

Jacek Petzel
University of Warsaw

SEARCHING LEGAL INFORMATION USING A NATURAL LANGUAGE

1. At present, one searches for legal information mainly through legal information computer retrieval systems. The first such systems were created in the United States in the 1950s, and this marked the beginning of the development of the applicatory side of legal informatics. Due to a growing information crisis in law¹, traditional legal information search systems, described as “manual” and based on hierarchical classifications or subject-headings ideas, were considered to be insufficient to serve as tools to acquire the information needed by professional lawyers. This led to the creation of the first ideas as to how computers could be used to gather full and up-to-date information about law. At present, there are as many as nearly 2,000 such systems in the world. They are mostly based on one of the two basic concepts²: they are either index systems³ or full text systems⁴. The primary difference between the two is that in the index systems a document is identified by several (no more than a dozen or so) terms that are assigned by the system-creator to specific documents as terms that adequately represent their contents⁵. In the full text systems, on the other hand, all terms that appear in the text are the text identifiers, except for words that are useless for the purpose of indexing⁶. This means

¹ At the time, the number of cases in the United States was 1.5 million; every year it grew by 20,000. Later on, European countries also suffered from an information crisis, so did Poland. About an information crisis in law see J. Petzel, *Informatyka prawnicza. Zagadnienia teorii i praktyki*, Warszawa 1999, pp. 13–14. See also S. Simitis, *Informationskriege des Rechts und Datenverarbeitung*, Karlsruhe 1970.

² At present, full-text solutions are mainly used, in particular in commercial databases, although many systems that can be described as mixed systems use index-based and sometimes also subject headings search tools.

³ These systems were the first to be created and based on the ideas of M. Taube.

⁴ The first full text system was the Pittsburgh Project System created at the end of 1950s by J. Harty; it covered statutes relating to healthcare. For details see C. Tapper, *Computers and Law*, London 1973, p. 79.

⁵ It is a standard that a system creator assigns indexes to documents, and this happens in most systems. Some systems, however, use in part for assigning indexes automatic methods.

⁶ Examples of words that are useless for indexing include “in”, “from”, “about” “to” “or” “not”; they carry no semantic meaning and their purpose is to organize an utterance. About index

that the number of the so-called “access points” to documents is much larger in the case of full text systems, and this helps one to search the database more effectively. In both systems, the users search for information by creating instructions in a given search language; the index systems and the full text systems accept the use of Boolean operators such as AND, XOR (exclusive OR), OR, NOT, which are presented in the various search languages in a variety of ways, while the full text systems also allow for the use of the positional operators which refer to the position of words in a text or to the distance between words in it⁷. Each system has its own search language, and the user must learn it. Initially, these search languages were complicated, and the users had to master them in order to be able to create a correct search instruction⁸. The users did not like that and considered those systems to be unfriendly and inadequate for their needs. The search languages were gradually simplified, first by adding menus (but this always caused the search to be less efficient), then by creating new simpler search languages, but that did not change the negative user opinions about how the systems worked, either⁹. Therefore, it became clear that effort had to be made to create such search methods that would relieve the users from putting any additional effort to formulate search instructions other than to verbalize their needs. The fact that not only professional but also lay users began to use the systems inspired the efforts to create such solutions. Lay users had to be given simple tools to communicate with the system.

2. It was considered that that aim could be achieved if the search instructions were formulated in a natural language. This notion requires clarification, mainly because in all systems (except for those that use semantic codes)¹⁰, the user uses the terms of that language to verbalize his or her needs when they create a search

and full text systems as well as about words useless for indexing see op. cit. J. Petzel, *Informatyka prawnicza...*, pp. 79–131.

⁷ Boolean operators, for example conjunction can be presented in the searching language as i.e. &, AND, +. Positional operators by using i.e.?, +, etc.

⁸ One example of such a complicated instruction is the instruction in the MISTRAL language created by BULL that was used to retrieve information in CELEX (Communitatis Europa Lex), an European law retrieval system (now replaced by Eur-Lex). In that system, in order to find information about EU subsidies for butter stored in private warehouses and used for the manufacture of ice-cream, one had to formulate the following instruction: “?oj aid? &pa butter &pa private &pa (ice? cream?) :title, :text”. See J. Petzel, *Komputerowe systemy wyszukiwania informacji prawnej wspólnot europejskich*, Bazy Europejskie, Biuletyn, Centrum Europejskie Uniwersytetu Warszawskiego. Ośrodek Informacji i Dokumentacji Rady Europy, Warszawa 1993, No. 9.

