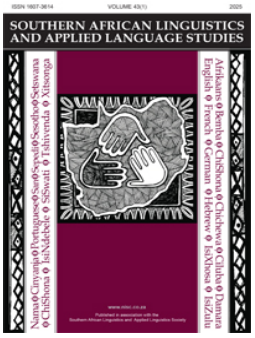


Routledge
Taylor & Francis Group

Southern African Linguistics and Applied Language Studies

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rall20

What makes poetic language different? An experiment in genre recognition using word co-occurrence networks in Afrikaans

Burgert Senekal & Eduan Kotzé

To cite this article: Burgert Senekal & Eduan Kotzé (29 Jan 2025): What makes poetic language different? An experiment in genre recognition using word co-occurrence networks in Afrikaans, Southern African Linguistics and Applied Language Studies, DOI: 10.2989/16073614.2024.2367436

To link to this article: https://doi.org/10.2989/16073614.2024.2367436

Published online: 29 Jan 2025.

Southern African Linguistics and Applied Language Studies 2025, 43(x): xxx–xxx
Printed in South Africa – All rights reserved

This is the final version of the article that is published ahead of the print and online issue

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607-3614 EISSN 1727-9461
https://doi.org/10.2989/16073614.2024.2367436

What makes poetic language different? An experiment in genre recognition using word co-occurrence networks in Afrikaans

Burgert Senekal* and Eduan Kotzé
Department of Computer Science and Informatics, University of the Free State,
Bloemfontein, South Africa
**Correspondence: burgertsenekal@yahoo.co.uk*

Abstract: Despite being one of the oldest literary genres, poetry is notoriously difficult to define. Nevertheless, various literary scholars draw attention to poetry's foregrounding of language itself, as opposed to using language to convey meaning. In the current study, we investigate whether there are structural differences between Afrikaans poetry and prose when texts are analysed as word co-occurrence networks from the perspective of network science. Specifically, we investigate average path lengths (L), clustering (C) and the small-world index (S). We find that poetry texts generally have higher average path lengths and lower clustering than prose texts, which results in lower scores on the small-world index on average. We also calculated these scores for 100-word text segments to eliminate the possibility that text lengths may affect our results, with the same results, albeit with a smaller margin. This means that poetry is a less densely connected genre than prose. Suggestions are also made for further research.

Introduction

Poetry is a unique form of literary expression that utilises language in a distinct and evocative manner. From the Greek *poiesis*, first attested in Herodotus, poetry literally means 'making', and by implication 'making with language' (Brogan 1993a: 938). Despite its widespread occurrence in most cultures worldwide and over millennia, poetry remains notoriously difficult to define.

Roxas and Tapang (2010) have shown that there are structural differences between poetry and prose when texts are analysed as word co-occurrence networks from the perspective of network science. However, their study was of only six English authors, which limits their study's generalisability and raises the question as to whether their findings can be replicated for other languages. In addition, while Roxas and Tapang (2010) use network measures such as the average path length (L) and average clustering (C), they do not use the exact definition of a small-world network by employing the small-world index developed by Humphries and Gurney (2008). The current study relates to previous word co-occurrence network studies and this particular study by investigating whether there are measurements in network science that can distinguish between Afrikaans poetry and prose. More specifically, we investigate the following research questions in relation to the study by Roxas and Tapang (2010):

RQ1: Is there a difference between the average path length (L) of word co-occurrence networks constructed for Afrikaans poetic and prose texts?

RQ2: Is there a difference between the average clustering (C) of word co-occurrence networks constructed for Afrikaans poetic and prose texts?

RQ3: Is there a difference between the small-world index (S) of word co-occurrence networks constructed for Afrikaans poetic and prose texts?

The article is structured as follows. We first provide an introduction to key aspects of poetry, as highlighted by eminent scholars in literary studies. We then provide an overview of related research regarding word co-occurrence networks. This is followed by a discussion of the methods used in the current study, including how data was gathered, processed and analysed. We then present and discuss our findings. The article concludes with summary remarks and suggestions for future research.

Southern African Linguistics and Applied Language Studies is co-published by NISC (Pty) Ltd and
Informa UK Limited (trading as Taylor & Francis Group)

Defining poetry

Many authors have written about the challenges of defining poetry (Brogan 1993a; 1993b; Pierce 2003; Eagleton 2006; 2007; Ribeiro 2007; Ambrosch 2018; Holyoak 2019). Eagleton (2007: 95), for instance, emphasises that it is exceedingly difficult to delineate poetry from other literary genres, since 'there is no one element or set of elements which poetry manifests and which nothing else does'. Pierce (2003), Eagleton (2006; 2007) and Ambrosch (2018) call attention to the fact that while rhyme, rhythm, metre and the use of symbols are all associated with poetry, these are not defining characteristics of poetry, since free verse – which lacks many of these features – is also considered to be poetry. Conversely, these elements of poetry are also found in some prose and in advertisements, which is not considered to be poetry (Brogan 1993b; Pierce 2003; Eagleton 2007). As Philip Sidney (quoted in Brogan (1993b: 1347)) recognised, '[i]t is not ryming and versing that maketh Poesie. One may be a Poet without versing and a versifier without Poetry'. Since rhyme, rhythm, metre and symbolism are absent in some forms of poetry, but present in some literary forms that are not considered to be poetry, it means that poetry cannot be strictly defined by its formal features.

Since formal features are inadequate to delineate poetry from other literary forms, Eagleton (2007: 95) formulates what he calls a 'pretty lame' definition of poetry: 'in poetry you, the poet, decide where the lines end, whereas in prose the typesetter does' (see also his similar definition in Eagleton 2006). This 'pretty lame' definition of poetry is, however, a sensible albeit superficial distinction, for as Brogan (1993a: 938) avers, '[l]ineation is … central to the traditional Western conception of poetry … Prose is cast in sentences; poetry is cast in sentences cast into lines'.

