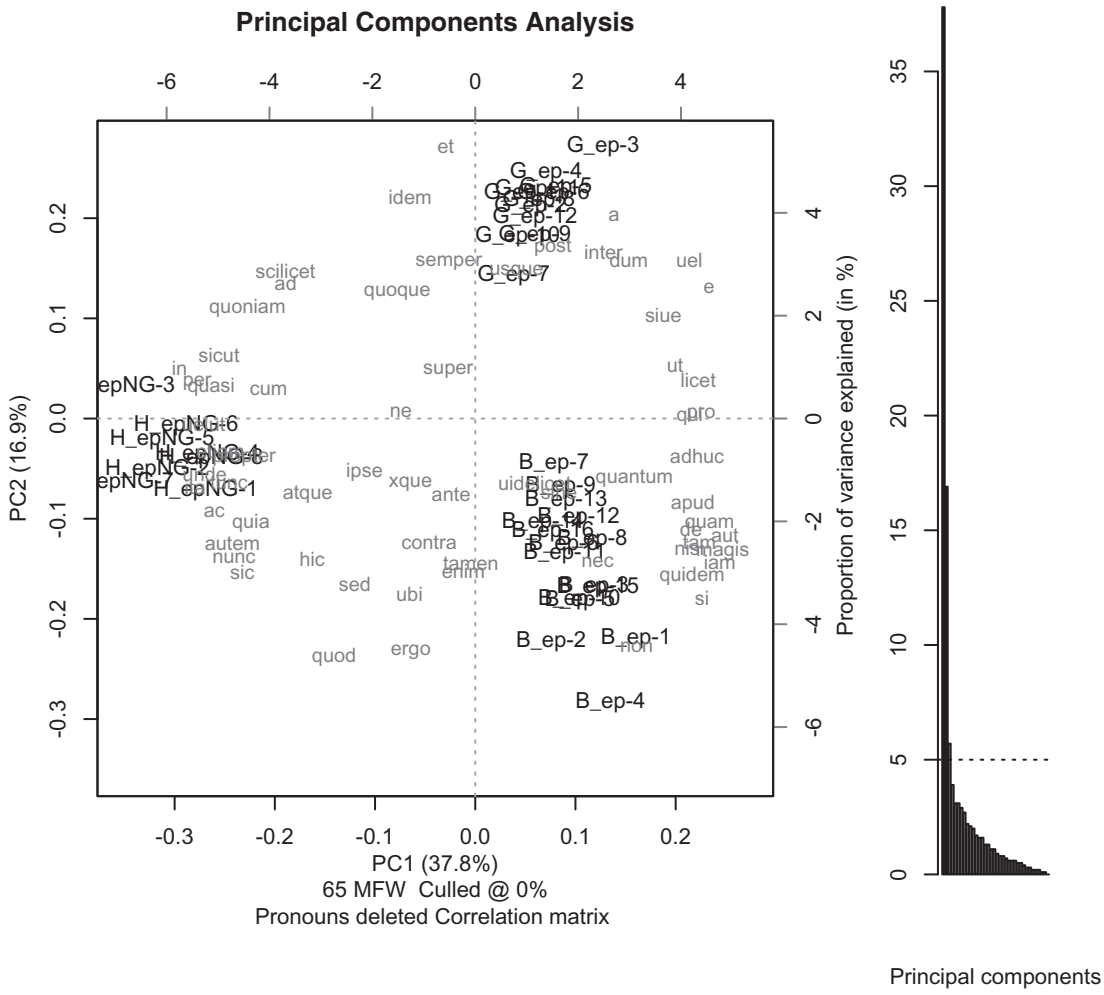

Fig. 3 PCA of the *epistolaria* by Hildegard, Guibert, and Bernard (10,000 lemmas/sample)

frequencies (medians, quartiles, etc.) for three interesting function words—*in*, *et*, and *non*—in samples of 2,000 words. In boxplot (a) concerning the use of *in*, the primary column refers to the counts in Hildegard; in the second boxplot (b) dealing with *et*, the left column concerns Guibert; and in boxplot (c), with the results for *non*, Bernard's results are displayed in first column. The secondary column in all three boxplots refers to the material by the two other authors, e.g. Guibert and Bernard in boxplot (a). These boxplots indeed reveal unmistakable differences between the respective *epistolaria* with respect to the frequency of these important function words. Interestingly, these differences coincide with stylistic observations that have been made in traditional philological research. Given the visionary discourse developed in much of her writings—even in her letters—it is not surprising to come across an intensive use of the preposition *in* in Hildegard's letters. She repeatedly sees things *in* divine visions; she continuously searches the allegorical meanings buried *in* the multitude of details that she discovers in her visions (Dronke, 1998). Guibert's writings are especially notorious for their all too inflated and artificial style, and Guibert's wearisome tendency to compose extremely long

sentences, full of coordinating conjunctions (see also Derolez, 1988, p. v and ix). Bernard's frequent use of *non* can be related to the didactic nature of his epistolary expositions in which he very often relies on an antithetical style to illustrate his thoughts (Mohrmann, 1958; Pranger, 2011, p. 222).

7 Testing Delta

For our PCA displayed in Fig. 3, we have been working with extremely generous sample sizes of 10,000 lemmas each. Because the ultimate goal of this article remains the attribution of the *Visio ad Guibertum missa* and the *Visio de Sancto Martino* of which the authorship seems very questionable, the problem of sample size needs to be put forward (Eder, 2010; Luyckx and Daelemans, 2011): while the first disputed *visio* at stake in this article still contains 7,489 lemmas, the latter only counts 3,301 words. The scatterplots in Fig. 5a and b show the results of the same procedure as in Fig. 3 but using sample sizes of 5,000 and 1,000 lemmas, respectively. This clearly illustrates the decrease in discriminatory performance of our PCA when we reduce the sample size in our experiments. Fig. 5b demonstrates that the authorial discrimination becomes less powerful, in particular between Guibert and Bernard in the vertical component.

To what extent will we be able to rely on PCA for a fairly solid attribution of a text, like the *Visio de*

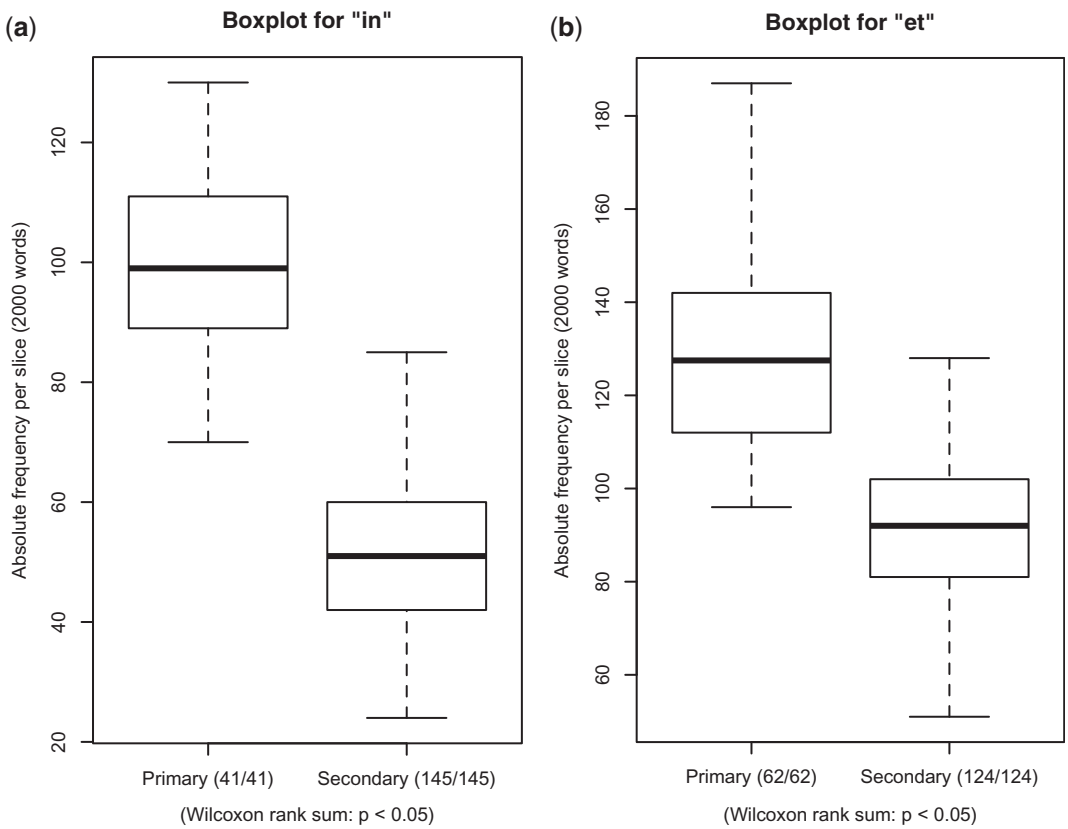

Fig. 4. (a–c) Boxplots of the absolute frequencies of *in*, *et*, and *non* in epistolary samples of 2,000 lemmas

