Active or descriptive: Textual activity and its dynamic changes of Ph.D. theses across disciplines

Shuyi Amelia Sun[1], Wei Xiao[2,3*]

[1] School of Foreign Language Education, Jilin University, China
[2] Research Center for Language, Cognition and Language Application, Chongqing University, Chongqing, China
[3] School of Foreign Languages and Cultures, Chongqing University, China
[*] Corresponding author's email: xiaoweiyyx@126.com

DOI: https://doi.org/10.53482/2023_55_411

ABSTRACT

As an innovative and systematic genre in the academic community, Ph.D. theses have been heatedly researched in the field of English for Academic Purposes. Although research on the functional and formal features of Ph.D. theses has been abundant, their stylometric traits regarding textual activity have not been explored. Accordingly, this study explored the textual activity of Ph.D. theses and its dynamic changes across natural sciences, social sciences and humanities. A total of 150 Ph.D. theses (50 from each discipline) were analyzed, and the Q and χ^2 values were calculated to determine the textual activity of theses as well as its dynamic changes with the progression of texts. The results showed that, although the theses were found to be active in general, significant differences across disciplines do exist, in that the theses in natural sciences and humanities were more active while those in social sciences were more likely to lean towards the descriptive mode. This study has implications for widening the scope of cross-disciplinary academic genre analyses from an innovative quantitative linguistic perspective.

Keywords: textual activity, stylometrics, Ph.D. theses, disciplinary academic writing.

1 Introduction

With the rapid development of English for Academic Purposes (hereinafter EAP), a large body of research has looked into Ph.D. theses and their disciplinary linguistic features (Paltridge and Starfield 2020). As an innovative and systematic academic genre, Ph.D. theses reflect the frontiers and trends of an academic community (Xiao and Sun 2020). Considering academic writing is specific to the discipline and manifests variations among different academic communities (Xiao et al 2022, 2023a; Hyland 2012; Jiang 2022), cross-disciplinary research on Ph.D. theses can shed light on the textual variations across different disciplinary communities, explore how knowledge is rhetorically constructed and negotiated within each academic community, and provide more empirical evidence to support the pedagogy and practice of Ph.D. theses.

To date, previous studies on Ph.D. theses have mainly explored their functional and/or formal features by scrutinizing particular sections (e.g. the 'introduction' section, see Kawase 2018). The functional perspective concentrates primarily on ways to achieve certain communicative goals with appropriate linguistic resources, and the formal perspective is generally devoted to lexical/syntactic features drawing on manual coding or text-mining approaches. In light of the consensus that knowledge is constructed and negotiated within each discipline (Hyland 2012), the above-mentioned perspectives have been gradually filtering through to (cross-)disciplinary research. For example, Bunton (2002) explored generic moves in Ph.D. thesis introductions and found variations on specific steps across fields of science and technology, humanities, and social sciences. Hyland (2008) studied the forms, structures and functions of four-word clusters, providing evidence for the distinctive discipline-specific idiosyncracies of clusters in Ph.D. theses. Xiao and Sun (2020) investigated the lexical features of Ph.D. theses across disciplines, suggesting significant differences regarding lexical diversity and richness between natural sciences and humanities.

Despite the fruitful findings, little is yet known about the discipline-specific stylometric features of Ph.D. theses. Style generally refers to linguistic characteristics that people tend to express via spoken and/or written communication (Popescu et al. 2014). In the field of quantitative linguistics, style is taken as a quantifiable trait of language that can be detected using statistical techniques, and the statistical measurement of style is referred to as stylometrics (Schreibman et al. 2008). Among the stylometric features, textual activity is an important one that depicts activity-descriptivity (dis)equilibrium, i.e. whether texts tend to be active (plotted with substantial verbs) or descriptive (embellished with rich adjectives) (Jiang et al. 2020). To date, most studies on textual activity have focused on political and literary texts (Kubát and Čech 2016; Melka and Místecký 2019; Zörnig and Altmann 2016). For example, Kubát and Čech (2016) analyzed 50 US presidential inaugural speeches, and found that presidential speeches were influenced by speaker's style and social affairs, such as wartime and financial crisis. Melka and Místecký (2019) explored the textual activity of Beam Piper's novelette *Omnilingual*. Their findings suggested that most chapters of the novelette were highly active, which could be accounted for by the author's stylistic preference, 20th-century fictions' common features and the sub-genre conventions.

Previous studies on textual activity have been confined mostly to political and literary topics, whereas the embodied regularities are expected to be figured out by exploring more genres (Čech and Kubát 2016; Chen and Liu 2018), such as Ph.D. theses. An investigation into the textual activity of Ph.D. theses across disciplines can reveal their stylometric features and shed light on the construction and negotiation of disciplinary discourse. Besides, it should be noted that previous studies on Ph.D. theses tended to concentrate on only selected section(s), probably due to the compromise made between manual coding/annotation and the sheer size of Ph.D. theses (Thompson 2013). Although looking into separate sections can be more focused, it would lead to fragmented knowledge of how they are constructed

in the entirety (Kanoksilapatham 2015). Only a full-length analysis of Ph.D. theses can capture their global features (Xiao and Sun 2020). In addition, among the handful of existing text-mining research on Ph.D. theses, little attention has been paid to the dynamic changes of quantitative properties as texts progress, which fails to reveal how the text as a system regulates itself as it develops (Zörnig and Altmann 2016). Investigating the dynamic development of Ph.D. theses' textual activity can reveal the whole picture as to how Ph.D. theses manifest itself from a macro-perspective and how the disciplinary academic discourse stylometrically governs itself into the complex adapted system (Liu et al. 2017).

