Rizun, N., & Revina A., Meister V. (2019), „Discovery of Stylistic Patterns in Business Process Textual Descriptions: IT Ticket Case", in Soliman, K.S. (Ed) Education Excellence and Innovation Management through Vision 2020, International Business Information Management Association (IBIMA), pp. 2103-2012

Discovery of Stylistic Patterns in Business Process Textual Descriptions: IT Ticket Case

Nina RIZUN
Gdańsk University of Technology, Gdańsk, Poland, nina.rizun@pg.edu.pl
Aleksandra REVINA
Technical University of Berlin, Berlin, Germany, revina@tu-berlin.de
Vera MEISTER
Brandenburg University of Applied Sciences, Brandenburg an der Havel, Germany, vera.meister@th-brandenburg.de

Abstract

Growing IT complexity and related problems, which are reflected in IT tickets, create a need for new qualitative approaches. The goal is to automate the extraction of main topics described in tickets in order to provide high quality support for the IT process workers and enable a smooth service delivery to the end user. Present paper proposes a method of knowledge extraction in a form of stylistic patterns in business process (BP) texts, here in incoming IT tickets texts. Hereby, the authors set an objective to predict their readability and perceived complexity for a process worker, what will influence further tasks execution. The results of experimental analysis of a data set of incoming ticket texts from an ITIL-based Change Management process showed that the specificity of stylistic patterns expressing the readability of a ticket and perceived complexity could be identified with the help of proposed measures of the ticket length, parts-of-speech distributions and wording style.

Keywords: IT Tickets, Text Mining, Stylistic Analysis, Readability

Introduction

Increasing digital channels and communication flows in the business context create a demand for new qualitative algorithms and techniques of identifying main aspects (topics, problems) described in potentially huge amounts of generated unstructured textual data as well as for effective instruments to support the process of interpreting the results obtained in this process (Rizun & Kucharska, 2018).

In an enterprise context, one of the units constantly dealing with unstructured textual data inputs is an IT department responsible for processing of the tickets on problems related to IT products or services. Many technologies, methods and tools evolve in order to address (among others) the trend of growing IT complexity and amount of work for IT departments (Wróblewska, Stanisławek, Prus-Zajaczkowski, & Garncarek, 2018). Hereby, the technologies of Text Mining and Natural Language Processing play an important role particularly in the area of IT ticket processing. The rapid pace of Text Mining technologies and their diverse application areas (Rizun & Kucharska, 2018; Chiu & Hsu, 2018; Rangu, Chatterjee, & Valluru, 2017) offer a broad spectrum of opportunities for future research work.

In the present paper, the researchers aim to apply recent approaches in the area of Text Mining to address the problems of the IT ticketing process. The suggested research is based on the state-of-the-art knowledge and focuses on the development of a new method for stylistic patterns retrieval that will enable novel capturing, extraction and interpretation of multi-aspect knowledge at the point of ticket entry, i.e. incoming ticket texts. Hereby, the goal is to predict ticket readability and perceived complexity allowing the differentiation between routine (simple) and cognitive (complex) tickets.

Related Work and Research Questions

As a background information for the present paper, the authors refer to the recent research (Rizun, Revina, & Meister, 2019) and suggest that in the incoming ticket texts there exist certain stylistic "patterns" expressing the readability of the ticket, under the influence of which the nature of the ticket in the context of its complexity and Decision-Making Logic (DML) level (routine vs cognitive) can be determined. This makes it possible to make statements about further ticket processing and propose appropriate AI-based support tools.

