The possibilities of text mining in the examination of the local society of the Ormánság region¹

Hornyák Miklós – Ragadics Tamás University of Pécs, Hungary

THE AIMS OF THE PAPER

This paper aims to represent the background of the consumer behaviour in families from underprivileged strata living in small settlements of the Ormánság region.

METHODOLOGY

Our diagnoses are based on interviews with mayors. We use text mining methods for deeper analysis of the interviews and present results with different visualization techniques. With the help of comparative word clouds, we performed the support of the examination of settlement clusters created through using the quantitative data.

MOST IMPORTANT RESULTS

According to our experimental work, the method is suitable for making the results of interview research more targeted, sophisticated and to demonstrate the results. Our experiences gathered in the course of the research indicated the importance of the size of the corpus, and the necessity of using specific knowledge when supporting sociological research with text mining.

RECOMMENDATIONS

Ormánság micro region (South Transdanubia, Hungary) is characterized by a frittered settlement structure, low economic figures, and disadvantageous social processes. Adult members of families living in the culture of poverty are not able to deliver the patterns and motivation for social and economic advancement for their children, neither the values or norms for the conventional organizing of everyday life. Illogical consumer habits ("Carpe diem" mentality, commercial credit, purchasing status symbols etc.) are deepening the social gap and segregation and confining the accidental chances for self-care. Complex supporting system is necessary with the elements of labour market incentives, consulting, mentoring and development of competencies.

Keywords: culture of poverty, data visualization, text mining, underclass

¹ Supported by the UNKP-17-3-IV-PTE-193 New National Excellence Program of the Ministry of Human Capacities.

INTRODUCTION

The subjects of the analysed interviews were the mayors of Ormánság, who were asked about the problems and opportunities of their settlements. The material of the 25 interviews prepared in 2010-12 provides insights into the state of the local communities, their cultural life and the consumption behaviour of the population, in addition to a general description of the villages. Our main questions were: How can the heterogeneity of the settlements, settlement clusters be detected in a uniformly disadvantaged region? Can text mining be used for the characterization of categories identified on the basis of statistical data? Can the consuming characteristics typical for each settlement category be presented with the help of text mining? Can text mining support the qualitative interview method through the results of the processing the textures concerning the villages?

The groups of villages were created on the basis of hard indicators characteristic for the settlements, thus sorting them into categories of development (Ragadics 2015). The comparison of thematically corresponding text sections belonging to the separate blocks was carried out with the method of text mining. We used a self-created Python programme for processing the corpus created from

the interviews. We demonstrate our results with different types of word cloud presentation methods, including the comparative word cloud used at the examination of the settlement groups. We support the results from the conventional processing of the interviews with the help of visualization.

VILLAGES IN THE ORMÁNSÁG REGION

The historic micro region of Ormánság is located in the south of Baranya county, on the floodplains of the Drava. The bordering of the area can mainly be defined on the basis of its ethnographic characteristics: traditionally, 45 Calvinist Hungarian villages were listed in the region (compare with Kiss 1937). Ormánság - demonstrating the ailment of the Hungarian society - has been in the focus of scientific interest on several occasions. First, it was in the focus in the first half of the 20th century due to the birth control leading to significant economic advantages, and in the second half of that century, due to the large increase of the proportion of the Roma population. Today, economic and social problems have become interlinked, and, crossing the borders of the region, have created a continuous crisis area.

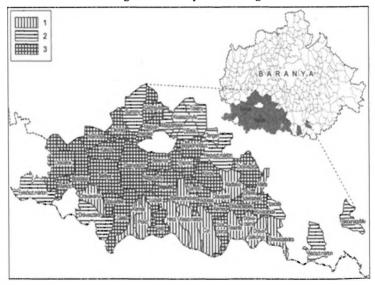


Figure 1: The map of Ormánság

Source: own construction

Legend: 1 -The settlements of the historic Ormánság (Kiss Géza 1937); 2 - Settlements involved in the research; 3 - The settlements of the historic Ormánság and the settlements involved in the research

The traditionally closed border region lies far from the hubs of development. The local social care system is underdeveloped and deteriorating, transportation is complicated due to the dispersed structure of the area. The number of local workplaces has dropped dramatically following the change of the regime, and the opportunities of communing have become limited as well. The structure of village societies has eroded: the educated and young residents have moved away, the proportion of disadvantaged population has increased, and ethnical segregation is high (Ladányi & Virág 2009). However, the fate of small settlements is different depending on their features, opportunities and social conditions. The extent of the local underclass2 is a key factor concerning the future of the villages.

