The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area

V B Barakhnin¹, A N Duisenbayeva²,³, O Yu Kozhemyakina¹, Y N Yergaliyev²,³ and R I Muhamedyev²,³,⁴

¹ Institute of Computational Technologies SB RAS, Russia
² Institute of Information and Computational Technologies MES RK
³ Satbayev University, Kazakhstan
⁴ ISMA University, Latvia

bar@ict.nsc.ru, a.duisenbayeva@gmail.com, olgakozhemyakina@mail.ru, erlan21@mail.ru, ravil.muhamedyev@gmail.com

Abstract. This work reviews the bibliometric indicators of a rapidly developing field of research as automatic text processing (Natural language processing). The differential indicators of speed and acceleration were used to evaluate the development dynamics of NLP domains. The evaluation was based on the data from the Science direct bibliometric database. The evaluation of the Russian research segment was conducted according to e-library data. The calculations for the following subdomains of NLP were performed: Grammar Checking, Information Extraction, Text Categorization, Dialog Systems, Speech Recognition, Machine Translation, Information Retrieval, Question Answering, Opinion Mining, Smart advisors and others. The areas with high growth rates (Grammar Checking, Information Extraction, Machine Translation and Question Answering) and the areas that have lost the previously existing dynamics of growth of the publication activity (Information Retrieval, Opinion Mining, Text Categorization) have been identified.

Keywords. Natural language processing, Machine Learning, Bibliometric Indicators, Scientometrics, Deep Learning, Neural Networks, Information Extraction, Text Categorization, Dialog Systems, Speech Recognition, Machine Translation, Information Retrieval, Question Answering, Opinion Mining, Smart advisors, D1, D2, semantic network.

1. Introduction
The field of research, combined terms of Natural language processing (NLP) or automatic text processing, causes a large and steadily growing interest of researchers since the 60’s of last century. Recently, the scientific research and the general increased level of calculations have led to a number of breakthrough results in the NLP among which it is possible to note the achievements in the area of machine translation, automatic summarization, information retrieval, question answering, sentiment analysis and information extraction.
However, it would be interesting to consider how the interest of researchers changes over time, and it is possible to identify those research areas that attract special attention.

There is an increase in the number of publications in almost all areas of Natural language processing. Therefore a simple statement about this increase, for example, of publications, will be insufficiently. The differential indicators [1] were introduced at this work. They are necessary to evaluate the speed and acceleration of changes in bibliometric indicators. In turn, the speed and the acceleration may point an increase or decrease in the interest of researchers to individual subdomains of NLP. In this paper, the number of publications and the number of citations are considered as the analyzed indicators.

The paper consists of following parts.
- The second part describes the applied analysis method and the corresponding bibliometric indicators.
- The third part describes the methodology of the study.
- The fourth part shows the obtained results.
- The results are briefly discussed in conclusion.

2. Bibliometric methods of the evaluation of publication activity
The idea of the evaluation of publication activity in the field of science by applying the citation index (SCI) was proposed by Eugene Garfield [2]. Later, the bibliometric indicators (number of publications, citation index, number of co-authors, etc.) were widely used to evaluate the productivity of scientists [3, 4], research areas [5], universities [6], the formation of a policy for management of scientific research [7] and forecasting [8, 9]. The last task is more interesting in the situation of the intensive technological changes. Because of that, the changes of the Hirsch index in time are considered in [10, 11] and the concept of the dynamic Hirsch index is proposed.

However, despite the wide usage of this indicator, its main purpose is the evaluation of personal contribution of a scientist. We have considered the indicators of citation and numbers of publications more reasonable as indicators in order to evaluate the development of research areas as a whole.

The evaluation of these indicators in given period of time, at least, of one year, does not seem to be a hard task. But the numerical increase indicators are important for the comparative analysis. With this aim, the differential metrics for evaluation of the increase of research areas were introduced in [11, 1].

The dynamics of increase of the subdomains was estimated on the base of the number of publications and citations for the field of the information and communication technologies.

For the evaluation of this dynamics there can be applied Compound Annual Growth Rate – CAGR, which is calculated according to the following equation:

\[
\text{CAGR} = \left( \frac{\text{Ending Value}}{\text{Beginning Value}} \right)^{\frac{1}{T}} - 1
\]

where \(T\) is the number of periods.

