DOI 10.2478/jazcas-2025-0058

DEEP CONTENT AND DEEP SENTIMENT ANALYSIS[1]

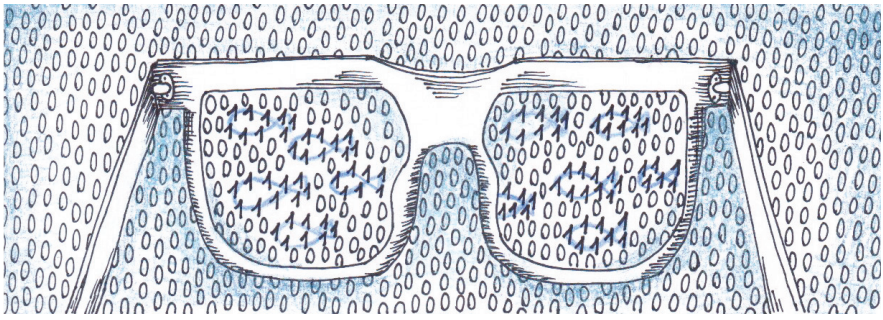

Figure 1: We realize that an individual's linguistic characteristics are hidden in the low-frequency layer. We see the individuality of an author by detecting uniqueness of their 1s in the sea of their 0s. 1s: the words occurring in the text produced by their author only once. 0s: words missing in the text compared to other texts produced by the same author.)

DAN FALTÝNEK – MARTINA BENEŠOVÁ – ONDŘEJ KUČERA[2]
Faculty of Arts, Palacký University Olomouc, Olomouc, Czech Republic

FALTÝNEK, Dan – BENEŠOVÁ, Martina – KUČERA, Ondřej: Deep context and deep sentiment analysis. Jazykovedný časopis (Journal of Linguistics), 2025, Vol. 76, No. 3, pp. 752–771.

Abstract: The objective of the article is twofold: first, to employ the knowledge of the recurrence of low-frequency words in authorial texts; and second, to prevent the misuse of this knowledge. Contrary to the prevailing authorship attribution theory and practice (Evert et al. 2017, Juola 2008), our research has revealed that the personal linguistic profile is not primarily composed of frequent words with grammatical functions. Instead, we have identified that a distinct set of full-meaning words defines an individual's linguistic profile (Faltýnek 2020, Faltýnek – Matlach 2021). An examination of these meanings reveals an individual's unconscious language habits and, consequently, their personality settings. Such personal profiling is referred to as "deep content" and "deep sentiment analysis". The innovation in question has the potential to facilitate a novel form of linguistic personalization in digital communication, one that has not been previously observed or utilized. The main aim of this article is to describe the algorithm to conduct single-person linguistic deep content and deep sentiment profiling and personalization. We will describe technical steps to provide such a form of digital communication processing and to facilitate the adjustment

[1] The work was supported by project Podpora zelených dovedností a udržitelnosti na UP (NPO_UPOL_MSMT-2118/2024-4).

[2] Author of illustrations: Klára Faltýnková.

of a text targeted at an individual, described as a *System and method for adapting text-based data structures to text samples* (Patent No.: US11797753B2, Faltýnek et al. 2023). This algorithm can be used to (a) produce a personal linguistic profile (analogically to psychometrics instruments such as NEO-FFI Big Five, Minnesota Multiphasic Personality Inventory (MMPI)), (b) target digital communication to an individual by "translating" a text to their language (i.e. linguistic habits) and stimulate desired feelings to a predetermined content. The algorithm is, however, also designed (c) to be used to avoid procedures (a) and (b) using any kind of digital communication platform by an individual. This algorithm is implemented in the software Cloakspeech (Faltýnek – Benešová – Kučera 2025), which provides personalization of AI-generated texts: AI speaks like a particular person.

Keywords: personalization, single-person personalization, marketing, influencing, people manipulation, mind engineering, PSYOPS, author identification, hapax legomena, low-frequency words, linguistic profile, deep content, deep sentiment, Bruntal effect, Cloakspeech

INTRODUCTION

Modern linguistics has traditionally rejected the existence of a text property which argues that there is a regularity of low-frequency words localized independently to the topic in any of an author's large texts. De Saussure (1966, p. 14) formulates that even a sentence shape is largely determined by chance and is not fully constrained by any specific kind of language structure. Bloomfield (1933, p. 70) argues that there is no other language structure above the sentence level. Halliday and Hassan (1976) show dependencies among close sentences, which do not extend into a wider text. De Beaugrande and Dressler (1981) define principles of text acceptability without formal principles of its structure. Van Dijk (2008, 2015) enunciates that the text is driven by its context, which is dependent on a communication situation, i.e., random. Overlooking of the low-frequency text quality was also attributable to the methodologies employed in data mining (Support Vector Machine, Principal Component Analysis) which do not take into account low-frequency structures among higher frequencies in an individual's text (Diederich et al. 2003; Koppel et al. 2009, p. 12, Binongo 2003).

Every text exceeding several thousand words tend to comprise a modest percentage of low-frequency words, especially hapax legomena representing single instances of words in a text (Fengxiang 2010, Jones 1972). Linguistics and data science assume that a list of words with a single occurrence in a text cannot reveal any structure (compare with Faltýnek et al. 2020). Just a few researchers noticed that low frequency words possess structural properties (see Hřebíček 1989, Amelin et al. 2018; Baayen et al. 1996; Lardilleux – Lepage 2007; Mikros 2009, Rybicki – Eder 2011 and Savoy 2012). A comparison of thousand-word-long text samples from an individual reveals the emergence of notable structural properties, with a focus on solely low-frequency components within these samples (Faltýnek et al. 2023). Such properties can tell us that the text on its large scope adheres to a set of principles, and the words are constrained to their designated textual locations, even over specific distances, which is hard to notice without the described insight.

The development of a comprehensive understanding of this phenomenon has the potential to expand the scope and depth of linguistic analysis by incorporating previously disregarded text-based information. The regularity of specific low-frequency words in the text written by a particular person can reveal the content characteristics of their language behavior. The list of personally preferred low-frequency words taken from the authorial clustering can, in addition, be misused to manipulate people. This category of words is distinct from an individual's ordinary vocabulary and likely occupies a specific position in their mental lexicon (Forster 1976). Alongside, these words are used (expressed or perceived) without conscious control. Their personal sensitive words can be (mis)used for direct influencing of the author because such presentations of the information are easier to be processed by them. Having a personal list of these words, a certain content can be communicated to an individual stressing the individual's sensitive topics and feelings relating to these topics, and also speaking their language. This kind of individual influencing is technically easily achievable by existing technologies: information for a particular person is "translated" according to their personal low-frequency list, additionally using wordnet entries for the choice of synonyms and hypernyms (performing finer lexical identification of the individual's language) and treebanks or neural network models ensuring fitting sentence construction.

