Exploring “Letters from the Future” by Visualizing Narrative Structure

Sytske Wiegersma[1], Anneke M. Sools[2], and Bernard P. Veldkamp[3]

1 Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands
s.wiegersma@utwente.nl

2 Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands
a.m.sools@utwente.nl

3 Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands
b.p.veldkamp@utwente.nl

Abstract

The growing supply of online mental health tools, platforms and treatments results in an enormous quantity of digital narrative data to be structured, analysed and interpreted. Natural Language Processing is very suitable to automatically extract textual and structural features from narratives. Visualizing these features can help to explore patterns and shifts in text content and structure. In this study, streamgraphs are developed for different types of “Letters from the Future”, an online mental health promotion instrument. The visualizations show differences between as well as within the different letter types, providing directions for future research in both the visualization of narrative structure and in the field of narrative psychology. The method presented here is not limited to “Letters from the Future”, the current object of study, but can in fact be used to explore any digital or digitalized textual source, like books, speech transcripts or email conversations.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Text visualization, NLP, Narrative structure

Digital Object Identifier 10.4230/OASIcs.CMN.2016.5

1 Introduction

In this paper “Letters from the Future”, a narrative-based instrument used by [29] to study the human capacity to imagine the future, is studied using a combination of quantitative analysis, Natural Language Processing and text visualization methods. Traditionally, in narrative psychology, qualitative methods for analysing narrative content and structure are predominantly based on hand-coded data. The underlying structure of a narrative is represented for example by defining clusters or counting word frequencies. A widely used approach in narrative studies is the componential analysis, which focusses on identifying and examining the structural elements that narratives consist of. The narrative framework of [18], who originally divided narratives into five structural units (orientation, complication, resolution, evaluation, and coda), is a prime example of the componential approach.

The many features and feature combinations that can potentially be extracted from narratives can quickly result in an overwhelming quantity of data to be processed and interpreted. In addition, as a consequence of the growing popularity of e-mental health

7th Workshop on Computational Models of Narrative (CMN 2016).
Editors: Ben Miller, Antonio Lieto, Rémi Ronfard, Stephen G. Ware, and Mark A. Finlayson; Article No. 5; pp. 5:1–5:18
Open Access Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

interventions, more and more digital narrative data becomes available for analysis. Processing and interpreting all this data by hand is a tremendous, if not impossible, task. However, the growing availability of digital narrative data also generates new opportunities. There now is sufficient narrative data available to scale up the study of narratives by applying Natural Language Processing (NLP) methods. In NLP computers are used to process and manipulate natural language [9], [19], “natural language” being any spoken or written language used by humans in everyday life [2]. The main benefits of NLP by computers over the manual processing of narratives is that it is far less time consuming and less error prone than human coders. Moreover NLP enables researchers to process and compare large datasets or very detailed textual data.

A recent systematic literature review on text mining applications in psychiatry [1] showed that the use of NLP and text mining methods is still in its infancy in the fields of psychology and psychiatry. NLP applications have also only recently found their ways in the field of humanities. From the humanities perspective, Computational Narratology [21] can be described as a methodological instrument to develop narratological theories, enabling researchers to extend and test their models on larger text corpora and to specify and apply concepts and models automatically and thus more consistently [24]. As described by [21], in computational narratology narratives and narrative structures are explored using computation and information processing methods.

[5] state that an efficient approach to explore the underlying mathematical structure of narratives is text visualization. The mathematical structure is generally captured using first-, second- and third-order statistics like word frequencies, clustering and natural language algorithms [37]. Visually representing this structure enables researchers to reveal and interpret differences and relationships within and between text documents that would have been difficult, or even impossible, to identify solely from the texts or from tables of numerical data extracted from these texts [5], [37], [33]. Contrary to graphs, visualizations are generally used as an exploratory tool to explore and analyse the data and not to present study results to the public [33].

The current study is a response to the suggestion of [29] that more in-depth insight into how and why narrative futuring works can be gained by combining traditional qualitative with quantitative methods. The principal aim is to gain more detailed understanding of the differences in letter content, specifically the distribution (sequential order) and proportion of narrative processes and grammatical elements, both within and between the letter types. The new insights from this study can be used not only to confirm the previous findings of [30] but also to develop new theories and hypotheses regarding the human capacity to imagine the future.

This paper is organized as follows. First the object of study, the instrument “Letters from the Future”, is described in detail, followed by a description of the dataset and data-processing steps in the Methods section. In this section, the two essential topics in the development of text visualizations are addressed as well: first an existing NLP package used to extract the mathematical structure of the narratives is described. Second it is investigated how these structures can be visualized in such a way they can be used to study differences both within and between the different types of letters. In the Results section, the developed text visualizations are compared to the previous findings based on qualitative methods by [30] and linked to the existing narrative framework of [18]. Finally, conclusions and implications for future work are described in the Discussion section.

Table 1 Letter structure and characteristics.

	Imagining/experiencing a future situation Extended core with imaginative component, information on events, places, persons, experience	**Generic letter** No/limited imaginative components. Possibly global descriptions of future situations at end of letter
Retrospective evaluation Look back from future or present to past	*Type 1* *Imagining and evaluating the futured past* Structure: - Narrative imagination of desired future situation (present tense) - Anticipated reminiscence of the future past (past tense) - Conclusion/insight from evaluated experiences and/or - Worldly wisdom (self-praising remarks) - Comments on implications for the future (moral advice/future promises)	*Type 4* *Reminiscing and evaluating the past without imagination of the future* Structure: Equal to structure of type 1. Recounted/ evaluated period in past instead of futured past, presented as current concern taking place before moment of writing.
Prospective orientation Look forward from present to future/ from future even further ahead	*Type 2* *Imagining and orienting to the futured present and futured past* Structure: - Statement about present position in life (present/past perfect tense) - Imaginary goals/purposes - Description of how to realize these objectives	*Type 5* *Intentional orientation with expression of emotions* No clear structure: No clear action orientation or path from present to future. Sometimes written from future instead of present. Much use of intentional time (hope/wish), future tense (shall/will) and hesitation.
Present-oriented Focus on moment in time (present/ future present) instead of period	*Type 3* *Expressive imagining of the futured present* No clear structure: No orientation/evaluation or path from present to future. Sometimes conclusions are drawn. Contains sensory details (hopes, wishes, gratitude and self-appraising). Imagined future described mainly in present-tense.	*Type 6* *Advisory letters about current practical and moral concerns* No clear structure: Consists mainly of general insights/ conclusions, generic (existential/moral) advices or worldly wisdoms. No path to origination of conclusions or insights.

1.1 Letters from the Future

In this study, computational narratology is used to explore "Letters from the Future", an online narrative-based mental health promotion instrument developed by [29]. The instrument is adapted from an earlier exercise by [4], in which storytelling groups are used to enhance mental health. Using a web-based tool, participants are asked to write a letter from a particular situation and moment in the future to someone in the present. [30] studied the human capacity to imagine the future by hand-coding narrative processes within each individual letter on sentence-level. They clustered these narrative processes into five

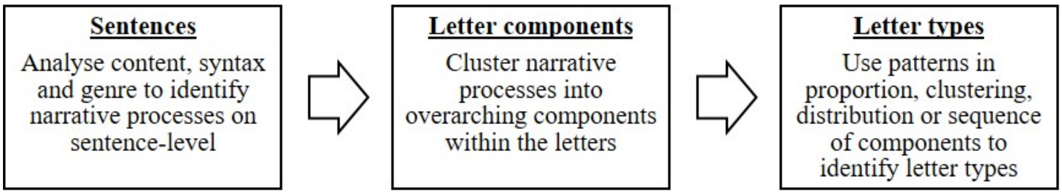

Figure 1 Schematic overview of procedure.

overarching components which were then used to identify six different letter types (see overview in Figure 1).