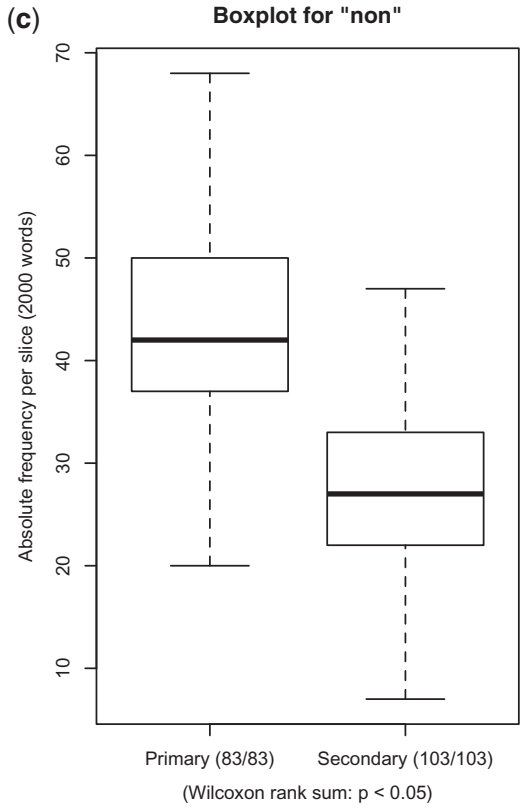

Fig. 4. Continued

Sancto Martino, of only ca. 3,000 words? Although the scatterplots in the previous section demonstrate the general validity of the stylometric approach for our corpus, it makes sense to apply a second attribution technique to our corpus to validate the outcome of the PCA more precisely. Because it is unfeasible to generate new scatterplots for every small change in parameter settings like e.g. sample size in our experiments, we additionally apply Burrows's Delta (2002) to the *epistolaria*.

In its traditional implementation, Delta offers a similarity metric to determine the authorship of anonymous works. Based on the frequencies of a small set of high-frequency items, Delta computes the stylistic distance between an unknown sample and a set of samples written by a series of candidate authors. It will attribute the anonymous sample to the author of the (single) sample in the data set to which it is closest in style according to the metric. As such, Delta uses a 'nearest neighbor' reasoning (Argamon, 2008). We can apply a 'leave-one-out validation' with Delta as follows. We can temporarily treat each sample in our collection as anonymous. Next, we can have Delta attribute the anonymized sample to one of the candidate authors and check whether the suggested attribution is successful or not. If at the end of this procedure, we divide the number of correct attributions by the total number of samples in the data set, we get a percentage that offers a useful approximation of the general effectiveness of our technique, should it, for instance, be applied to real-world samples of unknown provenance.

Fig. 6 shows the result of this leave-one-out validation for various sample sizes (multiples of 100 lemmas, ranging from 500 to 4,000). It is obvious that larger sample sizes invariably lead to higher accuracies in cross-validation. Yet, whereas the initial accuracies are fairly low (even < 85%), the attribution success quickly rises above the psychological barrier of 95% (sample sizes > 1,500 lemmas) and becomes entirely flawless when dealing with sample sizes of ca. 3,000 lemmas or more. For a text counting 3,301 lemmas, like the *Visio de sancto Martino*, we might well reach an attribution accuracy of about 99%. Moreover, because these numbers are in line with earlier reports concerning modern languages (Eder, 2010; Luyckx and Daelemans, 2011), Fig. 6 again demonstrates that even a highly inflected language like Latin contains a satisfying amount of useful stylistic information in its grammatical lemmas alone.

By now, we can assume that, when applied cautiously, PCA should offer enough solid ground to make conjectures about the authorship of the visions in the corpus traditionally attributed to Hildegard. Following a nearest neighbor reasoning (Argamon, 2008), we can plot unseen, anonymous texts together with the works of established authorial origin and investigate to which of the authorial clusters the unseen work is most similar in style. However, before moving on to the analysis of the visions, we have first tested this attribution procedure. In the PCA scatterplot in Fig. 7, we have added a new, 'anonymous' sample (amounting to 3,706

lemmas) by author 'X' to equal-sized samples from the aforementioned *epistolaria*. The new sample turns out to be stylistically much more similar to Bernard's samples than to those by Hildegard or Guibert. Should this sample have been truly anonymous, the analysis would have offered firm grounds for conjecture that the text from which the sample is derived is actually authored by Bernard of Clairvaux. In this specific case, this reasoning would have led to a historically sound attribution, as the anonymous text we have questioned is in reality the *Sermo in festo sancti Martini*, written by Bernard around 1150. An interesting fact about this example is that even though the topic and genre of this text are perhaps quite different from the epistolary material of our candidate authors (*viz.* a sermon about the aforementioned Saint Martin), it is clear that our PCA procedure allows for solid conclusions. Although one should perhaps not always expect such clear-cut stylistic, authorial differentiation in historical corpora, this promising example clearly illustrates the benefits of the present methodology for (future) research.

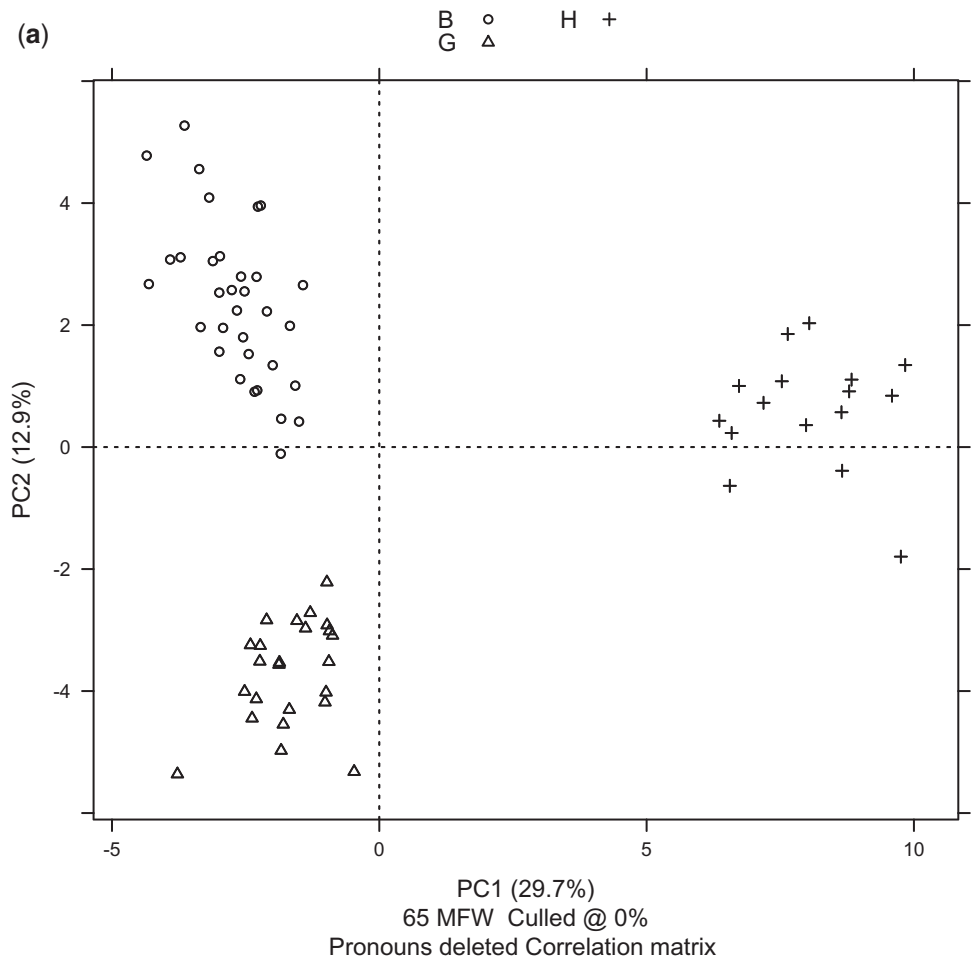

Fig. 5 (a and b) PCAs with reduced sample sizes (5,000 and 1,000 lemmas/sample)

