Detecting dementia using linguistic analysis: Terry Pratchett's Discworld tells a more personal story

Melody Pattison[a], Ahmet Begde[b,c], Thomas DW Wilcockson[b]*

[a.] School of English, Communication and Philosophy, Cardiff University, Cardiff, UK
[b.] School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, UK
[c.] Department of Psychiatry, Oxford University, Oxford, UK

* Correspondence
Thomas Wilcockson,
School of Sport, Exercise and Health Sciences,
Loughborough University,
Loughborough, LE11 3TU, UK
Email: t.wilcockson@lboro.ac.uk

Running title: Detecting dementia using linguistic analysis

Keywords: Dementia; Alzheimer's disease; Linguistics; Linguistic analysis; Terry Pratchett;

Detecting dementia using linguistic analysis: Terry Pratchett's Discworld tells a more personal story

Abstract

Dementia, characterised by cognitive decline, significantly impacts language abilities. While the risk of dementia increases with age, it often manifests years before clinical diagnosis. Identifying early warning signs is crucial for timely intervention. Previous research has demonstrated that changes in language, such as reduced vocabulary diversity and simpler sentence structures, may be observed in individuals with dementia. This study investigates the potential of linguistic analysis to detect early signs of cognitive decline by examining the writing of Sir Terry Pratchett, a renowned author diagnosed with Posterior Cortical Atrophy (PCA), a form of dementia caused by Alzheimer's disease. This study analysed 33 Discworld novels by Terry Pratchett, comparing linguistic features (e.g., vocabulary size, lexical diversity, word class distribution) before and after a potential turning point identified through analysis of adjective type-token ratios (TTR). A significant decrease in lexical diversity (TTR) was observed for nouns and adjectives in later works. Total word count increased, while lexical diversity decreased, suggesting a shift towards simpler language. This shift coincided with a decrease in adjective TTR below a defined threshold, occurring approximately ten years before Pratchett's formal diagnosis. These findings suggest that subtle changes in linguistic patterns, such as decreased lexical diversity, may precede clinical diagnosis of dementia by a considerable margin. This research highlights the potential of linguistic analysis as a valuable tool for early detection of cognitive decline. Further research is needed to validate these findings in larger cohorts and explore the specific linguistic markers associated with different types of dementia.

Detecting dementia using linguistic analysis:
Terry Pratchett's Discworld tells a more personal story

1 Introduction

Dementia is an umbrella term for neurodegeneration that causes cognitive decline, the most common of which is Alzheimer's disease[1]. While the risk of developing dementia becomes more common with age, it is not an inevitable part of ageing[2]. Around a third of people over 85 may experience dementia, but it can also occur in younger individuals, though this is fortunately rare. The cognitive decline observed in people with Alzheimer's disease is caused by amyloid beta and tau proteins[3], which are both naturally occurring in the brain, but in Alzheimer's disease become toxic and start causing damage to the brain[4]. The underlying reason why amyloid beta and tau become toxic in some people and not others is currently under debate, however, identifying those people with signs of Alzheimer's-related cognitive decline as early as possible would enable interventions to be utilised to delay or even prevent some of the damage cause by the toxic proteins.

People with dementia may first notice they have an issue when they experience increased episodes of confusion or issues with memory or language[5]. However, Alzheimer's pathology likely begins many years and perhaps decades before the onset of symptoms[6]. Indeed, research has shown that there are earlier warning signs of dementia which may be too subtle for a patient to be aware of, for example, problems with attention[7]. Further, research suggests that it is currently possible to predict who will experience dementia 12 years prior to formal diagnosis[8]. Begde, et al.[8] observed that reduced complex visual processing speed is significantly associated with a higher likelihood of a future dementia diagnosis and risk/protective factors. Therefore, it may be possible to use tasks which measure these visual and attentional difficulties to help identify at-risk people before their cognitive decline worsens.