The letter types were defined based on a comparative analysis of the following elements: 1) the dominant narrative process (imagining, evaluating, orienting, expressing emotions or engaging in dialogue); 2) the use of certain grammatical elements like past, past imperfect, present and future tense, modals ("would", "could", "should"), intentional time ("hope", "wish", "want"), or the imperative ("go!", "remember!"); 3) the presence and clearness of a path between present and future; and 4) the level of detail of the imagination. Table 1 gives an overview of the six letter types and the corresponding structures found by [30].

As shown in Table 1, [30] found a clear distribution and sequence of narrative processes and grammatical elements for half of the letters (letter types one, two and four). However, these structures are not always uniformly applicable to all letters of the corresponding type. For example, type one letters generally consist of five elements, but the order of these elements can differ; letters can start either with narrative imagination of the desired future, anticipated reminiscence of the future past or an evaluative part preceding narrative imagination. The same goes for letters of type two, about which [30] write: "The orienting function could be prominent from the first sentence, in letters starting with goal-setting or value orienting phrases rather than with a situation (but the order could be reversed as well)." (p. 19). Another remark on type two letters was the finding that hope, a prominent feature in those letters, could occur either at the beginning or end of a letter.

In the following section the dataset is described in more detail, followed by a description of the methods used to pre-process the dataset in order to capture and visualize the narrative structure and content of the different letter types.

2 Methods

2.1 Dataset

An existing dataset of 492 letters collected for a previous study by the Storylab, the Dutch expert centre for narrative psychology and mental health promotion at the University of Twente, was used (see [29], [30] for more information on the data collection process). Informed consent to re-use these letters for on-going research by the Storylab was obtained. The letters were written by a relatively diverse, mainly Dutch (70%) and German (27%) participants. The letters were manually categorized into six categories by three independent raters (interrater reliability score = 0.672). Table 2 shows an overview of the number of letters and mean text length per category. In the current study only Dutch letters that were clearly categorized in one of the six letter types are used, resulting in a dataset of 351 letters.

Table 2 Dataset characteristics.

	Imagination letter	Generic letter
Retrospective evaluation	Type 1: N(letters) = 137 Mean N(words/text) = 324	Type 4: N(letters) = 19 Mean N(words) = 292
Prospective orientation	Type 2: N(letters) = 47 Mean N(words) = 303	Type 5: N(letters) = 9 Mean N(words) = 196
Present-oriented	Type 3: N(letters) = 94 Mean N(words) = 289	Type 6: N(letters) = 45 Mean N(words) = 270

2.2 Pre-processing

Salutations, recipient and sender names, location and dates at the beginning and end of the letters are removed. This is done because these elements are considered non-informative and may cause difficulties when splitting and concatenating the letters into segments, distorting the results of subsequent analyses and visualizations. After that the narratives are split into equally sized segments, for which word frequencies can be plotted along the horizontal axis. In a previous study by [10], document streamgraphs were created for the book "Tom Sawyer" by splitting the text into ten segments. Although using ten segments is suitable for long text documents like books, the narratives used in the current study are much shorter (see Table 2 for mean number of words per letter type). Therefore a smaller number of segments may be more appropriate. To decide on the number of segments to use, three different splits were made and the resulting visualizations were compared.

First, following [10], the narratives were split into ten segments, which resulted in very dynamic and detailed visualizations. However, these results were too fine-grained, making it difficult to use the visualizations for their initial purpose; to confirm previous findings and develop new hypotheses. Second, the narratives were split into three segments (representing the beginning, middle and end of the story, a structure often used in the formation and analysis of narratives, [14]). It was expected that the three segments would result in more interpretable visualizations revealing major trends. The resulting visualizations were however very global and flat, making it difficult to draw conclusions or gain new insights. Therefore third, based on the framework of [18], widely used to represent narrative information and analyse personal narratives, the narratives were split into five segments: orientation, complication, resolution, evaluation, and coda.

Although five segments may still seem too fine-grained for short narratives like the letters used in the current study, the 'narrative clause' used by [18]) as the basic unit of narrative can be as short as one sentence. This framework therefore is very suitable (and widely used) for analysing short narratives like daily life stories or therapeutic interviews [17]. In addition, splitting the narratives into five segments is in line with the five narrative processes used by [30] to identify the different letter types and letter structures. The five segments resulted into well-interpretable visualizations, showing the same trends as the visualizations for ten segments but then for larger-grained sections more inherent in personal letters.

Since the aim is to develop visualizations per letter type, for each type the letters are split into five equal segments and concatenated in one new text file per segment. This results in five new text files for each letter type, as shown in Figure 2. The five segments are analysed and visualized for each letter type separately.

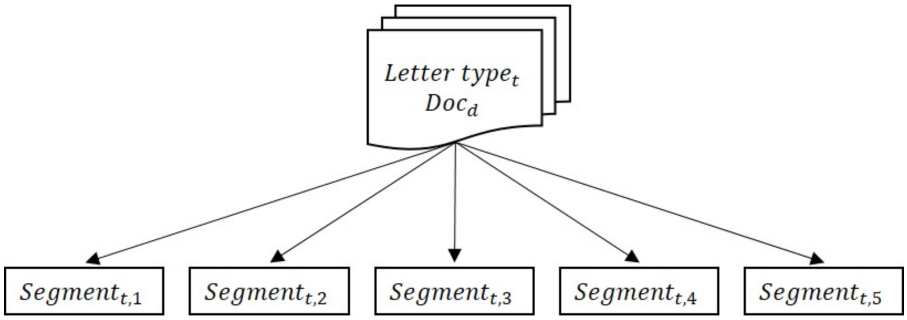

Figure 2 Splitting text documents into segments for each letter type.

2.3 Mathematical structure

To explore the differences in letter content and structure, plotting word-frequencies within each text segment for each letter type seems appropriate. However, since plotting frequencies for all used words will probably not lead to legible and interpretable visualizations, generally a sub selection of the occurring words is included in the visualizations. [10] for example only used words starting with capital letters or only the most prominent words as series in his graph. Another way to reduce the number of series is by categorizing them into word classes, as [36] did. In the current study words are categorized hierarchically using the text analysis program Linguistic Inquiry and Word Count (LIWC, [25]). LIWC is a structured, knowledge-rich method, relying on tight structures from existing software and dictionaries. LIWC processes texts on word level, comparing each word to a dictionary files for each category. It is a validated, ready-to use efficient and effective method to study a range of cognitive, emotional and structural components in spoken and written narratives [26].

In order to process Dutch texts, the Dutch LIWC dictionary developed by [38] was used. Contrary to the more complete English dictionary, the Dutch dictionary contains variables for the grammatical tenses past, present and future, but not for modals, intentional time or the imperative. The Dutch dictionary is based on the English LIWC dictionary (2001 version) and, as shown in Table 3, consists of 66 word categories divided over five dimensions. The words can be assigned to one or more categories, scoring the occurrences as percentages. The 66 LIWC categories are organized into a hierarchy of eleven main categories and 55 subcategories, which, when applied to the range of letter segments, results in a set of hierarchical additive time series.

2.4 Analytical procedure

As stated earlier, in this paper visualizations are used to explore differences in letter content both within the letters of the same type as between different types of letters. Two separate analyses were used to find the most informative categories. First, to find which categories best visualize the differences in category occurrence *between* the letter types a one-way analysis of variance (ANOVA) is used. The ANOVA is used to determine if there are significant differences between the means of multiple groups [22]. The mean category occurrence is calculated for each letter type by summing up the category scores for all segments and then dividing the sum by five (the total number of segments). The mean occurrences were compared using Welch's statistic [31].