Nevertheless, as outlined below, this rate demonstrates a rather smooth change, hindering the identification in the patterns of dynamics changes, and also calculated individually for publications and for citations. In this regard, it seems to be useful to introduce certain integrated indicator for more expressed evaluation of dynamics change. For this purpose, we will determine the following differential indicators of a scientific area [1], defined by the given key search term

\[
D_1 = f_1(n, \frac{dn}{dt}, \frac{dc}{dt})
\]

\[
D_2 = f_2(n, \frac{d^2n}{dt^2}, \frac{d^2c}{dt^2})
\]
In other words, the indicator $Dl_i$ is a function depending on the publication count $n_i$, speed of change of the publication count $\frac{dn_i}{dt}$ (time derivative) and the citation count $\frac{dc_i}{dt}$ in the given area of research.

The indicator $D2_i$ is a function depending on the publication count $n_i$, acceleration of change of the publication count $\frac{d^2n_i}{dt^2}$ and the citation count $\frac{d^2c_i}{dt^2}$ in the given area of research.

The functions, $f_1$ and $f_2$, aggregate the contribution of citations and publications by some means or other. In a particular case, the aggregation can be performed using the weighted summation. Then, for an individual scientometric database $j$, the indicators of the scientific research perspectivity in a field $i$ at the moment of time $t_k$, can be determined with following equations

$$Dl_i(t_k) = \alpha \times n_i(t_k) + \beta \times \frac{dn_i(t_k)}{dt} + \gamma \times \frac{dc_i(t_k)}{dt}$$

$$D2_i(t_k) = \alpha \times n_i(t_k) + \beta \times \frac{d^2n_i(t_k)}{dt^2} + \gamma \times \frac{d^2c_i(t_k)}{dt^2}$$

where $\alpha,\beta,\gamma,\alpha',\beta',\gamma'$ – are certain empirical coefficients which the" weight " of the contribution of the number of publications, the speed and acceleration of the change in the number of publications $n_i$, and the speed and acceleration of the change in the number of citations $c_i$, respectively.

For more evident estimation of the dynamics of change in publishing activity, especially in the fields with large amount of publications, the indicators $\alpha$ and $\alpha'$ can be made equal to 0. In this case,

$$Dl_i(t_k) = \beta \times \frac{dn_i(t_k)}{dt} + \gamma \times \frac{dc_i(t_k)}{dt}$$

$$D2_i(t_k) = \beta \times \frac{d^2n_i(t_k)}{dt^2} + \gamma \times \frac{d^2c_i(t_k)}{dt^2}$$

The equations 1, 5 and 6 were used for further calculations.

In addition to the dynamics of development of research areas, an important aspect of analysis is the connection of domains to each other. The evaluation of this connection allows us to identify the semantically close areas of research and estimate their impact on each other. The idea about the evaluation of the connection between the domains of the information and communication technologies proposed for the first time in [12], was developed in [13], and a semantic network of the domains of the information and communication technologies is also developed in [14].

3. Research methodology
As part of this work, we have defined the following directions of the analysis. First of all, we consider the NLP area from the point of view of the solved problems, which, based on [15–21], include the following (group «tasks NLP»): grammar checking, extraction of information, text categorization, dialogue systems, speech recognition, machine translation, information search, Question Answer, Question Answer, opinion, development and analysis of moods, smart advisors, automatic abstracting.

On the other hand, the NLP area is characterized by a rapid increase of technologies and methods that contribute to the solution of the above-mentioned problems. As the methods we classified the follows (the group "scientific methods of NLP": Machine Learning, Neural networks, Deep Learning,
Fuzzy logic, First order logic, Knowledge representation, Evolutionary computation & Genetic programming, Rule based system, Unsupervised learning, Clustering, Supervised learning, Statistical methods, Bayesian networks, Semantic networks, Keyword Spotting, Lexical affinity, Ontology, Information fusion, Taxonomy.

Considering that NLP uses the methods of machine learning (machine learning – ML) and is often a provider of complex tasks that are preferably solved by ML methods, we have additionally investigated the areas of publication activity located at the junction of NLP and ML (group «NLP&ML»).

In this paper we present the data of bibliometric analysis of the group of research areas «NLP tasks».

For providing this analysis we used Science Direct bibliometric database, which contains about 2500 scientific journals and 26,000 e-books [22] and which is one of the largest.

For studying the dynamics of publications in the natural language processing field (OEJA) in Russian, we used the leading Russian scientific electronic library eLIBRARY.ru. In the middle of 2018 the database eLIBRARY.ru contains more than 30.7 million articles, also include the publications of the last decades of the XX century. According to our requests, some publications of the mentioned period were found, but their number was very small, so the statistics in the tables are given since 2005. We emphasize that although in the database eLIBRARY.ru there are many English language publications from foreign editions, we limited to queries only in Russian language, since the study of English language publications was provided for the database Science Direct separately.

