

# Information Retrieval and Terminology Extraction in Online Resources for Patients with Diabetes

Sanja Seljan<sup>1</sup>, Maja Baretić<sup>2</sup> and Vlasta Kučič<sup>3</sup>

<sup>1</sup> University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences, Zagreb, Croatia

<sup>2</sup> University of Zagreb, University Hospital Center Zagreb, Division of Endocrinology, Department of Internal Medicine, Zagreb, Croatia

<sup>3</sup> University of Maribor, Faculty of Arts, Department of Translation Studies, Maribor, Slovenia

## ABSTRACT

*Terminology use, as a mean for information retrieval or document indexing, plays an important role in health literacy. Specific types of users, i.e. patients with diabetes need access to various online resources (on foreign and/or native language) searching for information on self-education of basic diabetic knowledge, on self-care activities regarding importance of dietetic food, medications, physical exercises and on self-management of insulin pumps. Automatic extraction of corpus-based terminology from online texts, manuals or professional papers, can help in building terminology lists or list of »browsing phrases« useful in information retrieval or in document indexing. Specific terminology lists represent an intermediate step between free text search and controlled vocabulary, between user's demands and existing online resources in native and foreign language. The research aiming to detect the role of terminology in online resources, is conducted on English and Croatian manuals and Croatian online texts, and divided into three interrelated parts: i) comparison of professional and popular terminology use ii) evaluation of automatic statistically-based terminology extraction on English and Croatian texts iii) comparison and evaluation of extracted terminology performed on English manual using statistical and hybrid approaches. Extracted terminology candidates are evaluated by comparison with three types of reference lists: list created by professional medical person, list of highly professional vocabulary contained in MeSH and list created by non-medical persons, made as intersection of 15 lists. Results report on use of popular and professional terminology in online diabetes resources, on evaluation of automatically extracted terminology candidates in English and Croatian texts and on comparison of statistical and hybrid extraction methods in English text. Evaluation of automatic and semi-automatic terminology extraction methods is performed by recall, precision and f-measure.*

**Key words:** health literacy, terminology, information extraction, diabetes mellitus type 1, documentation online, language barriers

## Introduction

Terminology use, as a mean for information retrieval or document indexing, plays an important role in one aspect of information literacy. Corpus-based terminology extraction can help in information retrieval by providing »browsing phrases« – N-grams in order to assess appropriate documents. Creation of specific terminology lists represents an intermediate step between the free text search and controlled vocabulary, which are used as indexes for document access.

Technology has entered most aspects of communication, education, business and medicine, not only from the

point of employers and users, but also from the point of patient and physician. Finding the proper information is crucial for the patient with type 1 diabetes since on the basis of information, important decisions are done many times, every day. If this key word is not understood enough, the consequences could be fatal. Use of adequate terminology in online materials, adapted according to users' needs, can raise information access to specific type of documents. Development of new technology put the patient, educator and physician in a demanding position: it is necessary to be acquainted not only with medical, but also with basic information literacy knowledge<sup>1</sup>.

According to multilingual EU policy, but specifically due to professional health care, a considerable amount of online material has been produced and translated, not only in English, but also in less spoken languages, especially in countries with significant number of patients with diabetes. Diabetes type 1 is intriguing disease and after 90 years of the discovering insulin, it still imposes many challenges to both patient and physician. Techniques how to use insulin or monitor glucose level are considered the »basic ones«. Glucometers, insulin pumps and continuous glucose monitors are not easy to use and techniques of use should be supported by the well written educational online material.

There are many tools that could be used to estimate the health literacy. One of them is REALM-R (Rapid Estimates of Adult Literacy in Medicine), a brief screening instrument used to assess an adult patient's ability to read common medical terms<sup>2</sup>. If a patient's score is very low (less than a 6th grade reading level) by the REALM, surveys sometimes use other instruments that are administered orally to ensure that the content of the survey questions were understood by the patient. The information from such studies can be used to adjust medical materials and instructions for patients<sup>3</sup>. There are also studies exploring instruments that measure health literacy in diabetes<sup>4,5</sup>. Some educational tools for diabetes have been even developed through social marketing approach<sup>3</sup>.