To address these issues, we would attempt to investigate the textual activity and its dynamic changes of Ph.D. theses across natural sciences, social sciences and humanities. The research questions are as follows:

(1) What are the texual activity features of Ph.D. theses? Is there any variation across natural sciences, social sciences, and humanities?

(2) How does the textual activity of Ph.D. theses change dynamically with the progression of texts? Is there any cross-disciplinary difference?

2 Material

Ph.D. theses were collected using the ProQuest (Clarivate 2023) search engine[1]. The selected Ph.D. theses satisfied the criteria that: (1) they were completed by doctoral candidates enrolled in the Ivy League universities in the U.S., (2) they were submitted to the universities within the recent ten years, (3) they were similar in length (30,000 words), and (4) they were organized in a typical 'Introduction-Literature Review-Methods-Results-Discussion-Conclusion' structure. The criteria were to ensure the validity and comparability of language material across disciplines. 50 theses were selected to represent natural sciences, social sciences, and humanities (Kagan 2009) respectively, and thus a total of 150 Ph.D. theses were enrolled.

These Ph.D. theses were first converted into plain texts using AntFileConverter (Anthony 2017) and then cleaned of the sections of abstract, acknowledgments, references and appendices. Details of the corpus are presented in Table 1. The one-way ANOVA test showed no significant difference in text length among the three disciplines (p>.05).

Table 1: Corpus information.

[1] ProQuest search engine for dissertations and theses can be accessed via the following link: https://about.proquest.com/en/dissertations/.

Discipline	Number of texts	Word count	Average text length
Natural sciences	50	1,552,615	31,052
Social sciences	50	1,641,342	32,827
Humanities	50	1,934,457	38,689
Total	150	5,128,414	34,189

3 Methodology

3.1 Indices and Formulas

The textual activity of Ph.D. theses was measured using Busemann's (1925) Q, rendered as:

(1) $$Q = V/(V + A)$$

in which V and A are sums of verbs and adjectives respectively and Q stands for textual activity. The indicator draws on the assumption that texts are remarkably characterized by either action or description. As such, a more narrative text (e.g. short stories or fairy tales) is usually higher in the value of activity than a more descriptive one (e.g. rhetorically picturing a scenery in a travel book).

Based on Formula (1), textual activity can be roughly classified as active, neutral and descriptive (Zörnig et al. 2015). To be more precise, a chi-square test (see below) is suggested to be employed in combination (Melka and Místecký 2019).

(2) $$\chi^2 = \frac{(V-A)^2}{A+V}$$

Based on the two indices, textual activity can be classified into five categories (cf. Table 2).

Table 2: Categories of textual activity.

Conditions	Textual activity
Q > 0.55 & χ^2 > 3.84	significantly active (SA)
Q > 0.55 & χ^2 < 3.84	active (AC)
0.45 < Q < 0.55	neutral (N)
Q < 0.45 & χ^2 < 3.84	descriptive (DE)
Q < 0.45 & χ^2 > 3.84	significantly descriptive (SD)

3.2 Data Analysis

We first calculated Q and χ^2 based on the full-length Ph.D. theses and accordingly identified the textual activity traits of full-length Ph.D. theses. Regarding dynamic changes, we calculated Q and χ^2 within each Ph.D. thesis upon accumulated text sizes that increase by 1000 words to figure out the textual activity of Ph.D. theses and the dynamic changes as texts progress. Then, we performed ANOVA tests

to examine whether the cross-disciplinary variations are statistically significant and further employed the TukeyHSD post-hoc analysis (Tukey 1949) to identify precisely where the significant difference lies[2].

4 Results

As to the full-length Ph.D. theses, the results show that the Q-value is relatively higher in natural sciences (M=0.587, SD=0.049) and humanities (M=0.591, SD=0.056), while it is lower in social sciences (M=0.567, SD=0.048).[3] The ANOVA suggests a significant effect of discipline on the Q-value ($F(2, 147)=3.130$, $p<.05$). A post-hoc test of multiple comparisons further shows significant variation between social sciences and humanities ($p<.05$, 95% CI [-0.044, -0.004]). The χ^2-value is higher in humanities (M=400.988, SD=394.341) compared with those in natural sciences (M=260.628, SD=293.543) and social sciences (M=197.562, SD=266.164). The ANOVA suggests a significant effect of discipline on χ^2 ($F(2, 147)=5.205$, $p<.01$). A post-hoc test of multiple comparisons shows noted variation between social sciences and humanities ($p<.01$, 95% CI [-0.044, -0.004]).