(continued)

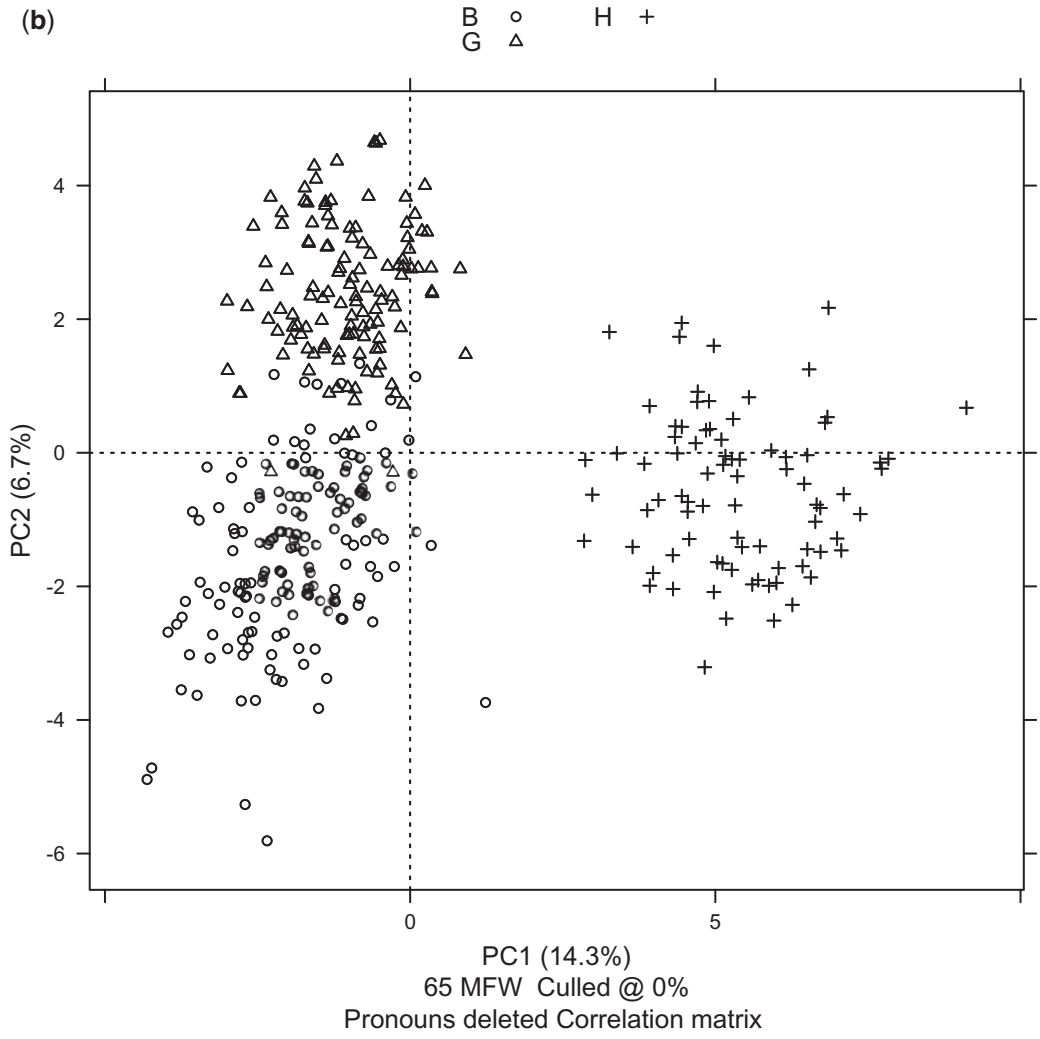

Fig. 5 Continued

8 Guibert's Secretaryship: Synergy and Beyond?

As discussed earlier, we have discerned two groups of letters in Hildegard's *epistolarium*: one that must have originated at the time when Volmar was still Hildegard's secretary and that bears no potential traces of Guibert's interference, and another containing the letters that are likely to have been revised by Guibert. If we confront samples of 5,000 lemmas from both portions, labeled here H_EPNG and H_EPG, respectively, in a PCA, we get the result in Fig. 8.

We notice that the first, horizontal PC captures an impressive 37% of the original variation in our data and primarily relates to the stylistic differentiation between Guibert's own letter collections (G_EP) and the anterior portion of Hildegard's *epistolarium* (H_EPNG). Interestingly, we see that the second PC in the right half of the plot (still capturing 9.4% of the original variation) discriminates between Hildegard's non-Guibertian letters and her letters that can be associated with Guibert's secretaryship.

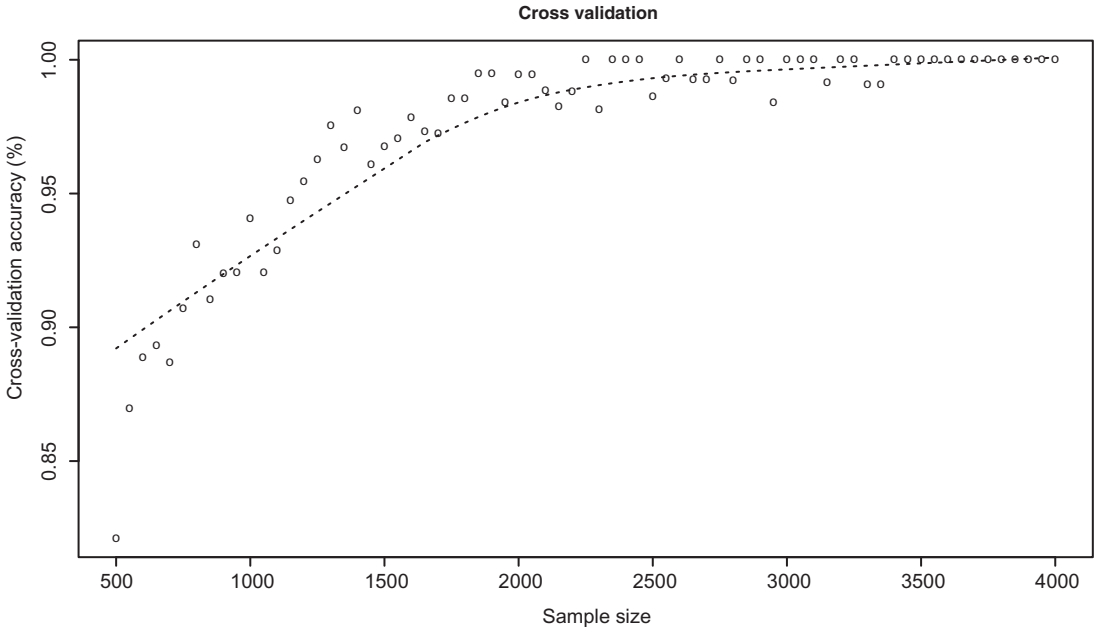

Fig. 6 Cross-validation using Delta (dotted lowess line fitted)

These results thus suggest that there do indeed exist stylistic differences between the oldest portion of Hildegard's *epistolarium* and the letters in which we expected to discern Guibert's editorial fingerprints. They also confirm what can be deduced from the surviving manuscript evidence. The so-called autograph copy of the *Liber divinorum operum* mentioned earlier offers unique insight into the way in which Hildegard's collaborators must have edited her texts under her supervision (Derolez, 1972). Fig. 9, showing a number of lines from the randomly selected page 370 of MS Ghent, University Library, 241, makes it clear that it was the function words in particular that were often altered by Hildegard's correctors; *tam* being erased, *quod* being replaced by *ut* or *quia*, *ad* being added, *etcetera*. A collaborator—especially Guibert, who is known to have had a great deal of freedom in his editorial work—may thus have had a notable impact on Hildegard's stylistic profile.