A recent systematic literature review analyzed fifty-six studies and found that the main focus in health literacy tools are reading and writing skills, not other important skills like verbal communication, health care system navigation, etc<sup>6</sup>. Information literacy<sup>7</sup>, i.e. health-information literacy is the condition *sine qua non* while using insulin pumps in type 1 treatment, in order to obtain, understand and apply the information. It is also necessary to improve communication between patients and health-care providers<sup>8</sup>. The specific type of user is patient with diabetes having great ability and interest in the use of technology<sup>9</sup>, but also being very responsible towards himself/herself. This type of patient needs every day self-management in the curative process, as well as constant education on basic diabetes knowledge, diet, medications, self-care behavior and disease-related distress<sup>10</sup>. Online services are trying to give support for self-education<sup>11</sup> (information on sugar quantities, medications, meal planning, importance of activities, etc.). Those services can be desktop-based or accessible on mobile application, all aiming to improve health care. When the general health literacy is poor in native language, the one in foreign language is mandatory worse. The question is how to perform adequate, meaningful, correct but understandable translation without losing the information. As manuals are intended to be written for everyday users, it is necessary to be written in clear sentences, using adequate terminology, logic order of examples and corresponding pictures, indicating clearly how to use the object, expected outcomes, possible inconveniences, etc. These demands are to be fulfilled in the source language,

but also in translated texts used by users on various markets. Such manuals are often accessible on Internet, so adequate translation is important for patients, doctors and pump manufacturers. Unfortunately, there are not many tools that evaluate the quality of translated text (instructions for patients, manuals, tutorials, etc.). Another dilemma is presented through language differences<sup>12</sup> interfering with linguistic barriers in diabetes self-management practices. The number of portals and web sites on less widely spoken languages has significantly grown up, and the use of medical technology asks for online accessible information through various usages of Internet resources on various languages<sup>13</sup>.

A paper from 2009 found that immigrants in Ireland had poorer glycaemic control compared with a matched population of native Irish patients. The phenomenon was associated with low health literacy<sup>14</sup>. It has been presented that number of environmental, social, personal factors, education level, as well as literacy and language barriers, influence diabetes self-management<sup>15</sup>.

While using Internet, all patients who are non-native English speakers in most web pages are »virtual immigrants«. They are certainly facing low health literacy. Similar situation happened while using written manuals and other educational tools written in English. The importance of appropriate use of key concepts and terms using diabetes terminology translated into non-widely spoken language was discussed in pilot project with Native American diabetes patients with limited English proficiency<sup>16</sup>. Need for adequate understandable information due to decrease in health literacy was elaborated in a study that evaluated the stability of health literacy in adults with diabetes over time<sup>17</sup>.

The prevalence of diabetes is increasing worldwide, estimating that 285 million people had diabetes in 2010 and by 2030 it will probably augment to 900 million<sup>18</sup>. Croatia as small country has a similar situation as two previous examples: it is a small nation of only 4.3 million inhabitants, with estimated 316000 diabetic patients in Croatia, with national prevalence in adults of 9.2%<sup>18</sup>, i.e. 6.1% in 18-65 ageing group<sup>19</sup> with a prevalence of obesity. According to<sup>20</sup>, diabetes mellitus is one of the main risk factors for development of coronary heart disease, being one of the significant health problems today.

Information and document retrieval through use of adequate »browsing phrases« and self-education through use of good quality online resources written in native language, may influence life-style habits of patients with diabetes problem. Use of proper terminology can improve search results and help to find valuable source of information important in everyday care.

In the paper, the extraction the most frequent phrases are given and evaluation process of corpus-based terminology extraction methods are given and suggested as a mean for information access for users with specific needs.

## Research

### Data set and methods

The research on evaluation of automatic terminology extraction is conducted on two types of texts:

- Official guide to insulin pump therapy (a guide originally written in English and translated by manufacturer into Croatian).
- Document set consisting of different medical texts collected from the web, i.e. online web pages in Croatian giving information on diabetes, its chronic complications, self-care, glucose monitoring, insulin pump therapy etc.