Ormánság – Local societies of deprived consumers

In addition to the above indicators, another significant factor of differentiation - in connection with the trends of migration due to negative causes - are the composition, quality, demographic characteristics and life opportunities of the population. By opportunities, in a consumer society we primarily mean the ability to access services and consumer goods: "In a consumer society, a 'normal life' is the life of consumers" (Bauman 2004, 38). Even for groups with low, irregular income the desire for possessing certain devices (e.g. smartphones) or luxury goods (e.g. well-known alcohol and tobacco products) is given as a norm. Veblen describes this phenomena as the presentation of "decency": "No class of society, not even the most abjectly poor, foregoes all customary conspicuous consumption. The last items of this category of consumption are not given up except under stress of direct necessity" (Veblen 1899/2001, 85)

The consumer habits of those living in permanent poverty – based on the interviews – correspond to the "culture of poverty" phenomenon described by Oscar Lewis (1966). Individuals and families who occasionally receive money but otherwise experience permanent poverty react with a strong consumer compensation, complying with the challenges of consumer values. Instead of a methodical budget management, in their case the rapid outflow of income can be observed. This behaviour makes recovery from poverty impossible and

consolidates the disadvantaged condition: "People in a culture of poverty produce little wealth and receive little in return. Chronic unemployment and underemployment, low wages, lack of property, lack of savings, absence of food reserves in the home and chronic shortage of cash imprison the family and the individual in a vicious circle" (Lewis 1966, 21).

Permanent unemployment leads to the significant increase of spare time in the world of lagging small settlements. Permanent inactivity deteriorates skills, undermines the motivation levels of the affected groups, decreases the skill of self-care and increases vulnerability. In addition, it results in severe frustration. Spare time is the time for consumption, and those living in poverty are unable to use this endless free time in accordance with the requirements of consumer values (Bauman 2004).

Even the low, hard-earned income flows out from the villages of Ormánság. This is facilitated by the mobile-vendors who appeared after the degradation of local retail trade, by the local governments organising shopping trips to nearby hypermarkets and taxing their own entrepreneurs. The culture of poverty facilitates this process. Complex training, the conscious development of the self-care skills would be necessary to break the vicious circles.

Several dozens of interviews were performed in the course of the research of the local society of Ormanság (Ragadics 2015). Currently the digital version of 25 interviews is available, where the interviews are organised along seven questions. The interviews were typically made with the leaders of the settlements. By using the tools of text mining we looked for a possibility of supporting the analysis of the texts and the clear presentation of the results.

TEXT MINING

Due to the availability of a large quantity of textual contents digitally, text mining is becoming increasingly widespread. Through the development of the technology, of algorithms, the mechanical analysis of the information contained by the text has become available for the experts of social sciences. The primary aim of text mining is to search for patterns using unstructured data: documents, articles, text flows, news, notes, products of social networks (short texts), multimedia contents. (Creese 2004)

² A status under the level of society, outside the structure of society, permanent, lifelong exclusion inherited from generation to generation (Wilson 1987).

Fan et al. (2005) provide a possible definition of text mining saying that it is the investigation of new, previously unknown information with computers, through automatically obtaining information from various written sources. As opposed to numerical data, in case of text-based inputs dependency on language is a real problem, as the texts have been written in a natural language (e.g. in Hungarian, English). In order to tackle this problem, semantic and taxonomic examinations of the input can constitute an essential element of text mining. We did not perform such examinations in our current study.

The fields of text mining

The primary aim of text mining is the analysis of the information, and the secondary aim is to carry out the search. Text mining provides an opportunity to look behind the information: to drill, analyse, compress and, through these, to support decision-making. These aims are facilitated by several software and hardware, whose development has enabled the examination of data of increasingly wider sources. These software and algorithms

mainly focus on learning from the text and from various data from other sources.

Textual data can be defined as the mass of letters and words, however, for deep examination the knowledge of semantics can be necessary. In special areas (medical databases, examination of the web) or when analysing natural languages, semantic examinations have an important role, especially in the case of obtaining data, through which the secondary meaning of the text can also be obtained (Aggarwal & Zhai 2012).