Based on the two indices, the majority of Ph.D. theses were found to be significantly active, and a minority were found to be neutral (cf. Figure 1). To be specific, the significantly active theses in natural sciences account for the largest proportion (82%), while neutral ones take up only 18%. In humanities, 74% of Ph.D. theses are significantly active, and 26% of them are neutral. Ph.D. theses of social sciences present a balanced distribution, where significantly active theses take up 56% and neutral ones account for 44%.

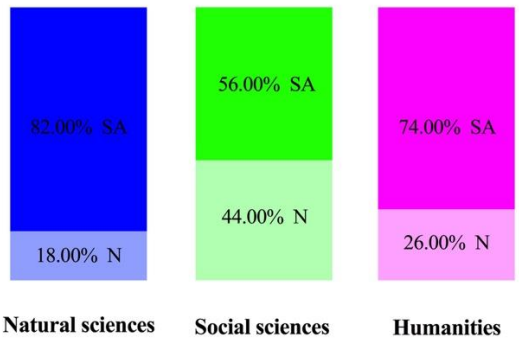

Figure 1: Textual activity of full-length Ph.D. theses. 'SA' stands for significantly active, and 'N' stands for neutral.

As to the dynamic changes, the mean Q-values alongside standard error of the mean (SEM, depicted as shadows) of each discipline are plotted in Figure 2, and the ANOVA and post-hoc results are shown in

[2] The procedure was adjusted by the Bonferroni correction.
[3] *M* and *SD* represent 'mean' and 'standard deviation' respectively.

Table 3. The Q-values are low when texts are not lengthy and become higher as texts progress. As to disciplinary variations, the Q-values of humanities are significantly higher at the beginning (Chunks 1-2, *p*s < .05). After that, the curves of humanities and natural sciences gradually overlap, while that of social sciences tends to diverge, with Q-values significantly lower (Chunks 14-17, 23-24 and 26-28, *p*s < .05).

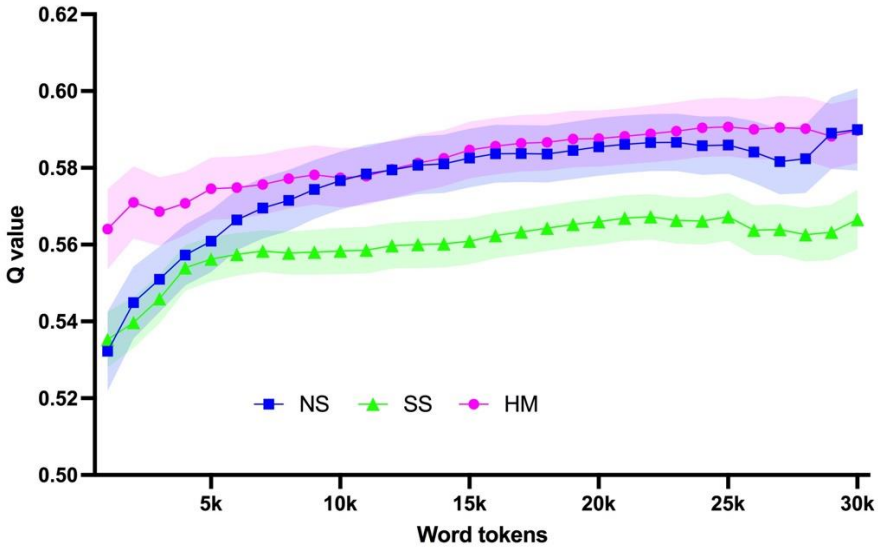

Figure 2: Q-value curves. 'NS', 'SS' and 'HM' represent natural sciences, social sciences and humanities respectively. The abbreviations have also been adopted in Figures 3 & 4 and Tables 3 & 4 below.

Table 3: Dynamic changes of Q-values across disciplines. *F* stands for the *F*-ratio. *F*-ratio would be close to 1 if the null hypothesis is true (i.e. no statistically significant variation lies across disciplines), and a larger *F*-ratio means that the variation among disciplinary groups is more than the possibility to see by chance (i.e. null hypothesis is rejected or statistically significant variation lies across disciplines). *p* stands for the *p*-value, which is to test the null hypothesis that data from all

disciplinary groups are drawn from populations with identical means. The two abbreviations are also adopted in Table 4. Asterisks (*) are intended to flag levels of significance. If the *p*-value is less than 0.05, it is flagged with a star (*).