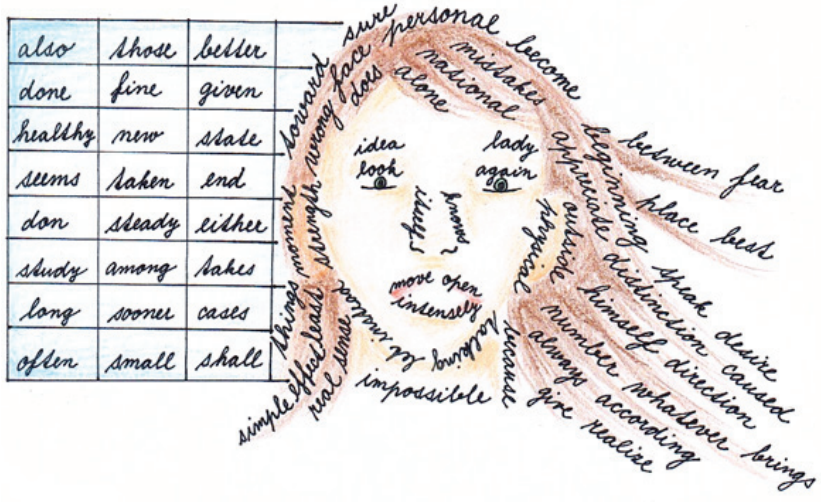

Figure 2: Regularly repeated low-frequency words manifest a fingerprint of an individual in the text produced by them. Such words are used unconsciously. Gaining such words may facilitate communication with such a person to a significant degree.

LANGUAGE STRUCTURE(S) OVERVIEW

Taking well-known linguistic levels (Sadilloevna et al. 2020) into account, we can notice specific organization principles of each. Qualities used for creation of speech sounds – acoustic characteristics like duration of a sound or tone frequency – are described as a system of *distinctive features* (Jakobson et al. 1952), which differ in the case of oral, written, etc. communication. Even though these substantial qualities are not consistent with the language categories based on expression and content relation, they are represented by a system of clearly definable distinctions. *Speech sounds* or *phonemes* are defined by their function in distinguishing words or word forms (Catford 2002; Trubetzkoy 1969). The alphabetical ordering, for example, already suggests the systematization of phonemes represented by letters. Word forming types show us principal components of words, *morphemes*, like prefixes, suffixes, or stems (Grzybek 2007; Haspelmath – Sims 2010; Benešová et al. 2015). *Words* are defined by their relations like synonymy, antonymy, or hyper/hyponymy. The *sentence* is defined by obligatory positions which should be occupied by specific word forms to construct a correct utterance of a specific language. Interconnection between neighboring sentences is ensured by word repetition, synonymy, or ellipsis, grammatical agreement, etc. Such *textual cohesion devices* are based on the principle of co-reference to a particular matter of speech. For this reason, the impact of interconnection principles does not extend more than a few sentences, mostly the connection between two neighboring sentences is covered by these cohesive devices.

Linguistics is brought into play to describe the structure of language and text in the scope from distinctive features to cohesive devices. An extensive text, from this point of view, is created by a certain author's intention and language knowledge (see the concept of not-knowledge-based-grammatical disposition in Dolník 2018, 2019, 2021), and in relation to the topic of the communication, i.e., a text reflects the actual context of the communication and thus we cannot define any formal principles for the construction of such an extensive text (van Dijk 2008; Fairclough 1995).

Following the recent research (Faltýnek – Matlach 2021; Faltýnek et al. 2023), we can add one specific principle engaged in the construction of an extensive text and enlarging the above-mentioned linguistic scope of language sub-systems. It is a habit of an individual to express their intentions by specific word forms and phrases (Faltýnek 2020). This author's specific regular low-frequency lexicon is also a decisive linguistic device regulating cohesion of an extensive text. In other words, an individual's low-frequency sensitive words determine grammatical and lexical properties of sentences encompassed in larger texts.

It seems that the text structure does not directly reflect the speaker's intended context and purpose in expressing specific matters. Considering intentions and the communication context at the same time, the text structure is determined by personal linguistic mannerisms. I.e., people use the same ways of expressing themselves.

These mannerisms are observed even in the low-frequency spectrum, which indicates that people regularly stress the same topics without any immediate connection to a particular situation (context) and without any overt intention. This is newly evidenced by a particular method of authorial clustering that utilizes solely low-frequency words: the individuality of a person emerges from a random selection of 2,500 low-frequency words from an extensive individual's text sample. We also realize that individuality identification based on a 6,000-word-long text sample is reliable (more or less dependent on a language type, see Faltýnek – Matlach 2021).

DEEP CONTENT AND DEEP SENTIMENT ANALYSIS AND PERSONALIZATION

In the forthcoming sections, we will provide description of a procedure for the mind-engineering procedure, which is based on the deep content and deep sentiment profile of an individual. Utilization of such an individual's profile facilitates "translation" of a text targeted towards that individual into their commonly and unconsciously repeated language behavior and stimulates a desired sentiment of this individual to the targeted text content according to the targeted-message client's intention, based on specific sentiment preferences of the individual derived from the individual's deep content and deep sentiment profile. Such a profile is also used as psychometrics of an individual, which contains detailed linguistics information related to tested characteristics.

The next methodological description enlarges and specifies the single-person personalization algorithm defined in the US patent *System and method for adapting text-based data structures to text samples* (Faltýnek et al. 2023, Patent No.: US11797753B2). The description of a personalization procedure listed below contains particular text processing techniques; sample sizes and text-part evaluations are not mentioned in the patent. This description also covers our experience with particular individual deep content and deep sentiment profiling (described also in Faltýnek et al. 2023).

Figure 3: The described technology enables preventing harm on an individual considering developed personalization techniques.