However, in Fig. 8, we see that the samples from Hildegard's *epistolarium* that bear the influence of Guibert's interference do not seek the company of Guibert's own writings in the scatterplot. After all, they continue to be somewhat more similar to Hildegard's style. This result is reminiscent of the Synergy Hypothesis, recently discussed by Pennebaker (2011).[12] Pennebaker puts forward three hypotheses concerning the stylistic effect of collaborations between different authors. Such projects can produce a language that is (1) similar to the one produced by a single person writing alone, (2) the average of the two writers, or (3) unlike either of one of the styles that the collaborating authors would produce on their own. Based on exploratory research on the Federalist papers and Beatles songs, Pennebaker ultimately argues in favor of the latter, so-called 'Synergy view' on collaborative authorship, not refuting however the possibility that one of the collaborating authors might have remained more influential with respect to the end product (cf. Petrie *et al.*, 2008). This Synergy Hypothesis thus might be applicable to a certain extent to the Hildegard–Guibert 'collaboration', where the result of the creative process does not fit in with the other letter samples written by Hildegard or Guibert individually, although the result is somewhat more similar to Hildegard.

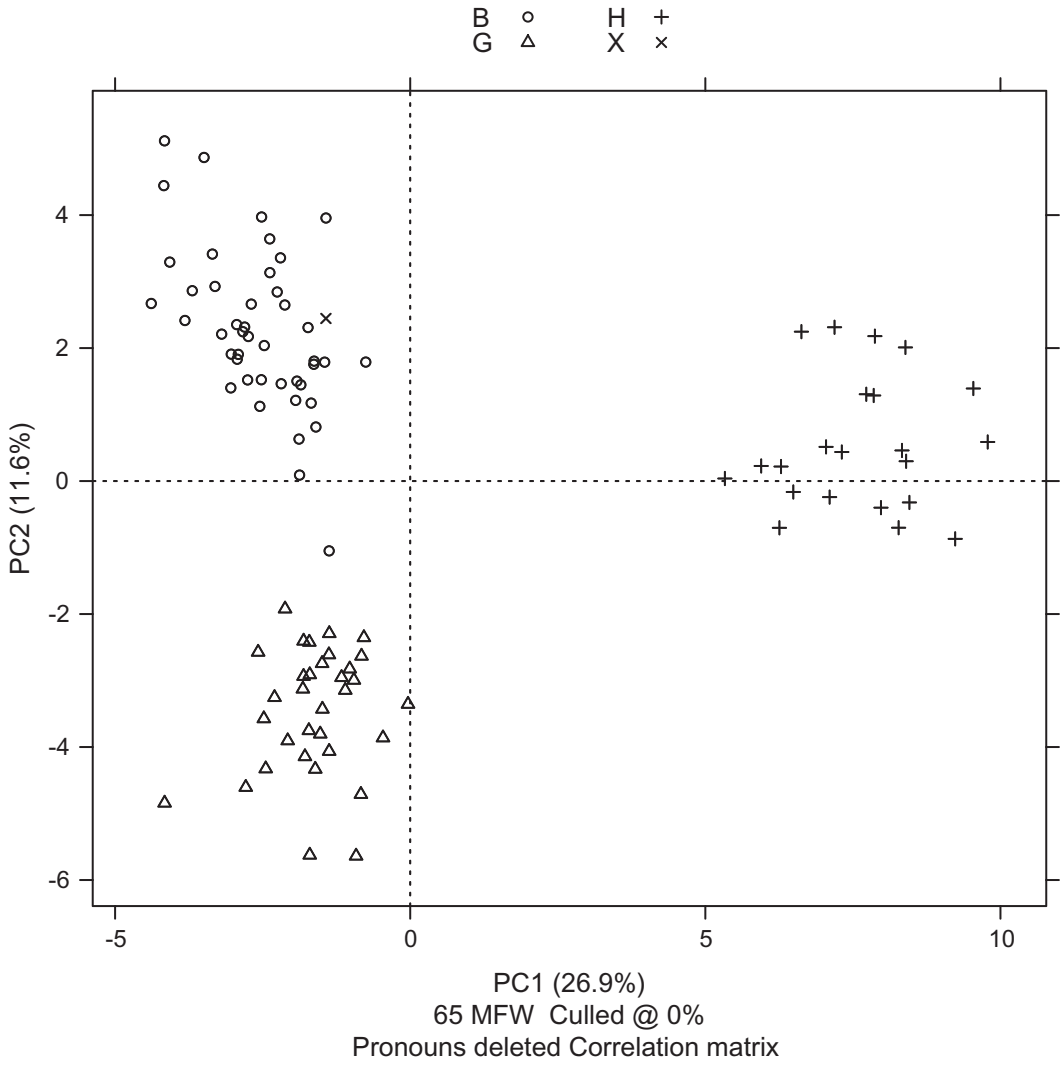

Fig. 7 Attribution of an anonymized *sermo* X to the Bernardian corpus

More can be learned about the stylistic dichotomy in Hildegard's *epistolarium* by applying a Mann–Whitney test to the lemmas occurring at least twice in 4,000 lemma samples. Here, we temporarily leave the realm of high-frequency lemmas and venture into the lower-frequency strata of the lexical spectrum. Hence, this test will not particularly emphasize the discriminatory power of high-frequency lemmas, as was the case with our other tests (Kilgariff, 2001). Fig. 10 contrasts the words that were predominantly used in the Hildegard's letters written under Volmar's secretaryship with those that become typical when Guibert took over the editorial work in the preservation of her letters. The lemmas have been ranked and plotted according to the *U* test statistic obtained for each lemma. Fig. 10 learns how the use of the relative pronoun *qui* ('who') for instance only becomes prominent in letters edited by Guibert, who is indeed notorious for constructing eloquent but complex sentences with a lot of embedded relative clauses. Moreover, this latter group of letters is also characterized by a