Second, to find which LIWC categories fluctuate the most *within* each letter type, the spread in category occurrence values for the segments was evaluated. The most commonly

Table 3 Categories Dutch LIWC dictionary (translated from [38]).

I. Linguistic processes	35. Family (daughter, husband)
1. Pronouns (I, them, our)	36. Humans (adult, baby, boy)
2. 1st person singular (I, me, mine)	**III. Relativity**
3. 1st person plural (we, our, us)	37. Time (end, until, season)
4. Total 1st person (I, we, me)	38. Verbs in past tense (went, ran)
5. Total 2nd person (you, your, thou)	39. Verbs in present tense (is, does)
6. Total 3rd person (they, their, she)	40. Verbs in future tense (will, going)
7. Negations (no, not, never)	41. Space (nearby, place, North)
8. Assent (agree, ok, yes)	42. Up (above, higher, top)
9. Articles (a, an, the)	43. Down (deeper, lower, bottom)
10. Prepositions (to, with, above)	44. Including (and, inclusive, too)
11. Numbers (second, thousand)	45. Excluding (unless, except, out)
II. Psychological processes	46. Motion (approach, walk, climb)
12. Emotional (happy, sad, down)	**IV. Personal concerns**
13. Positive emotions (happy, pleased)	47. Occupation (achieve, promote)
14. Positive feelings (fun, love, smile)	48. School (student, exam)
15. Optimism (proud, passionate)	49. Work (job, career, colleague)
16. Negative emotions (hurt, hostile)	50. Achievement (earn, hero, win)
17. Anxiety (nervous, fearful, worried)	51. Leisure (cook, bike, movie)
18. Anger (hate, annoyed, threat)	52. Home (kitchen, home, garden)
19. Sadness (grief, disappointment)	53. Sports (game, fitness, work-out)
20. Cognitive (cause, know, ought)	54. Television (film, video, tv)
21. Causation (because, effect, hence)	55. Music (sing, song, guitar)
22. Insight (think, know, consider)	56. Money (profit, cash, owe)
23. Discrepancy (should, would, could)	57. Metaphysical (altar, church)
24. Inhibition (block, constrain, stop)	58. Religion (pray, honour, bless)
25. Tentative (maybe, perhaps, guess)	59. Death (bury, mourn, kill)
26. Certainty (always, never)	60. Physical (ill, faint, appetite)
27. Perceptual (observe, heard, feeling)	61. Body (vital, thirsty, cramp)
28. See (view, saw, seen)	62. Sexual (flirt, love, kiss)
29. Hear (listen, hearing)	63. Ingestion (drink, hungry, dish)
30. Feel (feel, touch)	64. Sleep (dream, wake, sleepy)
31. Social (share, talk, help)	65. Groom (shower, make-up)
32. Communication (interview, rumour)	**V. Experimental dimensions**
33. Other references (we, them, they)	66. Swear words
34. Friends (buddy, friend, neighbour)	

used measure of spread in a set of values is the standard deviation (SD). As low SD values indicate that all data points are close to the mean, LIWC categories with low SD values can be presumed to show little to no fluctuation in occurrence within the letter. LIWC categories with high SD values can be presumed to be highly fluctuating and thus showing more differences in occurrence within the concerning letter type. Since there are big differences in the means of the occurrence categories, to be able to compare the variation each SD is normalized with respect to its mean by: SD/Mean. The resulting value is known as the coefficient of variation (CV, also known as relative standard deviation), which shows the amount of variability in relation to the mean [20]. A major limitation of the CV is that when the mean is very small, a small variation in the dataset will already result in a large CV value [6]. Therefore the LIWC categories with Mean < 1 were excluded. Then for each letter type the ten most fluctuating LIWC categories (thus the ten categories with the highest CV scores) were selected and included in the letter type specific visualizations.

2.5 Text visualization design

Time series analysis, the study of changes in variables over time, can focus on one given variable, or the change of a specific variable compared to others over a certain time period. The time series consists of a sequence of measurements over a continuous, equal distanced time interval [28]. Time series are often visualized using simple line graphs, which works well when comparing a small number of series since the line graph shows direct values for each series at each time point. Another way to visualize time series are stacked graphs, where the series, represented by coloured layers, are stacked on top of each other, showing not only the individual values for each layer but also the total value at certain time points on the horizontal axis [7]. Stacked graphs are very useful to visualize hierarchical time series. However, both line and stacked graphs become illegible when using a large number of series [7], [10]. To overcome this problem [13] created ThemeRiver, a smooth, continuous graph stacked symmetrically around the x-axis, which is situated at the centre of the graph instead of at the bottom. ThemeRivers later became known as Streamgraphs, thanks to a popular visualization in the New York Times by [11]. Streamgraphs differ mainly from ThemeRivers in the design and layout decisions (like colour, interaction or geometry) made to make the graph visually attractive and more organic. Although originally applied to music, movies [3] and (baby) names [34], streamgraphs have also been applied to text documents and seem very suitable for the visualization of narratives.

The smooth, continuous lines that distinguish between the layers are the main advantage of a streamgraph, since this visualizes the data in an intuitive and easily interpretable way [13], [7]. Continuous data is required to generate such smooth, curving lines. However, splitting the texts into five separate segments results in a discrete dataset with different values for y at the data points x_1, x_2, …, x_5. This problem is solved by interpolating between the discrete data points as suggested by [13]. By using interpolation, intermediate values between data points are estimated from the neighbouring data points [12]. This results in smooth, continuous lines connecting the discrete data points [23]. There are different interpolation methods, of which the Cubic Splines model based on third degree polynomials results in the smoothest curve fits [12] and is therefore used in the current study.

There are two final notes with regard to the graph design. First, for all visualizations counts that since the layers are stacked symmetrically at the centre of the graph, the values on the y-axis are of no added value and are therefore not included in the plots, as in [13] and [7]. Second, sequential colour palettes were used to visualize the hierarchical structure of the time series, as was done by [35]. Each main category is assigned its own colour with a range colours with slightly different shades to reflect the corresponding subcategories.

3 Results

In this section first the results of the quantitative analysis are described, followed by the resulting visualizations.

3.1 Selected LIWC categories

To determine for which LIWC categories there are significant differences in mean occurrence between the different letter types, a one-way ANOVA was used. For 31 of the 66 LIWC categories (indicated by an asterisk (*) in Table 4) significant differences between the means ($p \leq 0.05$) were found. The coefficient of variation (CV) was used to measure the amount of variability in category occurrences throughout the letters. Table 4 contains the mean proportion (in %) and standard deviation of each LIWC category over all segments. For each letter type, the values for the most fluctuating categories are printed in bold. These categories are selected for the visualizations in Figures 4, 5 and 6.