Chunk	Mean			*F*		*p*	Mean Difference					
	NS	SS	HM				NS-SS		NS-HM		SS-HM	
1	0.5322	0.5353	0.5641	3.432	*	0.035	-0.0031		-0.0318	*	-0.0288	
2	0.5449	0.5397	0.5710	3.809	*	0.024	0.0052		-0.0261		-0.0313	*
3	0.5510	0.5459	0.5686	2.208		0.114	0.0051		-0.0176		-0.0227	
4	0.5573	0.5540	0.5708	1.430		0.243	0.0033		-0.0135		-0.0168	
5	0.5609	0.5562	0.5746	1.654		0.195	0.0047		-0.0137		-0.0184	
6	0.5664	0.5575	0.5749	1.431		0.242	0.0089		-0.0085		-0.0174	
7	0.5696	0.5583	0.5757	1.494		0.228	0.0113		-0.0061		-0.0174	
8	0.5715	0.5578	0.5772	1.930		0.149	0.0137		-0.0057		-0.0194	
9	0.5744	0.5580	0.5782	2.254		0.109	0.0164		-0.0038		-0.0202	
10	0.5767	0.5583	0.5774	2.298		0.104	0.0184		-0.0007		-0.0191	
11	0.5784	0.5585	0.5778	2.526		0.083	0.0199		0.0006		-0.0193	
12	0.5795	0.5598	0.5796	2.619		0.076	0.0197		-0.001		-0.0198	
13	0.5807	0.5600	0.5813	2.925		0.057	0.0207		-0.0006		-0.0213	
14	0.5810	0.5602	0.5825	3.135	*	0.046	0.0208		-0.0015		-0.0223	
15	0.5825	0.5609	0.5846	3.467	*	0.034	0.0216		-0.0021		-0.0237	*
16	0.5836	0.5623	0.5857	3.401	*	0.036	0.0213		-0.0021		-0.0234	
17	0.5837	0.5633	0.5864	3.313	*	0.039	0.0204		-0.0027		-0.0231	
18	0.5836	0.5643	0.5866	3.048		0.050	0.0193		-0.003		-0.0223	
19	0.5845	0.5653	0.5875	3.006		0.053	0.0192		-0.003		-0.0222	
20	0.5855	0.5660	0.5876	2.902		0.058	0.0195		-0.0021		-0.0216	
21	0.5861	0.5669	0.5882	2.867		0.060	0.0192		-0.0021		-0.0213	
22	0.5866	0.5673	0.5889	2.920		0.057	0.0193		-0.0023		-0.0216	
23	0.5867	0.5664	0.5896	3.197	*	0.044	0.0203		-0.0029		-0.0232	
24	0.5858	0.5661	0.5904	3.290	*	0.040	0.0197		-0.0046		-0.0243	*
25	0.5860	0.5673	0.5907	2.987		0.054	0.0187		-0.0047		-0.0234	
26	0.5841	0.5638	0.5900	3.377	*	0.037	0.0203		-0.0059		-0.0262	*
27	0.5816	0.5640	0.5905	3.079		0.050	0.0176		-0.0089		-0.0265	*
28	0.5824	0.5626	0.5902	3.198	*	0.044	0.0198		-0.0078		-0.0276	*
29	0.5890	0.5632	0.5882	2.915		0.059	0.0258		0.0008		-0.0250	
30	0.5900	0.5666	0.5898	2.039		0.136	0.0234		0.0002		-0.0232	

The mean χ^2-values of each discipline are plotted in Figure 3, and the χ^2-values results are shown in Table 4. The χ^2-values are low when texts are not lengthy and become higher as texts progress. The increase of the χ^2-values is in fact due to the property of the indicator which generally increases as the sample size becomes larger (Mačutek and Wimmer 2013).

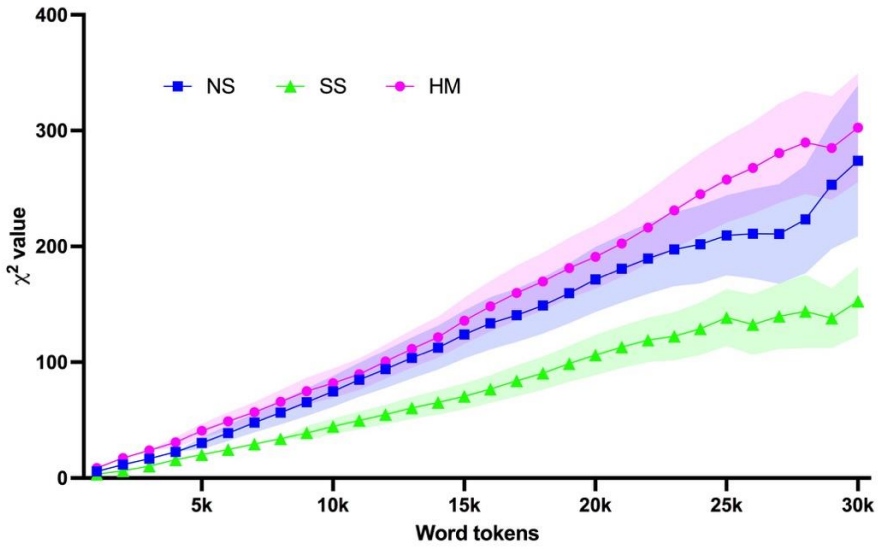

Figure 3: χ^2-value curves.

Table 4: Dynamic changes of χ^2-values.