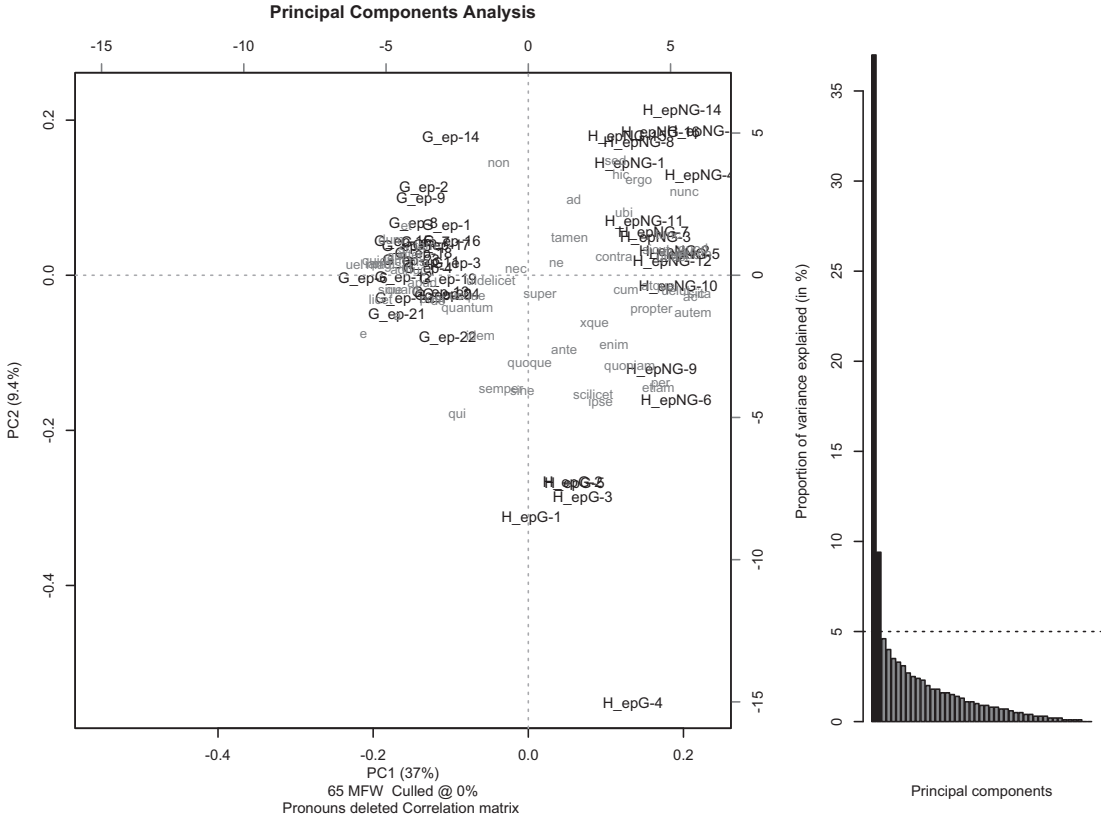

Fig. 8 PCA of the *epistolarium* of Guibert, of the letters of Hildegard transmitted without Guibert's editorial assistance, and of the Guibertian letters in Hildegard's *epistolarium* (5,000 lemmas/sample)

more dry and stereotypical ecclesiastical vocabulary (*omnipotens*, *sanctus*, *spiritus*, *verus*, . . .), whereas the letters not influenced by Guibert betray a more direct and lively narrative style (*sed*, *tunc*, *nunc*, *dico*, *ergo*, *deinde*, . . .), possibly more true to Hildegard's own preferred way of expressing herself. We might thus be inclined to agree with Newman (1987, p. 24) when she stated: 'Purists can at least rejoice that the collaboration [between Guibert and Hildegard] began only after the seer's major works were completed'. From the methodological point of view, these results also show that the discriminatory effects in lower-frequency strata correspond with the stylistic dichotomy present in the high-frequency vocabulary, thus corroborating the performance of the latter methodology.

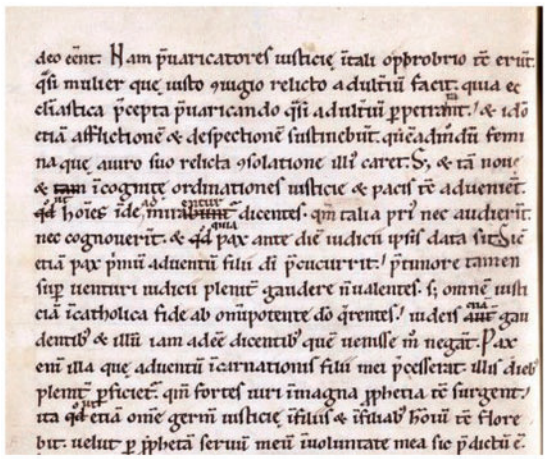

Fig. 9 MS Ghent, University Library, 241, p. 370 (detail). Reproduced with permission

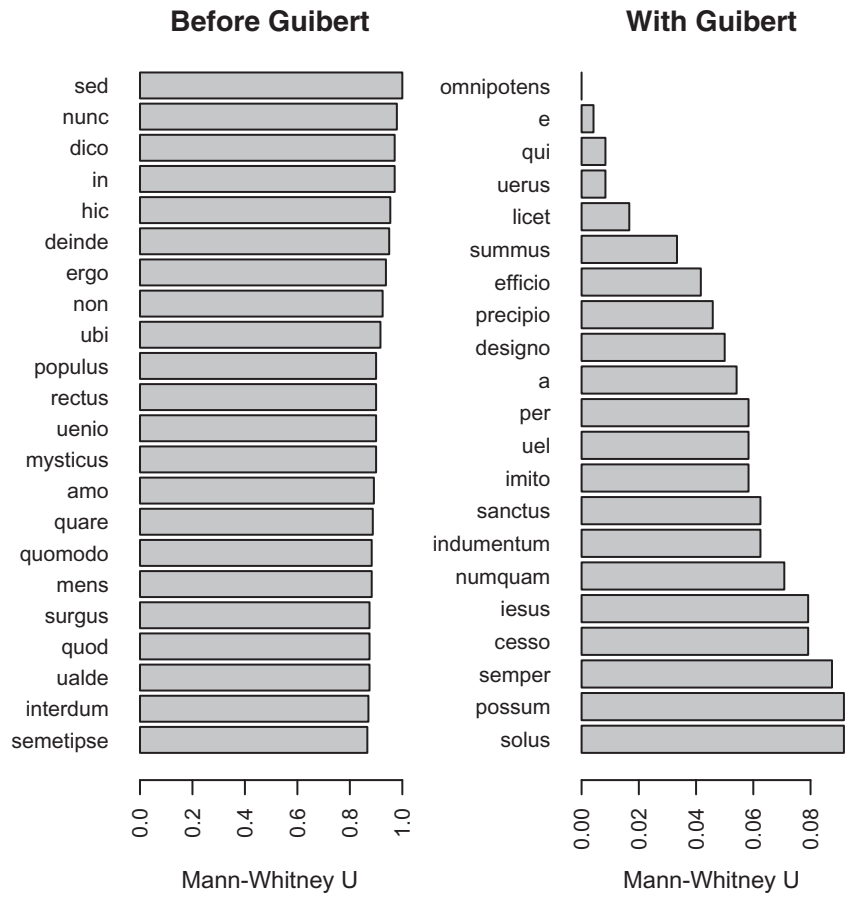

Fig. 10 Results of Mann–Whitney test (*U* statistic) comparing the vocabulary in Hildegard's *epistolarium* before and during Guibert's secretaryship

Let us finally turn to the original incentive for the present article, namely, the authorship discussion concerning two texts of dubious provenance: the relatively short *Visio de Sancto Martino* about Saint Martin (3,301 lemmas) and the somewhat longer *Visio ad Guibertum missa* (7,492 lemmas). Fig. 11 offers the result of three PCAs in which we have confronted both '*dubia*' (hence D_Mart and D_Missa) with the previously discussed epistolary collections, again using the same 65 lemmas and a sample size of 3,301 lemmas. Fig. 11a considers all texts by all authors; Fig. 11b excludes Bernard's texts; Fig. 11c only considers Guibert's *epistolarium* and the 'anonymous' visionary texts.

All subplots in Fig. 11 clearly show that both visions tightly cluster with Guibert's *epistolarium*, instead of with Hildegard's. This effect is perhaps least prominent in Fig. 11a, where D_Mart and D_Missa display modest similarities to some of the epistolary samples from the portion of Hildegard's *epistolarium* that was revised by Guibert. In all three plots, however, the visions are generally speaking far more similar to Guibert's writings than to Hildegard's. Significantly, most samples resulting from the combined authorial voices of Hildegard and Guibert again do not display any significant rapprochement to the *epistolaria* of the individual authors. These observations seem to reinforce the Synergy Hypothesis. Moreover, the visions' quasi-

random position in the final subplot (Fig. 11c) reveals no pronounced stylistic differences with Guibert's letters, regarding the high-frequency lemmas analyzed. They invariably cluster with Guibert's epistolary oeuvre, making him a much more plausible author than Hildegard—at the very least, from a stylistic point of view.

An important, yet inconspicuous, last feature of Fig. 11a is that it includes the *Sermo in festo sancti Martini*, even though it can hardly be spotted among Bernard's other samples. This sermon deals, just like the *Visio de Sancto Martino*, with Saint Martin. Both texts were even clearly influenced by the same late Antique hagiographical narratives concerning this saint, namely, the works of his first hagiographers Sulpicius Severus (c. 363–425) and Gregory of Tours (538–594). It is interesting to note that despite their interwovenness within the same intertextual tradition, they are still clearly distinguished and therefore demonstrate that topic-related stylistics hardly interferes with the author-related differences. The visionary texts under investigation thus betray Guibert's stylistic influence to such an advanced extent that we could wonder whether we should not entirely attribute these texts to Guibert, instead of arguing for any form of 'synergetical collaboration', as was still possible for the portion of the *epistolarium* over which both Hildegard and Guibert labored.