Table 4 Means and standard deviations for each letter type

LIWC category	**Type 1** Mean (SD)	**Type 2** Mean (SD)	**Type 3** Mean (SD)	**Type 4** Mean (SD)	**Type 5** Mean (SD)	**Type 6** Mean (SD)
1. Pronoun*	12.75 (0.66)	11.63 (1.14)	10.56 (0.47)	13.41 (1.11)	13.67 (2.44)	13.19 (0.82)
2. I*	5.07 (0.80)	**3.42 (1.14)**	4.84 (0.62)	6.60 (0.36)	6.95 (1.93)	**3.55 (0.69)**
3. We*	0.46 (0.07)	0.56 (0.15)	0.66 (0.18)	0.43 (0.29)	0.45 (0.73)	0.12 (0.11)
4. Self*	5.54 (0.81)	**3.99 (1.00)**	5.50 (0.71)	7.04 (0.58)	7.40 (1.32)	**3.67 (0.74)**
5. You*	5.50 (0.75)	5.93 (0.76)	**3.11 (0.84)**	4.69 (0.84)	4.01 (1.16)	8.05 (0.30)
6. Other	0.67 (0.19)	0.78 (0.26)	**1.02 (0.29)**	0.65 (0.32)	0.96 (0.51)	0.46 (0.15)
7. Negation	**1.42 (0.41)**	**1.27 (0.35)**	**1.52 (0.38)**	1.51 (0.28)	**1.42 (0.53)**	2.09 (0.31)
8. Assent	0.17 (0.05)	0.14 (0.09)	0.15 (0.06)	0.13 (0.10)	0.00 (0.00)	0.26 (0.20)
9. Article	7.37 (0.76)	7.68 (0.74)	8.21 (0.91)	7.27 (0.94)	7.18 (0.92)	7.16 (0.51)
10. Prepos.*	11.03 (0.72)	11.40 (0.76)	11.13 (0.08)	10.54 (0.95)	10.96 (2.10)	10.24 (0.44)
11. Number	**1.24 (0.60)**	**1.15 (0.63)**	**1.21 (0.58)**	**1.26 (0.64)**	**1.25 (0.82)**	0.83 (0.51)
12. Affect	4.20 (0.66)	3.72 (0.34)	3.90 (0.63)	4.02 (0.94)	4.64 (1.43)	**4.18 (0.76)**
13. Pos. emo.	**2.94 (0.62)**	**2.62 (0.49)**	**2.89 (0.50)**	**2.54 (0.64)**	3.84 (1.18)	**2.67 (0.81)**
14. Pos. feel.	0.77 (0.27)	0.55 (0.17)	0.83 (0.23)	0.52 (0.21)	0.73 (0.47)	0.56 (0.14)
15. Optimism*	0.55 (0.17)	0.56 (0.17)	0.48 (0.17)	0.50 (0.21)	**1.64 (0.54)**	0.51 (0.23)
16. Neg. emo.*	1.15 (0.09)	0.98 (0.16)	0.92 (0.19)	**1.44 (0.34)**	0.73 (0.33)	1.45 (0.14)
17. Anxiety	0.21 (0.07)	0.22 (0.11)	0.11 (0.06)	0.25 (0.12)	0.00 (0.00)	0.27 (0.10)
18. Anger	0.11 (0.03)	0.13 (0.07)	0.10 (0.04)	0.05 (0.08)	0.06 (0.13)	0.21 (0.12)
19. Sadness*	0.28 (0.07)	0.15 (0.04)	0.34 (0.11)	0.45 (0.16)	0.28 (0.20)	0.33 (0.09)
20. Cognitive*	5.72 (0.86)	5.53 (0.28)	5.11 (0.71)	5.99 (1.25)	8.14 (1.39)	7.60 (0.34)
21. Causation	0.57 (0.13)	0.63 (0.07)	0.56 (0.14)	0.52 (0.21)	0.56 (0.34)	0.72 (0.10)
22. Insight	2.10 (0.29)	1.83 (0.18)	1.71 (0.28)	2.35 (0.41)	**2.04 (0.99)**	2.90 (0.24)
23. Discrep.*	**2.33 (0.47)**	2.46 (0.24)	**2.19 (0.37)**	**2.35 (0.65)**	5.26 (0.75)	2.99 (0.24)
24. Inhibition	0.06 (0.03)	0.03 (0.04)	0.06 (0.03)	0.07 (0.10)	0.00 (0.00)	0.08 (0.04)
25. Tentative	**1.50 (0.28)**	1.57 (0.15)	1.49 (0.17)	**1.71 (0.65)**	**2.65 (1.14)**	1.72 (0.18)
26. Certainty	1.58 (0.23)	1.20 (0.14)	**1.32 (0.23)**	1.62 (0.37)	**1.53 (0.62)**	**1.60 (0.46)**
27. Senses	1.27 (0.13)	1.23 (0.16)	1.27 (0.14)	**1.55 (0.39)**	0.68 (0.51)	**1.46 (0.35)**
28. See	0.48 (0.05)	0.39 (0.11)	0.52 (0.17)	0.41 (0.15)	0.28 (0.28)	0.53 (0.12)
29. Hear*	0.44 (0.06)	0.51 (0.15)	0.46 (0.09)	0.76 (0.15)	0.23 (0.37)	0.57 (0.25)
30. Feel	0.34 (0.07)	0.34 (0.09)	0.26 (0.10)	0.38 (0.20)	0.17 (0.25)	0.35 (0.08)
31. Social*	9.61 (1.12)	10.24 (0.65)	8.04 (0.57)	8.99 (0.95)	8.47 (1.52)	11.29 (0.12)
32. Comm.*	0.81 (0.13)	0.80 (0.11)	0.77 (0.12)	1.03 (0.21)	0.23 (0.24)	1.00 (0.11)
33. Others*	6.71 (0.88)	7.33 (0.61)	4.94 (0.42)	5.97 (0.88)	5.48 (1.50)	8.74 (0.19)
34. Friends*	0.24 (0.05)	0.24 (0.05)	0.20 (0.10)	0.14 (0.10)	0.39 (0.16)	0.24 (0.10)
35. Family*	0.89 (0.05)	0.80 (0.19)	0.80 (0.09)	0.88 (0.24)	1.02 (0.33)	0.46 (0.09)
36. Humans*	0.65 (0.16)	0.56 (0.18)	0.86 (0.10)	0.54 (0.25)	0.96 (0.43)	0.69 (0.17)
37. Time	**7.10 (1.23)**	**6.61 (1.97)**	6.82 (0.81)	7.13 (1.16)	**6.04 (2.25)**	**6.73 (1.30)**
38. Past*	**4.46 (0.83)**	**3.87 (0.97)**	**2.91 (0.52)**	5.83 (1.26)	0.96 (0.42)	**2.61 (0.93)**
39. Present*	12.61 (1.23)	12.57 (0.85)	13.44 (0.92)	12.18 (1.35)	14.12 (1.11)	15.05 (0.83)
40. Future*	0.93 (0.19)	**1.21 (0.20)**	0.83 (0.17)	0.90 (0.49)	2.71 (0.82)	1.42 (0.13)

Table 4 Means and standard deviations for each letter type (Continued)