Chunk	Mean		
	NS	SS	HM
1	5.702	3.387	8.716
2	11.516	6.362	17.253
3	16.663	10.297	23.785
4	22.528	15.718	30.841
5	30.299	20.277	40.764
6	38.896	24.657	49.149
7	47.934	29.266	56.993
8	56.630	33.838	65.913
9	65.509	39.046	75.027
10	74.745	44.768	82.014
11	84.961	49.786	89.674
12	94.000	54.998	100.612
13	103.675	60.604	111.571
14	112.682	65.416	121.538
15	123.908	70.553	135.859
16	133.675	76.901	148.336
17	140.601	83.964	159.753
18	148.956	90.584	169.661
19	159.648	98.706	181.283
20	171.447	106.343	191.071
21	180.746	113.133	202.493
22	189.570	119.266	216.336
23	197.489	122.594	231.119
24	201.755	128.856	245.169
25	209.485	138.685	257.769
26	210.953	132.572	267.733
27	210.696	139.751	280.554
28	223.476	143.855	289.773
29	253.300	137.983	284.917
30	274.061	152.669	302.592

Based on the joint conditions of Q and χ^2, we can determine the dynamic changes of textual activity as texts progress. First, we assigned each chunk in each text a category of textual activity. We then calculated the percentages of texts in each category at each chunk. For example, at the first chunk in natural sciences, 28% of the texts are significantly active, 12% active, 50% neutral, 2% descriptive, and 8% significantly descriptive, and so forth (see Figure 4). It should be noted that the dynamic changes of textual activity in Figure 4 were not counted cumulatively in itself. Instead, as stated in Section 3.2, textual activity was determined by Q and χ^2 which were calculated within each Ph.D. thesis upon accumulated text sizes that increase by 1000 words.

As shown in Figure 4, at the beginning of theses, natural sciences are the least active and humanities are the most active, as is shown by the proportions of significantly active theses in each discipline. As texts progress, humanities remain active and natural sciences become even more active. Although social sciences drift towards the active mode, the change tendency is rather slow. Such tendencies last till the end of theses in that the significantly active theses in natural sciences (c.a. 80%) and humanities (c.a. 70%) far outnumber the neutral ones, suggesting an obvious active trend, whereas a considerable proportion (c.a. 40%) of theses in social sciences are neutral, suggesting a shift to the descriptive mode compared with the other two disciplines.

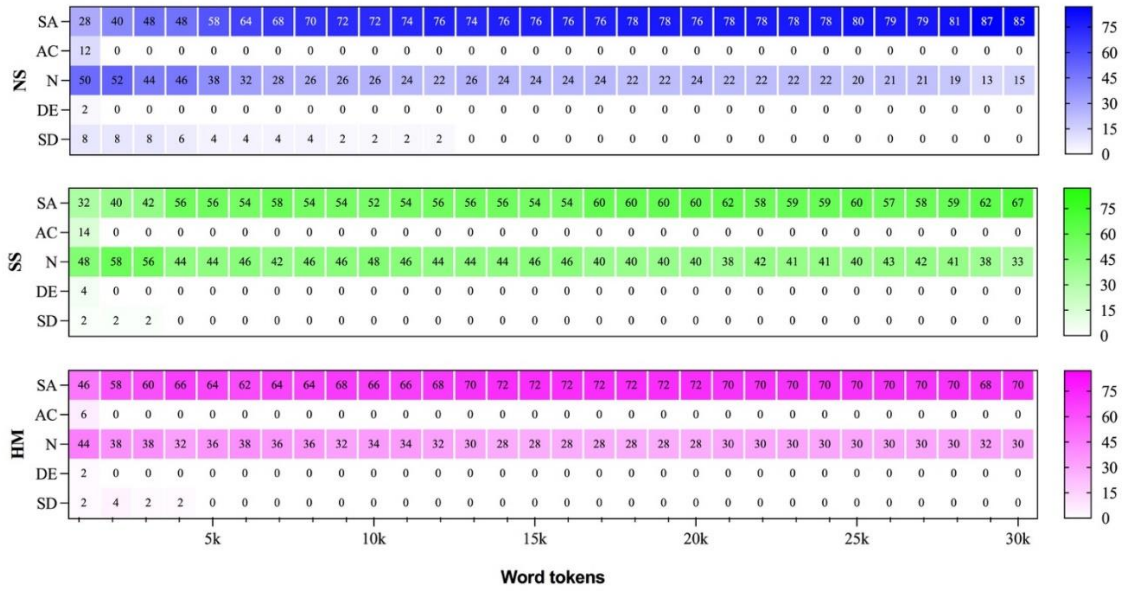

Figure 4: Dynamic changes of textual activity across disciplines.

5 Discussion

The present study aims to investigatethe textual activity and its dynamic changes of Ph.D. theses across the natural sciences, social science and humanities. To this end, full-length texts were split into 1000-word chunks, and the Q and χ^2 values were calculated upon accumulated text sizes that increase by 1000 words to figure out the textual activity of Ph.D. theses as well as the dynamic changes with the progression of texts. Accoridng to our results, the Ph.D. theses were found to be active in general. However, disciplinary variations could still be witnessed, in the way that the theses in natural sciences and humanities were more active than those in social sciences. As for the dynamic changes, natural sciences are the least active and humanities are the most active at the beginning of theses. As texts progress, humanities remain active and natural sciences become even more active. Although social sciences drift towards the active mode, the change tendency is rather slow.