The research is divided into three interrelated parts aiming to detect the role of terminology in online resources:

- i) comparison of professional and popular terminology use in English and Croatian manuals and in Croatian online texts
- ii) detection of key terminology in manuals and online texts by comparing results of automatic terminology extraction method with three types of reference lists and evaluation by recall, precision and f-measure
- iii) comparison of two types of approaches of terminology extraction (statistical and hybrid) performed on English manual, and evaluation by recall, precision and f-measure.

Automatically extracted terminology candidates are compared with three types of reference lists:

- Reference list RL1 – consisting of 11 frequently used terms in diabetes created by doctor specialist involved in diabetes education
- Reference list RL2 – consisting of 15 professional terms in diabetes created according to the official list of key terms found on URL: <http://www.nlm.nih.gov/mesh/MBrowser.html>.
- Reference list RL3 – consisting of 9 popular terms in diabetes, created as intersection of 15 lists made by non-medical persons

There are five common terms between the lists RL1-created by diabetologist and RL3-created by non-medical person. The rest of the list RL1 is more specific containing professional terminology, and the rest of the list RL3 is more general and practically oriented containing more popular terminology. The list RL2 contains only highly professional terminology consisting of specialized controlled vocabulary.

### Research A

The first part of the research is related to the use of professional and popular terminology in English and Croatian manuals and Croatian online texts for patients with diabetes. Professional and popular terminology used in online diabetes resources, often considered as synonyms, can be used in information retrieval, in document indexing, in education or in cross-language information access. Use of adequate terminology can play the key role

in the success of information access when looking for information on self-care in curative process, behavior, use of insulin pumps, diet, medications, etc. While the professional base, such as e.g. PubMed contains citations, papers and abstracts using highly professional terminology, online pages, instructions for use and manuals contain more popular terminology, frequently used by broader public, sometimes translated or written in national (Croatian) language.

The difference between the use of professional and popular terminology has been analysed in English and Croatian manuals and in the collection of Croatian web documents. The analysed professional terminology included the following: diabetes (Cro. dijabetes), blood glucose (Cro. glukoza u krvi), retinopathy (Cro. retinopatija), nephropathy (Cro. nefropatija), neuropathy (Cro. neuropatija), cardiovascular (Cro. kardiovaskularni), hyperglycemia (Cro. hiperglikemija) and hypoglycemia (Cro. hipoglikemija).

The analyzed popular terminology included the following: sugar disorders (Cro. poremećaj šećera u krvi), blood sugar (Cro. šećer u krvi), eye damage (Cro. oštećenje oka), kidney damage (Cro. oštećenje bubrega), nerve damage (Cro. oštećenje živaca), blood system damage (Cro. oštećenje krvnih žila/ srčano-žilnog sustava), high level of sugar (Cro. povišena vrijednosti/razina šećera), low level of sugar (Cro. niska razina šećera).

Latin expressions included *diabetes mellitus*, *dijabetes mellitus* or *dijabetes melitus* and the laboratory term HbA1c – glycosylated hemoglobin A1c assay (Cro. glikozirani hemoglobin).

### Results

All three types of texts contain professional and popular terminology, although in different proportions. In English manual the ratio is 5.8 in favor of professional terminology, in the Croatian manual the ratio is 2.92 and in online web Croatian text, the ratio is 1.42 (Figure 1). The results indicate more frequent use of professional terminology in English manual than in Croatian (e.g. diabetes, blood glucose, hyperglycemia, hypoglycaemia) and more in Croatian manual than in Croatian online texts. There is almost equal distribution of popular and professional terminology on Croatian web pages, although some popular terminology is more used (e.g. eye/kidney/