LIWC category	**Type 1** Mean (SD)	**Type 2** Mean (SD)	**Type 3** Mean (SD)	**Type 4** Mean (SD)	**Type 5** Mean (SD)	**Type 6** Mean (SD)
41. Space	1.91 (0.28)	**1.95 (0.44)**	1.93 (0.28)	**1.62 (0.45)**	**2.03 (0.94)**	1.40 (0.20)
42. Up	1.11 (0.16)	0.96 (0.22)	1.21 (0.17)	**1.06 (0.35)**	0.73 (0.51)	0.97 (0.19)
43. Down	0.04 (0.03)	0.02 (0.02)	0.05 (0.05)	0.05 (0.05)	0.06 (0.13)	0.02 (0.02)
44. Incl.*	8.66 (0.21)	9.13 (0.84)	8.52 (0.19)	8.12 (0.54)	10.00 (0.63)	7.71 (0.77)
45. Excl.*	3.92 (0.54)	3.16 (0.24)	**3.64 (0.66)**	**4.15 (1.00)**	**4.01 (1.54)**	4.71 (0.51)
46. Motion	**1.87 (0.33)**	**2.15 (0.39)**	1.91 (0.19)	**2.04 (0.53)**	**1.75 (0.83)**	**1.98 (0.40)**
47. Occup.*	**2.04 (0.38)**	1.84 (0.27)	**1.33 (0.39)**	0.90 (0.22)	0.96 (0.71)	**1.60 (0.28)**
48. School*	0.76 (0.23)	0.72 (0.23)	0.38 (0.10)	0.40 (0.08)	0.40 (0.47)	0.72 (0.11)
49. Job*	**1.01 (0.26)**	0.91 (0.25)	0.75 (0.30)	0.45 (0.17)	0.51 (0.36)	0.49 (0.14)
50. Achieve*	0.32 (0.09)	0.25 (0.13)	0.22 (0.11)	0.11 (0.08)	0.11 (0.15)	0.41 (0.14)
51. Leisure*	0.59 (0.15)	0.91 (0.37)	0.87 (0.21)	0.95 (0.38)	0.73 (0.74)	0.29 (0.12)
52. Home*	0.49 (0.18)	0.73 (0.30)	0.68 (0.20)	0.45 (0.25)	0.45 (0.37)	0.22 (0.13)
53. Sports*	0.05 (0.02)	0.17 (0.07)	0.14 (0.07)	0.27 (0.14)	0.23 (0.37)	0.05 (0.04)
54. TV	0.03 (0.03)	0.01 (0.02)	0.01 (0.01)	0.07 (0.08)	0.06 (0.13)	0.00 (0.00)
55. Music	0.02 (0.02)	0.01 (0.03)	0.04 (0.03)	0.22 (0.14)	0.06 (0.13)	0.02 (0.02)
56. Money*	0.25 (0.12)	0.32 (0.12)	0.39 (0.12)	0.07 (0.08)	0.34 (0.12)	0.27 (0.05)
57. Metaphys.	0.07 (0.01)	0.06 (0.06)	0.08 (0.03)	0.13 (0.12)	0.00 (0.00)	0.04 (0.04)
58. Religion	0.04 (0.02)	0.04 (0.04)	0.07 (0.03)	0.07 (0.08)	0.00 (0.00)	0.04 (0.04)
59. Death	0.03 (0.02)	0.03 (0.03)	0.02 (0.01)	0.05 (0.08)	0.00 (0.00)	0.00 (0.00)
60. Physical	0.66 (0.10)	0.54 (0.14)	0.83 (0.09)	0.74 (0.56)	0.73 (0.32)	0.66 (0.14)
61. Body	0.30 (0.03)	0.24 (0.06)	0.33 (0.05)	0.34 (0.39)	0.34 (0.12)	0.40 (0.09)
62. Sexual	0.06 (0.03)	0.06 (0.03)	0.07 (0.03)	0.02 (0.04)	0.06 (0.13)	0.10 (0.08)
63. Eating	0.07 (0.02)	0.07 (0.02)	0.20 (0.11)	0.11 (0.12)	0.17 (0.16)	0.05 (0.03)
64. Sleep	0.24 (0.08)	0.19 (0.12)	0.23 (0.07)	0.32 (0.19)	0.23 (0.13)	0.14 (0.07)
65. Groom	0.00 (0.00)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
66. Swear**	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.02)

Note:
Bold values: ten most fluctuating LIWC categories for each letter type.
* Significant differences ($p \leq 0.05$) between means of different letter types
** Only occurred in one letter type so means could not be compared

3.2 Visualizations

The figures below contain the streamgraphs for each letter type. The mean proportion of each LIWC category (over all the letters of the concerning letter types) is plotted per segment (s_1, s_2, ..., s_5) on the x-axis. The panel in Figure 3 contains six streamgraphs, one for each letter type. These visualizations show differences in occurrence proportions of LIWC categories throughout each letter type. All 66 LIWC categories are included in these graphs. The darkest shades of every colour show the main (overarching) categories, followed by the corresponding sub categories. The categories are plotted in the same order for each graph. These graphs can be used to find central themes within the letters and overall differences *between* the letters. In the legend, the asterisk (*) behind LIWC categories indicates that there are significant differences between the mean occurrences of the letter types for these categories. The visualizations in Figures 4, 5 and 6 show the ten most fluctuating LIWC categories for each letter type. These graphs can be used to find specific patterns and shifts in the occurrence proportions of LIWC categories *within* the letters.

The streamgraphs in Figure 3 show some clear similarities and differences between the six letter types. An interesting finding is that the visualizations for types 1–3 do not seem to differ as much as was expected based on the previous findings of [30]. Overall, the imagination

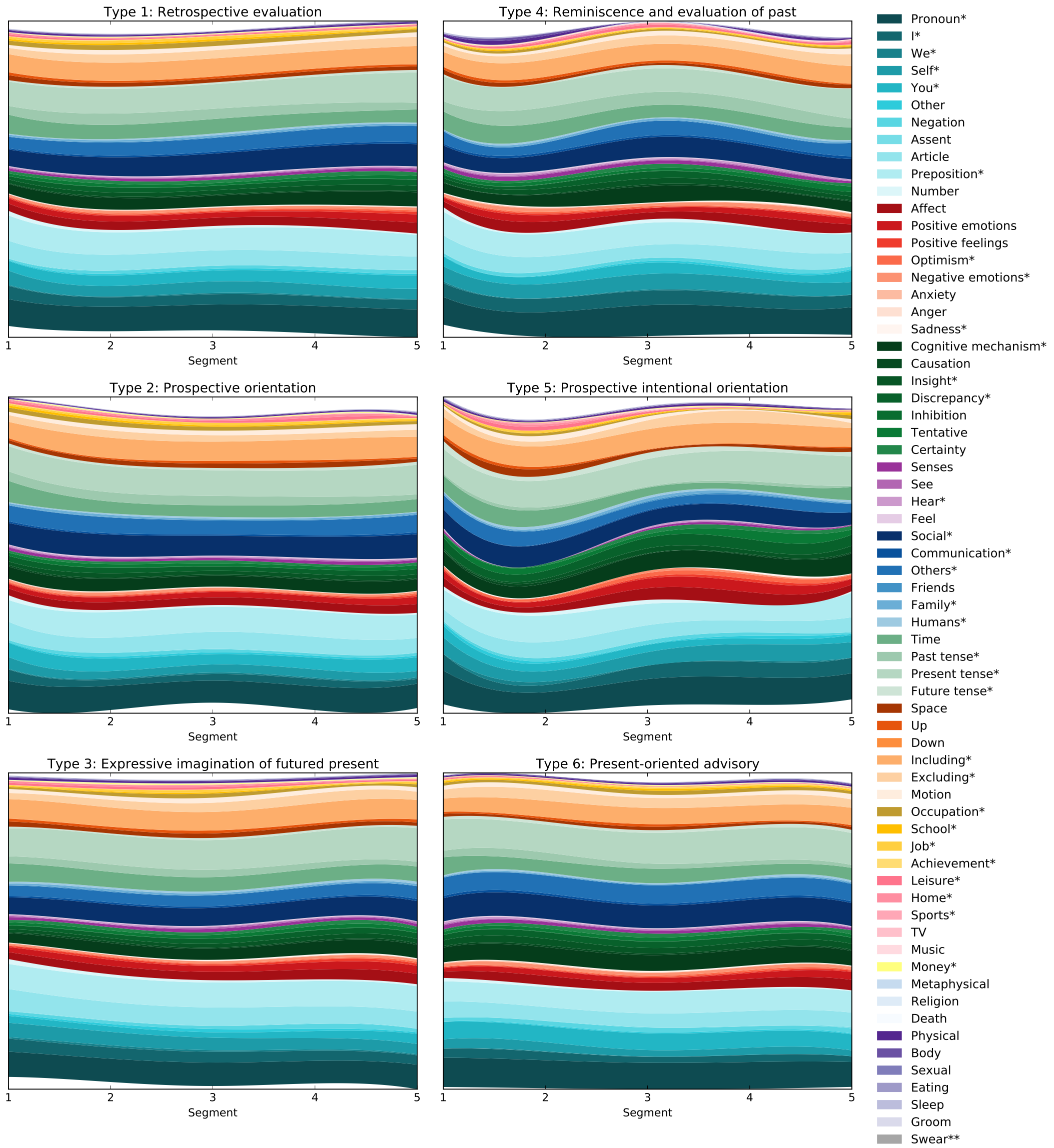

Figure 3 Overview LIWC categories per letter type.

letters (type 1–3) seem to have a calmer flow than the general letters (type 4–6), which show bigger differences in the proportions of the LIWC categories over the five segments. The differences between and within the streamgraphs will now be described in more detail and compared pairwise for the retrospective letters (types 1 and 4), prospective letters (types 2 and 5) and present-oriented letters (types 3 and 6).