Dementia also has a negative effect on both speech and writing[9,10]. Therefore measuring the first signs of decline in these functions may also provide an early biomarker for dementia. In early-stage Alzheimer's, researchers have observed impairments in both producing and understanding words and sentences[11]. These studies have also revealed a breakdown in the complex network of knowledge that gives meaning to objects and words, a system known as semantic memory[12]. Such research studies have used standardised tests like word fluency and picture naming to assess language abilities[13]. These tests have shown that factors such as how

often a word is used, how familiar a person is with a word, and the age at which a word was learned can all affect language performance[14]. Therefore, by looking for changes in how someone uses language, then this may provide an early warning sign for dementia. The complexity of sentence structure, as measured by factors like the number of clauses per utterance, decreases with age in both spoken and written language[15]. Older adults struggle more with complex sentence structures, such as those with left-branching clauses, compared to younger adults[16]. A longitudinal study by Kemper, et al.[17]revealed a significant decline in grammatical complexity in older adults, particularly in those with dementia. Additionally, Bates, et al.[18] demonstrated that grammatical production is impaired in individuals with Alzheimer's. While patients do not typically produce overt grammatical errors, they encounter difficulties in selecting the most appropriate grammatical forms to convey their intended meaning. This suggests that the underlying issue may lie in accessing and retrieving the optimal fit between meaning and grammatical expression.

Overall, linguistic changes are to be expected as people age. However, these changes become more profound within people with cognitive decline. If a patient's writing history is available, then linguistic analysis techniques could be used to supplement clinical assessments or as a standalone early detection tool. Recent studies have done exactly this by measuring individual writers' publications over their careers to analyse how their language use has evolved. Garrard, et al.[19] studied the works of Iris Murdoch, a renowned English author who was diagnosed with Alzheimer's posthumously. Her final novel, published shortly before her diagnosis, is widely considered to exhibit signs of cognitive decline. While Garrard found minimal differences in overall structure and syntax, they observed significant and consistent variations in lexical diversity and word choice between the final book and control books from earlier in Murdoch's career. These results provide evidence that Alzheimer's may indeed be measured using linguistic analysis. Le, et al.[20] explored this further by including additional authors, numbers of books, and improved analysis techniques. Le et al analysed two authors believed to have Alzheimer's disease during their careers, Iris Murdoch and Agatha Christie, as well as P.D. James to act as a control participant, who published until the age of 88 without experiencing evidence of cognitive decline. They included twenty of Murdoch's twenty-six novels, published between ages 35 and 76, sixteen of Christie's novels written between ages 28 and 82, and fifteen of the novels of P.D. James. They then produced an analysis of the novels at the lexical level, using a variety of measures, including vocabulary size, lexical repetition, lexical specificity, word-class deficits, and fillers. Type-token ratio (TTR) calculates the proportion of

unique words to the total word count, and the word-type introduction rate (WTIR), which measures the rate at which new words are introduced in the text, calculated every 10,000 words. Lexical repetition, while intentional repetition can be a stylistic device, an increasing rate of repeated words may suggest a limited vocabulary or difficulty accessing words. To examine this, they conducted two analyses: a global analysis and a local analysis. Lexical specificity is calculated by the frequency of indefinite nouns and high-frequency, low-imagery verbs in each text. A higher proportion of these generic words suggests lower overall lexical specificity. Word class deficit (WCD) is an analysis of the distribution of word classes across each text, examining both the total number of words and the number of unique words. This allows for identification of potential deficits or overreliance on specific word classes and to measure the vocabulary size of open classes. Filler words are a measure of the proportion of interjections and filler words. While these words often appear in dialogue, fiction authors strive for natural-sounding conversations. However, this measure may be influenced by stylistic choices rather than cognitive decline and should be interpreted with care. Le, et al observed that TTR and WTIR were associated with cognitive decline and a decline in vocabulary led to an increase in repetitions in content words, and a word-class deficit can be seen in noun-token proportion, with a compensatory increase in verb-token proportion. They also observed a deficit in noun tokens that is significantly correlated with a rise in verb and pronoun tokens. Syntactic-complexity results were also found to fluctuate in a relatively wider range. Interestingly, they also report that deficits in Murdoch's writing appeared in Murdoch's late 40s and early 50s, which suggests that language deficits are observed many years before a formal diagnosis and indicates that Alzheimer's disease has a long preclinical period. Therefore, linguistic analysis would appear to show promise in identifying whether an author has experienced cognitive decline and may even indicate when the preclinical phase of dementia has begun.