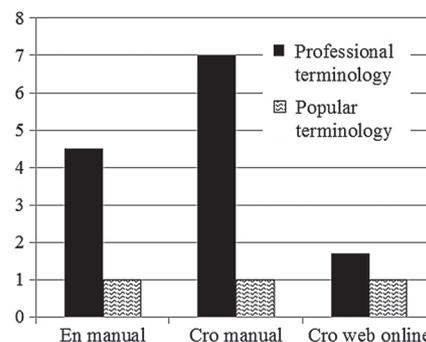


Fig. 1. Ratio of professional and popular terminology use in manuals and online texts.

nerve damage). The Latin expression *diabetes mellitus/diabetes melitus* is not at all used in English manual, in Croatian once, and on the web 21 times more, suggesting various levels of terminology used in online text ranging from popular, professional, strictly professional in the form of controlled vocabularies, up to Latin expressions and formulas. The term *HbA1c* is equally represented in all types of texts.

### Research B

The purpose of the second part of the research is detection and evaluation of automatically extracted terminology in English/Croatian manuals and Croatian online texts. Automatically extracted list is compared with three types of reference lists (RL1, RL2, RL3) and evaluated by measures of recall, precision and f-measure. Automatic extraction is made by the statistical extraction tool offering various options, such as extraction from monolingual or bilingual documents or from translation memories. When extracting from translation memory the user can define parameters such as minimal and maximal term length (in this case min=2, max=10), minimal translation frequency, and maximal number of translations. In this research, terminology candidates consist of multi word units of two or more words, since one word lists would be too long and not adequate for this analysis. In the next step, a stop-list is used, containing list of functional words (prepositions, articles, conjunctions, etc.) in order to refine the suggested list of terminology candidates.

### Evaluation

In order to perform evaluation, the automatically created lists are compared with three types of reference lists (RL1, RL2 and RL3) in order to detect the quality of automatically extracted terminology and user's expectations. Evaluation is performed using measures of recall, precision and f-score.

Recall is defined as the proportion between valid computer extracted terminology and expert (reference) extracted terminology. The perfect recall score of 100% indicates that all valid terms are automatically extracted, but does not say anything about the fact how many irrel-

evant terms are also extracted. As the user decides on relevant terminology according to his information needs, the measure of recall is more relevant comparing to other two measures.

Precision is defined as the proportion between valid computer-extracted terminology and all computer-extracted terminology candidates. The perfect precision score of 100% indicates that every extracted term is relevant but it does not indicate whether all relevant items are extracted.

F-measure (or F-score) allows adjusting the relationship between recall and precision. The F-measure is the weighted harmonic mean between precision and recall.

### Results

When comparing English and Croatian manuals, the results are better for the Croatian manual especially for the recall, for the list RL3 created by non-medical person, indicating more frequent use of general practically-oriented terminology. The English manual receives slightly lower scores in comparison with the list RL3. The list RL1 created by diabetologist is equally represented in English and Croatian manuals, indicating proportional use of professional and popular terminology in both languages. The list RL1 containing terminology suggested by professional healthcare person, receives the best scores for online Croatian texts, especially for the measure of recall which reflects the user's needs. When analyzing results for the English manual, scores are very close for lists RL1 and RL3, created by diabetologist and non-medical person, suggesting more frequent use of professional terminology than in Croatian manual.

The list RL2 obtained the worst result, indicating that highly professional terminology almost does not appear in manuals. Measure of recall, evaluated by comparison with specific reference list, gains better scores than recall for lists RL1 and RL3 and is more relevant for the user. Precision is lower indicating overlapping with all computer-extracted terminology candidates, and very similar for lists RL1 and RL3. The reference list RL2 has very low recall and precision for manuals but gains higher scores for online Croatian texts, indicating presence of highly professional terminology besides popular terminology presented on Croatian web pages.

### Research C

The third part of the research is related to comparison and evaluation of two types of terminology extraction approaches – statistical and hybrid, performed on English manual. Terminology candidates are evaluated through measures of recall, precision and f-score. Besides statistical terminology-extraction method presented in the research B, the hybrid method based on statistical and language approaches, is used. It offers possibility to choose single word unit which is not within the scope of this research, and extraction of multiword units, but only for the English language. For both tools the limitation to 100 terms was set up.