Although lineation is a sensible typographical distinction between poetry and prose, most scholars and poets have also focused attention on poetry's unique use of language. Samuel Coleridge emphasised the primacy of the word in his concise definition of poetry, 'the best words in their best order' (in Holyoak 2019: 34). Brogan (1993a: 938) writes that a poem 'conveys heightened forms of perception, experience, meaning, or consciousness in heightened language'. Similarly, Eagleton (2007: 98) regards poetry as more than just a typographical form, arguing that poetry is concerned 'with the weight, texture, shape, sound and density of the signifier, of the word'. Eagleton (2006: 41) elaborates that '[p]oetry is often characterised as language which draws attention to itself, or which is focused upon itself, or (as the semiotic jargon has it) language in which the signifier predominates over the signified' (see also Pierce 2003). In other words, poetry is primarily concerned with language itself, rather than with using language as an instrument with which to communicate. Above all, '[p]oetry liberates language from any purely functional, instrumental or utilitarian goal. It allows communication for its own sake, not for the sake of something else' (Eagleton 2007: 103). This foregrounding of language is also found in Ambrosch's (2018: 14) definition of poetry, as 'the art of making something with words, as opposed to merely using language as a means to an end – coding and conveying information – as we do in everyday situations'.

For Ribeiro (2007: 191), the difference between poetry and prose lies also in the foregrounding of language, but more specifically in repetition,

> [r]hyme schemes, stanza forms, metre, alliteration, anaphora, parallelism, and the numerous other poetic devices are all patterns of recurrence that began with literature but have remained central to poetry alone.

Using examples from different periods and from different languages, she goes on to show that some type of repetition is found in most poems, leading her to incorporate repetition in her definition of a poem, and since this repetition occurs through the use of language, language itself is foregrounded.

Since poetry and prose both involve the use of language, and language has a set syntax and grammar, we can on the one hand expect that there will be little difference between poetic and prose language from a structural point of view. On the other hand, however, poetry's foregrounding of language may lead to a different network structure than what is found in prose when texts are represented as word co-occurrence networks. The rest of this article will discuss how these assumptions were tested by studying word co-occurrence networks from the perspective of network science.

Related research

Networks, also known as *graphs* in graph theory, consist of entities (called *nodes* or *vertices*) and ties (called *edges* or *links*). What those entities and ties are is determined by the network under consideration, for instance it may be airports and flights, websites and hyperlinks, people and social ties, or academic papers and citations. In word co-occurrence networks, nodes are words and an edge is indicated if words occur adjacent to each other in a sentence or text. For instance, in the sentence, *John kicks the ball*, edges or ties will be indicated in the following manner: John → kicks → the → ball. If the same is done with the sentence, *The ball flies through the air and John scores a goal*, a network is created, and that network can be analysed using measures developed in network science.

Word co-occurrence networks have been studied in a variety of contexts from the perspective of network science. Many languages have been viewed as complex networks, including Afrikaans (Senekal and Geldenhuys 2016; Senekal and Kotzé 2017), Chinese (Zhou et al. 2008; Liang et al. 2009; Sheng and Li 2009; Chen et al. 2018; Jiang et al. 2020), Croatian (Margan et al. 2014), English (Dorogovtsev and Mendes 2001; Ferrer i Cancho and Solé 2001; Masucci and Rodgers 2006; Akimushkin et al. 2017), Mongolian (Bao and Dahubaiyila 2022), Portuguese (Antiqueira et al. 2007) and Slovenian (Markovič et al. 2019). In addition, some other applications of the study of word co-occurrence networks include authorship attribution (Mehri et al. 2012; Amancio 2015; Akimushkin et al. 2017; Marinho et al. 2018), determining the quality of texts (Antiqueira et al. 2007), determining the readership of texts (Markovič et al. 2019), studying language change (Chen et al. 2018), topic modelling (Benabdelkrim et al. 2020), and comparing languages (Liang et al. 2009; Liu and Cong 2013; Senekal and Geldenhuys 2016; Senekal and Kotzé 2017). Of particular interest for our current study is Roxas and Tapang's (2010) attempt to differentiate between prose and poetry based on complex network measurements, namely clustering (*C*), average path length (*L*) and degree distributions (see below). The authors studied 60 poems and 34 prose works written by six different English authors, namely Edgar Allan Poe, James Joyce, Oscar Wilde, Rudyard Kipling, Robert Louis Stevenson and William Butler Yeats, and showed that there are statistical differences between prose and poetry, although they limited their study to English literature and to only this handful of authors. Their findings are discussed below in relation to the findings of the current study.

While our aim is not to compare languages, it should be noted that although Afrikaans and English are obviously different languages, both belong to the West-Germanic branch of Indo-European languages, and previous research has found no significant difference between these two languages in terms of small-world-ness when analysed as word co-occurrence networks (Senekal and Kotzé 2017).

Methods

Data gathering

We first had to compile a diverse dataset containing samples from different authors, and in an electronic format. Since we needed a diverse Afrikaans dataset, we gathered example texts from LitNet's creative writing portal, with 300 texts gathered from the poetry section and 300 texts from the prose section:

- Poetry (https://www.litnet.co.za/category/nuwe-skryfwerk-new-writing/gedigte/)
- Prose (https://www.litnet.co.za/category/nuwe-skryfwerk-new-writing/fiksie/)

Along with the texts themselves, the title of the text was recorded, together with the author. This allowed us to ensure that we do not only use texts written by a small number of authors, which could conflate individual stylistic features with genre. Table 1 shows the twenty authors who contributed the highest number of texts in this dataset. From the names on this list, it can be seen that a variety of authors were included, including some well-known contemporary Afrikaans poets such as Joan Hambidge, Hilda Smits and Daniel Hugo, as well as lesser known and even anonymous authors. However, even the author who contributed the highest number of texts, Joan Hambidge, only contributed 9.6% of the total number of poetry texts in this dataset.