The current research further explores the idea of using linguistic analysis in dementia by studying the works of Sir Terry Pratchett. Terry Pratchett was an English author, humourist, and satirist, best known for his Discworld series of 41 comic fantasy novels published between 1983–2015. Terry Pratchett was diagnosed with Posterior Cortical Atrophy (PCA) in December 2007. This diagnosis came at a time when he was still actively writing and publishing his beloved Discworld series. Despite the challenges posed by his condition, Pratchett continued to write and advocate for dementia awareness until his passing in 2015. PCA is a rare form of Alzheimer's disease that primarily affects visual processing and spatial awareness[21]. It affects areas in the back of the brain responsible for spatial perception, complex

visual processing, spelling, and calculation[22]. Given Terry Pratchett's prolific writing career and the fact he continued writing after his diagnosis, a linguistic analysis of his novels could provide valuable insights into the potential early signs of cognitive decline. By comparing his earlier works to his later ones, particularly those written closer to his dementia diagnosis, it would be possible to identify subtle changes in linguistic patterns, such as decreased lexical diversity, increased reliance on simpler sentence structures, and a decline in the use of specific and descriptive language. Such an analysis could contribute to our understanding of the linguistic markers of Alzheimer's disease and potentially aid in the development of early detection tools.

2 Method

The methodology of this study enabled the analysis of the lexical diversity of Terry Pratchett's writing pre- and post-PCA diagnosis, and explored whether this measure can be used as a predictor of dementia. In order to explore this lexical diversity, SketchEngine[23,24] was used to establish the TTR of content lexical items, as well as vocabulary repetitions, in 33 out of the 41 of Terry Pratchett's 'Discworld' novels. Several titles were excluded from the analysis due to them being either shorter than the other full-length novels[†1] , or because they are part of his titles for younger readers[†2]. They were excluded due to the fact that for SketchEngine to accurately measure TTR, it is important for texts to be of roughly the same length and/or aimed at similar levels of readership.

Word classes of interest included nouns, verbs, adjectives, and adverbs, in addition to the numbers of unique lemmas to determine the overall vocabulary size. SketchEngine was chosen to retrieve this information due to its ability to work with a large sample of text and for the researchers to be able to specify lemmas and word classes of interest for analysis.

In order to retrieve the information, plain text files of the 33 Terry Pratchett books of interest were imported into SketchEngine[†3]. Analysis of TTR to measure the WCD and overall vocabulary size, as well as analysis of repetitions (lines where an individual word is repeated

[†1] *'Eric' (1990), 'The Last Hero' (2001).*
[†2] *'The Amazing Maurice and His Educated Rodents' (2001), 'The Wee Free Men' (2003), 'A Hat Full of Sky' (2004), 'Wintersmith' (2006), 'I Shall Wear Midnight' (2010), 'The Shepherd's Crown' (2015).*
[†3] For ethical considerations, it should be noted that all books have been individually purchased by at least one of the three authors of this paper.

within ten surrounding words) was then conducted for each book. To calculate the TTR in relation to vocabulary size, the number of individual lemmas was divided by the total number of words for each book. To calculate the TTR to determine the WCD, the same method was applied to each word class. Most of the tokens could be retrieved automatically; however, some manual analysis of values was required in some instances, such as where SketchEngine returned the character 'Death' and the noun 'death' as the same value when considering the top ten repetitions of each book. Percentages and TTRs were also calculated manually. These data were then statistically analysed to compare the linguistic measures before and after Pratchett's PCA diagnosis in 2007. Thus, 29 books[†4] were analysed as pre-diagnosis, and 4 books[†5] as post-diagnosis.

The advantages of using WCD, vocabulary size and repetitions can be exemplified through the work of Le[25] and Le, et al.[20], as mentioned in the introduction of this paper. Despite using a different method to extract lexical tokens, their findings indicated that a loss of TTR in vocabulary size, an increase in repetitions, and a deficit in nouns, alongside other measures, could reliably predict cognitive linguistic decline in the works of Murdoch and Christie, which were not evident in the works of James. It is therefore our intention to replicate such linguistic analysis as part of the present study.

2.1 Data availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

2.2 Statistical Analyses

Descriptive statistics were calculated to summarise the key linguistic features used in Pratchett's books. Independent t-tests were used to compare linguistic measures released before and after the dementia diagnosis. Linear regression analyses were conducted to investigate the relationship between various TTR types and age. Additionally, Receiver Operating Characteristic (ROC) curve analyses were performed to evaluate the accuracy of TTR measures

[†4] From *'The Colour of Magic' (1983)* to *'Thud!' (2005)*.
[†5] From *'Making Money' (2007)* to *'Raising Steam'* (2013).

in distinguishing between pre- and post-diagnosis phases. All statistical analyses were performed using SPSS (version 29), with the significance level set at $p < 0.05$.