⁹ It seems that such opinions are not justified and result from some unwillingness on the part of lawyers to learn any search language. A EU-sponsored research project carried out by Butler and Cox has shown that instructions formulated by lawyers in the LEXIS system in many cases consisted of only one word (!), despite the fact that the search language of the system allows for the use of operators.

¹⁰ There are only a few such systems. The only such system that retrieves information about law is the WRU (Western Reserve University) system created in the US.

instruction¹¹. However, in the case of “systems based on the use of a natural language”, the user does not need to employ any of the operators expressed in the search language, i.e. neither the Boolean operators, the positional operators, nor any other expressions required by the search language used by the system¹². The user can formulate a search instruction in the manner in which he or she asks a question. For instance, they can introduce the following instruction into the system: *Find all documents pertaining to the termination of rights based on statutory liability for physical defects*. The method used in these systems to search for information is based on solutions other than those applied in searches that use the Boolean operators (i.e. those in which a document is considered to be relevant only if it meets all of the criteria stated in the instruction, which is called “exact match”). The search in natural language systems is based on the idea that every single word from the instruction is used as separate searching criterion. Because of that the system will find any document that includes at least one of such words. To continue with the example above, the search will retrieve all documents that include the terms *termination*, *rights*, *statutory liability*, *defects*, *physical*, and their all grammatical forms. As a result, a document that would not have been identified by a system that uses the Boolean operators can appear as a search result in a retrieval system based on a natural language.

In order to build a system of this type, one should employ solutions other than those employed in standard systems. In particular, one must design an effective method for the transformation of a search instruction used by a user of the system into an instruction that is ultimately carried out by the system; also one must use other search options, i.e. such that are not based on the use of operators. That latter task requires that a computer performs an operation, based on a method that employs a statistical analysis of the vocabulary of texts of documents, aimed at, first, identifying the weight of particular words or expressions in the entire documentary base, and, next, assigning appropriate weight to each document, so that such weight signifies a certain degree of relevance for a user’s needs. Thus, it is possible to present the user with search results in an orderly format, namely the list of documents ordered according to their supposed relevance. This distinguishes these systems from the classic Boolean-type search systems in which search results are usually presented as a chronological list of documents¹³.

¹¹ There is some simplification in saying that these are expressions of a natural language, as in order to formulate a search instruction in a legal search the users usually use the lexicon that is typical for the legal language which in turn is part of a natural language and its register. For details see J. Petzel, *Język prawny w świetle lingwistycznej teorii rejestru językowego*, “Studia Iuridica” 2006, Vol. XLV, pp. 153–163. On language registers see also K. Brodziak, *O lingwistycznym statusie języka prawnego*, “Studia Prawnicze” 2004, No. 1.

¹² In some systems the search instruction must define the database in which the search is to be carried out or to identify the lexicon to be used in the search.

¹³ In some classic systems the option to organize documents showing such results in a reverse chronological order is also available.

3.1. The first attempts to employ a natural language search method were made in the 1960s. In 1967, Cherie B. Weil built a system to assist in the search for bibliographic references. Later, several other similar standard systems were built, such as Personal Librarian¹⁴ and Dow Quest¹⁵. Also the Canadian QUIC/LAW system used the natural language search method¹⁶. At present, WESTLAW and LEXIS are the most popular systems using that method¹⁷. As early as in 1992, West Publishing Company implemented the WESTLAW IS NATURAL (WIN) technique designed by H. Turtle and J. Olson that allows one to conduct a search in a natural language as an alternative to the search technique used in the classic version of WESTLAW which is a full text system with the Boolean operators. Later Lexis Nexis designed a similar solution and introduced FREESTYLE SEARCHING. Solutions used in the WESTLAW WIN system will serve as an example in a more detailed presentation of natural language-based search methods that follows; references will be made to FREESTYLE SEARCHING whenever the two systems substantially differ.