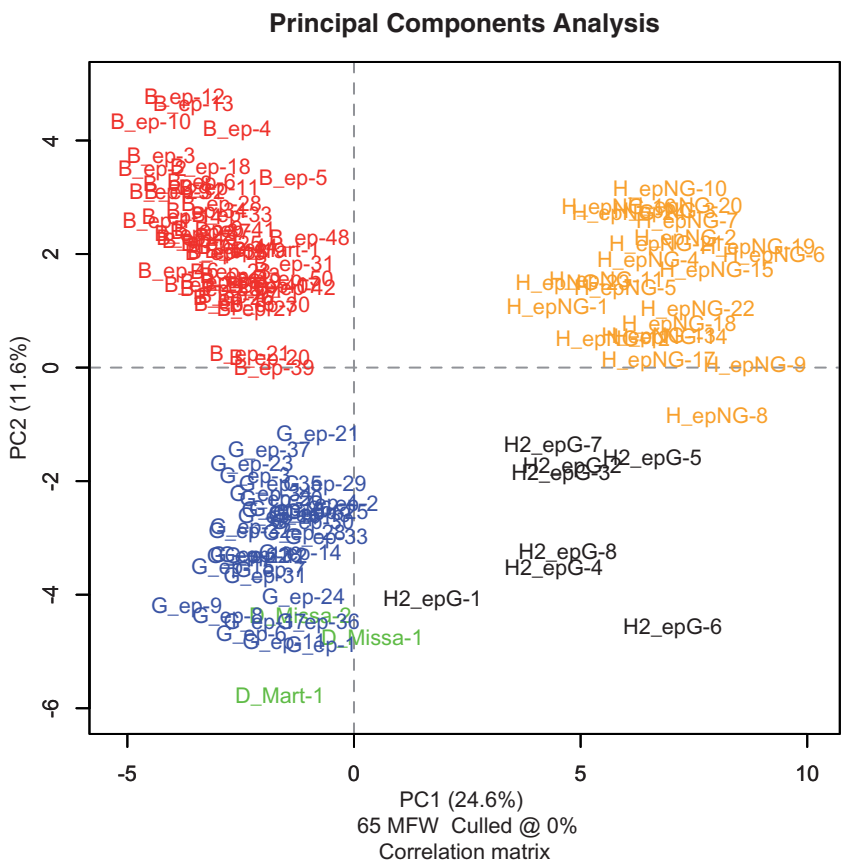

Fig. 11 PCAs including the *Visio de Sancto Martino* and the *Visio ad Guibertum missa*

(continued)

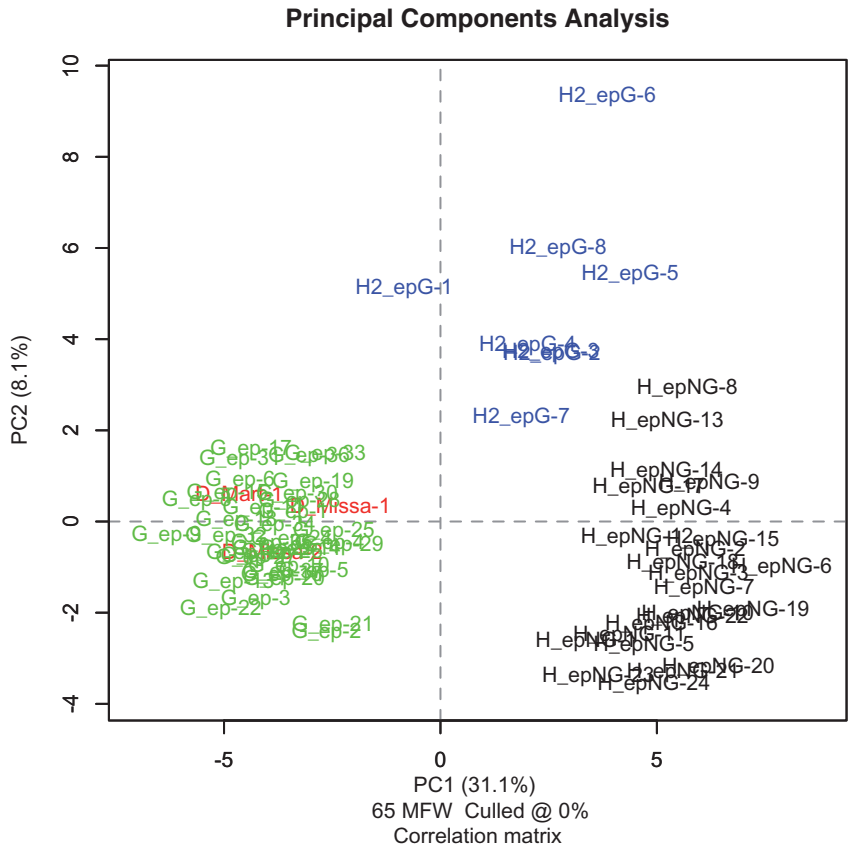

Fig. 11 Continued

9 Conclusions

It is obvious that the experiments reported in this article only touch the tip of the iceberg of the research on Hildegard's complicated authorship, to say nothing of the exciting, broader topic of twelfth-century Latin writing. As stated in our Introduction, individuality and authorship remain complex issues when it comes to medieval literature. Even an authoritative and highly idiosyncratic author like Bernard of Clairvaux is known to have been assisted by a team of collaborators. It is moreover clear that medieval scribes often gradually introduced errors and deviations when successively copying exemplars, thus possibly altering the original authors' style in the surviving copies of texts. Nevertheless, we hope to have demonstrated that these issues do not need to imply that stylometry, when applied cautiously, cannot yield valid research results in the field of medieval philology.

First we showed that authorial discrimination was possible in the corpus studied. Although samples had to be big enough to yield correct attributions, stylometric methods were generally able to model the overall differences in writing style. This suggests that superficial interference from scribes (or even later editors) can be by-passed to a certain extent, for instance through lemmatization. Interestingly, we obtained satisfying results with a word-level approach, notwithstanding the fact that Latin is a highly inflected language. Although other strategies might increase attribution accuracies in the future, this shows that even in highly inflected

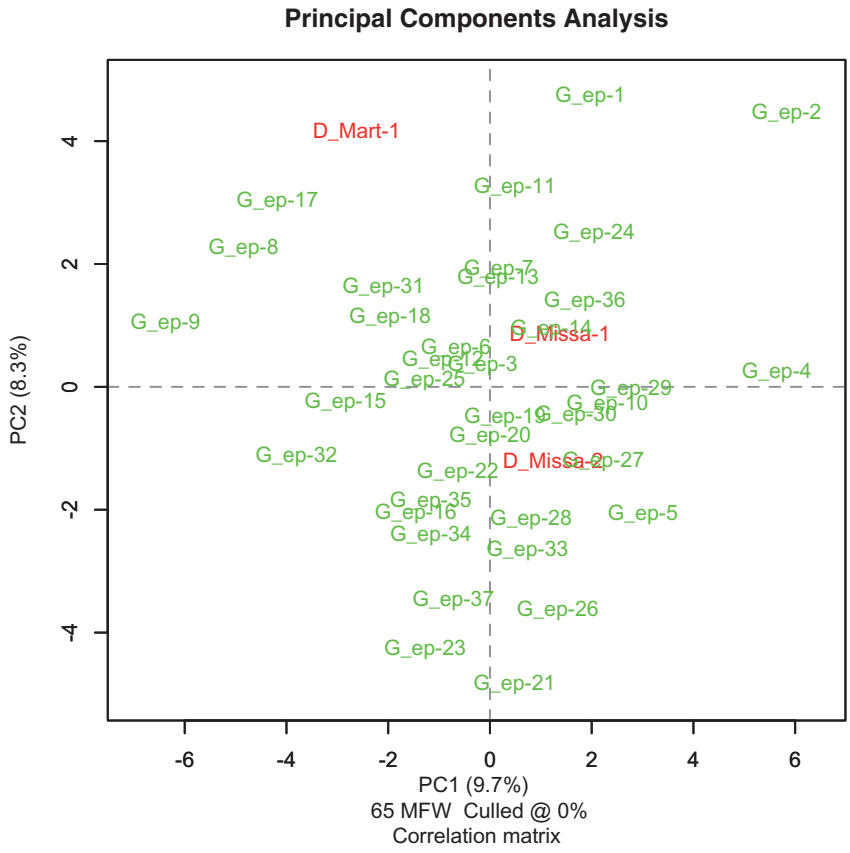

Fig. 11 Continued

languages, plenty of stylistic information can already be harvested at the word-level.