Table 1: Authors who wrote the highest number of texts in the current dataset

Poetry		Prose	
Author	Publications	Author	Publications
Joan Hambidge	29	Anonymous	14
Anoniem	21	Rentia Bartlett-Mohl	7
Hilda Smits	18	Ockert J du Preez	6
Daniel Hugo	17	Alex J Coyne	5
Marius Crous	17	Carla Kargaard	4
Izak van Rensburg	13	Christina van Deventer	4
Merwe Scholtz	12	Elza Smal	4
Carina Stander	11	Frederick Botha	4
Johann de Lange	10	Hannes Steyn	4
Andries ‘Roof’ Bezuidenhout	7	Jaco Fouché	4
Hennie Aucamp	7	Marlise Joubert	4
Charl-Pierre Naudé	6	Celia Claase	3
Danie Marais	6	Fransa van Mazijk	3
Hennie Nortje	6	Hennie Aucamp	3
Louis Esterhuizen	6	Joanita Erasmus-Alt	3
Fanus Rautenbach	5	Marike van der Watt	3
Jaco Fouché	5	Nini Bennett	3
Loftus Marais	5	Riekus van der Westhuizen	3
Jelleke Wierenga	4	Adriaan Coetzee	2

Data processing

Some amount of text preprocessing is always necessary in quantitative analyses of text, but in this case pre-processing was kept to a minimum. Following the majority of previous word co-occurrence studies (Ferrer i Cancho and Solé 2001; Grabska-Gradzińska et al. 2012; Senekal and Geldenhuys 2016; Senekal and Kotzé 2017; Marinho et al. 2018), punctuation marks were removed from the texts, although other researchers, such as Masucci and Rodgers (2006), Sheng and Li (2009) and Chen et al. (2018), included punctuation in their analyses. In the case of full stops, we followed Roxas and Tapang (2010) in indicating a link between the last word of one sentence and the first word of a new sentence, since even though these words belong to different sentences, they co-occur in the text itself. Stop words are common words in a language, such as articles, pronouns, prepositions and conjunctions, that are often filtered out in natural language processing tasks because they are considered to have little meaning for understanding the overall content. However, no stop words were removed from the corpus in the current study, unlike in the study by Akimushkin and colleagues (2017), because we needed the whole sentence to calculate network metrics. All text was also converted to lowercase, as was done by Marinho et al. (2018).

We split the texts into edge lists of word pairs that co-occur in both poetry and prose texts. Using Python, a function called *create_bigrams* was designed to process the textual input and create a collection of bigrams. The function begins by splitting the input text into individual words, storing them in a list named ‘words’. This initial step effectively tokenises the input text, allowing for the subsequent analysis of word pairs. Next, the function creates bigrams from the list of words. This is achieved by utilising the *zip* function, which pairs each word in the ‘words’ list with its immediate successor. The resulting pairs, or bigrams, are stored in a list named *bigrams*. Finally, the function returns the list of generated bigrams as its output. The output (collection of bigrams) was written to a text file. A function called *write_bigrams* was designed that iterates over each bigram in the list of bigrams, with the *enumerate* function used to obtain both the index and the corresponding bigram. The elements of each bigram are then joined together into a string format, where the words are separated by a space character. If the current bigram is not the last element in the list of bigrams, the code appends a newline character to the file to ensure that each bigram is written on a new line. Finally, the code ensured that each file contained the bigrams extracted from the corresponding input file. By organising the bigrams in this manner, we facilitated the subsequent analysis of sequential word relationships within poetry and prose texts.

During initial experiments, we discovered a correlation between one of the measures discussed below, the small-world index (*S*), and the number of words in the text, which has not been reported in previous word co-occurrence network studies. We subsequently split all texts into equal parts of 100 words in order to be able to compare texts without taking the number of words into account. We chose 100 words because the average number of words per poem in this dataset is slightly larger than 100 words, making 100 words a sensible cut-off point. In the results discussed below, we provide statistics for both the original texts and the 100-word segments.

Figure 1 shows an example of a word co-occurrence network for a poetry text, 'Witwyn kwatryn' by Charl-Pierre Naudé. The words with the highest number of co-occurrences are larger.