Our finding that Ph.D. theses were mostly found to be significantly active is in line with Xu and Jiang (2021) who also found that the academic genre is "generally active" (p. 118). Such a finding could be accounted for by the observation that verbs are central to the overall structure of sentences and play a pivotal role in sentences (Baker 2003). In the construction of sentences, verbs arguably carry the largest amount of syntactic and semantic information (Baker 2003; Goldberg 1995; Liu 2009), while adjectives are comparably dispensable in syntax and more likely to work just as modifiers (Jia and Liang, 2020; Zhou et al. 2022). In academic writing, although writers may adopt appraisal resources (e.g. *significant*, *satisfying*) to construct authorial stances and engage with readers (Hood 2006; Martin and White 2005), these adjectives usually occur alongside verbs (e.g. *it is significant to*), and writers would avoid an overuse of adjectives for it is the trustworthy contents rather than rhetorics that determine the quality of PhD theses (Sun and Crosthwaite 2022a, 2022b; Xiao et al. 2023b).

Regarding disciplinary variations, we found that the natural sciences and humanities, which use entirely different methodologies and discuss scientific evidence differently, are counter-intuitively close together in terms of textual activity. This closeness may be accounted for by their narrative nature. In natural sciences, knowledge is taken as a plain matter of facts and the procedures of uncovering knowledge depend on the accumulation of empirical inquiry (Kuteeva and Airey 2014). Theses in natural sciences would put more emphasis on the report of operating procedures, statistical/empirical results, strategies and activities. The language style, then, could be regarded as a typical narrative one that avoids rich adjectival embellishments (Jiang et al. 2020), giving rise to the rapid increase of activity in natural sciences. In humanities, knowledge is regarded as constructed interpretations due to the complicated nature of human beings (Kuteeva and Airey 2014). Thesis writers in humanities tend to resort to a wide range of multi-dimensional perspectives (Xiao et al. 2023a; Zhao et al. 2023; Coffin and Hewings 2003). For example, in English studies, students are generally required to interpret the message or themes of a literary text and support their interpretation by referring to the text as well as to literary

critics. In history, students are frequently expected to evaluate the plausibility of an interpretation of past events and to draw on documentary sources as evidence for their proposition (Coffin and Hewings 2003). The feature of multi-dimensionality requires the incorporation of a variety of "external facts" (Jiang et al. 2020, p. 10), which may result in the overtly narrative nature and the active style of Ph.D. theses in humanities.

In addition, we also found that Ph.D. theses in social sciences are more descriptive (less active) than the other two disciplines. The possible explanation may be that both natural sciences and humanities have a long tradition and are highly developed, while social sciences, as a combination of methods as in natural sciences and objects as in humanities, are lately emerging ones, and thus do not feature such a long tradition. From this perspective, the mid-way of social sciences can be regarded as in sharp contrast to natural sciences and humanities. The above-mentioned uniqueness of social sciences has been documented in some previous studies (Coffin and Hewings 2003; Flowerdew 2015; Paltridge and Starfield 2020). For example, Coffin and Hewings (2003) found that, as a result of empirical approaches and the compilation of social statistics, Ph.D. theses written by doctoral students from social sciences might feature quantitative data, which may appear in texts in the forms of tables, graphs and maps. Students have to organize the pictorial/numerical data, understand how to incorporate them convincingly, and eventually depict the complicated multimodal information in clear and logical words. Paltridge and Starfield (2020) also found that social sciences generally pay special attention to rhetorical issues, persuading the audience of the validity of authorial arguments. This argumentative trait requires writers to draw on substantial interpersonal resources (e.g. *clear*, *important*) to develop a convincing authorial voice (Martin and White 2005). Some scholars further argue that, in social sciences, writers' abilities to use interpersonal strategies, introduce authorial voices, engage with alternative views and establish solidarity with disciplinary communities are generally perceived as key features of successful thesis writing (Flowerdew 2015). Therefore, the special trait of social sciences may tune the textual activity of Ph.D. theses in social sciences to the descriptive mode.

6 Conclusion

This study investigated the textual activity of Ph.D. theses and dynamic changes across natural sciences, social sciences, and humanities from a stylometric perspective. The results show that in general, Ph.D. theses are significantly active, despite the fact that the theses in natural sciences and humanities are more active while those in social sciences are more likely to lean towards the descriptive mode. As to the dynamic changes, noted cross-disciplinary differences were also found. Similar trends of pro-activity were found in natural sciences and humanities, as opposed to the trend in social sciences that leans towards the descriptive mode. The findings could be accounted for by the different roles of verbs and adjectives in sentences (e.g. Baker 2003; Xu and Jiang 2021) as well as the features of academic/thesis

writing across disciplines (e.g. Xiao and Sun 2020; Sun and Crosthwaite 2022a, 2022b; Hyland 2012; Jiang 2022).