3.2.1 Retrospective letters

[30] found that imagination and general retrospective letters generally have the same structure. This is also reflected by the streamgraphs in Figure 3, which show that for the majority of the LIWC categories the distribution of the category proportions over the segments is quite similar for both retrospective letters.

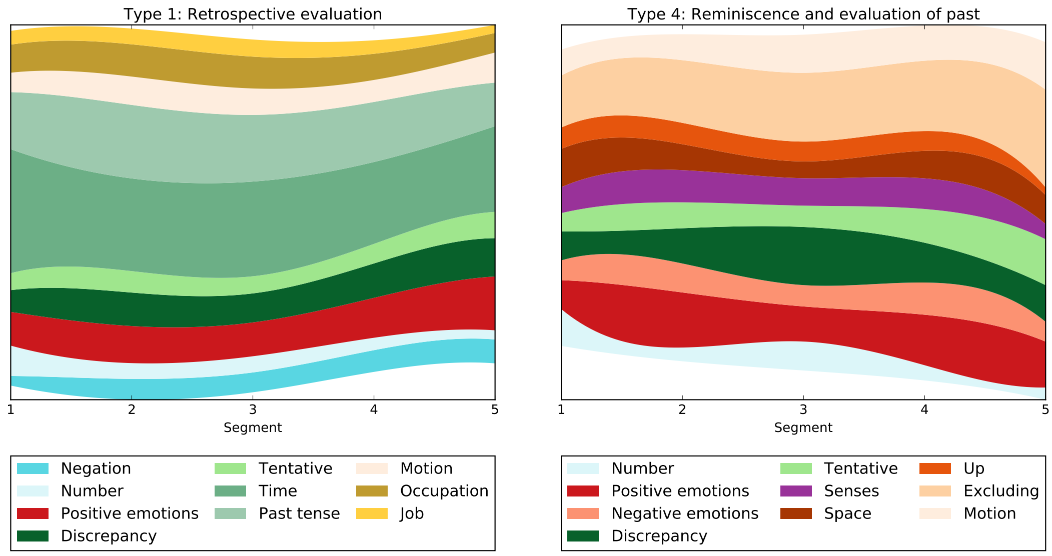

Figure 4 Ten most fluctuating categories retrospective letters.

According to [30], the main difference between both retrospective letters would be the verb tenses and the sequence in which these are used. The type 1 letters start with an imaginative future situation in the present tense followed by reminiscence of the future past in the past tense, whereas in the type 4 letters the recounted period actually lies in the past instead of the futured past and is described as a present concern. Based on these findings, one would expect to observe differences in the proportions of used verb tenses both between (Figure 3) and within (Figure 4) the letters. However, Figure 3 shows no observable differences in the proportions of the LIWC categories that regard verb tenses (past, present and future) between letters 1 and 4. Figure 4 does show “past tense” as one of the ten most fluctuating categories for letter type 1: the use of past tense slightly increases towards the middle of the letter and then decreases towards the end. The use of present tense does not seem to differ much throughout the type 1 letter as it is not amongst the ten most fluctuating categories included in the graph. None of the used tenses fluctuates much throughout the type 4 letters, as they are not amongst the ten categories included in the graph in Figure 4.

Overall it can be observed from both Figure 3 and Figure 4 that the imagination letters (type 1) contain more words regarding occupation and job, combined with motion words and positive emotions. The words related to occupation and job could be linked to the narrative element “orientation”, the first narrative element distinguished by [18]. The motion and positive emotion words could be used to describe the (path towards) the desired future situation or a period of personal growth (“complicated action”, [18]). The graphs further show an increasing use of discrepancy words (e.g. should, could, would) from the middle to the end. This supports the findings of [30], who state that towards the end of the letters conclusions or insights are drawn (pointing towards the narrative elements “evaluation” and “resolution”, [18]), followed by statements of worldly wisdom self-praising remarks (which could be defined as the “coda”, [18]).

For type 4 letters, Figure 3 and Figure 4 show an increase in the use of words from the categories “physical” and “body” combined with both positive and negative emotion words at the beginning and end of the letter. This could indicate that the element “orientation” from the framework of [18], contains mainly physical characteristics in letter type 4, as opposed to

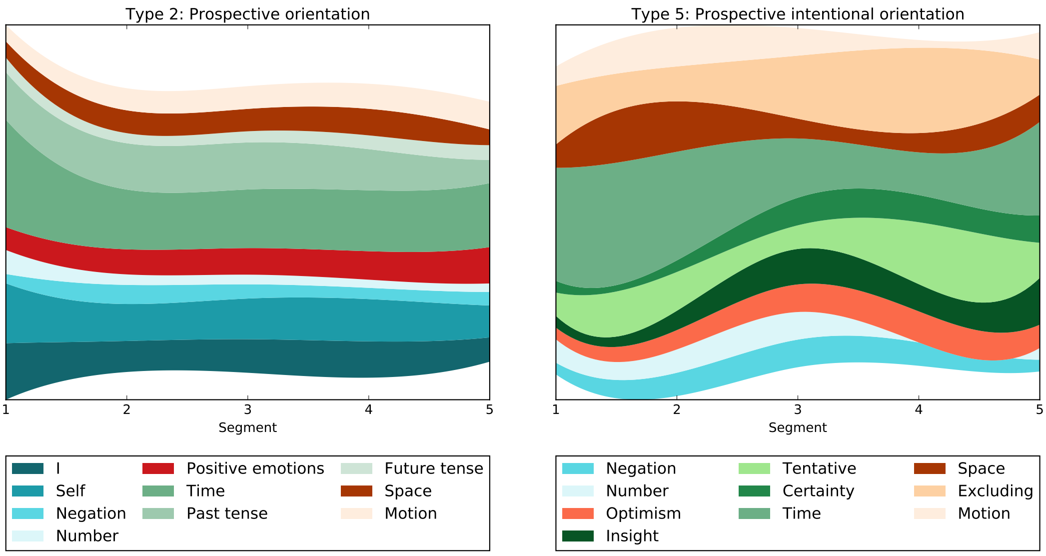

Figure 5 Ten most fluctuating categories prospective letters.

the professional characteristics used in letter type 1. Cognitive mechanisms are used more from the middle (insight and discrepancy) to the end (tentative) of the letters. This could be because these writers are still in the process of reminiscing and evaluating past events (pointing towards the elements "evaluation" and "resolution" of [18]). It could be that these letters start with a description of physical or emotional complaints or by a recollection of a happier past, which is then processed and evaluated, followed by moral advice or a tentative promise for a better future (distinguished by [30] as the "coda"). Finally the type 4 letters also contain more words related to senses and leisure. Overall, the general letters seem to be more sensitive, expressive and detailed than the type 1 letters.