In the typical future directions of usage and research, due to the huge increase in the quantity of textual data, the development of new methods is expected. With the use of interdisciplinary data, text mining can be the bridge between researcher groups working in different areas of science.

THE APPLIED TEXT MINING PROCESS

Figure 2 shows the process of text mining applied in the research.

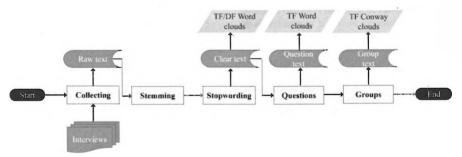


Figure 2: Process flow of our method

Source: own construction

The prerequisite of text mining is an existing text base, which can be analysed after various steps of preprocessing and purification. In the first step, it is necessary to gather the text base that will be the subject of the analysis, which, in the case of the present research, was the digitalised text material of the sociological interviews previously prepared in the Ormánság region. The raw corpus resulting from this step was the basis of further processing. The next step we performed on the raw corpus was to purify it from incomplete, incorrect data. This stage requires significant human labour, especially due to the fact that this activity can

only be performed with the help of an expert proficient in the topic. The sociologist expert participating in the research created a new version of the corpus especially relevant to the given domain by removing the parts irrelevant to the content, while the main text does not suffer any loss of information concerning its content. This stage also included a significant step which is essential in the case of the Hungarian language: the examination of the character encoding of the digitalized documents. In the case of the Hungarian language checking the encoding forms used for the correct presentation of the accented letters (e.g. ISO-8859-2) and the

avoidance of anomalies resulting from encoding conversions during the processing of the text are of special importance. As a result, a textual corpus suitable for mechanical processing was created.

In the first step of the stage supported by computers, with the help of tokenization, as opposed to breaking it down into sentences which happens in natural languages, we created smaller elemental units which can relatively clearly be determined in texts. In the case of the present research, these units were individual words. The group of tokens, that is, character series of identical content are called types, which are the components of the dictionary created from the corpus. Language-specific solutions have been developed for the performance of the task (Tikk 2007). In the present research, we carried out the automatic tokenization with the help of the Hungarian dictionaries and algorithms of the Hunspell programme package. As a result of breaking down the text down into words, the units that do not carry any content information became visible, and were removed from the text through the step of removing stop words. As a result, the size of the corpus to be analysed can be reduced significantly. The further reduction of the size without loss of information can be solved through stemming. Its aim is to trace back the modified versions (inflection, suffixes) of the words identified (tokenized) in the corpus to their stems, while its aim is to reduce the vector space, as the identical words are contracted in a common canonical form pl. büszkeség / büszkék / büszkén → büszke.

The scale of reduction is different depending on the language, the grammatical structure of the given language: in the case of the Hungarian language it can be up to 90%, as opposed to the 50% value of the English language (Tikk 2007). The result of stemming, as opposed to linguistic lemmatization, is not always a meaningful stem form. Various methods of stemming are known (e.g. NLTK Snowball, Hunmorph). In the present research the stemming algorithm for Hungarian language of the Hunspell programme package was used.

The main challenge of analysing texts lies in the fact that documents created in human (natural) languages can only be processed with a computer, that is, by converting the text into an artificial language. Documents can be represented through three, fundamentally different approaches: set theory, algebraic, probability. In the case of the set theory approach, the similarity of documents is approached through set operations, in the algebraic approach it happens through matrix operations, while in the case of the probability approach, it is performed through probability estimations (Tikk

2007). In the case of the present research the similarity (distance/proximity) of the documents could be determined with the help of the vector space model created through the algebraic approach. The basis of the model is the vocabulary of the examined documents, while the level of their similarity lies in the extent of the overlapping of this vocabulary.

That is, if, where vector D is identical to the document collection of our corpus, in which the d elements refer to the individual documents. In the vector space, we represent the individual documents with the vector created from the words contained in the documents, , where vector d consists of M number of words of document i (the frequency of those words). We can summarize the corpus compiled from the documents with the help of the Term Document Matrix (TDM).

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & & & A_{2N} \\ \vdots & & & \vdots \\ A_{M1} & A_{M2} & \cdots & A_{MN} \end{bmatrix}$$

In the D TDM matrix the number of the lines corresponds to the number of unique words (M), while the number of columns is identical to the number of documents constituting the corpus (N). The elenent of the TDM matrix contains the relevance of word i in document j. In the TDM matrix the position of the words within the document is lost, thus this technique is also called "bag of words" model (Tikk 2007).