3 Results

The analysis included 33 books, with 29 books written before the dementia diagnosis and 4 book written after the diagnosis. Independent t-tests were conducted to compare linguistic features between these periods (see Table 1).

Table 1: Comparisons of Linguistic Features Pre- and Post-Dementia Diagnosis

Linguistic Features	**Pre-Diagnosis Mean (SD) (n=29)**	**Post-Diagnosis Mean (SD) (n=4)**	**p-value**
Total Nouns	21,819 (4,042)	29,195 (1,696)	0.001*
Type Token Ratio (nouns)	0.197 (0.025)	0.170 (0.008)	0.044*
Total Verbs	20,015 (3,758)	26,405 (2,398)	0.003*
Type Token Ratio (verbs)	0.078 (0.014)	0.067 (0.005)	0.138
Total Adjectives	5,422 (759)	7,394 (426)	<0.001*
Type Token Ratio (adjective)	0.241 (0.024)	0.213 (0.010)	0.028*
Total Adverbs	7,206 (1,356)	10,082 (1,227)	<0.001*
Type Token Ratio (adverb)	0.079 (0.017)	0.063 (0.004)	0.076
Total Words	94,379 (16,646)	129,226 (10,325)	<0.001*
TTR	0.071 (0.010)	0.063 (0.004)	0.114
Lemmas	6,604 (623)	8,166 (250)	<0.001*

Note: * indicates statistical significance ($p < 0.05$), TTR = Total-Type Token Ratio

The analysis showed significant changes in several linguistic features following Pratchett's dementia diagnosis. Books written after diagnosis showed significantly higher total word counts across all word classes (nouns: $p=0.001$; verbs: $p=0.003$; adjectives: $p<0.001$; adverbs:

$p<0.001$, see Table 1) and overall text length ($p<0.001$). However, this increase was accompanied by a decrease in lexical diversity, as shown by significant lower type-token ratios for nouns ($p=0.044$) and adjectives ($p=0.028$). No significant changes were observed in the type-token ratios for verbs and adverbs, or the overall TTR, suggesting that some aspects of linguistic complexity remained stable after the diagnosis. While total word usage increased, the variety of words used (measured by type-token ratios) either decreased or remained unchanged across different word classes. The total number of unique word forms (lemmas) increased significantly ($p<0.001$).

Figure 1 shows the changes in TTR across different word classes (adjectives, adverbs, nouns, and verbs) in Pratchett's writing over time. All word classes demonstrated a declining trend in TTR with age, indicating a general decrease in lexical diversity. Linear regression analyses were conducted to examine the relationship between various TTR and age. All models showed significant predictive relationships ($p < 0.001$). The TTR for adjectives emerged as the strongest predictor ($F(1,31) = 73.101$, $p < 0.001$), followed by adverb TTR ($F(1,31) = 53.694$, $p < 0.001$). Noun TTR ($F(1,31) = 45.728$, $p < 0.001$) and verb TTR ($F(1,31) = 39.413$, $p < 0.001$) also showed significant relationships with age. The overall TTR demonstrated a comparatively lower but significant, predictive relationship ($F(1,31) = 31.407$, $p < 0.001$). These findings suggest that lexical diversity might be a strong indicator of age-related linguistic changes.