PRACTICAL STEPS OF DEEP CONTENT AND DEEP SENTIMENT ANALYSIS AND PERSONALIZATION

1 Purpose, individual and output identification

1.0 In the first step, identification of a text-analysis and processing **purpose** is required. Also, identification of an **affected individual** is required. The technology is not applicable to text samples encompassing any other individual's text, i.e., it is based on a single individual's linguistic characteristics which cannot be derived from a common language. Some examples of processing **outputs**: (a) a **personalized text** – advertisements, direct marketing, web-page content (news, web pages of an interest), digital communication using chatbots; (b) a **deep content and deep sentiment profile**. */The identification of a text's purpose is needed, as different purposes require different approach to processing of text samples and extracted text parts. In general, we distinguish two kinds of purposes: (a) text personalization by an individual's language habits and amendment of a text by the selected sentiment in the selected topics of an individual; (b) personality profiling. To achieve purpose (a), we use all of an individual's specific features; personality profiling in purpose (b) is, on the other hand, based on the text features with high significance to a particular individual's language behavior. For such reasons, we take different ways when proceeding through the personalization algorithm./*

2 Preprocessing steps

2.0 Further automated processing requires preparing a text sample. The aim of the preprocessing steps is to gain **a cleaned text sample (from different speakers' text production and any formatting)**. The text production by the author remains unaltered and the format is preprocessed.

2.1 **We specify the type of data (speech, digital communication, personal writing).** /*The way text is produced has a direct influence on the distribution of an author's key salient characteristics, as determined by their profiling. Depending on natural dialogical speech, personal writing, chat communication etc., the vocabulary richness of the text, i.e., the recurrence of specific word forms in conjunction with their respective contexts, varies. For this reason, the word-length threshold for a text sample can be changed to identify an author's key words adequately (see point 3.0).*/

3 Text sample processing and superhapax identification

The objective of text-sample processing is to extract unique words that occur precisely once in a minimum of two of the segments that an individual's corpus (text sample) is split into. Such words are referred to as **superhapaxes**. /*The number of extracted words depends on the purpose (1.0) and is limited by the requirement of supertopic identification and weights (6.6).*/

3.0 The text sample is at least 12,500 words long. /*The original 13,125-word-long source data was processed and corrected for typographical errors, additional type characters etc.*/ If the text sample is less than 13,125 (not yet cleaned source data) / 12,500 (a cleaned text) words long, one needs to find more source data – a spoken incomplete text sample is – for example – completed with another spoken sample. If the length of a cleaned text sample is less than 12,500 words and no other data is available, the procedure goes to the **high-frequency superhapax take** (follow to point 5.5). /*However, we recommend employing a text sample longer than 12,500 words*.

Threat: In the event that the text sample is less than 12,500 words in length, supertopic affirmation is achieved by dividing it into 3 sections, with each section containing less than 4,166 words. High-frequency hapaxes with stable topics are considered superhapaxes (they are listed in the superhapax list). Such superhapaxes could not, then, be identified in any additional text by the same author, i.e. the deep content profile and deep sentiment of the person are not defined by them./

3.1 **The text sample is divided into multiple segments** of the same length. / *12,500-word-long text sample is divided into two segments, each of 6,250 word forms*./

3.2 If the text sample is shorter than 18,750 word forms (three segments of a minimum length of 6,250 word forms), it is divided into two text segments with the same or maximally similar lengths. Limits for lengths of text samples gained in further **segment additions** are multiples of the minimum segment length of 6,250 words with maximum text sample length of 62,500 word forms.

3.3 Text segments are represented by the respective lists of their word forms. **Lemmatization is not needed for the purposes** of this analysis.

3.4 **The list of word forms** in each segment is supplied **with their frequencies**. /*3.3 and 3.4 describe the extraction of the bag-of-word representation of segments*/

3.5 **Text segments are represented solely by their hapax legomena**, i.e., the word forms with the exact frequency of 1 (word forms used in a segment just once).

3.6 **All segments** of a text sample, **represented by their hapax legomena** lists, **are compared**.

3.6.1 In instances where **two segments** must be compared, the list of hapax legomena occurring in either is mined. This relates to the 12,500-word-long text sample (or longer) divided into two 6,250-word-long segments (or longer adequately; 3.2) which produce a list of hapax legomena occurring in each of the two segments.

3.6.2 In the case of **three or more** 6,250-word-long **segments** (3.2) originating from a text sample to be compared, the list of hapax legomena occurring in each is mined. Therefore, the list of hapax legomena occurring in three or more segments is achieved.

3.7 Hapax legomena (3.5) occurring in at least two segments are gathered.

4 Superhapax takes

The superhapax is defined as a word form that reoccurs with low-frequency and homogeneous dispersion in a text by a single individual. Various definition of *low freqency* and *homogenous dispesion* depending on the analysis purpose result in different lists of superhapaxes. These lists are used in different product lines and are combined to ensure complex text mining which depicts deep content and deep sentiment of an individual.

4.1 **Identification of a complete set of superhapaxes**. The procedure follows the steps as described in article 3. The superhapax list contains every word form occurring in at least two text segments as a hapax legomenon (3.5).

4.2 **Sorting of superhapaxes in the superhapax list**. The superhapaxes in the superhapax list are supplied with the number of text segments they originate from and are subsequently arranged in accordance with this information. The list starts with superhapaxes present in all text segments (if applied, i.e. if not present in all text segments, we continue as follows). Followingly, we pick hapaxes present in a lower number of text segments (if applied). The list is closed with hapaxes appearing just in two text segments.

4.2.1 For example: In a 62,500-word-long text sample split into ten text segments, we pick hapaxes present in all ten text segments (if applied). Subsequently, we continue checking for the hapaxes present in nine, eight, etc., down to two text segments. */With a decrease in the number of text samples under consideration, the number of hapaxes increases. Also, the average word length decreases (the lower the absolute frequency of a word in a text sample, the longer the word – the brevity law). With a decreasing number of text segments containing an increasing number of hapaxes, we cover larger and larger amount of the text – in the case of a 30,000-word-long text sample, it is around 10% of the total text./*

played: 11	only: 10	creating: 9	monthly: 9	in-game: 9	talented: 9	element: 9	particularly: 9
investing: 9	successful: 9	based: 9	100: 8	keep: 8	global: 8	series: 8	between: 8
central: 8	something: 8	closely: 8	earlier: 8	consumption: 8	though: 8	took: 8	foundation: 8
wide: 8	begin: 8	a: 8	single: 8	quickly: 8	happening: 8	3: 8	regardless: 8
double: 8	for: 8	call: 8	mean: 8	planning: 8	hard: 8	linear: 8	drives: 8

Table 1: Exemplary embodiment of an individual's superhapaxes sorted in the superhapax list. The text of an individual was segmented into 16 text samples. The text samples possess a similar length with an approximate word count of 6250 words. We can see that superhapax *played* was present in 11 text samples (i.e. the word *played* had a frequency of 1 (not 0 or higher than 1) in 11 text samples). We can find the word *only* in the role of a superhapax in 10 text samples, superhapax *certain* in 9 text samples etc.