While emphasizing the term complexity at the current research stage, the researchers differentiate two levels: (1) perceived complexity and (2) objective processing complexity. The first one is related to the initial contact of a process worker with the incoming IT ticket text and is measured with its readability indicator. Thus, under perceived complexity, the authors understand ambiguity (number of various possible interpretations of the same piece of information) and uncertainty (lack of clarity about action-outcome relations) (Wood, 1986; Nadkarni & Gupta, 2007) that individuals face while reading the incoming ticket text. The second complexity level is associated with the objective complexity in the tasks processing and is suggested to be measured by already mentioned DML based on the incoming IT ticket text. Under DML, the researchers understand "cognition" level of a decision-making process, i.e. processing complexity of tasks to be executed within the process by a process worker (Rizun, Revina, & Meister, 2019). In the mentioned paper, the authors conducted a semantic analysis and presented the method of DML discovery in the BP textual data, whereby the concept of RTCC semantic tagging framework (nouns as ***R**esources*, verbs and verbal nouns as ***T**echniques*, adjectives as ***C**apacities*, and adverbs as ***C**hoices*) was introduced. The research was inspired by studies in the area of quantitative text analysis and the related conceptual framework building that discuss text analysis in a combination of such academic orientations as linguistics, computer science and social sciences (Roberts, 2000; Lasswell, 1948; Ordenes, Burton, Theodoulidis, Gruber, & Zaki, 2014). In line with the semantic meaning identified based on suggested contextual variables (for example, problem processing level, information, experience), the RTCC elements, parsed descriptive key words extracted from the same case study data set, are classified into three DML levels – routine, semi-cognitive and cognitive (Rizun, Revina, & Meister, 2019). The RTCC semantic tagging approach is used in the current paper while conducting stylistic analysis.

In the present paper, the authors focus on the readability indicator and perform a specified analysis in order to find stylistic patterns and measure the readability or comprehension difficulty of a text. There is a number of traditional approaches devoted to readability measures (Woloszyn, dos Santos, & Wives, 2017). The authors suggest a specific stylistic patterns analysis for measuring the readability, which is detailed next. The stylistic patterns analysis provides an opportunity to better understand texts and authors. Hereby, style can be defined as the "manner" in which a message is written. Stylistic variation depends on author's preference, competence, genre, and untold other factors. As a rule, it is expressed through subtle frequency variations of otherwise insignificant measures of a text that, taken together, are understood as stylistic indicators by a particular reader community (Shlomo, Karlgren, & Shanahan, 2005). In the present paper, as measures of stylistic patterns (also styles) influencing the readability of incoming ticket texts, the researchers propose the following: *Syntactic Structure and Wording Style*.

The first measure suggested by the authors is the *Syntactic Structure* (SynS) of a ticket text. Under syntax, one understands the set of rules or regularities that govern how words are put together to form phrases (Koopman, Sportiche, & Stabler, 2013). There are relevant studies related to the analysis of syntactic structure in various domains (Kaljahi, et al., 2015; Huadhom & Trakulkasemsuk, 2017; Niederdeppe, 2005). To the best of the researchers' knowledge, the area of SynS of IT ticket texts is not well studied yet. Therefore, the first research question posed in the paper is:

RQ1: Does the knowledge about SynS of the BP description texts, here in incoming IT ticket texts, provide an additional value for the prediction of their readability and perceived complexity?

The second measure proposed by the researchers is the *Wording Style* (WS), which was analyzed with the help of Zipf's Laws (Zipf, 1932). There is certain research work dedicated to the value of using Zipf's Laws to give a good description of data (KiBaek, Bernhardsson, & Minnhagen, 2011). The novelty of the present approach is its application domain of IT tickets, results interpretation, Zipf's Laws application as a WS measurement, and a novel interpretation of the WS itself, i.e. speed of new words appearance in the ticket texts (Rizun & Taranenko, 2018). Thus, the second research question would be:

RQ2: Does the knowledge about WS of BP description texts, here in incoming IT ticket texts, provide an additional value for the prediction of their readability and perceived complexity?

Research Methods and Hypotheses Development

As mentioned above, the present paper is focused on the stylistic patterns measures, such as SynS and WS, obtained while performing a computational analysis of ITIL (Axelos, 2019) Change Management ticket texts. Before conceptualizing the measures, the researchers conducted a short qualitative survey with the case study process workers in order to evaluate the dependency of perceived ticket processing complexity from the suggested list of criteria, such as ticket length, level of details, wording style among others, based on the scale from 1 (irrelevant criterion) to 10 (highly relevant criterion). The following criteria received relatively high ratings: ticket length being with the second highest score of 6.3 (placed after the criterion of standard business vocabulary usage) followed by the level of details (6.0) and wording style (3.7). Furthermore, the researchers discovered while performing the survey that case study process workers usually receive short texts in case of simple, explicit and already familiar requests. To a certain extent, this fact is also supported by the theory of the least effort (Zipf, 1949).