Visualizing text mining results

The wide-spread solution for the presentation of text mining results are word clouds, which enable the concise, compact visualization of the information included in the texts and the data connected to them. When the text is processed, the word clouds created from the natural text present the words most typical for the text or appearing most frequently therein.

The word cloud was first applied in the work of Stanley Milgram in the mid-1970s, used for a presentation based on quantitative values connected to words (Milgram & Jodelet 1976). The computerised support and presentation of word clouds became wide-spread from the second half of the 1990s, and this technique has become part of today's data visualization (Viégas & Wattenberg 2008). During the past years, various solutions have been developed, the most significant are the

following: Wordle (Feinberg et al. 2009), Prefix (Burch et al. 2013), SparkClouds (Carpendale et al. 2010). The shortcoming of the word clouds that can be considered conventional is that they are only suitable for the presentation of one single corpus, but are unable to present further background information. The comparative word cloud was created for the simultaneous examination of several corpuses, and also provides a more comprehensive image of the differences and similarities of two corpuses.

We applied two techniques in our research: conventional word clouds and comparative word clouds. The combination of these two techniques and their application in the characterisation of clusters was developed by Kruzslicz and colleagues. In their method, the scale of separateness of two document clusters is determined by the quality of individual words and the "butterfly shape" of words concentrated on the two edges, which is visualized with a comparative word cloud. The cohesion of the clusters is expressed by the proportion and position of large- and small-sized words. Thus, we can receive a visual confirmation of the extent the objects included in the two clusters are grouped around each concept. The method determining the quality of clustering is based exactly on this observation (Kruzslicz et al. 2016). Process of the cluster evaluation through labelling, applied in our research is the following:

- Clustering the objects which are the settlement groups created in a previous research on the basis of quantitative data.
- Assigning related documents, that is, the interviews performed in the settlements, to the objects.
- The weighted list of the potential labels of the document sets assigned to the clusters, which are based on the TDM matrixes created from the corpuses.

- 4. Preparing comparative word cloud from the words of the labels by cluster pairs.
- Forming an opinion concerning the delimitation and definiteness of each cluster with the help of comparative word clouds.

RESULTS

In our research, we carried out three different examinations on our corpus containing the text of the interviews. In the first step we generated Term Frequency (TF)/Document Frequency (DF) word clouds for the entire corpus, that is, for all of the interviews, then with the help of TF word clouds generated from the seven corpuses created through summarising the interview by questions, we attempted to drill deeper in order to explore hidden information in the responses, and finally, through the text mining analysis of the interviews assigned to the clusters (settlement groups) created by a sociologist expert in a previous research, we aspired to provide a more accurate description of the clusters.

Rase statistics

The 25 interviews involved in the corpus were organised along seven questions, to which responses of various sizes (length) were given. The text mining analysis of the responses given to the questions is determined by the size of the available text material. Table 1 shows the heterogeneous distribution of the size of the responses, ranging from 14 characters (question 3) to 4813 characters (question 1). Thus, it was not possible to analyse the questions for each interview, however, by contracting the responses of the interview questions, the aggregated text content of the seven questions made the examination possible.

Table 1: Statistics of interview question's answers in numbers of characters

Question ID	Question texts	Avg. Length	Min. length	Max. length
1	Introduction of village	1043	113	4813
2	Relations to settlements	749	98	3482
3	Local community	1278	14	3732
4	Local traditions	868	174	3332
5	Local opinion leaders	427	65	1573
6	Plans for the future	1160	147	3619
7	Consumption	490	162	2864

Source: own calculation

The examined settlements were divided into three groups on the basis of a number of indicators³. The first group indicates a high level of population decline (1990-2011), low qualification, high unemployment and crime rates, and the stronger presence of the Roma population. The second represents the mid-range in Ormánság. The third group shows lower than average population decline, higher employment and qualifications and lower crime rate and a lower proportion of the Roma population. For the appropriate interpretation of the above, a number of factors must be taken into consideration:

- The absence of population decline does not necessarily mean a stable, developing society, in the slums of the poor a population an increase was observed in the recent period.
- A low level of education can indicate a high number of children.
- 3. The high unemployment rate can indicate a high rate of inactive people.
- Criminal statistics only include the reported crimes, legal awareness and the proportion of using police assistance can be lower in the case of the vulnerable population.
- Belonging to the Roma population is based on self-declaration, willingness to response is uncertain, and the proportion of refusing to admit a Roma identity is high.