3.2.2 Prospective letters

[30] found a clear structure for the imaginative letters (type 2), but not for the general letters (type 5). The imaginative letters were expected to start with a statement about one's present position in life (in present or past tense). This is reflected in the high occurrence of words in the categories "I" and "Self" and words related to "Time" (e.g. end, until) and numbers at the beginning of the letters (see Figure 5), which could be used to describe one's present position in life (narrative element "orientation" [18]). The increase in the use of words regarding space (e.g. nearby, places, directions), in the middle can reflect concrete imaginary goals and purposes. The path towards the futured situation (possibly the "complicated action", [18]) could be indicated by the increasing use of motion words and positive emotions.

With regard to the used tense, Figure 3 further shows that, in addition to the present tense, more past tense is used in the type 2 letters, whereas more future tense is used in the type 5 letters. This is in line with the findings of [30]. Overall the general letters contain more affect and emotions, and more cognitive mechanisms towards the end, which could point towards encouraging oneself to realize their goals, as described by [30].

The intentional element, the major characteristic of the type 5 letters, is clearly reflected in Figure 3 by the use of future tense and the high occurrence of tentative words (like "hope", "believe", "try", "possible") and discrepancy words ("must", "wish","want"). Figure 5 further

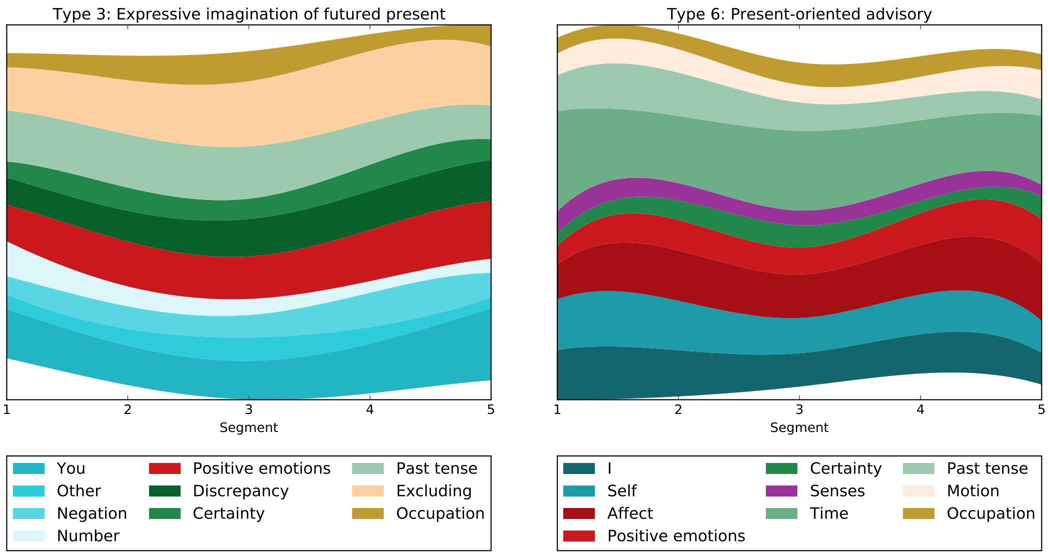

Figure 6 Ten most fluctuating categories present-oriented letters.

shows that the type 5 letters start and end more tentative, alternated with insight in the middle and end (pointing towards "evaluation", [18]). This letter is increasingly optimistic and certain, combined with an increasing use of excluding words. This might point towards an increasing insight in desired versus non-desired situations or future aspects, which may lead to more a more positive and concrete vision for the future in the "coda" [18]. However, the increasing use of excluding words combined with the high use of tentative and hesitative words could also reflect the doubt and uncertainty regarding the future related to prospective intentional orientation.

3.2.3 Present-oriented letters

[30] found no specific sequential order in narrative processes for the present-oriented letters. Figure 3 shows that the present-oriented letters are quite similar for both categories. However, the imagination letters (type 3) do contain more words regarding family, leisure, more superlatives (category "up") and slightly more positive emotions and feelings. This is in line with the findings of [30], who found that type 3 letters are positive, content, and joyful letters.

The letters generally end with hopes and wishes (shown by the increase in discrepancy words) and contain a lot of self-praising remarks (shown by the high increase in the use of "you" in the middle and end). This could point to the narrative elements "resolution" and "coda" [18]. The low use of cognitive mechanism and insight words supports the findings of [30] that the letter contains almost no orientation or evaluation, two of the five narrative elements distinguished by [18]. The high use of excluding words could point towards a breach with the past, without describing the current situation or the path from past to future (no "complicated action" [18]). The additional increase in the use of certainty towards the end indicates that the letters become more stimulating and convincing at the end (indicating "result/resolution" or "coda" [18]). It seems that the confidence of the writer increases by imagining the future situation. Finally, regarding the used tense, the type 3 letters are written mainly in the present tense, although Figure 6 shows that in both letters the past tense is used more in the beginning than in the middle and end of the letters.

In the general letters (type 6), more insight and discrepancy words are used. These letters also contain more negative emotions and feelings and slightly more sensory words. This supports the findings of [30], who state that the function of these letters is mainly to provide insight in and guidance for current problems or concerns, followed by statements of worldly wisdom. The finding of [30] that these letters do not contain a clear path or clarification of how and where certain knowledge or insights have been gained is supported by the fact that these letters contain almost no causation words. The high use of certainty words in the middle of the letter may be explained by the statements of wisdom and moral advice, combined with the fact that these letters do not contain evaluative aspects, which introduce more uncertainty. Apart from the elements "resolution" and "coda" it is difficult to link the letter characteristics from the visualizations to the narrative elements of [18].

4 Discussion

In this paper, a combination of Natural Language Processing, quantitative analysis and visualization techniques was used to explore differences in letter content, specifically the distribution (sequential order) and proportion of narrative processes and grammatical elements, both within and between the different types of "Letters from the Future". The visualizations could be used for two purposes; to confirm findings of previous studies on the content of the letters and to explore the letters in a broader sense to come to new insights or theories. Two essential topics in the development of text visualizations – capturing the underlying mathematical narrative structure and choosing a suitable format to visualize changes in letter content throughout the letter – were addressed. In general, the use of text visualizations proved to be a good method to globally explore and compare the underlying structures and differences in contents within and between the letter types. Thanks to the shape of the streamgraphs and the use of sequential colour palettes, the hierarchical time series plots of the letters were easily interpretable and comparable. By combining the visualizations with quantitative analysis of variance and the coefficient of variation, more specific insights in the distribution and proportion of narrative processes and grammatical elements throughout the letters was gained.

All in all, the visualizations were found to be very usable to at least partially confirm the previous findings of [30]. Finding strong additional characteristics or differences between and within the letters turned out to be more challenging. An interesting finding is that the proportional distributions of the LIWC categories, especially those of letter types one, two and three do not differ as much as expected based on the previous findings of [30]. The visualizations for those types look very similar, as opposed to the visualizations for letter types four, five and six. An explanation for this may be that the LIWC categories used as underlying structure are too global or do not directly apply to the current dataset. A more specific categorization system developed especially for the "Letters from the Future" dataset might perform better. A possibility is to develop a new LIWC dictionary based on the previous findings of [30] and the visualizations generated in this study, and apply this to a new dataset. Potential features to include in this dictionary could be the most informative features that discriminate between the six letter types. These most informative features have been extracted from the current dataset for a different study by the authors in which supervised text classification algorithms are used to automatically categorize the letters to their corresponding classes. It would be interesting to visualize the occurrence of these features within the letters.