In WIN WESTLAW, the system first identifies which words or expressions that appear in the instruction adequately describe the object of the search. This happens in a series of operations. The system first removes from the instruction (phrased in a natural language) expressions such as “how can I find”, “what is”, “find all documents concerning” that the users typically place at the beginning of the instructions they create. Then it checks whether any of the expressions that appear in the instruction is misspelt; to do this, the system compares them with a list of expressions that is installed in the system and includes misspelt and correctly spelled terms. At this stage, the system corrects mistakes made by the user. It also removes from the instructions the words that carry no information (such as “and”, “or”, “that”, “when”, “in order to”). It then identifies the words that are relevant for the search by confronting them with a legal dictionary being part of the system. The dictionary makes possible to, i.e., find out that two or more words create a phrase even if the user does not indicate in the instruction (and he or she can do so by using inverted commas) that the user searches for

¹⁴ That system was used by the Polish system LexPolonica.

¹⁵ For details see S. A. Weyer, *Questing for “DAO”: DowQuest and Intelligent Text Retrieval*, “Online” 1989, Vol. 13, No. 5. That system was used to find information in press articles published in 185 journals. At the time it was created, the database covered approximately 1 GB of text.

¹⁶ The system provided the user with 11 methods for ranking documents by assigning weights. For details see op. cit. J. Petzel, *Informatyka prawnicza...*, p. 123.

¹⁷ WESTLAW (created by West Publishing Company) and LEXIS (created by LexisNexis) dominate on the American legal databases market. There even appeared the name WEXIS to reflect the dominance of these two firms. Some American firms use also other databases such as LoisLaw, BlombergLaw or Pacer CourtLink, but usually only alongside WESTLAW and LEXIS. On computer legal information databases in the United States other than LEXIS and WESTLAW see L. K. Justiss, *A Survey of Electronic Alternatives to LexisNexis and Westlaw in Law Firms*, “Law Library Journal” 2011, Vol. 103, No. 1.

a phrase rather than an individual word. On the other hand in some cases when the user formulates a phrase in the search instruction, that phrase is split into individual words and the search is carried out based on each of such words, except for the so-called hard phrases¹⁸. Next, the system allows one to join other words into the instruction, first of all synonyms, using the WESTon-line thesaurus¹⁹. Words that are grammatical forms of the words already used are also added to the instruction. This is done by a stemming programme that derives other forms from the stem form. The system can add the plural form of nouns and irregular verb forms as part of the operation²⁰. The instruction created in such a manner is then presented to the user and the user can correct it. The operation of broadening the instruction cannot be performed automatically and is always performed under the supervision of the user.

Once the instruction is entered into the system, the system looks for documents that include at least one of the words stated in the instruction. Therefore it will cause the situation that some of the search results will be irrelevant. Because of that statistical analysis of the words in all documents from the database of the system is carried out which allows the assigning to each document a specific weight. This weight should reflect the degree of the relevance of each retrieved document.

In the WIN system the analysis is based on the following assumptions. First, the frequency with which each of the terms appears in the entire document base is checked. It is assumed that a greater weight should be assigned to those phrases that appear less frequently, which is correct, as such words are unique and, therefore, when they appear in a document, they identify the contents of the document as more relevant to the user's needs when the user uses that unique term in the search. Secondly, the weight assigned to each term depends on the frequency with which it appears in a document, considering also its proximity to other terms that are also part of the search instruction, and considering the length of the document. This allows one to assign a greater weight to shorter documents that include the specific terms than to documents that also include the same terms but are longer²¹. Thirdly, the weight of each document is calculated by adding the weight

¹⁸ For instance, *res ipsa loquitur* is a hard phrase that cannot be split.

¹⁹ Thesauri which are present in many retrieval systems are dictionaries that show mutual relations between lexical units, such as synonymity, hypo- and hipernimity relations and associations. For details see J. Petzel, *Tezaurusy systemów wyszukiwania informacji prawnej*, Materiały konferencji "Prawo i język", Warszawa 2009, pp. 99–111.

²⁰ The methods for generating grammatical forms applied in American systems such as WESTLAW and LEXIS are not perfect. Nowadays in many systems, including those available in Poland, such as LEX and Legalis, grammatical forms are added to the instruction through thesauri of grammatical forms, which is the best method to solve the problem of inflection forms in computer retrieval.