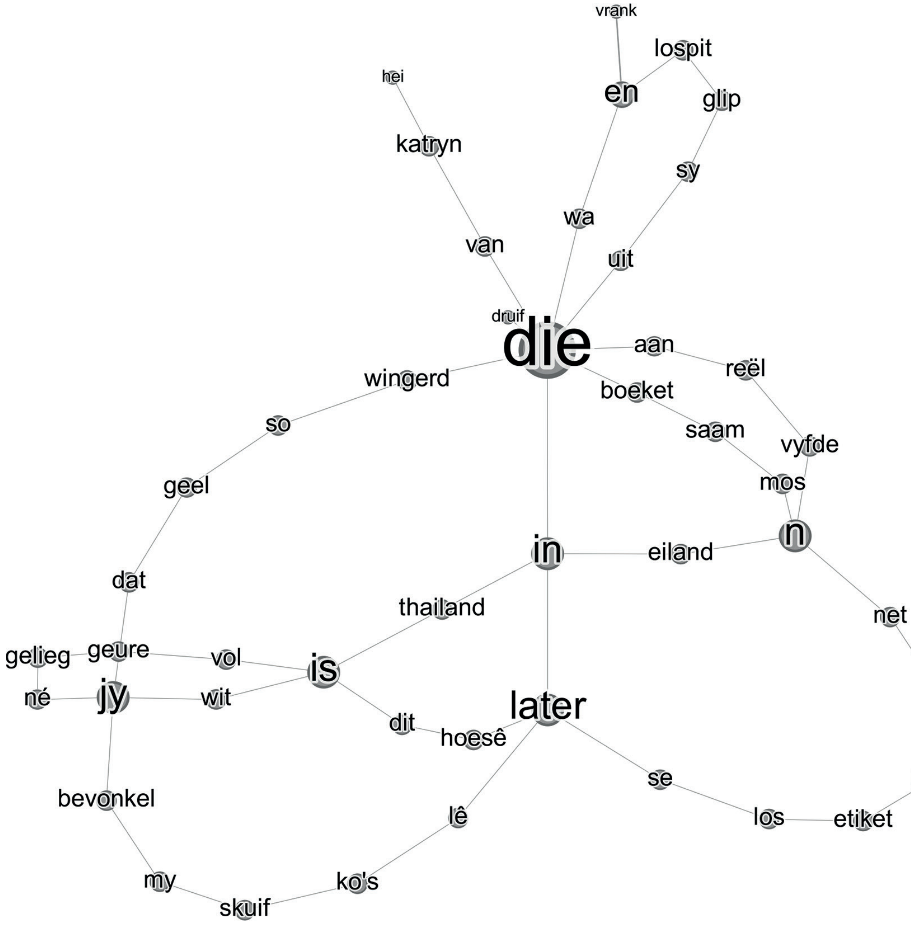


Figure 1: An example of a word co-occurrence network for a poetry text using 'Witwyn kwatryn' by Charl-Pierre Naudé

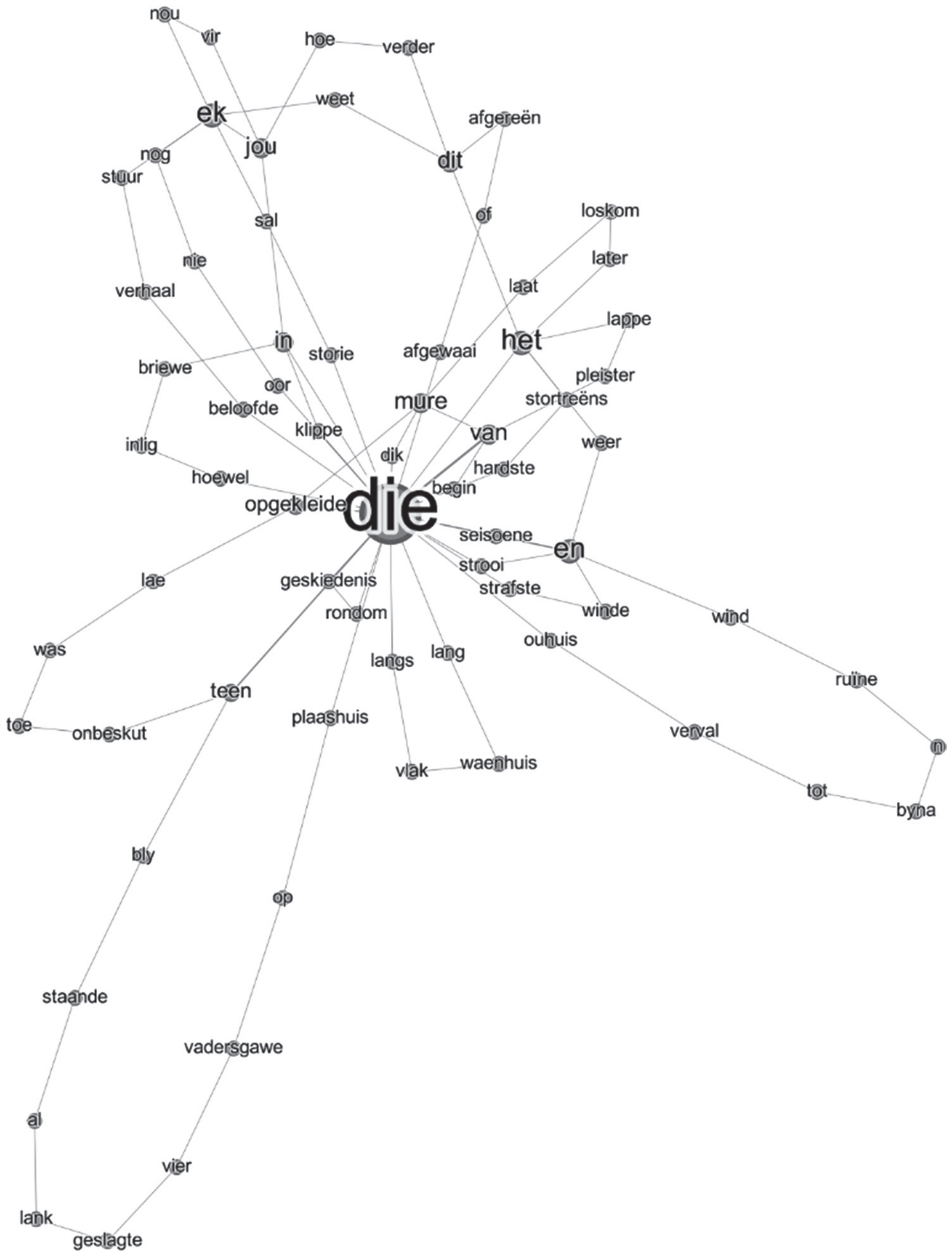


Figure 2: An example of a word co-occurrence network for a poetry text using “Woede” by Abraham H de Vries

Figure 2 shows an example of a word co-occurrence network for a prose text using “Woede” by Abraham H de Vries. As in Figure 1, words with a higher number of co-occurrences are shown larger.

Data analysis

A variety of measures have been developed in network science that facilitate the comparison between networks (see e.g. Estrada 2011; Barabási 2016). For the current study, we focus on three measures: average path length (*L*) and clustering (*C*), as well as how the latter two are used to determine the small-world index (*S*).