In the course of our research, we have also touched on collaborative authorship, an issue that recently has raised considerable interest in stylometry (Reynolds *et al.*, 2012). Our methodology enabled us to discover clear stylistic differences in Hildegard of Bingen's epistolary work between those letters for which she had relied on the modest assistance of her first collaborator Volmar and the letters that have been compiled and copy-edited by Guibert of Gembloux. Interestingly, the letter samples influenced by the collaboration between Hildegard and Guibert formed an isolated cluster that did not display advanced stylistic similarities to Hildegard's former epistolary oeuvre, nor to that of Guibert. These results argue in favor of what Pennebaker (2011) has called the Synergy Hypothesis: when two authors are involved in the same texts, the end result need not resemble the writing style of one of the two individually; the result might rather resemble that of a 'new', third author. The evidence offered in this particular case study is valuable in this light, but at the same time still too scant to come to a final verdict on this fascinating topic.

Finally, with respect to our initial research question, we hope to have convincingly disputed the authorship of two texts allegedly attributed to Hildegard: the *Visio de Sancto Martino* and the *Visio ad Guibertum missa*. We argued that these visions are stylistically speaking completely in line with the writing style of Guibert de Gembloux, Hildegard's last secretary. These results offer

quantitative support to suspicions voiced in earlier, traditional philological research: if Guibert is not to be considered their original author altogether, it is clear that he reworked these texts so profoundly that hardly anything of Hildegard's writing style is still discernible in them. In fact, it is noteworthy that our analyses could not offer any stylistic evidence at all that Hildegard once authored (even a preliminary or simply oral version of) these texts, although this remains of course an interesting historical possibility.

Acknowledgements

We thank the Corpus Christianorum Library & Knowledge Centre of Brepols (Turnhout) and in particular Luc Joqué for generously putting at our disposal the corpora analyzed in this article. Marco Passarotti (Università Cattolica del Sacro Cuore, Milan) generously provided us with the *IT-TB*, while Helma Dik (University of Chicago) provided the word list from the *Perseus Project* (Tufts University). We are moreover very grateful for the valuable feedback from Albert Derolez, Wim Verbaal, Antoon Bronselaer, and Guy De Tré. In addition, we thank the anonymous reviewers of the *Digital Humanities Conference 2013* for their helpful comments on this research project, as well as the anonymous reviewers of this journal, in particular, for their extensive feedback on the normalization procedures described. Mike Kestemont developed the stylometric methodology for this article. Sara Moens brought in her domain expertise concerning Guibert of Gembloux and medieval epistolography. Jeroen Deploige, who took the initiative for this collaborative research, contributed from his involvement with Hildegard scholarship. All three authors contributed equally to the end result.

Funding

This work was supported by the Research Foundation – Flanders, of which both Sara Moens and Mike Kestemont are fellows, and by the Flemish Hercules Foundation, which finances the project 'Sources from the Medieval Low Countries (SMLC)', directed by Jeroen Deploige.

References

Argamon, S. (2008). Interpreting Burrows's Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Binongo, J. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, **16**(2): 9–17.

Binongo, J. and Smith, W. (1999). The application of principal components analysis to stylometry. *Literary and Linguistic Computing*, **14**(4): 446–66.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. Sebastopol: O'Reilly.

Burrows, J. (1987). *Computation into Criticism. A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Cerquiglini, B. (1999). *In Praise of the Variant: A Critical History of Philology*. Baltimore: JHU Press.

Chrupała, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with Morfette. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010*. Marrakech, Morocco: European Language Resources Association, pp. 2362–7.

Coakley, J. (2006). *Women, Men and Spiritual Power: Female Saints and Their Male Collaborators*. New York: Columbia University Press.

Constable, G. (1976). *Letters and Letter-collections*. Turnhout: Brepols.

Delehaye, H. (1889). Guibert, abbé de Florennes et de Gembloux, XIIe et XIIIe siècles. *Revue des Questions Historiques*, **46**: 5–90.

Deploige, J. (1998). *In Nomine Femineo Indocta. Kennisprofiel en Ideologie van Hildegard van Bingen (1098-1179)*. Hilversum: Verloren.

Deploige, J. (2005). Anonymat et paternité littéraire dans l'hagiographie des Pays-Bas Méridionaux (ca. 920 - ca. 1320). Autour du discours sur l''original' et la 'copie' hagiographique au Moyen Âge. In Renard, E., Trigalet, M., Hermand, S., and Bertrand, P. (eds),

Scribere Sanctorum Gesta. Recueil d'études d'hagiographie médiévale offert à Guy Philippart. Turnhout: Brepols, pp. 77–107.

Deploige, J. and Moens, S. (eds), Visio de Sancto Martino et Visio ad Guibertum missa. In Deploige, J., Embach, M., Evans, C., Gärtner, K., and Moens, S., *Hildegardis Bingensis opera minora. Pars secunda*. Turnhout: Brepols, forthcoming.

Derolez, A. (1972). The genesis of Hildegard of Bingen's *Liber divinorum operum*. The codicological evidence. In Gumbert, J.P. and De Haan, J.M. (eds), *Litterae Textuales. Essays Presented to Gerard I. Lieftinck. II: Texts & Manuscripts*. Amsterdam: Van Ghent, pp. 23–33.

Derolez, A. (ed.) (1988–1989). *Guiberti Gemblacensis epistolae: quae in codice B.R. BRUX. 5527-5534 inveniuntur*. Turnhout: Brepols.

Derolez, A. and Dronke, P. (eds), (1996). *Hildegardis Bingensis Liber Divinorum Operum*. Turnhout: Brepols.

Dronke, P. (1998). The allegorical world-picture of Hildegard of Bingen: revaluations and new problems. In Burnett, C. and Dronke, P. (eds), *Hildegard of Bingen: The Context of Her Thought and Art*. London: The Warburg Institute.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Digital Humanities 2010. Conference Abstracts.* King's College London, pp. 132–5.

Eder, M., Kestemont, M., and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013. Conference Abstracts.* University of Nebraska-Lincoln, pp. 487–89.

Embach, M. (2003). *Die Schriften Hildegards von Bingen*. Berlin: Akademie Verlag.

Ferrante, J. (1998). *Scribe quae vides et audis*. Hildegard, Her Language, and Her Secretaries. In Townsend, D. and Taylor, A. (eds), *The Tongue of the Fathers. Gender and Ideology in Twelfth-Century Latin*. Philadelphia: University of Pennsylvania Press, pp. 102–35.

Herwegen, I. (1904). Les collaborateurs de Ste. Hildegarde. *Revue Bénédictine*, **21**: 192–204; 302–15; 381–403.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Kestemont, M. and Van Dalen-Oskam, K. (2009). Predicting the past: memory-based copyist and author discrimination in medieval epics. In Calders, T., Tylus, K., and Pechenizkyi, M. (eds), *Proceedings of BNAIC 2009*. Eindhoven: Benelux Association for Artificial Intelligence, pp. 121–8.

Kestemont, M., Daelemans, W., and De Pauw, G. (2010). Weigh your words—memory-based lemmatization for middle Dutch. *Literary and Linguistic Computing*, **25**(3): 287–301.

Kilgariff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**(1): 49–66.

Klaes, M. (ed.) (2001). *Hildegardis Bingensis Epistolarium. Pars III*. Turnhout: Brepols.

Köhler, R. (2005). Synergetic linguistics. In Köhler, R., Altman, G., and Piotrowoski, R. G. (eds), *Quantitative Linguistik/Quantitative Linguistics. Ein Internationales Handbuch/An International Handbook*. Berlin, New York: Walter de Gruyter, pp. 760–75.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.

