UDC 811.111’33
DOI https://doi.org/10.17721/folia.philologica/2022/3/5

Liudmyla VLASIUK
PhD student at the Department of Theory, Practice and Translation of the English Language, Lecturer of The Department of English for Engineering 1, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Peremohy Avenue, 37, Kyiv, Ukraine, 03056
ORCID: 0000-0003-1020-0076
LyudmylaVlasyuk@ukr.net

Olga DEMYDENKO
PhD in Education, Associate Professor at the Department of Theory, Practice and Translation of the English Language, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Peremohy Avenue, 37, Kyiv, Ukraine, 03056
ORCID: 0000-0002-0643-5510
olga.demydenko80@gmail.com

To cite this article: Vlasiuk, L., Demydenko, O. (2022). Linguistic indexation as way of media text clusterization. *Folia Philologica*, 3, 37–41, doi: https://doi.org/10.17721/folia.philologica/2022/3/5

LINGUISTIC INDEXATION AS WAY OF MEDIA TEXT CLUSTERIZATION

Linguistic indexation is a highly complex phenomenon due to collecting, sorting and storing data aimed at providing high-speed, high-quality and accurate search of information. That is why it has quickly turned into one of the core problems for researchers, especially when it comes to examining it in the media text, which complicates this process for scholars engaged in linguistic studies. Profound research of the linguistic text’s indexation is predetrmined by the necessity to structure of the information system. The stated phenomenon has gained high relevance in linguistics over the recent years.This can be explained by the fact that improvement of data structure is of an utmost importance, because it contains information about diverse texts and documents. The increase in the data structure level of effectiveness is caused by the speed of search for relevant documents. It can only be achieved by qualitative linguistic indexation of the texts. This article deals with the research of media text from the standpoint of applied linguistics. Aimed at identifying main features of linguistic indexation phenomenon and analyzing it as a method of text clusterization, the article reveals this notion taking into account diverse approaches to its study, examines types of linguistic indexation, as well as methods and typology. The article is also concerned with defining the main principles of identification of the effective keywords for the document and determining the extent to which the chosen keywords are relevant to both the text itself and research results. This, in turn, will allow to identify how keywords contribute to carrying out effective search for information and any objects contained in the information resources. Thus, they influence on the potential readers and serve as an instrument of text’s popularization.

Key words: *linguistic indexation, media text, keywords, classification and clustering of the text.*

Людмила ВЛАСЮК
аспірантка кафедри теорії, практики та перекладу англійської мови, викладач кафедри англійської мови технічного спрямування № 1, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», проспект Перемоги, 37, м. Київ, Україна, 03056
ORCID: 0000-0003-1020-0076
LyudmylaVlasyuk@ukr.net

Ольга ДЕМИДЕНКО
кандидат педагогічних наук, доцент кафедри теорії, практики та перекладу англійської мови, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», проспект Перемоги, 37, м. Київ, Україна, 03056
ORCID: 0000-0002-0643-5510
olga.demydenko80@gmail.com

Бібліографічний опис статті: Власюк, Л., Демиденко, О. (2022). Лінгвістична індексація як спосіб кластеризації медіатексту. *Folia Philologica*, 3, 37–41, doi: https://doi.org/10.17721/folia.philologica/2022/3/5