For each bibliometric database we generated the relevant search queries, including the mentioned above terms in combination with NLP, NLP and ML, etc. The results of the queries were the annually amount of publications and the citations since 2005. The data of 2018 is not used due to its incompleteness.

On the base of the obtained results for each search query, we calculated the indicators CAGR, D1, D2 were performed. The constants $\beta^*, \gamma^*$, $\beta^*, \gamma^*$ are assumed equal to be 0.95.

4. Results

The tables 1 and 2 show the annual number of publications and citations, respectively, concerning the main tasks being solved in the field of NLP according to Science Direct.

**Table 1.** Change dynamics of publication count in «NLP tasks» by ScienceDirect data.

| Search terms                        | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| NLP & Grammar Checking             | 2    | 3    | 1    | 2    | 1    | 0    | 0    | 2    | 0    | 2    | 2    | 2    | 1    |
| NLP & Information Extraction       | 41   | 79   | 50   | 55   | 80   | 68   | 77   | 104  | 120  | 128  | 135  | 153  | 168  |
| NLP & Text Categorization          | 8    | 14   | 10   | 20   | 27   | 16   | 27   | 28   | 28   | 47   | 45   | 49   | 56   |
| NLP & Dialog Systems               | 22   | 31   | 11   | 11   | 17   | 18   | 13   | 11   | 11   | 10   | 14   | 24   | 13   |
| NLP & Speech Recognition           | 50   | 71   | 37   | 39   | 49   | 43   | 33   | 54   | 58   | 72   | 95   | 112  | 138  |
| NLP & Machine Translation          | 17   | 65   | 18   | 28   | 35   | 28   | 36   | 53   | 65   | 78   | 81   | 117  | 129  |
| NLP & Information Retrieval        | 74   | 141  | 110  | 111  | 142  | 105  | 154  | 168  | 211  | 247  | 302  | 339  | 314  |
| NLP & Question Answering           | 20   | 45   | 46   | 36   | 39   | 35   | 45   | 53   | 75   | 66   | 77   | 86   | 114  |
| NLP & Opinion Mining               | 0    | 0    | 0    | 0    | 3    | 5    | 17   | 21   | 30   | 61   | 68   | 97   | 94   |
| NLP & Smart advisors               | 16   | 18   | 19   | 30   | 28   | 34   | 26   | 32   | 30   | 37   | 60   | 57   | 91   |
| NLP & Automatic summarization      | 3    | 15   | 10   | 2    | 12   | 5    | 12   | 16   | 15   | 14   | 23   | 17   | 20   |
| NLP & Sentiment analysis           | 0    | 0    | 0    | 1    | 11   | 9    | 17   | 43   | 46   | 107  | 130  | 182  | 234  |
Table 2. Change dynamics of citation count in «NLP tasks» by ScienceDirect data.

| Search terms                                      | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|--------------------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| NLP & Grammar Checking                           | 1    | 2    | 1    | 2    | 0    | 0    | 0    | 1    | 0    | 1    | 2    | 1    | 1    |
| NLP & Information Extraction                    | 21   | 42   | 31   | 34   | 46   | 48   | 47   | 64   | 66   | 69   | 61   | 83   | 108  |
| NLP & Text Categorization                        | 5    | 10   | 5    | 15   | 18   | 11   | 17   | 19   | 18   | 33   | 31   | 30   | 34   |
| NLP & Dialog Systems                             | 19   | 19   | 9    | 5    | 9    | 10   | 6    | 7    | 7    | 4    | 8    | 14   | 7    |
| NLP & Speech Recognition                        | 38   | 52   | 32   | 31   | 38   | 35   | 23   | 43   | 42   | 52   | 74   | 85   | 116  |
| NLP & Machine Translation                       | 14   | 40   | 15   | 20   | 17   | 22   | 20   | 34   | 46   | 48   | 48   | 61   | 66   |
| NLP & Information Retrieval                      | 49   | 90   | 73   | 76   | 92   | 67   | 99   | 118  | 138  | 167  | 205  | 236  | 228  |
| NLP & Question Answering                         | 15   | 32   | 26   | 22   | 20   | 18   | 22   | 37   | 44   | 41   | 45   | 49   | 58   |
| NLP & Opinion Mining                             | 0    | 0    | 0    | 0    | 2    | 3    | 9    | 9    | 12   | 37   | 36   | 57   | 50   |
| NLP & Smart advisors                             | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    |
| NLP & Automatic summarization                    | 2    | 7    | 10   | 2    | 5    | 4    | 8    | 11   | 9    | 8    | 13   | 10   | 12   |
| NLP & Sentiment analysis                         | 0    | 0    | 0    | 10   | 3    | 9    | 25   | 18   | 49   | 67   | 96   | 102  |

Currently, the number of publications and citations shows a steady increase in almost all considered domains. The usage of differential indicators D1 and D2 allows us to imagine more clearly the dynamics of this increase. The usage of expressions 1, 4, 5 is illustrated in figure 1.