Based on the collected information, the researchers introduce the following Assumptions (A):
A1: As one parameter, allowing the differentiation of the ticket text readability and perceived complexity, the ticket text length will be accepted.
A2: The ticket text readability and perceived complexity can be largely understood as functions of the suggested stylistic measures SynS and WS (among other measures not in scope of the present paper).

The researchers suggest the following criteria as a measure of SynS (assesses the criterion level of details in a qualitative survey):
- *Distribution of PoS* in relation to all words in the ticket text;
- *Distribution of unique* (not repeating) *PoS* in relation to all words in the ticket text.

The WS measure analysis is performed using Zipf's First "Rank-Frequency" and Second "Quantity-Frequency" Laws (Zipf, 1932; Rizun & Taranenko, 2018), which bring in relation the frequency of words with their importance rank and the frequency of words with the quantity of words with the same frequency.

Based on the information presented above, the following hypotheses (H) were formulated:
H1: The SynS specificity of stylistic pattern depends on the length of the ticket text.
H2: The WS specificity of stylistic pattern depends on the length of the ticket text.

To sum up, in (Rizun, Revina, & Meister, 2019), using the DML method and RTCC semantic tagging framework, the authors extracted the knowledge out of the incoming IT ticket texts in order to assess the objective ticket processing complexity. The stylistic analysis-based readability indicator accompanied with the perceived complexity prediction complements and justifies the conducted research in the following: (1) SynS stylistic measure (distribution of PoS and unique PoS) provides information on the presence of RTCC elements organized as per three DML levels in the incoming ticket texts; (2) WS stylistic measure accounts for the density (condensed vs dispersed) of the RTCC elements in the text. Thus, the authors presume a positive correlation between SynS, WS and RTCC elements in the DML levels.

Data and Results

For the development of the suggested stylistic patterns measures, the ticket data set from the ticketing system of the IT Change Management (according to ITIL Framework) department of a big enterprise with more than 200,000 employees worldwide was processed and converted into a CSV-formatted text corpus with more than 1,000,000 documents (text entries) of English, German and English-German ticket texts created in the period of 2015-2018. After removing duplicates and selecting English texts, the final case study data sample comprised 28,157 entries.

Following the assumption A1, the researchers introduce the parameter *Length (L)* of a ticket text. Using Python 3.4.3 and the Microsoft Office Excel Data Analysis tool, the researchers identified the average, minimum and maximum length of the ticket texts sample based on the word count and their distribution. Applying these statistical characteristics, the authors formed three main clusters: 1) tickets with under (or equal to) 25 words (40% of the sample); 2) tickets with 26 to 100 words (38% of the sample); 3) tickets with more than 100 words (22% of the sample).

Syntactic Structure

As mentioned above, the researchers consider the following criteria as a characteristic of the SynS measure of the incoming ticket text readability: distribution of PoS and unique (not repeating) PoS in relation to all words in the ticket text. In general, with the growing length of a ticket, the following listed below trends regarding the mentioned criteria have been identified (see Fig 1 and 2).

The relative mean values of *nouns* verified by corresponding relative distribution are steadily decreasing from 55.2% to 45.2%. According to the RTCC Framework, nouns are ***R****esources* indicating the specificity of business process task items. Short texts in the first cluster count high relative mean values of ***R****esources* (nouns) with also high uniqueness 96.5%, which decreases to 76.0 % in the third cluster. Another characteristic of this cluster is that it counts 7% of tickets with no RTCC PoS element identified. In this case, these elements are very specific names of servers, databases, applications. These trends can be interpreted as exact *telegraphic* way of expression and intensity of naming process resources, i.e. exact names of servers, databases, configuration items related to the ticket, in the short texts if compared to the long ones. The readability is expected to change from *transparent* and *comprehensive* implying *low* perceived complexity with straightforward interpretation of information and clarity in action-outcome relation towards *less transparent* and *not directly comprehensive* with *medium or high* perceived complexity.