ЛІНГВІСТИЧНА ІНДЕКСАЦІЯ ЯК СПОСІБ КЛАСТЕРИЗАЦІЇ МЕДІАТЕКСТУ

Лінгвістична індексація є надзвичайно складним явищем з огляду на збір, сортування та зберігання даних із метою забезпечення швидкого, якісного та точного пошуку інформації. Дослідження лінгвістичної індексації є однією з актуальних мовознавчих проблем, особливо у випадку її вивчення в медіатексті. Детальне дослідження питання лінгвістичної індексації зумовлене необхідністю структуризації інформаційного простору текстів. Це пояснюється тим, що покращення структури даних є вкрай важливим, адже в ній зберігається інформація про різноманітні тексти та документи. Підвищення рівня ефективності структури даних зумовлюється збільшенням швидкості пошуку релевантних документів, що може бути досягнено насамперед шляхом якісної лінгвістичної індексації текстів. Статтю присвячено дослідженню медіатексту в контексті прикладної лінгвістики. Мета статті полягає у визначенні особливостей явища лінгвістичної індексації як способу кластеризації тексту. Зокрема, досліджуються види й методи лінгвістичної індексації, а також її типологія. У статті також проаналізовано основні принципи вияву ключових слів і визначено ступінь їхньої релевантності. Ключові слова є релевантними як до конкретного тексту, так і до результатів дослідження. Окрему увагу зосереджено на їхній ролі при проведенні ефективного пошуку інформації або будь-яких об'єктів, які наявні в інформаційних ресурсах і, таким чином, впливають на потенційних читачів та слугують популяризації тексту.

Ключові слова: *лінгвістична індексація, медіатекст, ключові слова, класифікація та кластеризація тексту.*

Problem's topicality. Over the recent years has arisen a necessity of media texts' information ecosystem structuring and clusterization as modern times are characterized by the huge amounts of ever-increasing, diverse information, which is accessible and represents an interest to the broad spectrum of social layers. Moreover, Internet technologies, programme and technical tools are also accessible to the majority of people and, thus, they enabled us to look for information at any time in any place. Out of this reason exists the need of improvement of text clusterization with the aim to provide more exact search results that are highly relevant to user's query. This can be achieved exclusively through the specific phenomenon named "linguistic indexation". However, the achievement of the aforementioned task is not as simple as it sounds. The problem lies in the fact that applied linguistics is considered to be relatively new and difficult approach, especially in the case of linguistic indexation, which is viewed as complex and intricate process itself. What is more, due to being poorly investigated, linguistic indexation is dismissed as a highly acute issue in the linguistic circles, what gives an explanation to the necessity of more profound investigation of linguistic indexation phenomenon and its features that allow it to serve as a method of text clusterization.

Analysis of recent research and publications. The phenomenon of linguistic indexation as a method of media text clusterization has been an object of investigation of both national and foreign researchers. As a theoretical base for the article served multiple works of prominent researchers who were focused on the study of media text in terms of applied linguistics, particularly, by using linguistic indexation phenomenon. Hence, M. Steinbach works enabled the profound research of the notion of clusterization as well as its types; J. May's ground-breaking research allowed to analyze linguistic indexation's nature and its typology; S. Ticher's outline of linguistic indexation's features has provided the basis for building schematic model of this process.

The aim and tasks of the article. The article is **aimed** at defining the key features of linguistic indexation and studying this phenomenon as a particular method of media text clusterization. The achievement of the aforementioned aim requires the fulfillment of the following **tasks**:

1) to define the notions of 'clusterization' and 'linguistic indexation';

2) to analyze the types of text clusterization;

3) to study media text and its key characteristics;

4) to identify the types of linguistic indexation;

5) to analyze the process of media text clusterization through linguistic indexation.

The outline of the main research material. Present-day information ecosystem is characterized by the abundance of diverse texts from multiple sources. This, in turn, complicates the search for information needed and overloads the user with the excess information, which is not always relevant to the queries. Therefore, the efforts of researchers engaged in the linguistic studies are targeted at normalizing the data structure by clustering the texts.

Text clusterization has been the subject of multiple investigations for years, but, nevertheless, it remains highly acute problem for modern linguistics. Clusterization consists in pointing out of the semantically connected texts in the multidimentional information space as well as in defining the centers of clusters, which represent thematic rubrics. Generally two stages of clusterization are distinguished: during the first stage occurs clusters formation that is based on the limited selection of texts; second stage is characterized by the conclusive division of the entire corpus of texts.