4.3. **Purposes** of individual analyses, as defined later in article 7, determine the appropriate number of the most frequent superhapaxes from its list (4.1, 4.2) to be chosen and further employed.

4.4 **Assessment of superhapaxes.** The usability of identified superhapaxes for specific purposes serves to sort the superhapax list. The superhapax is taken to create a personalized text or to identify personal sentiment that is contingent upon a stable vicinity, i.e., the presence of a stable content and sentiment is concomitant with the superhapax occurrence in a text. A stable vicinity is then a characteristic decisive for the **final sorting of the superhapax list**. Word forms with a higher rank possess superior features for an intended goal in personalization. In the production phase, word forms are processed as ordered in the superhapax list.

4.4.1 **Degree of superhapax topic bond**. The order of superhapaxes in the final superhapax list is given by the degree of their topic bond. The **topic bond** of a superhapax is the percentage of the topic occurrence in this superhapax's vicinity, where the topic is defined as a set of content word forms contained in the vicinities of at least two occurrences of the superhapax (5.0).

4.4.1.1 For example: Let us assume we have a 31,250-word-long text sample. We split it into 5 text segments. If a chosen topic occurs in 3 of total 5 vicinities of a superhapax, the superhapax possesses the degree of topic bond 60%.

The degree of topic bond is applied separately in two parts of the superhapax list. These parts are defined as Superhapax list part (a): hapaxes present in three segments and more, and Superhapax list part (b): hapaxes present in two segments. The degree of topic bond of Superhapax list part (a) is defined in 4.4.1. The degree of topic bond of superhapaxes from Superhapax list part (b) is defined as percentage of the topic occurrence in this superhapaxe's vicinity divided by 2; i.e., superhapaxes from part (b) assume values of 50% and 0%, respectively.

4.4.1.2 For example: Let us assume we have a 43,750-word-long text sample. We split it into 7 text segments. The hapax list is divided into two parts – the first, upper part contains superhapaxes present in 7, 6, 5, 4 and 3 text segments and the other, lower part is composed of superhapaxes solely from 2 text segments. /*The number of superhapaxes recurring in multiple text segments differentiates dramatically; their amount increases exponentially with the decreasing number of text segments where they are present. The set of superhapaxes originating from solely 2 text segments always contains a far higher number of superhapaxes compared to any higher. At the same time, superhapaxes originating from a higher number of text segments than 2 ensure identification of vicinity stability, i.e. of the topic bond and its scalability.*/

4.4.2 **Sentence length**. Due to the evidence that superhapaxes affect their syntactic vicinity, the sentence length is used for their evaluation too. (a) Superhapaxes are contained in sentences with more than an average length, which was tested on several datasets, however, it has not yet been published by the research team. The length of sentences containing superhapaxes that exceeds the mean sentence length (a) is accompanied by additional features that distinguish these sentences from the rest of the text. I.e., some sentences in the superhapax vicinities

exhibit (b)/ lower type-token ratio (called vocabulary richness, which is the proportion of unique words in the text compared to repeated words. i.e. these sentences exhibit a higher frequency of repeated lexicon; (c)/ higher word length in average. Sorting hapaxes using criteria A, B, C is at the disposal and is used as an is an auxiliary criterion.

4.4.2.1 The process of text processing and superhapax identification (article 3) can be supplanted or completed through the utilization of sentence length as an identifier of an author's sensitive lexicon (4.4.2). The low-frequency lexicon is present in sentences of greater length than is average. A disadvantage of this approach is that (a) in comparison to the average sentence length, which can be a decimal number, the length of a particular sentence is a natural number; consequently, to exceed the average it must surpass the immediate higher natural number (sentences shorter than this higher natural number contain the author's sensitive lexicon); (b) the sentence structure is consistent with both the communication needs and the rules of language system, and is not determined only by the authorial repeated low-frequency lexicon and, related to this, by the author's preferred content and sentiment either. Our described technology leading to single-person personalization (article 3) is due to that based on low-frequencies and homogenous dispersion in conjunction with the Bruntal effect (vicinity stability). The sentence length (plus TTR and word length) is, thus, an auxiliary criterion of superhapaxes sorting in analyses (Production phase, article 7).

4.5 Additional superhapax approaches

4.5.1 **High-frequency superhapax take** is performed if (a) a text sample is shorter than 12,500 words (3.0), (b) the Bruntal effect imposed by higher-frequency words is present – i.e., words with the frequency higher than 1 (hapax legomena) possess a stable vicinity. Under the Bruntal effect influence, the affected words are added to the superhapax list. This kind of superhapax take is obligatory in the case of a text sample shorter than 12,500 words and facultative for any longer, under the condition of stable vicinity identification (i.e., the vicinity is similar in both cases). Higher-frequency superhapaxes are added to Superhapax list part (b) (4.4.1), in which the degree of topic bond is divided by 2.

4.5.1.1 **High-frequency superhapaxes** are defined by their frequency of 2 and 3 in text segments when the text sample is split into two and three text segments at the same time. Such words, at the same time, have to meet the Bruntal effect (4.5.1) to be listed in Superhapax list part (b). /*This text processing enables words repeated within a closer proximity in compliance with syntactic rules to be included into the analysis. Such a close repetition breaks the rules of absolute frequency and of the frequency of a word within a text segment. In contradiction to that, such words can comply with the conditions of low-frequency and homogeneous dispersion but are not taken into account due to the threshold of absolute frequency.*/