The relative mean values of verbs along with their distribution ranges are steadily increasing from 5.0% to 10.3%. In the RTCC Framework, verbs (and verbal nouns) are ***T****echniques* of knowledge and information transformation activity by which the decision-making process affects existing ***R****esources*. The identified trend can be interpreted as a growing number of options in ***T****echniques* describing the decision-making activities affecting the ***R****esources*. The uniqueness of ***T****echniques* is also growing from a mean of 49.0% to 84.6%, however with the highest value in the second cluster. This can be explained by an observation that very long ticket texts also appear in the third cluster due to repeating tasks for different IT objects, i.e. configuration items or services, like "perform update server X, perform update server Y, perform update server Z". In this case, the relative average uniqueness drops. Overall, the complexity of the ticket ***T****echniques* is increasing (except for the tickets with repeating tasks). One can identify a transition tendency from an exact *telegraphic* style (short texts) containing *high* relative distribution of unique ***R****esources* and *low* relative distribution of ***T****echniques* with *low* uniqueness values (a lot of repeating verbs) to more *verbose* and *ambiguous* styles. The readability is predicted to become *less transparent* and *not directly comprehensive* with resulting *medium or high* perceived complexity due to the growing number of options for tasks execution.

Together with the ***T****echniques*, the indicators for ***C****apacities* (*adjectives*) are also growing. The relative mean values of adjectives increase from 4.0% to 5.4% confirmed by increasing distribution ranges. Their uniqueness rises from 41.0% to 86.0%. The longer the text is, the more unique ***C****apacities* describing the situational specificity of decision-making ***T****echniques* can be identified. This is another indicator of a transition from rather simplistic *telegraphic* style towards more complex *verbose* one. This indicator verifies a *decreasing* readability *transparency* and *comprehension* and *medium* or *high* perceived complexity with ambiguous information interpretations and unclear action-outcome relations.

The same trend is detected in case of ***C****hoices* (*adverbs*). Here, a particularly dramatic increase from mean 0.5% with uniqueness 6.6% to mean 1.8% with uniqueness 76% can be observed. This increasing number of unique ***C****hoices* in the ticket text determining the selection of a required set of decision-making ***T****echniques* indicates *verbosity* of the text style, *decreasing* readability *transparency* and *comprehension* what complicates further ticket processing (*medium* to *high* perceived complexity).