The tables 3 and 4 show similar results for eLIBRARY.ru data. Let’s note that due to the peculiarities of search queries, the data presented in the table probably reflect only some of the available publications in the Russian-language segment.

Table 3. Change dynamics of publication count in NLP field by e.library.ru data.

| Search terms                                      | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|--------------------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Natural language processing                      | 9    | 6    | 13   | 24   | 19   | 19   | 35   | 44   | 58   | 76   | 97   | 118  | 132  |
| NLP & Machine learning                           | 0    | 0    | 0    | 1    | 1    | 0    | 3    | 0    | 2    | 4    | 2    | 4    | 9    |
| NLP & Neural networks                            | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 2    | 5    |
| NLP & Deep learning                              | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    |
| NLP & Grammar checking                           | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Information Extraction                    | 3    | 0    | 0    | 0    | 2    | 0    | 2    | 2    | 4    | 6    | 9    | 5    | 9    |
| NLP & Text Categorization                        | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    |
| NLP & Dialog Systems                             | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 1    | 0    | 1    | 3    | 2    | 1    |
| NLP & Speech Recognition                        | 0    | 0    | 0    | 0    | 2    | 3    | 0    | 2    | 0    | 3    | 4    | 6    | 6    |
| NLP & Machine Translation                        | 0    | 0    | 1    | 0    | 0    | 0    | 1    | 2    | 2    | 9    | 2    | 6    | 2    |
| NLP & Information Retrieval                      | 1    | 0    | 0    | 1    | 0    | 0    | 3    | 0    | 0    | 1    | 2    | 7    | 3    |
| NLP & Question Answering                         | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 2    | 1    |
| NLP & Opinion Mining                             | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Smart advisors                              | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| Grammar checking                                  | 2    | 1    | 2    | 4    | 4    | 13   | 9    | 6    | 9    | 14   | 13   | 11   |
| Information Extraction                           | 67   | 66   | 61   | 91   | 91   | 98   | 153  | 160  | 153  | 192  | 249  | 325  | 298  |
| Text Categorization                              | 0    | 1    | 0    | 0    | 2    | 3    | 1    | 4    | 10   | 6    | 7    | 8    |      |
| Machine Translation                              | 10   | 23   | 20   | 18   | 31   | 39   | 61   | 54   | 57   | 78   | 83   | 112  | 144  |
| Information Retrieval                            | 152  | 190  | 223  | 280  | 326  | 326  | 432  | 487  | 540  | 797  | 1005 | 1209 | 1215 |
| Opinion Mining                                   | 0    | 0    | 1    | 1    | 2    | 0    | 3    | 8    | 6    | 22   | 38   | 44   | 47   |
Figure 1. CAGR, D1 (speed), and D2 (acceleration) for some domains of the «NLP tasks» group.
The usage of expressions 1, 4, 5 for eLibrary.ru data.

5. Conclusion
The increasing of D2 indicates an increase the rate of increase of the number of publications and citations in the research field. The negative dynamics, in turn, shows a slowdown in publications compared to previous periods. The data show an interesting picture of change for most domains (figure 1), when the period of initial increase is followed by a decrease, and then there is a re-acceleration (question answering, speech recognition). It can be assumed that this dynamics characterizes the intensity of development in the field of researches, and, besides, the acceptance of the new concept by researchers and its application in researches. Only one of the presented fields (grammar checking) is characterized by constant increase of citations and publications (indicator D2 has only positive values for the whole period under review).

The domains (machine translation, information extraction), as mentioned above, are characterized by repeated acceleration (indicator D2, which had a negative value for several periods, again becomes positive), whereas for domains information retrieval, opinion mining and text categorization period of the fall is the same.

In general, the analysis of the NLP research shows significant increase dynamics of publication activity in the domains of Grammar Checking, Information Extraction, Machine Translation and Question Answering.