It could also be that the way the letters are split into five segments influences the proportional distributions. For example, when a certain narrative process starts at the end of the first segment and finishes at the beginning of the second segment, the characteristics for this process are evened out between the first to segments. This may cause a blur in the resulting visualization. It would be interesting to see if splitting the letters manually into five segments, based either on the narrative elements of [18] or the five narrative processes distinguished by [30] would lead to more distinctive variations both between the letter segments and the letters as a whole.

Splitting the narratives into the structural elements distinguished by [18] also opens up to a new avenue for future research, namely to investigate variations in the narratives that depend on the characteristics of the writer. The framework of [18] has already been used to investigate differences in narrative content between classes [17], [15], gender [16], [8], age [27], [32] and geography [16]. Visualizing the narrative structures for groups with different characteristics may lead to new insights or hypotheses for further research on these topics.

As a final note, although the current focus is on visualizing the content of “Letters from the Future”, the resulting method can in fact be used to explore any available digital text document or corpus. The methods and results described in this paper can be seen as a first step in an ongoing study by the authors and the Storylab to study therapy-related textual features in e-mental health interventions. By using methods like NLP and text visualization to analyse patterns in therapy-related textual features, extracted for example from written narratives or the linguistic interaction between counsellor and client, more insight can be gained in what happens within therapy, when progress is made, or for which persons a certain type of therapy is more effective. This could greatly improve e-mental health interventions and advance therapy change process research. Future research will therefore include expanding the time series to include more letters written by the same person, studying changes between subsequent narratives and analysing counsellor-client interaction.

References

1 Adeline Abbe, Cyril Grouin, Pierre Zweigenbaum, and Bruno Falissard. Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research*, 2015. doi:10.1002/mpr.1481.

2 Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python.* O'reilly Media, Inc, Sebastopol, 2009.

3 M. Bloch, L. Byron, S. Carter, and A. Cox. The ebb and flow of movies: Box office receipts 1986-2007. *New York Times*, 2008. URL: http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html?_r=0.

4 Ernst Bohlmeijer. *De Verhalen die we leven. Narratieve psychologie als methode.* Boom, Amsterdam, 2007.

5 J. Bradley and G. Rockwell. What scientific visualization teaches us about text analysis. In *ALLC/ACH Conference*, Paris, 1994.

6 Charles E. Brown. Coefficient of Variation. In *Applied Multivariate Statistics in Geohydrology and Related Sciences*, pages 155–157. Springer Science & Business Media, 2012.

7 Lee Byron and Martin Wattenberg. Stacked graphs – Geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008. doi:10.1109/TVCG.2008.166.

8 Jenny Cheshire. The telling or the tale? narratives and gender in adolescent friendship networks. *Journal of Sociolinguistics*, 4(2):234–262, 2000.

9 Gobinda G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89, 2005. doi:10.1002/aris.1440370103.

10 Jeff Clark. Tom Sawyer Character StreamGraph, 2008. URL: http://www.neoformix.com/2008/TomSawyer.html.

11 A. Cox and L Byron. The Ebb and Flow of Box Office Sales, 1986-2007,. *The New York Times, February 23, 2008*, 2008.

12 O. A. de Carvalho, Renato Fontes Guimarães, Roberto Arnaldo Trancoso Gomes, and Nilton Correia da Silva. Time series interpolation. In *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, pages 1959–1961. IEEE, IEEE, 2007.

13 Susan Havre, Beth Hetzler, and Lucy Nowell. Themerivertm: In search of trends, patterns, and relationships. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

14 P. C. Hogan. Continuity and change in narrative study. Observations on componential andfunctional analysis. *Narrative Inquiry*, 16(1):66–74, 2006.

15 Barbara Horvath. Text on conversation: Variability in storytelling texts. In K. Denning, S. Inkelas, F. McNair-Knox, and Rickford. J., editors, *Variation in Language*. Department of Linguestics, Stanford, 1987.

16 Barbara Johnstone. Variation in discourse: Midwestern narrative style. *American Speech*, 65(3):195–214, 1990.

17 William Labov. Some further steps in narrative analysis. *Journal of Narrative and Life History*, 7:395–415, 1997.

18 William Labov and Joshua Waletzky. Narrative analysis: oral versions of personal experience. In J. Helm, editor, *Essays on the verbal and visual arts*, chapter Narrative, pages 12–44. Washington University Press, Seattle, 1967.

19 E. D. Liddy. Natural language processing. In M.A. Drake, editor, *Encyclopedia of library and information science*. Marcel Decker, New York, 2nd edition, 2001.

20 Pat Lovie. Coefficient of Variation. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd:, Oxford, UK, 2005.

21 Inderjeet Mani. Computational Narratology. In Peter Hühn, Christoph Meister, Jan, John Pier, and Wolf Schmid, editors, *Handbook of Narratology*, pages 84–92. De Gruyter, 2014.

22 Scott E. Maxwell, Harold D. Delaney, and Ken Kelley. *Designing Experiments and Analyzing Data*. Taylor & Francis Group, New York, 2003.

23 Duncan M. McGreggor. *Mastering Matplotlib*. Packt Publishing, 2015.

24 Jan C. Meister and Alastair Matthews. *Computing Action*. De Gruyter, Berlin, 2003.

25 J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic Inquiry and Word Count (LIWC): LIWC2001, 2001.

26 James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. The development and psychometric properties of LIWC2007, 2007.

27 C. Peterson and A. McCabe. *Developmental psycholinguistics: Three ways of looking at a child's narrative*. Plenum Press, New York, 1983.

28 Robert H. Shumway and David S Stoffer. *Time series analysis and its applications. With R examples*. Springer Science & Business Media, New York, 2006. `doi:10.1016/j.peva.2007.06.006`.

29 Anneke Sools and Jan Hein Mooren. Towards Narrative Futuring in Psychology: Becoming Resilient by Imagining the Future. *Graduate Journal of Social Science*, 9(2):203–226, 2012.

30 Anneke M. Sools, Thijs Tromp, and Jan H. Mooren. Mapping letters from the future: Exploring narrative processes of imagining the future. *Journal of Health Psychology*, 20(3):350–364, 2015. `doi:10.1177/1359105314566607`.

31 Andrew J. Tomarken and Ronald C. Serlin. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, 99(1):90–99, 1986. `doi:10.1037/0033-2909.99.1.90`.

32 M. J. Toolan. *Narrative: A critical linguistic introduction*. Routledge, London, 1988.

33 P. Valéry, G. Rockwell, and J. Bradley. Printing in Sand; Scientific Visualization and the Analysis of Texts. *Geoffreyrockwell.Com*, pages 1–16, 1999.

34 Martin Wattenberg. Baby names, visualization, and social data analysis. In *IEEE Symposium on Information Visualization. INFOVIS 2005.*, pages 1–7. IEEE, 2005.

35 Martin Wattenberg and Jesse Kriss. Designing for social data analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 12(4):549–557, 2006.

36 Wibke Weber. Text visualization-what colors tell about a text. In *Information Visualization, 2007. IV'07. 11th International Conference*, pages 354–362. IEEE, 2007.

37 James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Visualizing the non-visual. Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Information Visualization Symposium 1995*, pages 51–58, Atlanta, Georgia, 1995. IEEE.

38 Hanna Zijlstra, Tanja Van Meerveld, Henriët Van Middendorp, James W. Pennebaker, and Rinie Geenen. De Nederlandse versie van de 'Linguistic and Word Count' (LIWC). Een gecomputeriseerd tekstanalyseprogramma. *Gedrag & Gezondheid*, 32(4):271–281, 2004.