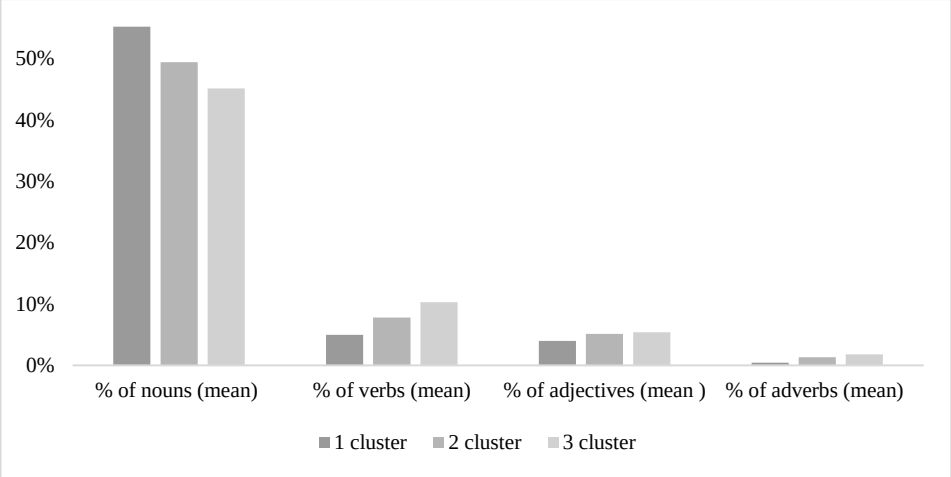

Fig 1. Relative PoS Mean Values in the Three Ticket Clusters

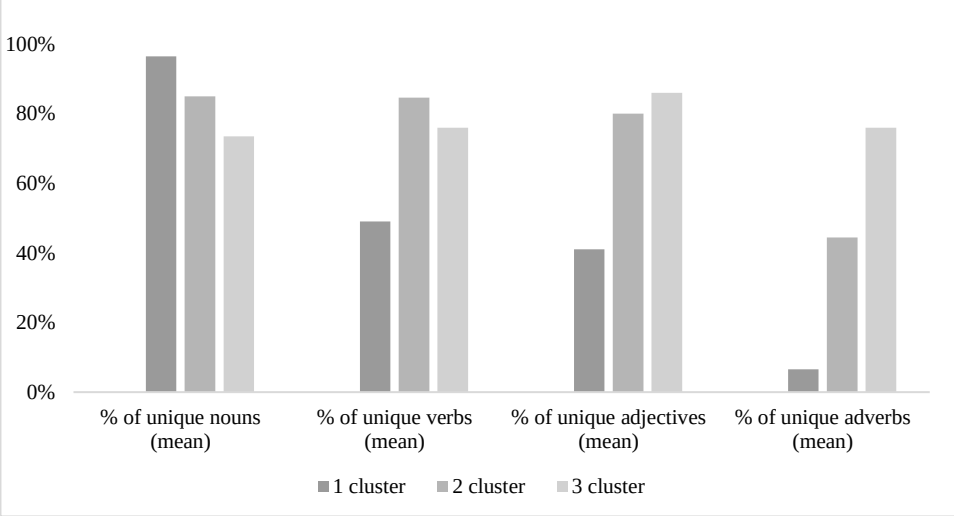

Fig 2. Relative Unique PoS Mean Values in the Three Ticket Clusters

Wording Style

In the WS measure, the researchers examine the internal structure of the ticket text with regard to the specific distribution of word usage. It is based on the combination of Zipf's First "Rank-Frequency" and Second "Quantity-Frequency" Laws (Zipf, 1932; Rizun & Taranenko, 2018), which bring in relation the frequency of the words with their importance rank and the frequency of the words with the quantity of the words with the same frequency. The application of these laws provides a characteristic of the ticket WS.

The researchers propose to use the approximation in the equation of Zipf's Laws $y = a + \frac{b}{x}$ based on the ordinary least squares method. For each ticket, the researchers calculated and built the approximated hyperbolic function *y* to illustrate the rank-frequency and quantity-frequency distribution in the ticket. The basic coefficients considered are the approximation of the average frequency of identified key words (coefficient *a*) and the slope of the hyperbolic function (coefficient *b*) as the approximated values of average speed of appearance of new words in the text. If there are very few words with high frequency in the text and many words with low, the speed of new words appearance is high (the slope of the hyperbolic function is steep with low values of coefficient *b*). And vice versa, if there are many words with high frequency and only few with low, the speed of new words appearance is low (the slope of the function is softly decreasing with high values of coefficient *b*).

Based on the obtained values for the coefficients *a* and *b*, the following ticket text characteristics per each of the three clusters can be inferred (see Fig 3). The first cluster is characterized by relatively high mean values of the approximated coefficient $a = 0.9$, what indicates a high average level of frequency of most part of the words. At the same time, the mean values of approximated coefficient *b* are low (0.5). The speed of new words appearance is very high. This indicates the *telegraphic* style of the first cluster with *condensed flow of information presentation* testifying to *transparent* and *comprehensive* readability and *low* perceived complexity.

The second cluster is characterized by slightly lower mean values of the approximated coefficient $a = 0.87$. Mean values of coefficient *b* increase dramatically to 2.6. This means, with the growing length of the ticket, the speed of new words appearance slows down. The wording style becomes more *verbose*, the same words (RTCC PoS elements) are used more often, while the rest of the words is used randomly, depending on the ticket context. The *flow of information presentation* becomes more *dispersed* and *redundant* decreasing the readability *transparency* and *comprehension* and *increasing* the perceived complexity.

The third cluster is featured by slightly higher mean values of the approximated coefficient $a = 0.97$. Mean values of coefficient *b* again strongly increase to 6.2., what points to a very low speed of new words appearance. In the third cluster, as already mentioned, one type of tickets represents *telegraphic* style with *repeating tasks* (*transparent* and *comprehensive* readability and *low* perceived complexity), another – *verbose* style with *highly dispersed information presentation flow* (*not transparent* and *not directly comprehensive* readability and *medium* to *high* perceived complexity).