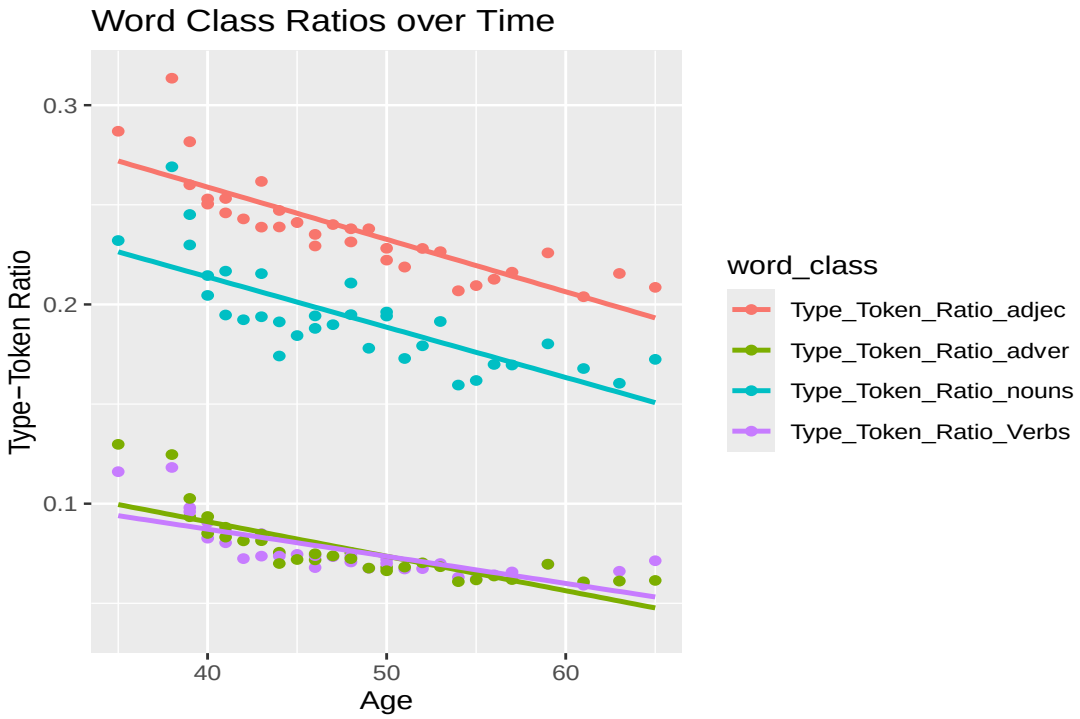

Figure 1: Type/token ratio over time in Pratchett's Discworld books

Receiver Operating Characteristic (ROC) curve analysis was performed to assess the accuracy of various TTR measures in detecting linguistic changes associated with dementia. All TTR measures demonstrated significant predictive ability ($p < 0.05$), with AUC values ranging from 0.80 to 0.91.

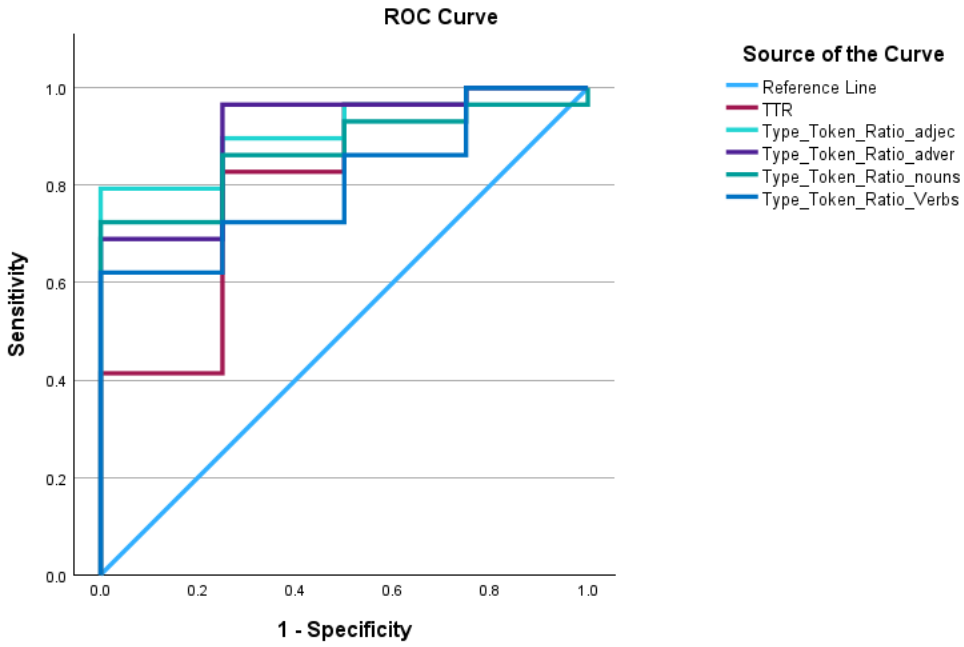

Figure 2. Receiver Operating Characteristic (ROC) curve analysis for the TTR variables

The TTR for adjectives showed the highest classification accuracy (AUC = 0.91, 95% CI: 0.80-1.02, $p < 0.001$), with a cut-off value of 0.227 showing 76% sensitivity and 100% specificity. This was followed by adverb TTR (AUC = 0.90, 95% CI: 0.76-1.04, $p < 0.001$), which demonstrated 96% sensitivity and 75% specificity at a cut-off of 0.616. Noun TTR also showed strong predictive performance (AUC = 0.87, 95% CI: 0.73-1.00, $p < 0.001$), with 86% sensitivity and 75% specificity at a cut-off of 0.172.