The average path length in graph theory is the average number of edges that need to be traversed to reach any two vertices in a graph, calculated by summing the shortest path lengths between all pairs of vertices and dividing by the total number of pairs. The average path in a graph (*L*) is calculated with Equation 1 (Liang et al. 2009; Roxas and Tapang 2010), where d_{ij} indicates the shortest path between nodes *i* and *j* and *n* indicates the total number of nodes in the network.

$$L = \frac{2\sum_{i>j} \quad d_{ij}}{n(n-1)} \qquad (1)$$

Clustering in graph theory refers to the measure of how closely connected a set of nodes are to each other, and it is calculated by determining the ratio of the number of actual connections between nodes in the set to the maximum possible connections. Clustering can be calculated using Equation 2 (Liang et al. 2009; Barabási 2016), where E_i refers to the number of edges between the neighbours of node *i*, and k_i refers to the number of edges of node *i*. The clustering of the entire network is then the average of C_i for the entire network.

$$C_i = \frac{2E_i}{k_i(k_i-1)} \qquad (2)$$

Networks are often compared with network models in order to study their structure, and one of the key network models in this respect is Erdös and Rényi's (1959) random network model (the ER model). When comparing a network to the ER model, an equivalent network is constructed, meaning a network with the same number of nodes and edges as the original network, but where edges are formed in a random manner. The average path length and average clustering in the real network are then compared with these values for the random network, respectively L_{ER} and C_{ER}. To calculate the average path length in an ER network (L_{ER}), Fronczak et al. (2004) suggest using Equation 3, where <*k*> indicates the average number of edges in the network, and *n* indicates the number of nodes in the network.

$$L_{ER} = \frac{ln(n)-\gamma}{ln\langle k\rangle} + \frac{1}{2} \qquad (3)$$

For calculating clustering for the random network (C_{ER}), Shen and Wu (2005) propose using Equation 4, again with <*k*> indicating the average number of edges in the network and *n* indicating the total number of nodes in the network.

$$C_{ER} \simeq \frac{\langle k\rangle}{n} \qquad (4)$$

Modelled by Watts and Strogatz (1998), small-world networks are networks that have a similar average path length (*L*) between nodes as an equivalent network constructed using Erdös and Rényi's (1959) random network model (L_{ER}), i.e. $L \approx L_{ER}$, but with a significantly higher level of clustering (*C*) than the clustering found in an equivalent ER network (C_{ER}), i.e. $C \gg C_{ER}$. Humphries and Gurney (2008) quantified the relationship between *L* and L_{ER} and between *C* and C_{ER} by proposing the small-world index (*S*), as calculated using Equation 5 (Humphries and Gurney 2008).

$$S = \frac{C/C_{ER}}{L/L_{ER}} \qquad (5)$$

According to Humphries and Gurney (2008), a network is a small-world network when $S \geq 1$, although cases where $1 \leq S \leq 3$ are considered borderline cases and hence $S > 3$ is the clearest indication that a network is a small-world network.

For the current study, network calculations were done using GraphCrunch2 (Kuchaiev et al. 2011), which is the successor to the original GraphCrunch (Milenković et al. 2008). GraphCrunch2 was developed in a C# framework and automates the process of building networks and comparing them to models such as Erdös and Rényi (1959), and GraphCrunch2 evaluates the fit of networks to models using local and global properties (Kuchaiev et al. 2011), for example, average path length (*L*) and clustering (*C*), as well as numerous other local measures that were not required for the current study. Although the original GraphCrunch was developed to analyse biological networks, it is also suitable for the study of other types of networks (Kuchaiev et al. 2011). Because word co-occurrence networks were compared to network models such as the ER model (which is an undirected binary graph model), GraphCrunch2 created networks in a binary (unweighted) and undirected manner. The following section discusses the results of the current study.

Results and discussion

Table 2 provides the results of the current study, with results from the original texts on the left, and results from the 100-word text segments on the right. Poetry networks have higher average path lengths than prose networks, indicating that the distance between words in terms of co-occurrence is larger for poetry than for prose. Note that while the difference is much smaller for 100-word segments than for the original texts, the average path length for poetry networks remains higher even with 100-word text segments. Roxas and Tapang (2010) also found that poetry texts have a slightly higher average path length than prose, although their difference is also small (3.81 for poetry and 3.29 for prose) and keeping in mind that their results are confined to English and with a much smaller dataset. Roxas and Tapang (2010) argue that since words are repeated more frequently in prose than in poetry because of longer text lengths, prose tends to have shorter path lengths. Additionally, prose would use more words that would co-occur with a large number of other words, such as the definite and indefinite articles. Poetry, on the other hand, would have longer average paths since it typically uses words that only occur once in the text.

Clustering, on the other hand, is significantly higher for the prose than for poetry networks, and even when 100-word segments are considered, prose networks have a higher clustering than poetry networks (although this difference is very small). This is similar to the finding by Roxas and Tapang (2010), who also found higher levels of clustering for prose (0.096) than for poetry (0.056). Roxas

Table 2: The results of the current study

Original texts				Texts with an equal number of words		
Poetry	Prose	Total	Variable	Poetry	Prose	Total
300	300	600	Number of texts	500	500	1 000
76	211	287	Number of authors	-	-	0
35 716	122 729	158 445	Total number of words	50 000	50 000	100 000
119.05	409.10	264.075	Average number of words	100.00	100.00	100
25 921	141 364	167 285	Total number of nodes	26 415	33 731	60 146
37 518	322 263	359 781	Total number of edges	33 797	45 502	79 299
4.44	3.39	3.91	Average *L*	4.46	4.24	4.35
0.06	0.15	0.10	Average *C*	0.05	0.06	0.06
2.32	22.50	12.41	Average *S*	1.15	1.45	1.30
0.85	0.17	0.31	Correlation words/S	0.00	0.00	0.00