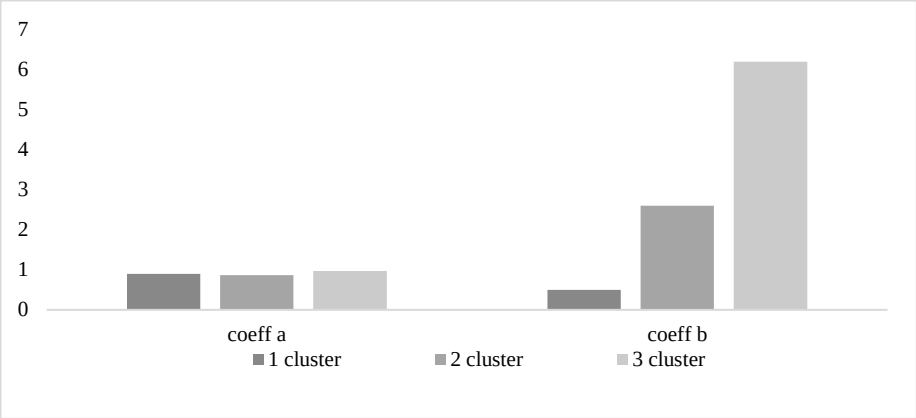

Fig 3. Mean Values of the Zipf's Law Coefficients

Discussion and Conclusion

Generalizing the results obtained in the analysis of the suggested stylistic patterns in the context of IT ticket texts readability and perceived dynamic complexity, the table with the developed characteristics for each pattern and ticket text length-based cluster and practical implications is proposed below.

Table 1: Summary of Stylistic Pattern Characteristics and Their Practical Implications

Measure	**1 Cluster**	**2 Cluster**	**3 Cluster**
SynS	Exact and intensive in process ***R**esources*	Wide in process ***T**echniques*, ambiguous in ***C**apacities* and ***C**hoices*	Mixture of clusters 1 and 2
WS	Condensed information presentation flow	Dispersed information presentation flow	Mixture of clusters 1 and 2
Practical Implications			
Conceptualized Styles	Telegraphic	Verbose	Ambiguous
Readability	Comprehensive	Not directly comprehensive	
Perceived Complexity	Low	Medium / High	

As shown in Table 1, using the identified characteristics of SynS and WS measures, the researchers conceptualized three basic styles or stylistic patterns in the context of incoming IT ticket texts. These are *telegraphic*, *verbose* and *ambiguous* (comprising mutual indicators of the first two) styles. Additionally, the researchers presented further practical implications of the findings: readability can be expected to be *transparent* and *comprehensive* in case of a *telegraphic* style and *not directly comprehensive* in case of *verbose* style with the third *ambiguous* cluster comprising the characteristics of both. The predicted perceived complexity is expected to be *low*, *medium / high* and *ambiguous* correspondingly. With the help of the SynS stylistic measure (distribution of PoS and unique PoS), the researchers extracted the information on the presence of RTCC elements in the incoming ticket texts. Furthermore, WS stylistic measure indicated their density (condensed vs. dispersed) in the text. Based on the introduced text length parameter, a positive correlation and described patterns of the SynS, WS and RTCC elements could be identified.

In order to illustrate future practical implications of the research in the context of the case study ticket processing, the researchers suggest a schematic design of a Recommender System for process workers. This system will adapt the recommendation based on identified stylistic patterns, predicted readability and perceived complexity (see Fig 4).