4.5.2 **Low-frequency superhapax take.** This take identifies the superhapaxes with the most homogeneous **dispersibility** – therefore, such superhapaxes are regarded as being less consciously controlled by an individual. On this basis, low-frequency superhapaxes give the superhapaxes (word forms) from a text sample split with a higher number of text segments the **potentiality to reach a higher degree of the topic bond**. The low-frequency superhapax take is applied in the cases of more splits of a text sample than 4. To get the low-frequency hapaxes, text segments are re-split as follows: 10 segments to 5 segments, 9 segments to 4 segments, 8 segments to 4 segments, 7 segments to 3 segments, 6 segments to 3 segments, 5 segments to 2 segments, 4 segments to 2 segments. The degree of topic bond is copied into the final superhapax list (4.4), respectively, to the low-frequency superhapaxes take. / *Due to the empirical test – in the case of 10 splits of a text sample changed to 5 splits, 30 percent of hapaxes give a chance to reach a higher degree of the topic bond based on high dispersibility and lower number of text splits answering to the same topic presence (5.0)*/

5 SUPERTOPIC IDENTIFICATION

The aim of topic identification is to determine the superhapaxes with a **stable vicinity (referred to as Bruntal effect)**, i.e., a stable text behavior (leading to predictability in an individual's text). The stable vicinity of superhapaxes identifies the **deep sentiment** of an individual (**unconsciously recurring attitudes**). The stable vicinity of superhapaxes specifies the content used by an individual to address a different context in language use and defines their consciously or unconsciously **preferred contents** referred to as **supertopics, described as an individual's content profile**. A stable topic reflects **different mental processing** of a word form by an individual and marks language phenomena used for text **personalization**, **influencing** an individual and their **mind engineering**.

5.0 **Supertopic identification**. The **supertopic** is defined as a content word form set contained in the vicinity of at least two occurrences of a superhapax.

5.1 **Vicinity definition**. The **superhapax vicinity** is defined as the sentence with the occurrence of a superhapax and the following sentence.

5.2 **Supertopic determination**. The supertopic is defined as word forms and their collocations occurring in a defined **superhapax vicinity**. **The frequency list of word forms**, word forms **bigrams** and **trigrams** from superhapax vicinities is prepared for topic identification in the following manner. (a) A supertopic relates to the superhapax which defines the vicinity (5.1) where the supertopic appears, i.e., it is to be found in the vicinities of all occurrences of a single superhapax (hereinafter the **single supertopic**). */In this case we can apply such a combination of a superhapax word form and the specific supertopics (content and sentiment) – it can result in creating a text fully personalized to the particular individual that activates their*

sentiment in relation to the purpose of the demand/. (b) The whole sum of all vicinities of all superhapaxes is used for identification of the complete set of supertopics (hereinafter the **general supertopics**). /*In this case the word forms from all vicinities of different superhapaxes strengthen the significance of specific single supertopics occurring in the vicinities of different superhapaxes.*/ To determine individual supertopics the frequency and relevance (according to a particular relevance measure) of word forms and word form n-grams occurring in superhapaxes' vicinities are used.

5.2.1 Priming of the sentiment by a particular superhapax is performed. This procedure is referred to as the **Fairy-tale effect**.

5.2.2 The supertopic is defined as word forms and word form bigrams or trigrams with a frequency higher than 1 in the superhapax vicinities (the single supertopic) or all superhapaxes' vicinities (the general supertopic).

5.2.3 The single supertopic (referred to as **single supertopic list (a)**) and general supertopic (referred to as **general supertopic list(b)**) lists supplied with the respective frequencies are created to get the desired single-person personalization (5.2.2).

5.2.4 The method used for supertopic identification on the whole of superhapaxes' vicinities can be Latent semantic analysis (Pritchard – Stephens – Donnelly 2000).

5.2.5 Word forms from superhapaxes' vicinities are also represented by the **wordnet** hypernyms and are grouped with supertopics based on their frequency (referred to as the **Yellow effect**).

5.3 **Supersupertopic.** Supertopic identification is facilitated by the fact that superhapaxes have direct impact on their vicinities. In case superhapax vicinities overlap, such conjunctions can be interpreted as text spots which possess a higher level of unconscious repetitions and topic compulsion (referred to as the **Attractor effect**). For the supersupertopic identification, the topic extraction span extends, i.e., one chooses a wider vicinity. /*This is performed to avoid missing superhapax content and sentiment interconnection.*/ The scope of supersupertopic identification is the sentence with the respective superhapax and 5 sentences before and 5 after. For the supersupertopics extraction, only such segments of the vicinities of different superhapaxes which overlap are considered.

Practically, in case superhapaxes' vicinities overlap and mark the same word forms or word form n-grams, supersupertopics are detected. Supersupertopics can emerge in the case of single as well as general supertopics. With respect to the supersupertopic being primed by more superhapaxes' word forms, it is considered with increased relevance. Supersupertopic word forms are, thus, presented in a supersupertopic list with the respective frequencies (referred to as **supertopic list (c),** as added to 5.2.2), in a supertopic list with some frequencies multiplied (**supertopic list (d)**) and in a supertopic list with original frequencies (**supertopic**

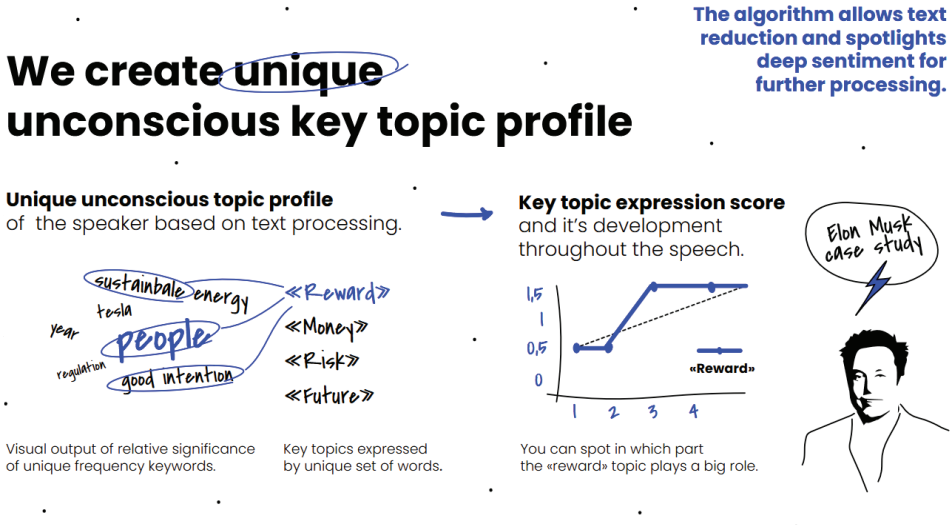

Figure 4: The technology provides the opportunity to mine a sentiment profile of an author from the text produced by them.