As an initial attempt, this study has methodological implications by showing the promising prospect of using textual activity as the stylometric method to unravel the stylistic features of Ph.D. theses, where traditional qualitative methods still prevail in the analyses of academic genres such as theses and research articles (Xiao and Sun 2020; Paltridge and Starfield 2020). The improved approach has increased the statistical soundness of results and may inspire EAP scholars to look into academic texts from an innovative quantitative linguistic perspective. In addition, from the theoretical perspective, our results confirm the active nature of the academic genre and complement previous disciplinary findings in a couple of ways. Such findings can be particularly vital to EAP and English for Research and Publication Purposes (ERRP) practitioners, who have to elaborate on such cross-disciplinary variations so as to equip green-hand students and novice academic writers with an awareness of the discipline-specific stylometric features in thesis writing.

Despite the meaningful findings, there remain some limitations. First, although a sample of 50 texts per disciplinary group has already exceeded the minimum requirement for the sample size (Roever and Phakiti 2017), the validity of the results could be improved with an enlarged sample. In addition, the scope of this study is but limited to textual activity of PhD theses. Future studies could measure more indicators (e.g. TTR, writer's view, Gini coefficient) to capture a wider picture of stylometric features of more academic genres. As stated in Section 1, previous research on textual activity has been confined mostly to political and literary topics, whereas the embodied regularities are expected to be figured out by exploring more genres (Čech and Kubát 2016; Chen and Liu 2018). It would be interesting to investigate textual activity of other academic genres such as research articles, which is also a key genre for knowledge creation and communication.

References

Anthony, L. (2017). AntFileConverter (Version 1.2.1) [Computer Software]. Tokyo: Waseda University. Retrieved from https://www.laurenceanthony.net/software (Accessed on Jan 1, 2022).

Baker, M. C. (2003). *Lexical categories: Verbs, nouns and adjectives*. Cambridge University Press.

Bunton, D. (2002). Generic moves in PhD theses introductions. In: Flowerdew, J. (Ed.). *Academic Discourse*, pp. 57-75. Harlow: Longman.

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik* [Youth's speech as an imprint of the rhythm of development]. Fischer.

Čech, R., Kubát, M. (2016). Text length and the thematic concentration of text. *Mathematical Linguistics*, 2(1), pp. 5-13.

Chen, R., Liu, H. (2018). Thematic concentration as a discriminating feature of text types. *Journal of Quantitative Linguistics*, 25(1), pp. 53-76. https://doi.org/10.1080/09296174.2017.1339441.

Clarivate. (2023). ProQuest [Database]. https://about.proquest.com/en/.

Coffin, C., Hewings, A. (2003). Writing for different disciplines. In: Coffin, C., Curry, M. J., Goodman, S., Hewings, A., Lillis, T., Swann, J. (Eds.). *Teaching Academic Writing: A Toolkit for Higher Education*, pp. 45-72. London: Routledge.

Flowerdew, L. (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, 20, pp. 58-68. https://doi.org/10.1016/j.jeap.2015.06.001.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Hood, S. (2006). The persuasive power of prosodies: Radiating values in academic writing. *Journal of English for Academic Purposes*, 5(1), 37-49. https://doi.org/10.1016/j.jeap.2005.11.001.

Hu, G., Cao, F. (2015). Disciplinary and paradigmatic influences on interactional metadiscourse in research articles. *English for Specific Purposes*, 39, pp. 12-25. http://dx.doi.org/10.1016/j.esp.2015.03.002.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), pp. 4-21. https://doi.org/10.1016/j.esp.2007.06.001.

Hyland, K. (2012). *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: Cambridge University Press.

Jia, H., Liang, J. (2020). Lexical category bias across interpreting types: Implications for synergy between cognitive constraints and language representations. *Lingua*, *239*. https://doi.org/10.1016/j.lingua.2020.102809.

Jiang, F. (2022). *Metadiscursive Nouns: Interaction and Persuasion in Disciplinary Writing*. New York: Routledge.

Jiang, F., Hyland, K. (2022). "The datasets do not agree": Negation in research abstracts. *English for Specific Purposes*, *68*, pp. 60-72. https://doi.org/10.1016/j.esp.2022.06.003.

Jia, H, Liang, J. (2020). Lexical category bias across interpreting types: Implications for synergy between cognitive constraints and language representations. *Lingua*, *239*. https://doi.org/10.1016/j.lingua.2020.102809.

Jiang, X., Jiang, Y., Hoi, C. (2020). Is Queen's English drifting towards common people's English? — Quantifying diachronic changes of Queen's Christmas messages (1952–2018) with reference to BNC. *Journal of Quantitative Linguistics*, 29(1), pp. 1-36. https://doi.org/10.1080/09296174.2020.1737483.

Kagan, J. (2009). *The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century*. Cambridge: Cambridge University Press.

Kanoksilapatham, B. (2015). Distinguishing textual features characterizing structural variation in research articles across three engineering sub-discipline corpora. *English for Specific Purposes*, 37, pp. 74-86. https://doi.org/10.1016/j.esp.2014.06.008.

Kawase, T. (2018). Rhetorical structure of the introductions of applied linguistics PhD theses. *Journal of English for Academic Purposes*, 31, pp. 18-27. https://doi.org/10.1016/j.jeap.2017.12.005.

Kubát, M., Čech, R. (2016). Quantitative analysis of US presidential inaugural addresses. *Glottometrics*, 34, pp. 14-27.