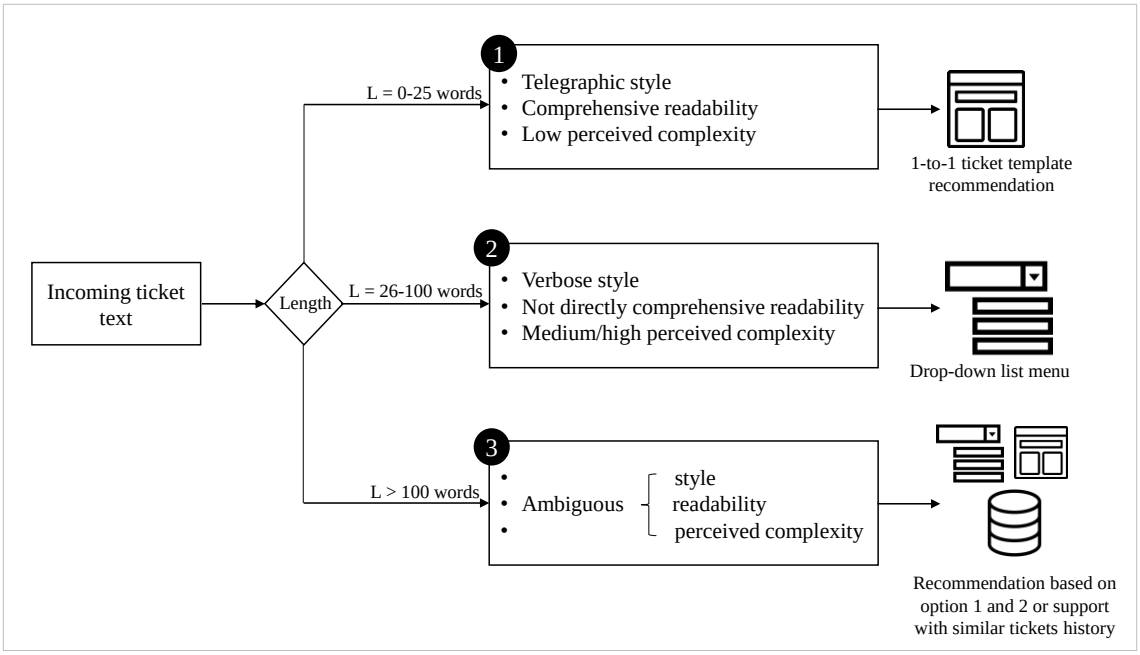

Fig 4. Schematic design of a Recommender System

The presented schematic design is based on the following process steps: (i) determine the number of words in a ticket; (ii) apply the presented rule set; (iii) provide the appropriate supporting tool for the process worker: (1) 1-to-1 ticket template recommendation in case of simple requests; (2) drop-down list menu recommendation in case of more complex tickets; (3) in the ambiguous (mixed) ticket category, all types of recommendation are possible depending on the presence or absence and its degree of the mentioned decision criteria, up to the minimal recommendation support with the history of similar tickets in case of the most complex tickets. Thus, in third cluster of tickets with L > 100 words, there is an additional step to determine which of the three options fits best.

To conclude, at the beginning of the experiments, the researchers introduced the length of the ticket text as an important parameter assuming that the process workers tend to write shorter texts in case of simple, explicit tickets (*telegraphic* style) and longer ones in case of more complex tickets (*verbose* style). Based on the identified characteristics of the stylistic patterns of SynS and WS, the transition from *telegraphic* (describing simple, explicit tickets) to *verbose* (describing complex tickets) style with the growing ticket length was successfully verified. These facts allow to state that H1 and H2 are accepted. The suggested measures of stylistic patterns serve to verify the narrow scoped H1 and H2 of the present paper and contribute to answering the wider scoped RQ1 and RQ2, i.e. prediction of the readability and perceived dynamic complexity.

However, the researchers experienced the following limitations (L) that will be addressed in the future work correspondingly:

L1: limitation related to other stylistic pattern measures, which might be valuable for the present research, such as sentiment or standard business vocabulary usage highly rated by the case study workers within the brief qualitative survey. Following the same research procedure, the researchers plan to assess a possible value of these measures for readability and perceived complexity.

L2: limitation related to other parameters helping to determine the overall ticket complexity (ticket objective processing complexity and ticket implementation complexity). After finalizing the research on

stylistic pattern measures, the researchers plan to consider other parameters, such as tasks breakdown, execution time, risk level, severity, involved configuration items and assigned groups.

L3: limitations related to the missing feedback loop towards the case study process workers. The researchers will develop detailed qualitative interview guidelines to assess the quality of the proposed method in a more comprehensive way while performing collective intelligence workshop with the case study process workers.

References

Axelos (2019). ITIL (IT Infrastructure Library) Certifications. [Online], [Retrieved March 17, 2019], https://www.axelos.com/certifications/itil-certifications

Chiu, S.-I. and Hsu, K.-W. (2018), 'Predicting Political Tendency of Posts on Facebook,' Proceedings of the 2018 7th International Conference on Software and Computer Applications (ICSCA), ISBN: 978-1-4503-5414-1, 08-10 February 2018, Kuantan, Malaysia, 110-114.