6 VISUALIZATION OF INDIVIDUAL'S DEEP CONTENT

To grab the relevant content of the text rapidly (perceiving by an individual), the outputs of the text processing analysis stage are visualized by the method of word cloud designing. The method facilitates demonstration of the sensitive content of the text by an individual and provides an overview of the text narrative. The word form and word form n-gram frequencies are represented by different sizes of the font used, which highlights their various significance. Color differentiation of individual word forms and word form n-grams enable the reader to distinguish them more easily. A deep content visualization advantage is based on highlighting the sensitive (i.e. manneristically repeated) content and on processing just a part of the text (see the superhapax take, 4.0). */It is recommended to grab 10 percent of an individual's text sample defined by superhapaxes with the highest degree of the topic bond/.*

6.1 The platform of the deep content visualization is topic list (b) and (d) (5.2.3), i.e., the general supertopic list (representing the stable context of superhapaxes) and supertopic list with multiplied supersupertopic frequencies. Word forms and word form n-grams from the topic lists are visualized with different font sizes (depending on their frequencies and relevance) and with different fonts and colors (to make them easier for grabbing).

6.1.1 In the event of the detection of sets of word forms and word form n-grams belonging under one content umbrella (e.g., via WordNet; 5.2.5), these are drawn in the same color with different font sizes (depending on their frequencies and relevance). In any other case the colors employed are arbitrary.

6.1.2 This is also applicable in instances supertopics are combined in particular superhapax vicinities or are interconnected in supersuperhapaxes vicinities (identification is based on extending the vicinity take and on the condition of overlapping different superhapax vicinities, 5.3).

7 UTILIZATIONS

7.1 The utilization we called "Inception". Superhapaxes of an individual are used in a single person targeted communication to activate their specific sentiment response to reach a client's target. An individual sentiment response is detected as a sentiment quality of a superhapax vicinity; superhapaxes with a higher degree of topic bond are selected for text transformation. An individually personalized text is utilized to be perceived by automatized linguistic processing of a person (a habitual way of language use) and to be grasped without reflection by an individual, i.e., the content of a message is received with less mental processing.

Superhapaxes are incorporated into a text skeleton representing a specific client's purpose (advertisement, internet content personalization). Prior to the entry of superhapaxes, the text skeleton is general, i.e., not yet targeted individually and not yet representing a client purpose. Inception applications are used as an automated tool of mind engineering, delivering attitudes and content characteristics on the unconscious level.

7.2 The utilization we called "A scanner darkly". Superhapaxes reveal a manneristic or compulsive language behavior of an individual and show the content which an individual reflects in their mental processing. Also, the relations of manneristic language devices present in a text are detected, i.e., we can see individual discursive strategies in individual perceiving the outer world, and by the linguistic analysis we can determine their sentiment profile which is not reflected on the surface of the text (willingly performed text).

The deep sentiment profile of an individual is used as an identifier of sensitive topics of hers or his to be used in psychoanalysis. Deep sentiment in a text reveal to a psychoanalyst hidden regular attitudes of a person, which is interpreted in conjunction with central psychological impairment. A deep sentiment analysis marks those parts of a text in which an individual exhibits a specific sentiment shift, or which are rich in individual sentiment.

A scanner darkly produces a list of superhapaxes and their related supertopics, i.e., a client's compulsive contents, feelings, attitudes towards the outer world consciously or unconsciously hidden under the text surface. The list of respective word forms and their vicinities is passed to a psychologist to be interpreted in a psychological manner. The psychologist is, thus, given parts of a client's text with a higher occurrence of their manneristic lexicon, i.e. with accumulated supertopics, to interpret them in a psychological manner. This serves for creating a personal

profile in the HR field. This personal profile can be interpreted via NEO-FFI Big Five, MMPI or some other psychometrics.

Figure 5: The described method may enable to highlight text parts which bear a specific significance for their author. It can reveal specifically important topics for the author together with their related sentiment. It also serves for detecting their compulsive language behavior.

7.3 Transformation of received digital content to avoid inception targeting. The respective procedure described as inception (7.1) is used for the transformation of received messages and internet content. The client prevents activation of their unconscious reaction to a content via manipulation by superhapaxes and their vicinities. Selected superhapaxes and supertopics are extracted from a received text and replaced by their synonyms with an unmarked style form.

7.4 Detection of automated text production (disinformation). Bruntal effect, i.e. recurrence of low-frequency phenomena in connection with a stable vicinity, provides a characteristic of a human-made text. In case a certain text does not manifest low-frequency occurrence in a certain vicinity (an expression, content and sentiment), an automatically produced text is detected. It also applies for a text originating from a collection of other texts (collected automatically or by a copywriter).

7.5 Highlighting text sections with specific significance for translators and interpreters. As seen in 7.1 and 7.2, superhapaxes and deep sentiment analysis are capable of identifying the text sections which bear specific importance also for interpreters. The signs of a specific author's language behavior are hard to grasp on the surface, i.e., interpreters perceive it difficult (or even impossible) to interpret such text sections while maintaining its original deep sentiment and hidden mannerisms. Interpreters are guided through the text to be interpreted while being notified of the superhapaxes, supertopics and sections charged by the author with special sentiment and manneristic burden.

An interpreter is informed of the phenomena with low frequencies which manifest the authorial style. They are prompted to copy the distinctive characteristics of the expression, content and sentiment of this phenomenon and its vicinity. This applies to approximately 30% of the text which is lexically ordered by superhapaxes; consequently their vicinities are shown relative to the highlighted supertopics selected for replication.

7.6 Automated text production assigned to a chosen author. Automated text generator utilizes a modulator defining the style of the text author. The modulator contains superhapaxes and supertopics; superhapaxes are designated within the text in the locations of their synonyms and supertopics are inserted into their vicinities in appropriate syntactic and semantic forms. The user of the generator selects the author in whose style to shape the text.

DISCUSSION

Our evidence is that people regularly repeat specific words even with minimal frequency and without any direct relation to the context of the communication. When utilizing this, people are profilable by extracting specific topics and their related sentiments that are repeated unconsciously (and infrequently). This kind of knowledge gives us an advantage when defining an individual's personality development and highlighting an individual's key drivers and aspirations.