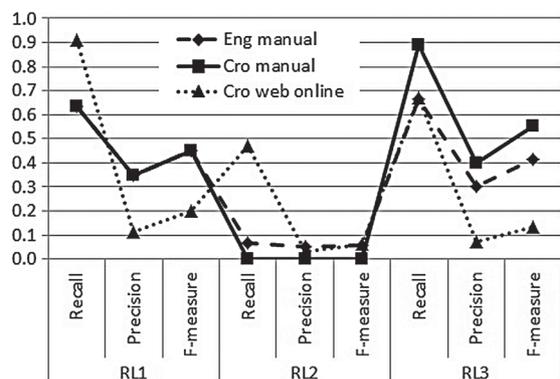


Fig. 2. Results of extracted terminology from three types of English and Croatian texts, compared with three reference sets and evaluated by metrics of recall, precision and f-measure.

## Results

Figure 3 presents evaluation results when comparing statistical and hybrid approaches. Results are generally better for recall indicating the user's perception, showing that more relevant terms are retrieved in the list RL3 when extracting by the hybrid approach. For the list RL1 the hybrid approach gives slightly better results than the statistical approach. Precision scores are higher in all cases for the statistical approach and slightly better for the list RL1 indicating that more consistent terminology appears more frequently in text.

When comparing 10% of the most frequent terminology candidates in English manual suggested by two types of approaches, better results are obtained for the hybrid approach, indicating the relevance of various language constructions (Figure 4.), but also due to smaller data set consisting of one manual.

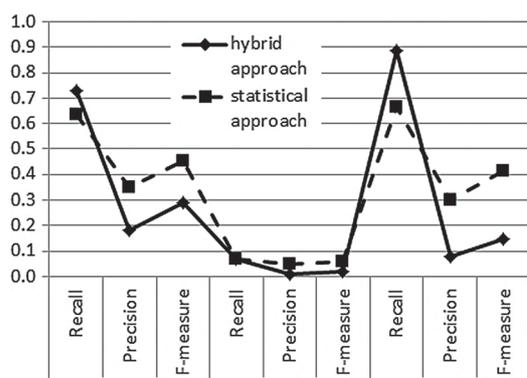


Fig. 3. Results of extracted terminology for hybrid and statistical approaches on English manual, compared with three reference sets and evaluated by metrics of recall, precision and f-measure.

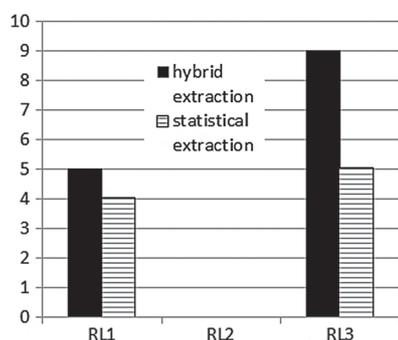


Fig. 4. Retrieved 10% of the most frequent terminology by two approaches.

## Conclusions

The paper presents results of automatic terminology extraction in manuals and online texts which could help in building terminology lists or list of »browsing phrases« useful information retrieval, in document indexing, in machine learning, in education, or extended to cross-language information access. In the research, three types of texts containing professional and popular terminology have been analyzed: English and Croatian manuals and

Croatian online web resources, relevant for patients with diabetes.

Detection of key terminology used shows that professional and popular terminology is used in all types of texts, but in different proportions. Manuals about insulin pump therapy, written for the specific type of users with diabetes type 1 problem, contain more professional terminology, supposing already good knowledge on self-management process. The professional terminology is more used in English manual than in Croatian.

Web sites written on Croatian, intended for broader public looking for adequate information on basic diabetes knowledge, diet, medications, self-care behavior, contain popular and professional terminology in proportion 1:1.42. Although there is more frequent use of professional terminology on Croatian web pages (e.g. diabetes, blood glucose, hyperglycemia, hypoglycaemia), some popular terminology is more used (e.g. eye/kidney/nerve damage).