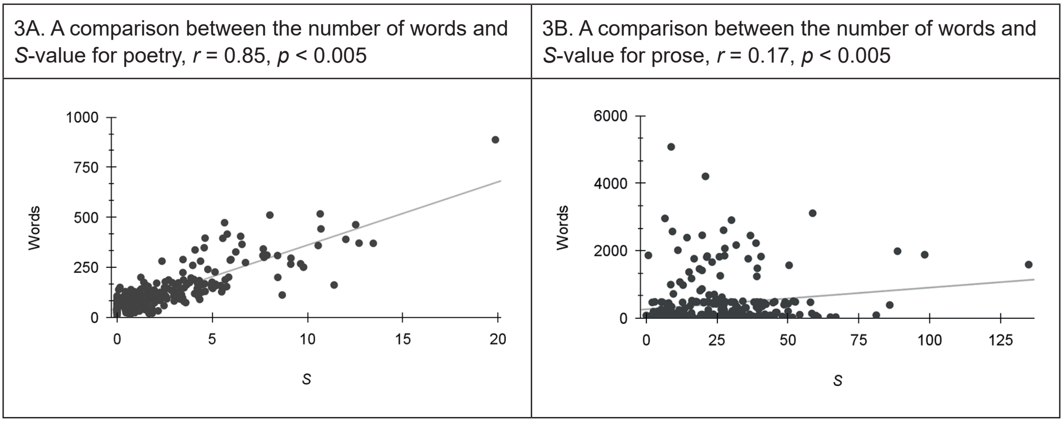
and Tapang (2010) explain that the higher clustering coefficient of prose texts is caused by the fact that words are utilised in different combinations more frequently in prose than in poems.

On average, poetry networks in our dataset have a much smaller small-world index than prose networks, and their average is even in the borderline area suggested by Humphries and Gurney (2008). For 100-word segments, both poetry and prose networks fall in this borderline category, although it is noteworthy that poetry again has a lower small-world index than prose. It is also surprising that although there is a strong correlation between the number of words in the text and that text's small-world index for poetry, there is only a weak correlation between these measures for prose. Unfortunately, Roxas and Tapang (2010) did not investigate whether their texts are small-world networks, and hence we cannot compare our results with theirs in this respect.

Figure 3 shows the correlation between the number of words and the small-world indexes of texts in both poetry (3A) and prose (3B) in more detail, as calculated using the original texts. It can be seen here that there is a strong correlation between these measures for poetry, but a weak correlation between these measures for prose. In the case of prose, there are a large number of outliers with a low small-world index score, but a high number of words, whereas for poetry the majority of texts cluster around the trend line. This is a visual illustration of the correlation shown in Table 2: For poetry, *S*-values increase with the number of words in the text, but this is not the case for prose.

Figure 4 shows the percentage of texts below thresholds of *S*-values, for the original texts (4A) as well as for the 100-word segments (4B). Figure 4A indicates that poetry tends to have a lower degree of interconnectedness compared to prose, and note that the percentage of texts below a certain threshold climbs more rapidly for poetry than for prose. For example, 46.67% of poetry texts fall below the first threshold ($S < 1$), while 97.00% fall below the last threshold ($S < 10$). This illustrates that poetry exhibits a lower level of small-world properties, with a lesser tendency for words to co-occur in a highly interconnected manner. On the other hand, prose shows a different pattern, and the percentage of texts falling below each threshold increases more gradually compared to poetry. This shows that prose texts have a higher level of interconnectedness compared to poetry. For instance, only 1.00% of prose texts fall below the first threshold ($S < 1$), while 32.33% fall below the last threshold ($S < 10$). Prose therefore exhibits a higher degree of small-world properties, with a greater tendency for words to co-occur in a highly interconnected manner.

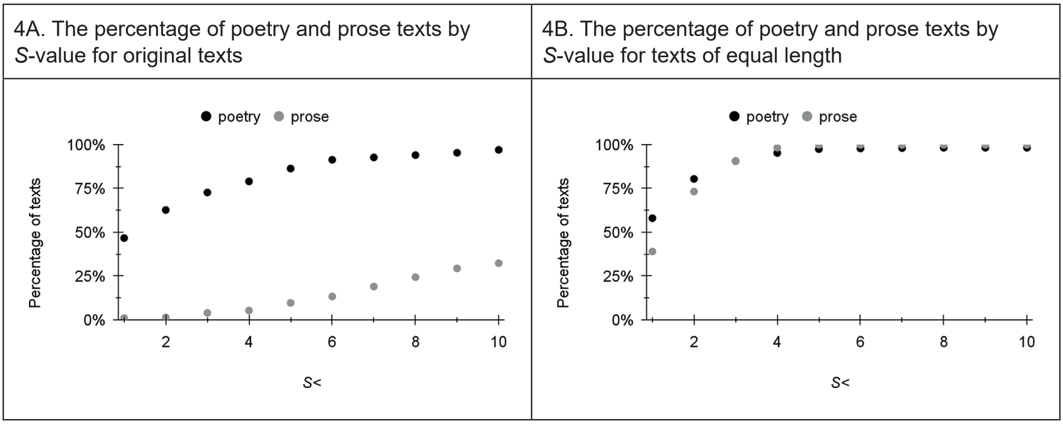
Figure 4B provides the same figures for the 100-word text segments. Here the difference between the percentage of texts below a certain threshold is much smaller, but still noticeable. For example, at the first threshold, poetry has a percentage of 58.00%, while prose has only 39.00% with *S*-values below 1. This suggests that poetry tends to have a lower degree of interconnectedness in its word co-occurrence network compared to prose, regardless of the number of words in the text.