Verb TTR and overall TTR show relatively lower but significant accuracy (both AUC = 0.80, $p \leq 0.015$). Verb TTR achieved highest specificity (100%) but lower sensitivity (62%) at a cut-off of 0.717, while overall TTR showed balanced performance with 82% sensitivity and 75% specificity at a cut-off of 0.065.

Table 2. Area Under the Curve results for each TTR variable

Measure	**AUC (95% CI)**	**Cut-off**	**Sensitivity**	**Specificity**	**p-value**
TTR-Adjectives	0.91 (0.80-1.02)	0.227	0.76	1.00	<0.001*
TTR-Adverbs	0.90 (0.76-1.04)	0.616	0.96	0.75	<0.001*
TTR-Nouns	0.87 (0.73-1.00)	0.172	0.86	0.75	<0.001*
TTR-Verbs	0.80 (0.62-0.98)	0.717	0.62	1.00	0.001*
Overall TTR	0.80 (0.55-1.04)	0.065	0.82	0.75	0.015*

Note: AUC = Area Under the Curve; CI = Confidence Interval; * indicates statistical significance ($p < 0.05$)

As TTR for adjectives showed the highest diagnostic accuracy, we can use the cut-off value of 0.227 to identify when in Pratchett's writing his TTR for adjectives started to fall below this cut-off score. It was found that eleven of Pratchett's works were found to have a TTR for adjectives lower than 0.227, with the earliest published being The Last Continent (Discworld 22), which was published in May 1998 – 9 years and 7 months before his formal diagnosis. All books published after this date were found to have a TTR for adjectives of less than 0.227 whilst all books published before this date were found to have a TTR for adjectives more than 0.227. The only outliers are Discworlds 23 (Carpe Jugulum) and 25 (The Truth) which both had scores of 0.228, which do not meet the cut-off despite being published after Discworld 22, but are only 0.001 outside the cut-off.

Because Discworld 22 was the first book to fall below the TTR for adjectives cut-offs, independent t-tests were performed on pre and post Discworld 22 to compare linguistic features between these periods (see Table 3).

Table 3: Comparisons of Linguistic Features Pre- and Post-Discworld 22

Linguistic Features	Pre-Discworld 22 Mean (SD) (n=20)	Post-Discworld 22 Mean (SD) (n=13)	p-value
Total Nouns	20,179 (3,548)	26,613 (2,769)	<0.001*
Type Token Ratio (nouns)	0.206 (0.024)	0.175 (0.013)	<0.001*
Total Verbs	18,388 (3,240)	24,484 (2,312)	<0.001*
Type Token Ratio (verbs)	0.083 (0.015)	0.067 (0.004)	<0.001*
Total Adjectives	5,098 (677)	6,528 (677)	<0.001*
Type Token Ratio (adjective)	0.251 (0.021)	0.217 (0.009)	<0.001*
Total Adverbs	6,595 (1,153)	9,032 (1,042)	<0.001*
Type Token Ratio (adverb)	0.085 (0.017)	0.065 (0.004)	<0.001*
Total Words	87,197 (14,372)	116,152 (12,303)	<0.001*
TTR	0.074 (0.010)	0.065 (0.004)	<0.001*
Lemmas	6,337 (511)	7,496 (583)	<0.001*

Note: * indicates statistical significance ($p < 0.05$), TTR = Total-Type Token Ratio

The analysis showed significant changes in several linguistic features following Pratchett's publication of Discworld 22. Discworld books written after Discworld 22 showed significantly higher total word counts across all word classes (nouns: $p<0.001$; verbs: $p<0.001$; adjectives: $p<0.001$; adverbs: $p<0.001$, see Table 3) and overall text length ($p<0.001$). However, this increase was accompanied by a decrease in lexical diversity, as shown by significant lower type-token ratios for nouns: $p<0.001$; verbs: $p<0.001$; adjectives: $p<0.001$; and adverbs:

$p<0.001$. Overall TTR ($p<0.001$) and lemmas also significantly differed ($p<0.001$), indicating overall that linguistic features differed on all aspects pre and post Discworld 22.