At the same time, the domains Information Retrieval, Opinion Mining and Text Categorization show a decrease in growth.

| Search terms                                   | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|------------------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Natural language processing                    | 2    | 6    | 5    | 6    | 13   | 27   | 369  | 101  | 113  | 164  | 258  | 274  | 301  |
| NLP & Machine learning                        | 0    | 0    | 0    | 0    | 0    | 0    | 29   | 0    | 0    | 2    | 0    | 0    | 0    |
| NLP & Neural networks                         | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Deep learning                           | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Grammar checking                        | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Information Extraction                  | 4    | 0    | 0    | 32   | 0    | 0    | 0    | 3    | 1    | 7    | 5    | 2    | 0    |
| NLP & Text Categorization                     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Dialog Systems                          | 0    | 0    | 0    | 0    | 0    | 0    | 187  | 0    | 8    | 0    | 7    | 7    | 0    |
| NLP & Speech Recognition                      | 0    | 0    | 0    | 0    | 0    | 0    | 2    | 0    | 2    | 0    | 1    | 2    | 0    |
| NLP & Machine Translation                     | 0    | 0    | 0    | 0    | 0    | 0    | 2    | 23   | 2    | 2    | 2    | 2    | 0    |
| NLP & Information Retrieval                   | 0    | 1    | 0    | 0    | 0    | 0    | 1    | 4    | 0    | 0    | 3    | 2    | 1    |
| NLP & Question Answering                      | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Opinion Mining                          | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| NLP & Smart advisors                          | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| Grammar checking                              | 21   | 15   | 20   | 27   | 32   | 27   | 32   | 36   | 45   | 46   | 41   | 62   | 46   |
| Information Extraction                        | 72   | 84   | 94   | 146  | 169  | 139  | 170  | 233  | 326  | 443  | 674  | 679  | 772  |
| Text Categorization                           | 0    | 0    | 1    | 1    | 0    | 2    | 3    | 0    | 2    | 17   | 17   | 24   | 20   |
| Machine Translation                           | 36   | 54   | 93   | 97   | 112  | 118  | 171  | 184  | 284  | 294  | 427  | 445  | 495  |
| Information Retrieval                         | 118  | 140  | 256  | 334  | 426  | 529  | 713  | 934  | 1246 | 1786 | 2645 | 3427 | 3922 |
| Opinion Mining                                | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 2    | 6    | 24   | 45   | 45   | 28   |
Figure 2. CAGR, D1 (speed), and D2 (acceleration) for eLibrary.ru data.

Acknowledgments
The work was funded by grant No. BR05236839 of the Ministry of Education and Science of the Republic of Kazakhstan.

References
[1] Muhamedyev R I et al 2001 Annals of Library and Information Studies (ALIS) 65.1 62–69
[2] Garfield E 1955 Science 122 122.3159 108–11
[3] Gauthier E 1998 Canada: Science and Technology Redesign Project (Statistics Canada)
[4] Van Raan, A FJ 2003 Technology Assessment-Theory and Practice 1.12 20–29
[5] Mokhnacheva Julia V and Mitroshin Ivan A 2014 *Information resources of Russia* 6 17–23
[6] Giovanni A, D’Angelo C A and Pugini F 2008 *Scientometrics* 76.2 225
[7] Koenraad D and Glänzel W 2004 *Scientometrics* 59.2 253–76
[8] Daim T U, Rueda G R and Martin H T 2005 *Technology management: a unifying discipline for melting the boundaries* 112–22
[9] Daim T U 2006 *Technological Forecasting and Social Change* 73.8 981–1012
[10] Leo E 2007 *Journal of the Association for Information Science and Technology* 58.3 452–54
[11] Ronald R and Fred Y Ye 2008 *Journal of the Association for Information Science and Technology* 59.11 1853–55
[12] Adilmanova A, Aligulyev R M and Mukhamediev R I 2016 *Cloud of Science* 3.3 366–79
[13] Mukhamediev R I, Aligulyev R M and Muhamedijeva J 2017 *International Conference on Digital Transformation and Global Society* (St Petersburg, Russia) 130–35
[14] Muhamedyev R I, Kalimoldaev M N, Uskenbayeva R K 2014 *Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia* (St Petersburg, Russia) 178–86
[15] Cilibrasi R L and Vitányi P M B 2007 *IEEE Transaction on Knowledge and Data Engineering* 19.3 370–83
[16] Sun S, Luo C and Chen J 2017 *Information Fusion* 36 10–25
[17] Goldberg Y 2016 *Journal of Artificial Intelligence Research* 57 345–420
[18] Dale R 2015 *Natural Language Engineering* 21.4 653–59
[19] Tablan V et al 2013 *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Sofia, Bulgaria) 19–24
[20] Conroy N J, Rubin V L and Chen Y 2015 *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (American Society for Information Science, St. Louis, Missouri, USA) 82
[21] Wikipedia.org *Natural language Processing*
[22] Wikipedia.org *Elsevier ScienceDirect*