3A. A comparison between the number of words and *S*-value for poetry, $r = 0.85$, $p < 0.005$

3B. A comparison between the number of words and *S*-value for prose, $r = 0.17$, $p < 0.005$

Figure 3: A comparison between the number of words and *S*-value for poetry and prose


4A. The percentage of poetry and prose texts by *S*-value for original texts

4B. The percentage of poetry and prose texts by *S*-value for texts of equal length

Figure 4: The percentage of poetry and prose text by *S*-value for original texts and texts of equal length

In conclusion, the key pattern observed in the above table and figures is the consistently lower small-world index scores in poetry networks compared to prose networks. These lower small-world index scores are the result of higher average path lengths and lower levels of clustering for poetry than for prose, which is consistent with the findings of Roxas and Tapang (2010) in terms of English poetry. While many of these differences are confounded by the length of texts, when text length is removed as a contributing factor, there remains a discernible difference in network structure between Afrikaans poetry and prose, although the difference is slight. In simple terms, this means that language use in Afrikaans poetry is more diverse than for prose, while prose is more interconnected, even if the differences between these two genres are small. This suggests that Afrikaans poetry tends to use a wider range of different words in the same text compared to prose writing in the same language. Prose includes a lot of repetition of words, most often so-called stop words, that provide links and short paths in a word co-occurrence network (thereby leading to higher *S*-values), and these occur less often in poetry. Poetry therefore uses language in a more disconnected manner. This less-connected way of using language can be considered a way of foregrounding language, which, as mentioned earlier, has been noted by Brogan (1993a), Pierce (2003), Eagleton (2006; 2007) and Ambrosch (2018) as an important characteristic of poetry. Since prose is the more common form of language use, the disconnected way that poetry uses language can be seen as a way in which poetry 'liberates language from any purely functional, instrumental or utilitarian goal' (Eagleton 2007: 103). In addition, regarding Ribeiro's (2007) argument that poetry is characterised by repetition, poetry does involve the repetition of rhyme schemes, metre, stanza patterns, parallelism, anaphora, alliteration and the like, but it is found here that it is the *lack of repetition* of stop words that leads to smaller *S*-values and a more disconnected use of language. Since stop words are considered less meaningful for the overall meaning of a text, it can be argued that poetry uses language in a more distilled manner than prose. Read against the backdrop of the findings by Roxas and Tapang (2010), this may be a distinguishing feature of poetry that transcends languages.

However, it should be emphasised that when the length of texts is removed as a factor that influences *S*-values, the difference between poetry and prose is very small. Despite the differences noted above, poetry and prose texts deliver highly similar word co-occurrence networks. This is in line with the mentioned attempts at delineating poetry from prose that found the task to be exceedingly difficult (Brogan 1993a; 1993b; Pierce 2003; Eagleton 2006; 2007; Ribeiro 2007; Ambrosch 2018; Holyoak 2019). While we did find that poetry uses language in a more disconnected manner than prose, one should not lose sight of the fact that the difference between these two genres is very small.

Conclusion

In this study, we compared word co-occurrence networks in Afrikaans poetry and prose to evaluate their network properties. Our findings revealed that poetry networks have higher average path lengths, suggesting that the distance between words in terms of co-occurrence is larger in poetry compared to prose. On the other hand, clustering is significantly higher in prose networks than in poetry networks, indicating that words in prose are utilised in different combinations more frequently. Moreover, the small-world index was consistently lower in poetry networks, suggesting a lower level of interconnectedness compared to prose networks. These differences in network properties between poetry and prose were observed even when comparing 100-word text segments, indicating that they are not solely dependent on the length of the texts. Overall, our study shows that there are structural differences (in the sense of a word co-occurrence network) between Afrikaans poetry and prose networks, as Roxas and Tapang (2010) found for English prose and poetry.

While we did not analyse texts in other languages, the findings by Roxas and Tapang (2010) for English texts suggest that our findings may extend to other languages. If this is the case, it would mean that poetry may be distinguished from prose through its less-connected use of language. Future studies can confirm whether the same distinctions are found in other languages, and if so, what the underlying reasons are for this lower level of connectedness, for instance, perhaps a greater use of connotation in poetry, while prose may be more explicit in its descriptions.

ORCIDS iDs

Burgert Senekal – https://orcid.org/0000-0002-1385-9258
Eduan Kotzé – https://orcid.org/0000-0002-5572-4319

References

Akimushkin C, Amancio DR, Oliveira ON. 2017. Text authorship identified using the dynamics of word co-occurrence networks. *PLoS One* 12(1): e0170527. https://doi.org/10.1371/journal.pone.0170527

Amancio DR. 2015. A complex network approach to stylometry. *PLoS One* 10(8): e0136076. https://doi.org/10.1371/journal.pone.0136076

Ambrosch G. 2018. *The Poetry of Punk: The meaning behind punk rock and hardcore lyrics*. London: Routledge.

Antiqueira L, Nunes MGV, Oliveira Jr ON, da F. Costa L. 2007. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications* 373: 811–820. https://doi.org/10.1016/j.physa.2006.06.002

Bao L, Dahubaiyila. 2022. Construction and analysis of Mongolian word co-occurrence networks. *2022 International Conference on Asian Language Processing (IALP)*, Singapore: 110–115. https://doi.org/10.1109/IALP57159.2022.9961279

Barabási A-L. 2016. *Network Science*. Cambridge: Cambridge University Press.

Benabdelkrim M, Levallois C, Savinien J, Robardet C. 2020. Opening fields: A methodological contribution to the identification of heterogeneous actors in unbounded relational orders. *M@n@gement* 23(1): 4–18. https://doi.org/10.37725/mgmt.v23.4245

Brogan TVF. 1993a. Poetry. In: Preminger A & Brogan TVF (Eds), *The New Princeton Encyclopedia of Poetry and Poetics*. Princeton: Princeton University Press. pp. 938–942.

Brogan TVF. 1993b. Verse and prose. In: Preminger A & Brogan TVF (Eds), *The New Princeton Encyclopedia of Poetry and Poetics*. Princeton: Princeton University Press. pp. 1346–1351.