4 Discussion

The current study aimed to explore the potential of linguistic analysis as a tool for early detection of cognitive decline, specifically focusing on the case of Terry Pratchett and his diagnosis of PCA due to Alzheimer's disease. Our analysis of Pratchett's Discworld series revealed significant changes in linguistic patterns over time. The most notable finding was a significant decrease in lexical diversity, as measured by TTR, for adjectives and nouns in Pratchett's later works. This suggests a decline in vocabulary richness and a reliance on simpler language structures. While the overall TTR remained relatively stable, the decrease in lexical diversity within specific word classes indicates a subtle but significant change in linguistic style. These findings align with previous research on the linguistic markers of Alzheimer's disease and other forms of dementia[20]. Studies have shown that individuals with dementia often exhibit reduced vocabulary diversity, simpler sentence structures, and increased reliance on clichés and formulaic language[26]. Our analysis of Pratchett's works suggests that similar linguistic changes could potentially also be observed in individuals with PCA due to Alzheimer's disease.

The high predictive accuracy of TTR measures, particularly for adjectives, suggests that linguistic analysis could be a valuable tool for early detection of cognitive decline. By identifying subtle changes in language use, it may be possible to detect the onset of dementia years before a formal diagnosis. This early detection could enable timely interventions and potentially slow the progression of the disease. The analysis revealed a significant shift in Pratchett's writing style following the publication of "The Last Continent" (Discworld 22). This book marks a potential turning point, as it was the first to exhibit a TTR for adjectives below the established cut-off value of 0.227, a threshold with a high diagnostic accuracy for potentially detecting linguistic changes related to dementia in the Discworld series. To further investigate this shift, we compared linguistic features in books published before and after "The Last Continent." The results revealed significant differences across various measures. Post-"Last Continent" books exhibited higher word counts across all word classes (nouns, verbs, adjectives, adverbs), increased overall text length, and a significant decrease in lexical diversity across all word classes as measured by TTR. Furthermore, the total number of unique lemmas

(vocabulary size) also increased significantly. These findings suggest a substantial shift in Pratchett's writing style after this point, characterised by an increase in word count but a decrease in lexical diversity and a potential shift towards simpler sentence structures. It should be noted that this book was published 9 years and 7 months before his formal diagnosis, indicating a long preclinical period prior to diagnosis. This observation is striking and highlights the potential for linguistic analysis to identify subtle changes in writing style that may precede clinical diagnosis of dementia by a considerable margin.

However, it is crucial to acknowledge the limitations of this study. Firstly, the analysis was based on a single case study, limiting the generalisability of the findings. Further research is needed to validate these findings in a larger sample of individuals with dementia, including those with PCA. Secondly, while PCA and Alzheimer's disease share some common features and PCA is thought to be caused by Alzheimer's disease, they are distinct conditions with different clinical presentations[21]. It is possible that the linguistic markers of PCA may differ from those of Alzheimer's disease. Furthermore, several methodological limitations should be considered. The precise chronology of Pratchett's writing is uncertain. We do not know the exact dates of writing for each book, whether he worked on multiple books concurrently, or if the publication order accurately reflects the writing order. These uncertainties could introduce potential biases into the analysis. Additionally, the observed changes in Pratchett's writing style after his diagnosis may not solely be attributed to PCA. Factors such as reduced writing time due to the disease, potential collaboration with other writers, or significant editorial alterations could also have contributed to these changes. Finally, it is crucial to emphasise that the observed changes in Pratchett's writing may not exclusively reflect cognitive decline due to PCA. Age-related changes in writing style are expected, and it is possible that some of the observed changes represent natural stylistic evolution rather than disease-related decline.

Despite these limitations, this study provides valuable insights into the potential of linguistic analysis as a tool for detecting early signs of cognitive decline. By analysing the works of Terry Pratchett, we have demonstrated how subtle changes in language use can be indicative of underlying cognitive impairment. Further research is needed to explore the full potential of linguistic analysis as a diagnostic tool for dementia, including the development of more sophisticated analytical methods and the investigation of larger and more diverse datasets. Future research could explore the use of more advanced linguistic analysis techniques, such as computational linguistics and machine learning, to identify additional linguistic markers of

cognitive decline. By developing more sensitive and accurate diagnostic tools, we may be able to improve early detection and intervention for dementia.

In conclusion, our study provides evidence that linguistic analysis can be a valuable tool for detecting early signs of cognitive decline. By analysing the works of Terry Pratchett, we have demonstrated how subtle changes in language use can be indicative of underlying cognitive impairment. The results also emphasises that language deficits may be observed many years before a formal diagnosis and indicates that Alzheimer's disease has a long preclinical period, in the case of Terry Pratchett, potentially almost ten years. Further research is now needed to explore the full potential of linguistic analysis as a diagnostic tool for dementia.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors report no competing interests.