 The two largest settlements of the Ormánság – in order to enable the easier comparison of the villages – were not included in the sample.

On the basis of the above factors, the micro centres of the Ormánság become visible, such as Drávafok and Kémes, which have above-average population and several local institutions (e.g. schools), in addition to some small settlements of more favourable social structure. The villages in the worst situation, thus obtaining low scores, included e.g. Kórós, Marócsa, Okorág, Ózdfalu and Sumony. The group of villages representing the middle level includes a wider range of settlements with mixed indicators.

Table 2 shows the statistics of the interview contents assigned to the triple cluster group created through the expert's grouping on the basis of this settlement typology. The heterogeneous composition of the text base of the three groups can be observed, which forecasts the limitations of the text mining support of cluster evaluation concerning each pair of groups. The significantly low average length of the interviews of group 1 makes examination with comparative word clouds more difficult, while in the case of the other two groups the average length of the interviews makes the examination easier.

Table 2: Statistics of interviews by village groups

Group ID	Villages	Avg. length	Min. length	Max. length
1	Drávafok, Drávasztára, Kémes, Drávacsepely	3444	1426	6119
2	Besence, Hegyszentmárton, Ipacsfa, Sósvertike, Adorjás, Diósviszló, Drávaszerdahely, Nagycsány, Páprád, Sámod, Sumony, Szaporca, Tengeri, Téseny	5884	2350	13821
3	Kákics, Marócsa, Ózdfalu, Cún, Kórós, Lúzsok, Okorág, Ózdfalu, Vejti	7036	1259	18023

Source: own calculation

³ For further details, see Ragadics 2015.

A common text base was created by contracting the responses given to the questions. In this text corpus, with the help of the Term Frequency (TF) value, those frequent words can be identified whose occurrence is high in the corpus, which, however may mean that the words occurred frequently within one interview, while did not occur in the other interviews. With the help of this method, it is impossible to identify the problems affecting the majority of the settlements, that is, to identify the words that occurred in several interviews. The calculation of Document Frequency (DF) provides a solution to this problem, where the frequent words are those whose occurrence is high

in the documents comprising the corpus, that is, they occur in several interviews. Thus, it can be said that these words refer to a significant, recurring problem or topic. When examining Figure 3, it can be stated that the word "önkormányzat" (local government) occurs several times (TF) and in several places (DF) thus it can be identified as a significant issue, however, the lower occurrence (TF) and higher DF value of the word "gyerek" (child) refers to an even more significant problem source. Due to the small size of the corpus, no significant differences can be identified in the results of the two ways of examination.

Figure 3: Word clouds of our whole corpus

Term Frequencies: important topics

Document Frequencies: important topics

Term Frequencies: importan

Source: own design

Examining responses to the questions

With the contraction of responses given to the questions, seven text bases were created for the seven questions. We received the results presented on Figure 4 through a Term Frequency-based word

cloud representation concerning the seven text materials. In the responses given to the questions the size of the words occurring with higher frequency is larger, while with the decrease of frequency, the size of the words also decreases.

település reg falu tudio: gyerek bal Hakos lakosságközség 🖼 nyugdíjas gyereknapkorocso karácsony = élelmiszer merkoz jone falus települes oreg nap phallaat falu reg polgármester O: dologember fiatal chetosege climegy > csalad kap vállalkozó 2 velemeny segítidos ebb penz ember 25 05 Elehtose Mer 25 000 Elehtose Mer 25 0000 szotud falu tsz szokt .

Figure 4: Term Frequencies word cloud of question's answers

Source: own design

Through studying the word clouds, the following conclusions can be drawn, presented in Table 3.