Evaluation of automatic statistically-based terminology extraction shows higher scores for the measure of recall which reflects the user's needs. Terminology contained in the list created by regular non-medical persons is more represented in Croatian manual especially for the measure of recall, while the English manual receives slightly lower scores. When analyzing results for the English manual, scores are very close for lists RL1 and RL3, created by diabetologist and non-medical person, suggesting more frequent use of professional terminology than in Croatian manual. Terminology contained in the reference list created by diabetologist is more represented in online texts, especially for the measure of recall which reflects the user's needs. The list created by diabetologist is equally represented in English and Croatian manuals, indicating proportional use of professional and popular terminology in both languages. The list of highly specialized terminology contained in official MeSH controlled vocabulary is almost not included in manuals, but gains higher scores in online texts where professional and popular terminology is more equally represented. Measure of recall gains better scores than precision for lists created by medical and non-medical persons and is more relevant for the user.

When comparing two approaches for terminology extraction performed on English text, results are generally better for recall, indicating that more relevant terms are retrieved by the list created by regular user when extracting by the hybrid approach, but also due to smaller data set consisting of one manual. The precision is higher in all cases for the statistical approach indicating that smaller number of terminology units appears more frequently. When comparing 10% of the most frequent terms, better results are obtained for the hybrid approach, indicating the importance of various language constructions.

Language barriers and highly professional terminology can represent a barrier to needed information access and understanding of diabetic problems as a number of patients with diabetic problems is augmenting, there is need for creation of adequate resources on diabetic self-care, self-education and self-management written on national language. Facing the need for health literacy, ter-

minology extraction methods could represent valuable support to improve information retrieval when using professional or popular terminology. Corpus-based terminology extraction could serve as a mean for matching of collaboration between online resources created by health-care providers, producers of technological devices and patients using adequate terminology in information retrieval process.

## REFERENCES

1. LEHMANN ED, *Diabetes Technol Ther*, 6 (2004) 755. — 2. BASS PF 3rd, WILSON JF, GRIFFITH CH, *J Gen Intern Med*, 18 (2003) 1036. — 3. DAVIS TC, CROUCH MA, LONG SW, JACKSON RH, BATES P, GEORGE RB, BAIRNSFATHER LE, *Fam Med*, 23 (1991) 433. — 4. KIRK JK, GRZYWACZ JG, ARCURY TA, IP E, NGUYEN HT, BELL RA, SALDANA S, QUANDT SA, *J Gen Intern Med*, 27 (2012) 534. DOI: 10.1007/s11606-011-1927-y. — 5. SCHILLINGER D, GRUMBACH K, PIETTE J, WANG F, OSMOND D, DAHER C, PALACIOS J, SULLIVAN GD, BINDMAN AB, *JAMA*, 288 (2002) 475. DOI: 10.1001/jama.288.4.475. — 6. AL SAYAH F, WILLIAMS B, JOHNSON JA, *Health Educ Behav*, 40 (2013) 42. — 7. HAWTHORNE K, *Health Educ Res*, 16 (2001) 373. — 8. WARD P, *Nurse Stand*, 26 (2011) 35. — 9. WILSON E, WARDLE EV, CHANDEL P, WALFORD S, *Diab Medicine*, 10 (1993) 177. — 10. KAVIN M, ANELTINGCO RM, MAUGER DT, GABBAY RA, *Diabetes Ther*, 1 (2010) 93. — 11. LYLES CR, HARRIS LT, LE T, FLOWERS J, TUFANO J, BRITT D, HOATH J, HIRSCH IB, GOLDBERG HI, RALSTON J, *Diabetes Technol*

## Acknowledgements

The research is made within the project »Digital document classification supported by machine translation methods« (eClas) funded by University of Zagreb – grant 1.