References

[1] Fymat AL. On dementia and other cognitive disorders. Clin Res Neurol. 2019;2(1):1-4.

[2] Fotuhi M, Hachinski V, Whitehouse PJ. Changing perspectives regarding late-life dementia. Nat Rev Neurol. 2009;5(12):649-658.

[3] Allsop D, Mayes J. Amyloid β-peptide and Alzheimer's disease. Essays Biochem. 2014;56:99-110.

[4] Breijyeh Z, Karaman R. Comprehensive review on Alzheimer's disease: causes and treatment. Molecules. 2020;25(24):5789.

[5] Knopman DS. The initial recognition and diagnosis of dementia. Am J Med. 1998;104(4):2S-12S.

[6] Lloret A, Esteve D, Lloret MA, et al. When does Alzheimer's disease really start? The role of biomarkers. Int J Mol Sci. 2019;20(22):5536.

[7] Wilcockson TD, Mardanbegi D, Xia B, et al. Abnormalities of saccadic eye movements in dementia due to Alzheimer's disease and mild cognitive impairment. Aging (Albany NY). 2019;11(15):5389.

[8] Begde A, Wilcockson T, Brayne C, Hogervorst E. Visual processing speed and its association with future dementia development in a population-based prospective cohort: EPIC-Norfolk. Sci Rep. 2024;14(1):5016.
[9] Calzà L, Gagliardi G, Favretti RR, Tamburini F. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Comput Speech Lang. 2021;65:101113.
[10] Luzzatti C, Laiacona M, Agazzi D. Multiple patterns of writing disorders in dementia of the Alzheimer type and their evolution. Neuropsychologia. 2003;41(7):759-772.
[11] Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci. 2015;7:195.
[12] Nebes RD. Semantic memory in Alzheimer's disease. Psychol Bull. 1989;106(3):377.
[13] Maseda A, Lodeiro-Fernández L, Lorenzo-López L, et al. Verbal fluency, naming and verbal comprehension: three aspects of language as predictors of cognitive impairment. Aging Ment Health. 2014;18(8):1037-1045.
[14] Ferrante FJ, Migeot J, Birba A, et al. Multivariate word properties in fluency tasks reveal markers of Alzheimer's dementia. Alzheimers Dement. 2024;20(2):925-940.
[15] Burke DM, Shafto MA. Language and aging. In: Craik FIM, Salthouse TA, eds. The Handbook of Aging and Cognition. 3rd ed. New York, NY: Psychology Press; 2008:373-443.
[16] Maxim J, Bryan K. Language of the elderly: a clinical perspective. London, UK: Whurr Pub Ltd; 1994.
[17] Kemper S, Thompson M, Marquis J. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. Psychol Aging. 2001;16(4):600.
[18] Bates E, Harris C, Marchman V, Wulfeck B, Kritchevsky M. Production of complex syntax in normal ageing and Alzheimer's disease. Lang Cogn Process. 1995;10(5):487-539.
[19] Garrard P, Maloney LM, Hodges JR, Patterson K. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain. 2005;128(2):250-260.
[20] Le X, Lancashire I, Hirst G, Jokel R. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Lit Linguist Comput. 2011;26(4):435-461.
[21] Mendez MF, Ghajarania M, Perryman KM. Posterior cortical atrophy: clinical characteristics and differences compared to Alzheimer's disease. Dement Geriatr Cogn Disord. 2002;14(1):33-40.

[22] Yong KX, Crutch SJ, Schott JM. Posterior cortical atrophy. In: Oxford Textbook of Neurologic and Neuropsychiatric Epidemiology. Oxford, UK: Oxford University Press; 2020:141.

[23] Sketch Engine. Sketch Engine website. http://www.sketchengine.eu. Published 2003. Accessed November 15, 2019.

[24] Kilgarriff A, Baisa V, Bušta J, et al. The Sketch Engine: ten years on. Lexicography. 2014;1(1):7-36.

[25] Le X. Longitudinal Detection of Dementia Through Lexical and Syntactic Changes in Writing [master's thesis]. Toronto, ON: University of Toronto, Department of Computer Science; 2010.

[26] Banovic S, Zunic LJ, Sinanovic O. Communication difficulties as a result of dementia. Mater Sociomed. 2018;30(3):221.