Table 3: Textual analysis of the responses given to the questions

Question ID	Question texts	Frequent words
1	Introduction of village	idős – segély – cigány – vég
2	Relations to settle- ments	ház – ivás – akar/tud – elmegy
3	Local community	dolog – rég – mer – egymás - irigység
4	Local traditions	hagyomány – bál – falunap – tud – nyugdíjas
5	Local opinion leaders	képviselő – hallgat – csinál – hisz
6	Plans for the future	pályázat – kormányzat – önerö – munka
7	Consumption	vállalkozó – hitel – kap – ünnep / TESCO / kocsma

Source: own compilation

The results of the text mining examination of the interview extracts grouped around the seven topics provide an insight of the situation of the examined small settlements. When we asked mayors to describe their settlements, statements concerning aging, marginality appeared in high numbers. They deemed it important to mention the high number of people living on benefits and the high proportion of the Roma population. This negative attitude appears in the relationship of the residents and the village. From the point of view of mayors, only those residents remain in the settlements who are unable to leave. It is difficult to sell residential homes, and it makes moving to more developed areas impossible. The most serious problem according to settlement leaders is the fragmentation, disintegration of local society and the lack of a local community. In the centre of this block of questions is the expressions "irigység" (envy). Local traditions and culture is appropriately reflected by the words "bál" (ball), "falunap" "nyugdíjas" (village festival), (pensioner) appearing in the word cloud. The elderly are the most active class in terms of preserving traditions, as there are pensioners' clubs in several settlements. In settlements possessing rich cultural values, today only a very small group of residents is connected to the heritage of the Ormánság (in several settlements these customs have completely disappeared). Local culture manifests itself in the village festival and the periodically organised balls. The population is characterised by a significant passivity, the majority of the locals are primarily attracted to the programmes by food, drink and music. Key expressions indicating activity only appear in the case of the local government representatives considered to be opinion leaders. The future of the settlements is determined by the scale of external help. This is the reason why this group of questions contain the words "kormányzat" (government) and "pályázat" (tender) and the concept of "önerő" (own contribution) which means a serious burden. The expressions appearing in the questions concerning consumption also demonstrate the characteristics connected to disadvantaged consumers very effectively: loan, as the basis of consumption, aid packages distributed at holidays, and, in addition to food donation, the name of the pub and that of a hypermarket appeared in the word cloud.

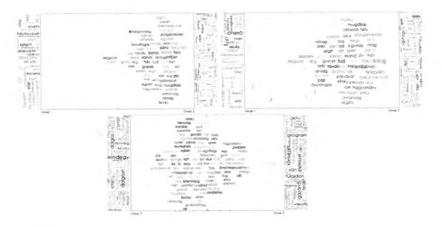
Examining village groups

We applied the method on the corpuses created from the groups of interviews divided by the expert. The textual transcript of the interviews performed in the settlements divided into three clusters was assigned as documents to the groups, thus we received three groups of documents. Thereafter the list of word frequency (TF) of the potential labels which was represented through a word cloud. Then the comparative word cloud of the document groups was prepared. The comparative word cloud created from the 100 common words shows the common words occurring the pairs of document groups, placing those words on the peripheries which occur in higher frequency in one or the other document group. The shift of the vertical axis determined by the set of words in the right or left direction

indicates the distribution of the quantity of the entire set of words belonging to the document group.

Figure 5 shows the comparative word clouds of the cluster pairs. Groups 1 and 2, as well as 1 and 3 are less distinguishable from each other on the basis of their text content. The lack of the butterfly shape cannot support the group division or assist the characterisation of these groups with extra information.

Figure 5: Comparative word cloud of village groups



Source: own design

As opposed to the two groups mentioned above, in the case of the cluster pair 2-3, the formation of butterflies is more pronounced, thus the two clusters are more distinguishable. The Conway word cloud (middle part) of group 3 is about work and activity, which is not only indicated by the word "dolog" (task) clearly belonging to group 3 among common words, but also the words "terv" (plan), "elmegy" (go away) and "csinál" (do) located on the edge of the butterfly's wing, thus more typical for group 3. However, these words can also be found in group 2 as well, but in lower frequency. The words of outstanding frequency located on the right edge of the figure, contained only in the word cloud of group 3: "környezet" (environment), "nép" (population), "program", also strengthen the connection with acts, activity. As opposed to this, the Conway word cloud, located in the middle of the figure, contains the words more often present in group 2, bearing a passive connotation: "próbál" (try), "pályázat" (tender), "kap" (receive), which are strengthened by the words only appearing in group 2, located on the left edge of the diagram: "enged" (let), "dolgozni" (work), "mindegy" (just the same).

To sum up, on the basis of the text of the interviews, these two groups can be characterised in separate ways: group 3 shows a rather active attitude, while group 2 shows a rather passive one, however, the central element of both groups is the activity carried out by the "család" (family), "polgármester" (mayor) and the "közösség" (community).