Chen H, Chen X, Liu H. 2018. How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. *PLoS One* 13(2): e0192545. https://doi.org/10.1371/journal.pone.0192545

Dorogovtsev SN, Mendes JF. 2001. Language as an evolving word web. *Proceedings of the Royal Society: Biological Sciences* 268(1485): 2603–2606. https://doi.org/10.1098/rspb.2001.1824

Eagleton T. 2006. *How to Read a Poem*. Malden: Wiley-Blackwell.

Eagleton T. 2007. What is poetry? *Dodoni: Scientific yearbook of the Philology Department of the Philosophy School of the University of Ioannina* 36: 95–103.

Erdös P, Rényi A. 1959. On random graphs. *Publicationes Mathematicae* 6: 290–297. https://doi.org/10.5486/PMD.1959.6.3-4.12

Estrada E. 2011. *The Structure of Complex Networks*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591756.001.0001

Ferrer i Cancho R, Solé RV. 2001. The small world of human language. *Proceedings of the Royal Society: Biological Sciences* 268(1482): 2261–2265. https://doi.org/10.1098/rspb.2001.1800

Fronczak A, Fronczak P, Hołyst JA. 2004. Average path length in random networks. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* 70(5): 056110. https://doi.org/10.1103/PhysRevE.70.056110

Grabska-Gradzińska I, Kulig A, Kwapień J, Drożdż S. 2012. Complex network analysis of literary and scientific texts. *International Journal of Modern Physics C* 23(7): 1250051. https://doi.org/10.1142/S0129183112500519

Holyoak KJ. 2019. *The Spider's Thread: Metaphor in mind, brain, and poetry*. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/11119.001.0001

Humphries MD, Gurney K. 2008. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS One* 3(4): e0002051. https://doi.org/10.1371/journal.pone.0002051

Jiang Z, Zhao D, Zheng J, Chen Y. 2020. A study on differences between simplified and traditional Chinese based on complex network analysis of the word co-occurrence networks. *Computational Intelligence and Neuroscience* 2020: 8863847. https://doi.org/10.1155/2020/8863847

Kuchaiev O, Stevanović A, Hayes W, Pržulj N. 2011. GraphCrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics* 12: 24. https://doi.org/10.1186/1471-2105-12-24

Liang W, Shi Y, Tse CK, Liu J, Wang Y, Cui X. 2009. Comparison of co-occurrence networks of the Chinese and English languages. *Physica A: Statistical Mechanics and its Applications* 388(23): 4901–4909. https://doi.org/10.1016/j.physa.2009.07.047

Liu H, Cong J. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin* 58(10): 1139–1144. https://doi.org/10.1007/s11434-013-5711-8

Margan D, Martinčić-Ipšić S, Meštrović A. 2014. Preliminary report on the structure of Croatian linguistic co-occurrence networks. *arXiv*. https://doi.org/10.48550/arXiv.1405.4433.

Marinho VQ, Hirst G, Amancio DR. 2018. Labelled network subgraphs reveal stylistic subtleties in written texts. *Journal of Complex Networks* 6(4): 620–638. https://doi.org/10.1093/comnet/cnx047

Markovič R, Gosak M, Perc M, Marhl M, Grubelnik V. 2019. Applying network theory to fables: complexity in Slovene belles-lettres for different age groups. *Journal of Complex Networks* 7(1): 114–127. https://doi.org/10.1093/comnet/cny018

Masucci AP, Rodgers GJ. 2006. Network properties of written human language. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* 74(2): 026102. https://doi.org/10.1103/PhysRevE.74.026102

Mehri A, Darooneh AH, Shariati A. 2012. The complex networks approach for authorship attribution of books. *Physica A: Statistical Mechanics and its Applications* 391(7): 2429–2437. https://doi.org/10.1016/j.physa.2011.12.011

Milenković T, Lai J, Przulj N. 2008. GraphCrunch: a tool for large network analyses. *BMC Bioinformatics* 9: 70. https://doi.org/10.1186/1471-2105-9-70

Pierce RB. 2003. Defining 'poetry'. *Philosophy and Literature* 27(1): 151–163. https://doi.org/10.1353/phl.2003.0030

Ribeiro AC. 2007. Intending to repeat: A definition of poetry. *Journal of Aesthetics and Art Criticism* 65(2): 189–201. https://doi.org/10.1111/j.1540-594X.2007.00249.x

Roxas RM, Tapang G. 2010. Prose and poetry classification and boundary detection using word adjacency network analysis. *International Journal of Modern Physics C* 21(4): 503–512. https://doi.org/10.1142/S0129183110015257

Senekal BA, Geldenhuys C. 2016. Afrikaans as ’n komplekse netwerk: Die woordko-voorkomsnetwerke van woorde in André P. Brink se Donkermaan in Afrikaans, Nederlands en Engels [Afrikaans as a complex network: The word co-occurrence network in André P. Brink’s *Donkermaan* in Afrikaans, Dutch and English]. *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 35(1): 1368. https://doi.org/10.4102/satnt.v35i1.1368

Senekal BA, Kotzé E. 2017. Die statistiese eienskappe van geskrewe Afrikaans as ’n komplekse netwerk. *LitNet Akademies Geesteswetenskappe* 14(1): 27–59.

Shen K, Wu L. 2005. Folksonomy as a complex network. *ArXiv*, doi.org/10.48550/arXiv.cs/0509072

Sheng L, Li C. 2009. English and Chinese languages as weighted complex networks. *Physica A: Statistical Mechanics and its Applications* 388(12): 2561–2570. https://doi.org/10.1016/j.physa.2009.02.043

Watts DJ, Strogatz SH. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393(6684): 440–442. https://doi.org/10.1038/30918

Zhou S, Hu G, Zhang Z, Guan J. 2008. An empirical study of Chinese language networks. *Physica A: Statistical Mechanics and its Applications* 387(12): 3039–3047. https://doi.org/10.1016/j.physa.2008.01.024

Received 26 October 2023; revised 21 May 2024, accepted03 June 2024