Leclercq, J. (1962). Saint Bernard et ses secrétaires. In *Recueil d'études sur Saint Bernard et ses écrits*, Vol. 1. Rome: Edizioni di storia e letteratura, pp. 3–25.

Leclercq, J. (1987). Lettres de S. Bernard: histoire ou litterature? In *Recueil d'études sur Saint Bernard et ses écrits*, Vol. 4. Rome: Edizioni di storia e letteratura, pp. 125–225.

Leclercq, J. and Rochais, H. (eds), (1974–1977). Epistolae In *Sancti Bernardi opera*, Vols 7–8. Rome: Editiones cistercienses.

Leclercq, J., Talbot, C. H., and Rochais, H. (eds), (1957–1977). In *Sancti Bernardi opera*. Rome: Editiones cistercienses.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, **26**(1): 35–55.

Moens, S. (2010). Twelfth-century epistolary language of friendship reconsidered. The case of Guibert of Gembloux. *Revue belge de Philologie et D'histoire*, **88**(4): 983–1017.

Mohrmann, C. (1958). Observations sur la langue et le style de saint Bernard. In *S. Bernardi opera*, Vol. 2. Rome: Editiones cistercienses, pp. IX–XXXIII.

Newman, B. (1987). *Sister of Wisdom. St. Hildegard's Theology of the Feminine*. LA: University of California Press.

Newman, B. (ed.) (1998). *Voice of the Living Light: Hildegard of Bingen and Her World*. LA: University of California Press.

Nichols, S. (1997). Why Material Philology? Some Thoughts. *Zeitschrift für deutsche Philologie*, **116**: 10–30.

Passarotti, M. and Dell'Orletta, F. (2010). Improvements in Parsing the *Index Thomisticus* Treebank. Revision, Combination and a Feature Model for Medieval Latin. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D. (eds), *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17–23 May 2010*. Valetta: European Language Resources Association, pp. 1694–71.

Pennebaker, J. (2011). *The Secret Life of Pronouns. What our Words Say About Us*. NY: Bloomsbury.

Petrie, K., Pennebaker, J., and Sivertsen, B. (2008). Things we said today: a linguistic analysis of the Beatles. *Psychology of Aesthetics, Creativity, and the Arts*, **2**(4): 197–202.

Pitra, J. B. (1882). *Analecta Sacra et Classica Spicilegio Solesmensi Parata*, Vol. 8. Paris: A. Jouby et Roge.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. California: Morgan & Claypool Publishers.

Pranger, B. (2011). Bernard the Writer. In McGuire, B.P. (ed.), *A Companion to Bernard of Clairvaux*. Leiden: Brill, pp. 220–48.

Reynolds, N., Schaalje, G., and Hilton, J. (2012). Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English Works. *Literary and Linguistic Computing*, **27**(4): 409–25.

Rigg, A. (1996). Orthography and pronunciation. In Mantello, F. and Rigg, A. (eds), *Medieval Latin: An Introduction and Bibliographical Guide*. Washington: The Catholic University of America Press, pp. 79–82.

Rybicki, J. and Eder, M. (2011). Deeper delta across genres and languages: Do we really need the most frequent words? *Literary and Linguistic omputing*, **26**(3): 315–21.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace & Co..

Schinke, R., Greengas, M., Robrtson, A. M., and Willett, P. (1996). A stemming algorithm for Latin text databases. *Journal of Documentation*, **52**(2): 172–87.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Schrader, M. and Führkötter, A. (1956). *Die Echtheit des Schrifttums der heiligen Hildegard von Bingen. Quellenkritische Untersuchungen*. Keulen–Graz: Böhlau Verlag.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

Van Acker, L. (1989). Der Briefwechsel der heiligen Hildegard von Bingen. Vorbemerkungen zu einer kritischen Edition. *Revue Bénédictine*, **99**: 118–54.

Van Acker, L. (ed.) (1991–1993). *Hildegardis Bingensis Epistolarium*. Turnhout: Brepols.

Notes

1. Among the letters written with the help of Volmar, we count those in MS Wien, Österreichische Nationalbibliothek, 963 (theol. 348), which offers a copy of a collection compiled by Volmar before 1173 (Van Acker, 1991, p. XXVI), and the limited number of letters that can be found distributed over MS Stuttgart, Württembergische Landesbibliothek, Cod. theol. phil. 4° 253; MS Wien, Österreichische Nationalbibliothek, 881; MS Berlin, Staatsbibliothek Preussischer Kulturbesitz, Cod. theol. lat. fol. 699; MS London, British Library, Cod. Add. 17292; MS Paris, Bibliothèque Nationale, Nouv. Acquis. Lat. 760; MS Trier, Stadtbibliothek, Cod. 771/1350 and MS Kynžvart, Cod. 40. Among the letters compiled and edited under Guibert's supervision, we count those in the Riesenkodex Wiesbaden, Landesbibliothek, 2 (dating from 1177-1179/1180), that are *not* also found in MS Wien, Österreichische Nationalbibliothek, 963 (theol. 348) (Van Acker, 1991, p. XXVII), as well as those copied in MS Berlin, Staatsbibliothek Preussischer Kulturbesitz, Cod. lat. 4° 674, which bear traces of Guibert's editorial assistance (Klaes, 2001, p. XVII). Among the letters contained in the latter group, compiled under Guibert's supervision, we obviously encounter all Hildegard's letters addressed to Guibert and the ones that have been written in the years in which he stayed in Rupertsberg.
2. MSS Brussels, Royal Library, 5397–5407 and 5527–5534 (both originating from Gembloux, early thirteenth century) and MS Brussels, Royal Library, 1510–1519 (originating from Sint-Maartensdal near Louvain, fifteenth century).
3. See www.brepolis.net. The critical editions of the works of both Hildegard of Bingen and Guibert of Gembloux are published in several volumes in Brepols's own *Corpus Christianorum* series. For the works of Bernardus, the Brepols *Library of Latin Texts* relies on Leclercq *et al.* (1957–1977).
4. Bernard's letters, edited by Leclercq and Rochais (1974–1977), contain the 'official' *epistolarium*,

compiled shortly after Bernard's death, as well as letters transmitted elsewhere. Guibert's letters were edited by Derolez (1988–1989) on the basis of MS Brussels, Royal Library, 5527–5534.

5. See note 1. Hildegard's letters are edited by Van Acker (1991–1993) and by Klaes (2001)
6. We supplemented this list with three words—*plerumque*, *utrumque*, and *quicumque*—yet did not allow any of these items into the restrictive set of function words we list below. We did not consider other, much less frequent clitics (e.g. *–ne* ('if') or *–ve/ue* ('or')), because it is difficult to automatically detect these using a simple rule-based approach and to distinguish them from e.g. the *–ne* in *deuotione* or the *–ue* in *serue*.
7. We have described our approach in a generic way for future reference. It should be noted, however, that there still remains a small number of possible spelling variants in medieval Latin that are hard to deal with but that were not relevant for the present research because we worked with critical editions that have already normalized orthography to a large extent. One can think here of the interchangeability of *–mqu–* and *–nqu–* in some words and the problem of single/double consonants (as e.g. in *litera* and *littera*). A lesser frequent, yet still important, orthographical variant that we leave unaddressed is (–)*exs–* versus (–)*ex–*, because it is difficult to automatically detect it using a rule-based approach. Nevertheless, this variant hardly affects any of the function words to which we have restricted our analyses.
8. In these training data too, we have substituted all *v*s for *u*s and all *ae*s/*oe*s for *e*s.
9. Note that *licet*, which strictly speaking derives from the impersonal verb *licere*, is considered a function word because it is primarily used as a subordinating concessive conjunction.
10. Other errors in the lemmatization displayed in Table 3 are 'hildegars', 'us', and 'ta'.
11. Note that from this point onwards, we will express the size of textual samples in terms of the number of consecutive lemmatized words they contain (a number which, after tokenization, need not be identical to the original number of surface forms in the original texts).
12 For the sake of conceptual clarity we shall keep Pennebaker's original terminology, although it should be stressed that our present use of the term 'Synergy Hypothesis' is completely unrelated to the concept of 'Synergetic Linguistics' in the field of quantitative linguistics (Köhler, 2005).