*Ther*, 13 (2011) 563. DOI: 10.1089/dia.2010.0200. — 12. HSU WC, CHEUNG S, ONG E, WONG K, LIN S, LEON K, WEINGER K, KING GL, *Diabetes Care*, 29 (2006) 415. — 13. LEHMANN E, *Diabetes Technol Ther*, 5 (2003) 75. — 14. THABIT H, SHAH S, NASH M, BREMA I, NOLAN JJ, MARTIN G, *Q J Med*, 102 (2009) 713. — 15. COLLOFF E, *Diabetes Technol Ther*, 2 (2000) 91. — 16. MCCABE M, GOHDES D, MORGAN F, EAKIN J, SCHMITT C, *The Diabetes Educ*, 32 (2006) 714. DOI: 10.1177/0145721706292101. — 17. MORRIS NS, MACLEAN CD, LITTEMBERG B, *Diabetes Educ*, 39 (2013) 638. DOI: 10.1177/0145721713496871. — 18. ŠEKERIJA M, POLJIČANIN T, METELKO Ž, *Coll Antropol*, 36 Suppl 1 (2012) 171. — 19. POLJIČANIN T, ŠEKERIJA M, BORAS J, VULETIĆ BS, METELKO Ž, *Coll Antropol*, 36 Suppl 1 (2012) 41. — 20. VRAŽIĆ H, LUCIČANIĆ T, ŠIKIĆ J, RAJČAN ŠPOLJARIĆ I, POLIĆ S, LJUBIČIĆ Đ, MATIĆ K, BOŽIN T, ŠUBJAK I, BERGOVEC M, *Coll Antropol*, 36 Suppl 1 (2012) 223.

S. Seljan

*University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences, Ivana Lučića 3, 10000 Zagreb, Croatia*  
e-mail: sanja.seljan@ffzg.hr

## PRETRAŽIVANJE INFORMACIJA KROZ EKSTRAKCIJU TERMINOLOGIJE U ONLINE IZVORIMA ZA BOLESNIKE S DIJABETESOM

### SAŽETAK

Primjena terminologije kao sredstva za pretraživanjem informacija ili indeksiranjem dokumenata, predstavlja važnu ulogu u zdravstvenoj pismenosti. Specifična vrsta korisnika, tj. pacijenti s problemima dijabetesa trebaju pristup online izvorima (na stranom i/ili materinjem jeziku) u potrazi za informacijama koje se odnose na samu naravi bolesti, samokontrolu i liječenje, na pravilan način ishrane i provođenje fizičke aktivnosti, ali i specifičnije informacije kao npr. samostalno korištenje inzulinske pumpe. Automatska ekstrakcija korpusne terminologije iz online tekstova, priručnika ili profesionalnih radova, može pomoći u izgradnji terminološke liste ili liste »frazu za pretraživanje« koje se mogu koristiti u pretraživanju informacija ili u indeksiranju dokumenata. Specifične terminološke liste predstavlja posrednički korak između slobodnog pretraživanja i kontroliranog vokabulara, između korisničkih zahtjeva i postojećih online izvora na materinjem i stranom jeziku. Istraživanje kojem je cilj identifikacija e evaluacija automatski ekstrahirane terminologije, provedeno je na engleskom i hrvatskom priručniku i na hrvatskim online tekstovima te podijeljeno u tri međusobno povezane faze: i) usporedba korištene profesionalne i popularne terminologije ii) evaluacija automatske ekstrakcije primjenom statističkog modela na engleskim i hrvatskim tekstovima iii) usporedba i evaluacija automatski ekstrahirane terminologije engleskog priručnika primjenom statističkog i hibridnog modela. Ekstrahirani termini uspoređeni su s tri vrste referentnih listi: s listom izrađenom od strane profesionalne medicinske osobe dijabetologa, listom visoko standardiziranog MeSH kontroliranog vokabulara i listom sastavljenom od strane nemedicinski obrazovanog osoblja, nastalom kao presjek 15 različitih listi. Rezultati pokazuju omjer primjene popularne i profesionalne terminologije u online izvorima, ishode evaluacije automatski ekstrahirane terminologije u engleskim i hrvatskim tekstovima i usporedbu statističkih i hibridnih metoda u ekstrakciji terminologije teksta na engleskom jeziku. Evaluacija automatski i poluautomatski ekstrahirane terminologije provedena je primjenom odziva, preciznosti i f-mjere.