According to M. Steinbach ***clusterization*** is defined as a process of multiple texts' division into groups based on the similarity of their content (Steinbach, 2011: 34–37). As a rule, the following **types** of text clusterization are distinguished (Corazza, 2004: 21–32):

1) *stemming*: allows to unify lexical units (for instance, by removing endings), what enhances the accuracy of clusterization process;

2) *stopword removal*: the most frequent words, that are used in the texts of all types, does not pose an interest to the clusterization due to their being dependent on the text's style rather than on the subject of this text. Consequently, the removal of these words can significantly boost clusterization accuracy;

3) *latent semantic indexing*: fusion of synonyms. It also allows to increase the accuracy of search results, though the problem of possible homonymy remains unsolved;

4) *linguistic indexation* (in other words, highlighting the key words): assumes the pointing out of main words in the text. This approach is applied to the clusterization of texts of large volumes. However, as a practical matter diverse algorithms of this method are used, which presuppose distinction of collocation. This, in turn, makes the process of clusterization even more effective.

Linguistic indexation is considered to be one of the most effective types of text clusterization, especially in the media texts' ecosystem. This can be explained by the fact that nowadays media is firmly entrenched in human society. Media text is a powerful tool of influencing the human mind. Being dynamic in nature, the language of the media responds most quickly to all the changes in public consciousness. It is in the language of the media that it is easy to notice new trends in approaches to language learning.

Media text is defined as the dynamic complex unit of the highest order with the help of which the information transmission occurs. The notion "media text" serves as a hyperonym to the entire range of terms: journalistic text, PR-text, publicistic text, newspaper text, radio text, TV-text and so on (Mytrofanov, 2020: 35-37).

To its main *categories* belong media character, mass character, integrity or polycoding, openness and intertextuality:

1) ***media character***: media text is determined by the communication channel. Each method of mass information can be characterized by a set

of media features, which exert influence on the text properties;

2) ***mass character***: mass communication specification lies in the fact that it is socially-oriented communication, where the figures of both author and recipient are subjected to changes;

3) ***integrity*** (polycoding): modern media texts according to the form of their creation and form of their representation are multimodal that is they are verbal, visual, audiovisual and other components which integrate in one sematic space;

4) ***openness***: media text cannot be characterized by the sematic completion since it is the structure, which is open to multiple interpretations: mass media text is a phrasal unity of endless hypertexts with the great variety of references and quotations;

5) ***intertextuality***: any media text is a part, fragment of the information continuum, which serves as a communicative background and provides sematic ambiguity of the message interpretation by the reader (Shapovalova, 2003; Ticher, 2009).

When the complex structure of media text and its features as well as the fact that media text represents a key source of information are taken into account, it becomes clear why it is needed to simplify the search for information in media text through the use of linguistic indexation.

J. May defines "**linguistic indexation**" as the process of pointing out the language peculiarities, which directly refer to the context where the utterance unfolds (May, 2009: 42–50). In other words, that's the search for words, collocations or phrases, which are connected with different meanings (referents). The most typical kinds of linguistic indexation are the following:

1) *direct indexation*: is the direct semantic interrelation between the language and indexed notions;

2) *indirect indexation*: which is based on distinction of key concepts instead of key words.

Linguistic indexation is closely connected with two notions: deixis and search index. *Deixis* is generally viewed as a word or collocation that possesses the referential meaning of indexation due to their ability of being connected with both language and context and reflecting the structure of the language. *Search index* is a data structure, which contains information about texts and enables the process of collecting, parsing and storing of data aimed at the provision of quick and acute information search (Cutting, 2002: 60–73). Generally speaking search index is a key notion, which exerts impact on the linguistic indexation phenomenon. Five major types of search indexes are distinguished (Giorgi, 2012: 34–48):

1) *suffix tree* – is based on storing the suffixes of words;

2) *inverted index* – data structure, where for each word of texts' collection in the corresponding list are outlined all the texts containing this word;

3) *citation index* – storing of citations or hyperlinks;

4) *n-gram* – fixed collocations;

5) *document-term matrix* – matrix, which describes the frequency of words used in the collection of texts.

Thus, the process of linguistic indexation of media text consists of five key steps:

1. *Step 1*: collection forming module and pointing out meta information out of texts.

2. *Step 2*: module of storing index structures of texts.

3. *Step 3*: module of forming and providing user with the list of texts ranked according to their relevance.

4. *Step 4*: module of forming markers containing meta information.