The regular recurrence of low-frequency words in our texts could elicit philosophical inquiry regarding the free will of humans or programmed behavior. An opposite and also commonly perceived impression is that we are used to recognizing individual language habits of people, their favorite words, topics and emotional reactions. These traits are the essence which makes up a person's personality. What has surprised us is how a set of these words is regular even if accompanied with a minimum frequency (Faltýnek – Matlach 2021; Faltýnek et al. 2023). This is also the reason why this phenomenon has not been recognized until now. Alongside, in certain circumstances, this phenomenon could be used as a tool to utilize a language personal profile in applications for various reasons: in the forensic use as an identifier of the person who produced the text (see Chaski 2001), in information retrieval as a tool to classify texts or as a textual definition of the person, specifying their individual influence traits.

In the field of digital communication, it is now common to encounter situations where a provider of digital communication (search engines, social media, web portals) tailors the digital text for the user in order to have a better impact. See sample ways to utilize this text tailoring:

Patent 1: Linguistic Personalization of Messages for Targeted Campaigns (patent number: US20160171560A1), inventors are: Rishiraj Saha Roy, J. Guna Prasaad, Aishwarya Padmakumar and Ponnurangam Kumaraguru;

Patent 2: Computer method and apparatus for targeting advertising (patent number: US11120476B2), inventors are Thomas Gerace and Russell Barbour.

As previously stated in the aforementioned patent examples, personalization of digital communication or internet content has not yet been targeted to an individual by their own language. People are currently being personalized as a demographic set, additionally by their search history, individual location, device features etc. It is not known how to define a person via a defined set of individual-sensitive keywords, topics or feelings. Stimulation of a person is provided using demographically defined vocabulary. For example: The demographic group is then represented by words which are commonly used by members of the demographic group, in a contradiction to members of a different demographic group. So, in the case of a young American the message contains word "cool", not words "fabulous" or "brilliant". It is in contradiction to an older Australian or a middle-aged British (for examples of social language diversity see Orgoňová et al. 2023, pp. 38–60). Using our technology described above, an individual is personalized by their own favorite words, topics and is stimulated by their individual feelings attributed to these words and topics. We assume that this is the future of digital communication amendment, with the objective to enhance its effectiveness and efficiency so that it is prepared to be easily understandable and to be received by an individual. Marketing is much more efficient when communicating with customers in their personal language. People are also directly persuaded in case of political marketing. It is also possible to purposefully change attitudes of people in certain critical positions. We can tackle extremism and terrorism using the above-described technique in surveillance. We can use it as a hybrid tool in PSYOPS activities. We can do mind engineering. In accordance with these efforts and knowing these techniques, we are capable of engineering an instrument to prevent unwanted personalization, as "where there is use, there is abuse as well".

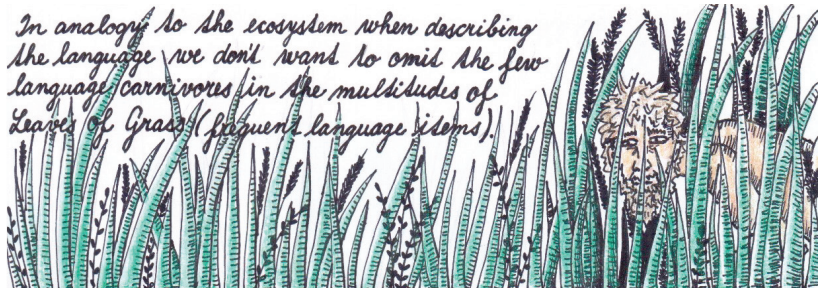

Figure 6: The new discovery of the role of low-frequency regularly-dispersed words occurring in a text by a single author is a game-changer in the field of mind engineering.

References

AMELIN, Konstantin – GRANICHIN, Oleg – KIZHAEVA, Natalia – VOLKOVICH, Zeev (2018): Patterning of writing style evolution by means of dynamic similarity. In: *Pattern Recognition*, Vol. 77, pp. 45–64.

BAAYEN, Harald – VAN HALTEREN, Hans – TWEEDIE, Fiona (1996): Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. In: *Literary and Linguistic Computing*, Vol. 11, No. 3, pp. 121–132.

BENEŠOVÁ, Martina – FALTÝNEK, Dan – ZÁMEČNÍK, Lukáš Hadwiger (2015): Menzerath-Altmann Law in differently segmented text. In: TUZZI, Ariuna, BENEŠOVÁ, Martina, MAČUTEK, Ján (eds): *Recent Contributions to Quantitative Linguistics.* Berlin/ Boston: De Gruyter Mouton, pp. 27–40.

BINONGO, José Nilo G. (2003): Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. In: *Chance*, Vol. 16, No. 2, pp. 9–17.

BLOOMFIELD, Leonard (1933): *Language*. New York: Henry Holt & Co. 466 p.

CATFORD, John C. (2002): *A Practical Introduction to Phonetics*. Oxford: Oxford University Press. 320 p.

CHASKI, Carol E. (2001): Empirical evaluations of language-based author identification techniques. In: *The International Journal of Speech, Language and the Law*, Vol. 8, No. 1, pp. 1–65.

DE BEAUGRANDE, Robert A. – DRESSLER, Wolfgang (1981): *Introduction to Text Linguistics*. London – New York: Longman. 286 p.

DE SAUSSURE, Ferdinand (1966): *Course in General Linguistics*. New York – Toronto – London: McGraw-Hill Book Company. 240 p.

DIEDERICH, Joachim – KINDERMANN, Jörg – EDDA, Leopold – PAASS, Gerhard (2003): Authorship attribution with support vector machines. In: *Applied Intelligence*, Vol. 19, No. 1–2, pp. 109–123.

DOLNÍK, Juraj (2021): Vertikálne a horizontálne jazykové potreby. In: *Jazykovedný časopis*, Vol. 72, No. 1, pp. 21–36.

DOLNÍK, Juraj (2019): Stratifikácia používateľov jazyka. In: *Jazykovedný časopis*, Vol. 70, No. 3, pp. 515–528.

DOLNÍK, Juraj (2018): Jazykové znalosti a ovládanie jazyka. In: *Jazykovedný časopis*, Vol. 69, No. 1, pp. 77–89.

EVERT, Stefan – PROISL, Thomas – JANNIDIS, Fotis – REGER, Isabella – PIELSTRÖM, Steffen – SCHÖCH – Christof – VITT, Thorsten (2017): Understanding and explaining Delta measures for authorship attribution. In: *Digital Scholarship in the Humanities*, Vol. 32, No. 2, pp. 4–16.