Kuteeva, M., Airey, J. (2014). Disciplinary differences in the use of English in higher education: Reflections on recent language policy developments. *Higher Education*, 67(5), pp. 533-549. https://doi.org/10.1007/s10734-013-9660-6.

Liu, H. (2009). *Dependency grammar: From theory to practice*. Science Press.

Liu, H., Xu C., Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, pp. 171-193. https://doi.org/10.1016/j.plrev.2017.03.002.

Mačutek J.,, Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3), pp. 227-240. http://dx.doi.org/10.1080/09296174.2013.799912.

Martin, J. R., White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Melka, T., Místecký, M. (2019). On stylometric features of H. Beam Piper's *Omnilingual*. *Journal of Quantitative Linguistics*, 27(3), pp. 1-40. https://doi.org/10.1080/09296174.2018.1560698.

Paltridge, B., Starfield, S. (2020). *Thesis and Dissertation Writing in a Second Language*. New York: Routledge.

Popescu, I., Čech R., Altmann, G. (2014). Descriptivity in special texts. *Glottometrics*, 29, pp. 70-80. Retrieved from https://www.ram-verlag.eu/journals-e-journals/glottometrics/ (Accessed on Feb 12, 2022).

Roever, C., Phakiti, A. (2017). *Quantitative Methods for Second Language Research: A Problem-Solving Approach*. New York: Routledge.

Schreibman, S., Siemens R., Unsworth, J. (Eds.). (2008). *A Companion to Digital Humanities*. New Jersey: Wiley-Blackwell.

Sun, S., Crosthwaite, P. (2022a). "Establish a niche" via negation: A corpus-based study of negation within the Move 2 sections of PhD thesis introductions. *Open Linguistics*, *8*(1), pp. 189-208. https://doi.org/10.1515/opli-2022-0190.

Sun, S. A., Crosthwaite, P. (2022b). "The findings might not be generalizable": Investigating negation in the limitations sections of PhD theses across disciplines. *Journal of English for Academic Purposes*, *59*, pp. 101155. https://doi.org/10.1016/j.jeap.2022.101155.

Thompson, P. (2013). Thesis and dissertation writing. In: Paltridge, B., Starfield, S. (Eds.). *The Handbook of English for Specific Purposes*, pp. 283-299. Chichester: John Wiley & Sons.

Tukey, J. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, *5*(2), pp. 99-114. https://doi.org/10.2307/3001913.

Xiao, W., Sun, S. (2020). Dynamic lexical features of PhD theses across disciplines: A text mining approach. *Journal of Quantitative Linguistics*, *27*(2), pp. 114-133. https://doi.org/10.1080/09296174.2018.1531618.

Xiao, W., Liu, J., Li, L. (2022). How is information content distributed in RA introductions across disciplines? An entropy-based approach. *Research in Corpus Linguistics*, *10*(1), 63-83. https://doi.org/10.32714/ricl.10.01.04.

Xiao, W., Li, L., Liu, J. (2023a). To move or not to move: An entropy-based approach to the informativeness of research article abstracts across disciplines. *Journal of Quantitative Linguistics*, *30*(1), pp. 1-26. https://doi.org/10.1080/09296174.2022.2037275.

Xiao, W., Guo, Y., Zhao, X. (2023b). Towards Positivity: A Large-Scale Diachronic Sentiment Analysis of the Humanities and Social Sciences in China. *Fudan Journal of Humanities and Social Sciences*. https://doi.org/10.1007/s40647-023-00380-2.

Xu, Z., Jiang, Y. (2021). Activity of translational Chinese: A study based on three online corpora. *Foreign Language Teaching and Research*, 53(1), pp. 113-124. https://doi.org/10.19923/j.cnki.fltr.2021.01.010.

Zhao, X., Li, L., Xiao, W. (2023). The diachronic change of research article abstract difficulty across disciplines: a cognitive information-theoretic approach. *Humanities & Social Sciences Communications, 10*, pp. 194. https://doi.org/10.1057/s41599-023-01710-1.

Zhou, H., Jiang, Y., Wang, L. (2022). Are Daojing and Dejing stylistically independent of each other: A stylometric analysis with activity and descriptivity. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqac042.

Zörnig, P., Altmann, G. (2016). Activity in Italian presidential speeches. *Glottometrics*, 35, pp. 38-48. Retrieved from https://www.ram-verlag.eu/wp-content/uploads/2018/08/glo35abstract.pdf (Accessed on Feb 12, 2022).

Zörnig, P., Stachowski, K., Popescu, I., Miyangah, T., Mohanty, P., Kelih, E., Chen, R., Altmann, G. (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences*. Lüdenscheid: RAM-Verlag.

Funding statement

This work was supported by the Social Science Foundation of Chongqing [grant number 2019QNYY51]; the Fund of the Interdisciplinary Supervisor Team for Graduates Programs of Chongqing Municipal Education Commission [grant number YDSTD1923]; the Fundamental Research Funds for the Central Universities [grant number 2021CDJSKZX07], and the Graduate Innovation Fund of Jilin University.