5. *Step 5*: module of text's linguistic analysis and forming of corresponding markers.

This model of linguistic indexation process not only enables text clusterization, but also determines the extent to which the chosen keywords are relevant to both the text itself and the research results. This, in its turn, helps to identify how keywords contribute to carrying out efficacious search of information and any objects contained in the information resources.

Conclusions and scope for further research. The results received show that linguistic indexation enables the process of text clusterization and, consequently, structures the extensive media texts ecosystem by pointing out the key words in the texts. Linguistic indexation is largely dependent on two notions 'deixis' and 'search index' as they influence the type of linguistic indexation and specify its process.

For modern linguistics it is of the utmost importance to study the phenomenon of linguistic indexation in media text. This is mainly due to the fact that media text encompasses an extensive variety of texts, which need to be structured and clusterized to regulate information ecosystem.

Moreover, despite being highly acute issue, media text is poorly investigated from the standpoint of applied linguistics and, therefore, gives rise to the necessity of its further investigation, particularly through the phenomenon of linguistic indexation.

BIBLIOGRAPHY:

1. Giorgi A. *About the Speaker: Towards a Syntax of Indexicality*. Oxford : Oxford University Press, 2012. 214 p.
2. Corazza E. *Reflecting the Mind: Indexicality and Quasi-Indexicality.* Oxford : Oxford University Press, 2004. 350 p.
3. Cutting J. *Pragmatics and Discourse*. London : Routledge, 2002. 187 p.
4. May J. *Concise Encyclopedia of Pragmatics.* Amsterdam : Amsterdam Press, 2009. 400 p.
5. Steinbach M. *A Comparison of Document Clustering Techniques.* Minnesota : Minnesota Publishing, 2011. 180 p.
6. Ticher S & Mejer M. *Methods for analyzing text and discourse*. Oxford : Oxford University Press, 2009. 288 p.
7. Митрофанов О.О. Особливості сучасного медіатексту. *Медіатекст в сучасному комунікативному дискурсі : матеріали науково-практичної конференції (м. Миколаїв).* Миколаїв. 2020. С. 35–38.
8. Шаповалова Г.В. *Інноваційні процеси в сучасному медіатексті (функціонально-лінгвістичні аспекти)* : дис. … канд. пед. наук. Київ, 2003. 197 с.

REFERENCES:

1. Corazza, E. (2004). *Reflecting the Mind: Indexicality and Quasi-Indexicality*. Oxford University Press.
2. Cutting, J. (2002). *Pragmatics and Discourse.* Routledge.
3. Giorgi, A. (2012). *About the Speaker: Towards a Syntax of Indexicality*. Oxford University Press.
4. May, J. (2009). *Concise Encyclopedia of Pragmatics*. Amsterdam Press.
5. Mytrofanov, O.O. (2020). *Osoblyvosti suchasnoho mediatekstu. Mediatekst u suchasnomu komunikatyvnomu dyskursi [Mediatext in modern communicative discourse].* (s. 35–38). Pryvatnyi zaklad vyshchoi osvity «Mizhnarodnyi klasychnyi universytet imeni Pylypa Orlyka». Retrieved from: https://mku.edu.ua/wp-content/uploads/2021/04/Zbirnyk-tez-2020-2021 Mediatekst-ostannij.pdf. [in Ukrainian].
6. Shapovalova, H.V. (2003). *Innovatsiini protsesy v suchasnomu mediateksti (funktsionalno-linhvistychni aspekty) [Innovative processes in the modern mediatext (functional and linguistic aspects)].* [Dys. kand. ped. nauk, Kyivskyi natsionalnyi universytet imeni Tarasa Shevchenka]. Repozytarii Natsionalnoi biblioteky Ukrainy imeni V.I. Vernadskoho. http://irbis-nbuv.gov.ua/ASUA/0104024. [in Ukrainian].
7. Steinbach, M. (2011). *A Comparison of Document Clustering Techniques*. Minnesota Publishing.
8. Ticher, S., & Mejer, M. (2009). *Methods for analyzing text and discourse.* Oxford University Press.