FAIRCLOUGH, Norman (1995): *Critical Discourse Analysis: The Critical Study of Language*. London: Longman. 265 p.

FALTÝNEK, Dan (2020): Celkom iste sa príde na to, že niektoré slová sa opakujú: Horeckého hypersyntax. In: *Jazykovedný časopis*, Vol. 71, No. 2, pp. 185–196.

FALTÝNEK, Dan – BENEŠOVÁ, Martina – MATLACH, Vladimír – KUČERA, Ondřej (2023): System and method for adapting text-based data structures to text samples. United States Patent and Trademark Office: US 17/809 130; patent US 11 797 753 B3.

FALTÝNEK, Dan – LACKOVÁ, Ľudmila – OWSIANKOVÁ, Hana (2020): Once again about the hapax grammar – Epigenetic Linguistics. In: *Linguistic Frontiers*, Vol. 3, No. 1, pp. 23–27.

FALTÝNEK, Dan – MATLACH, Vladimír (2021): Hapax remains: Regularity of low-frequency words in authorial text. In: *Digital Scholarship in the Humanities*, Vol. 37, No. 3, pp. 693–715.

FALTÝNEK, Dan – BENEŠOVÁ, Martina – KUČERA, Ondřej (2025): Cloakspeech. Olomouc: Palacký University. (software).

FENGXIANG, Fan (2010): An asymptotic model for the English hapax/vocabulary ratio. In: *Computational Linguistics*, Vol. 36, No. 4, pp. 631–637.

FORSTER, Kenneth I. (1976): Accessing the mental lexicon. In: F. Wales – E. Walker (eds): *New Approaches to Language Mechanisms*. Amsterdam: North Holland, pp. 257–287.

GERACE, Thomas – BARBOUR, Russell (2017): Computer method and apparatus for targeting advertising. United States Patent Office, US 11120476 B2.

GRZYBEK, Peter (2007): History and methodology of word length studies: The state of the art. In: P. Grzybek (ed): *Contributions to the Science of Text and Language*. Dordrecht: Springer, pp. 15–90.

HALLIDAY, Michael A. K. – HASAN, Ruqaiya (1976): *Cohesion in English*. London: Longman. 390 p.

HASPELMATH, Martin – SIMS, Andrea (2010): *Understanding Morphology*. London: Routledge. 384 p.

HŘEBÍČEK, Luděk (1989): The Menzerath-Altmann Law on the semantic level. In: *Glottometrika*, Vol. 11, pp. 47–55.

JAKOBSON, Roman – FANT, C. Gunnar M. – HALLE, Morris (1952): *Preliminaries to speech analysis.* Technical Report, No. 13: Acoustics Laboratory, M.I.T. 58 p.

JONES, Sparck K. (1972): A statistical interpretation of term specificity and its application in retrieval. In: *Journal of Documentation*, Vol. 28, pp. 11–21.

JUOLA, Patrick (2008): Authorship attribution. In: *Foundations and Trends in Information Retrieval*, Vol. 1, No. 3, pp. 233–334.

KOPPEL, Moshe – SCHLER, Jonathan – ARGAMON, Shlomo (2009): Computational methods in authorship attribution. In*: Journal of the Association for Information Science and Technology*, Vol. 60, pp. 9–26.

LARDILLEUX, Adrien – LEPAGE, Yves (2007): The contribution of the notion of hapax legomena to word alignment. In: *The 3rd Language and Technology Conference (LTC'07).* Poznań, pp. 458–462.

MIKROS, George K. (2009): Content words in authorship attribution: an evaluation of stylometric features in a literary corpus. In: R. Köhler (ed): *Studies in Quantitative Linguistics 5*. Lüdenscheid: RAM, pp. 61–75.

ORGOŇOVÁ, Oľga – BOHUNICKÁ, Alena – KAZHARNOVICH, Marina (2023): *Sociálna inklúzia a používanie jazyka.* Bratislava: Comenius University. 209 p.

PENNEBAKER, James W. – BOYD, Ryan L. – JORDAN, Kayla – BLACKBURN, Kate (2015): *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin. 26 p.

PRITCHARD, Jonathan K. – STEPHENS, Matthew – DONNELLY, Peter (2000): Inference of population structure using multilocus genotype data. In: *Genetics*, Vol. 155, No. 2, pp. 945–959.

ROY, Rishiraj S. – PRASAAD, Guna J. – PADMAKUMAR, Aishwarya – KUMARAGURU, Ponnurangam (2014): *Linguistic Personalization of Messages for Targeted Campaigns*. United States Patent Office, US 20160171560 A1.

ROY, Rishiraj S. – PADMAKUMAR, Aishwarya – JEGANATHAN, Guna P. – KUMARAGURU, Ponnurangam (2015): Automated linguistic personalization of targeted marketing messages mining user-generated text on social media. In: A. Gelbukh (ed): *Computational Linguistics and Intelligent Text Processing, CICLing 2015. Lecture Notes in Computer Science.* Cham: Springer, pp. 203–224.

RYBICKI, Jan – EDER, Maciej (2011): Deeper Delta across genres and languages: do we really need the most frequent words? In: *Literary and Linguistic Computing*, Vol. 26, No. 3, pp. 315–321.

SADILLOEVNA, Dilfusa F. – ANVAROVNA, Mavlyuda M. – FATTOXOVICH, Furgat D. – BAXRONOVICH, Barat N. (2020): Dimensions and levels of linguistic analysis. In: *International Journal of Psychosocial Rehabilitation*, Vol. 24, No. 3, pp. 394–403.

SAVOY, Jacques (2012): Authorship attribution based on specific vocabulary. In: *ACM Transactions on Information Systems*, Vol. 30, No. 2, Article No. 12, pp. 1–30.

TRUBETZKOY, Nikolaj S. (1969): *Principles of Phonology*. Berkeley – London: University of California Press. 344 p.

VAN DIJK, Teun A. (2015): Critical discourse analysis. In: D. Tannen – H. E. Hamilton – D. Schiffrin (eds): *The Handbook of Discourse Analysis*. 2nd ed. West Sussex: John Wiley & Sons, Inc., pp. 466–485.

VAN DIJK, Teun A. (2008): *Society in Discourse. How Context Controls Text and Talk.* Cambridge: Cambridge University Press. 287 p.