Future work

Several problems were identified during the examination. The majority of these are related to the difficulties of processing concerning the Hungarian language (e.g. using a Hungarian sentiment dictionary), while others require the further sophistication of the examination method (the examination of n-grams for the identification of negative meanings). Through the analysis of the word clouds, even in the case of this small-sized corpus it can be seen that with the use of topic-specific dictionaries (in order to improve stop wording) the size of our vector space can be further decreased without a loss of information.

SUMMARY

On the whole, it can be said that word mining applied as a complement to qualitative interview examination can provide the analyst with a form of quantitative confirmation. Naturally it does not promote the deep interpretation of the text, it does not replace in-depth interview techniques, but it can make the analytic work more sophisticated and - by indicating the appropriate key expressions -- focuses it on the essential content. Besides, it presents the examined scope of issues in an understandable way. Contrasting the different settlement groups vielded less success. The important concepts characterizing the given cluster also appear in this comparison. however the description of the character of the different village groups is not expressive, and it does not converge with the other results of the research. A more accurate work can possibly be carried out with a larger text and the detection of individual collocations. The key expressions only appearing in the case of one or the other settlement type can provide important, useful information complementing the technique of in-depth interviews. The further exploration of the possibilities of applying text mining in social science requires further work and interdisciplinary cooperation.

REFERENCES

- Aggarwal, C. C. and Zhai, C. X. (2012), *Mining Text Data*, Springer
- Bauman, Z. (2004), Work, Consumerism and the new Poor, New York: McGraw-Hill Education
- Burch, M., Lohmann, S., Pompe, D. and Weiskopf D. (2013), "Prefix Tag Clouds", *Information Visualisation (IV)*, 17th International Conference, 45-50
- Carpendale, S., Karlson, A. K., Lee, B. and Riche, N. H. (2010), "SparkClouds: Visualizing Trends in Tag Clouds", Visualization and Computer Graphics, 16 6, 1182-9
- Creese, G. (2004), "Duo-Mining: combining data and text mining+, *DM Reviews*, September
- Fan, W., Wallace, L., Rich, S., Zhang, Z. (2005), "Tapping into the Power of Text Mining", Communications of the ACM, 49 9, 76-82
- Feinberg, J., Viégas, F. B., Wattenberg, M. (2009), "Participatory Visualization with Wordle", Visualization and Computer Graphics, 15 6, 1137-44
- Kiss G. (1937): Ormányság. Budapest: Sylvester Rt. Kovács T. (2005): A sellyei (siklósi-) KSHtérség rövid leírása. In. Baranyi Béla (ed.):

- Roma szegregációs folyamatok a csereháti és dél-baranyai kistérségekben. Budapest: Gondolat – MTA Etnikai-nemzeti Kisebbségkutató Intézet. 29-52
- Kruzslicz F., Kovács B. and Hornyák M. (2016), "Összehasonlító klaszterjellemzés külső, szöveges források bevonásával", Statisztikai Szemle, 94 11-12, 1123-48
- Ladányi J. Virág T. (2009), "A szociális és etnikai alapú lakóhelyi szegregáció változó formái Magyarországon a piacgazdasági átmenet időszakában", Kritika, Jul-Aug, 2-8
- Lewis, O. (1966), "Culture of Poverty", Scientific American, 215 4, 19-25
- Milgram, S., Jodelet, D. (1976), "Psychological Maps of Paris", Environmental Psychology, 104-24
- Ragadics T. (2015), Helyi társadalom, lokális közösségek az ormánsági kistelepüléseken, PhD értekezés, Pécsi Tudományegyetem Bölcsészettudományi Kar Interdiszciplináris Doktori Iskola
- Tikk D. (2007), Szövegbányászat, Budapest: Typo-Tex
- Veblen, T. (/1899/2001): The Theory of the Leisure Class. New York
- Viégas, F., Wattenberg, M. (2008), "Tag Clouds and the Case for Vernacular Visualization", *Interac*tions, 15 4, 49-52
- Wilson, W. J. (1987), The Truly Disadvantaged. The Inner City, the Underclass, the Public Policy, Chicago, IL: The University of Chicago Press

Hornyák Miklós, Assistant Lecturer hornyakm@ktk.pte.hu University of Pécs Faculty of Business and Economics Department of Quantitative Management

Tamás Ragadics PhD, Assistant Professor ragadics.tamás@pte.hu
University of Pécs
Faculty of Humanities
Institute of Social and Media Studies
Department of Sociology