Huadhom, N. and Trakulkasemsuk, W. (2017), 'Syntactic Analysis of Online Tourism Slogans: Frequency, Forms and Functions,' *Journal of Language Teaching and Learning in Thailand*, 53, 182-213.

Kaljahi, R., Foster, J., Roturier, J., Ribeyre, C., Lynn, T., and LeRoux, J. (2015), 'Foreebank: Syntactic Analysis of Customer Support Forums,' Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), ISBN: 978-1-941643-32-7, 17-21 September 2015, Lisbon, Portugal, 1341-1347.

KiBaek, S., Bernhardsson, S. and Minnhagen, P. (2011), 'Zipf's Law Unzipped,' *New Journal of Physics*, 13(4), 1-22.

Koopman, H., Sportiche, D. and Stabler, E. P. (2013) An Introduction to Syntactic Analysis and Theory, Wiley-Blackwell, Los Angeles.

Lasswell, H. D. (1948) The Structure and Function of Communication in Society, in L. Bryson, The Communication of Ideas, 37-51, Harper & Row, New York.

Nadkarni, S. and Gupta, R. (2007), 'A Task-Based Model of Peceived Website Complexity,' *MIS Quarterly*, 31(3), 501-524.

Niederdeppe, J. D. (2005) 'Syntactic Indeterminacy, Perceived Message Sensation Value-Enhancing Features, and Message Processing in the Context of Anti-Tobacco Advertisements,' *Communication Monographs*, *72*(3), 324-344.

Ordenes, F. V., Burton, J., Theodoulidis, B., Gruber, T., and Zaki, M. (2014), 'Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach,' *Journal of Service Research*, 17 (3), 278-295.

Rangu, C., Chatterjee, S. and Valluru, S. R. (2017), 'Text Mining Approach for Product Quality Enhancement: (Improving Product Quality through Machine Learning),' Proceedings of the 7th International Advance Computing Conference (IACC), ISBN: 978-1-5090-1560-3, 05-07 January 2017, Hyderabad, India, 456-460.

Rizun, N. and Kucharska, W. (2018), 'Text Mining Algorithms for Extracting Brand Knowledge. The Fashion Industry Case,' Proceedings of the 31st International Business Information Management Association (IBIMA), ISBN: 978-0-9998551-0-2, 25-26 April, Milan, Italy, 1972-1983.

Rizun, N. and Taranenko, Y. (2018), 'Methodology of Constructing and Analyzing the Hierarchical Contextually-Oriented Corpora,' Proceedings of the Federated Conference on Computer Science and Information Systems, ISBN 978-83-949419-5-6, 09-12 September, Poznan, Poland, 505–514.

Rizun, N., Revina, A. and Meister, V. (2019), 'Method of Decision-Making Logic Discovery in the Business Process Textual Data,' Business Information Systems (BIS) 2019 Proceedings, 26-28 June, Sevilla, Spain. Accepted for publication.

Roberts, C. W. (2000) 'A Conceptual Framework for Quantitative Text Analysis,' *Quality & Quantity*, 34 (3), 259-274.

Shlomo, A., Karlgren, J. and Shanahan, J. G. (2005), 'Stylistic analysis of text for information access,' Papers from the workshop held in conjunction with the 28th Annual International ACM Conference on Research and Development in Information Retrieval, ISSN: 1100-3154, 15-19 August, Salvador, Brazil, 1-69.

Woloszyn, V., dos Santos, H. D. and Wives, L. K. (2017), 'The Influence of Readability Aspects on the User's Perception of Helpfulness of Online Reviews,' *Revista de Sistemas de Informacao da FSMA*, 18, 2-7.

Wood, R. (1986) 'Task Complexity: Definition of the Construct,' *Organizational Behavior and Human Decision Processes*, 37(1), 60-82.

Wróblewska, A., Stanisławek, T., Prus-Zajaczkowski, B., and Garncarek, Ł. (2018), 'Robotic Process Automation of Unstructured Data with Machine Learning,' Position Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), ISBN 978-83-949419-8-7, 09-12 September, Poznan, Poland, 9-16.

Zipf, G. K. (1932) Selected Studies of the Principle of Relative Frequency in Language, Harvard University Press, Cambridge.

Zipf, G. K. (1949) Human Behaviour and The Principle of Least Effort, Addison-Wesley, Reading.