²¹ The fact that documents occurring in a document database vary in length makes a proper search more difficult. The frequency of occurrence of a particular word in a longer document

assigned to each of the terms, multiplied by the number of its appearances in the document. It should be added that WIN has a special feature, namely one can find a relevant document on the basis of analysis of the part of this document. This method includes a stage where parts of documents are identified, i.e. those that include several of the terms stated in the instruction, and 40 other words that are not the search terms but appear before and after such terms. Therefore, a document that has 20 sections and only one of the sections covers exactly the subject matter of the search is assigned a greater weight than the entire document would have if it were evaluated as a whole²². As a result, a list of search results is produced that is ordered by their assumed relevance to the user's needs. At the top of the list there are documents with the greatest weight possible, those that are considered to be most relevant. Documents with a smaller weight are placed down the list. As a standard, WIN shows a 100 search results, which seems not to be a good idea. The list of documents presented to the user is too long which means the user must put an extra effort to select those documents that satisfy the user's information needs from the list. The user can limit the number of search results to 20 but then the risk is that a relevant document may not be shown.

The primary format of presentation of the results of search in the WIN system is list of documents ordered according to the assumed relevancy; however, it can also be presented in the form of a reversed chronology list. This is noteworthy as that feature enables the user to assess relevancy of the documents that have been added to the database most recently. The users often consider such most up-to-date documents as the most relevant ones. Once the user has received and analysed the search results, he or she can formulate new corrected search instruction and mark the phrases that the user does not wish to appear in the documents the system will retrieve. It allows performing a new search and obtaining new search results.

Despite the fact that the theoretical foundations of the system, including the assumptions as to how weights are to be assigned to particular words that appear in documents, are largely correct, tests of the WIN system have shown that the search results are not satisfactory²³. Searching performed in the system does not allow the identification of all relevant documents. Even more so, in extreme cases, the search results may even include no relevant documents. The WIN system is designated to provide the user with an answer that is a list of documents, irre-

is greater. As a consequence, ranking documents on the basis of the statistical analysis will not always lead to satisfactory results.

²² For details on weighting documents see E. M. McKenzie, *Natural Language searching: How WIN Works in Westlaw*, "Legal Reference Services Quarterly" 2001, Vol. 18, No. 4, pp. 39–47.

²³ For details see S. E. Desert, *WESTLAW Is Natural v. Boolean Searching: A Performance Study*, "Law Library Journal" 1993, Vol. 84, issue 4. Critical opinion about functioning of WIN but also about other systems based on ranking algorithms is presented in: E. Schweighofer, *Legal Knowledge Representation. Automatic Text Analysis in Public International and European Law*, The Hague–London–Boston 1999, pp. 55–58.

spective of whether or not the list includes a document that adequately satisfied the user's needs. Also the ordering of the documents according to their assumed relevance leaves much to be desired in some cases, and it is often the case that the most relevant documents are down the list even if they should appear on the top²⁴.

3.2. Natural language-based retrieval is also a search option offered by LEXIS as part of the so-called FREESTYLE SEARCHING, a system more advanced when compared with the WIN system²⁵. Similarly to WIN, weights are assigned to individual documents, based on the weights of particular words. As in WIN, the weight of the terms is determined by their frequency in a document. The greater the frequency of a given term, the greater the likelihood that it better identifies the contents of a document than the terms that appear less frequently. Secondly, inverse document frequency is taken into account. The assumption is that the greater the number of documents are identified with the specific terms, the lower is discriminatory ability²⁶ of these terms to identify particular documents. To the terms that appear in numerous documents lower weights are assigned than to the terms that appear in a smaller number of them. The simplest way to weight documents is to multiply the number of terms in a document that are included in the search instruction by their weight in the entire document database. It is noted, however, that this simple method is not sufficient because longer documents will probably have greater weight than shorter ones, and methods that allow correcting the results due to the length of the document are also applied. However, the evaluation of the results obtained with such methods is not unambiguous.

In FREESTYLE SEARCHING the so-called terms vector, composed on particular words or phrases, is created for each document. The words are selected on the basis of the statistics of their appearance and with the assistance of a phrase-recognition programme. The search operation uses the so-called *relevance feedback* that allows the users to modify the search instruction. Once the user has seen the search result, he or she can broaden the search instruction. This can be done by adding new search terms to terms being present in the initial

²⁴ As S. E. Desert demonstrates, after she asked a search question in the WIN system, while she knew no relevant case existed for the subject matter of her request, 20 cases were retrieved, none of which was relevant. Other searches she describes, run by experienced users and librarians, have shown many errors in the ordering of the documents. See *op. cit.* as above.

²⁵ FREESTYLE SEARCHING was launched in 1993, a bit later than Westlaw Is Natural. There is competition between West Publishing Company and Lexis Nexis, and soon after one of the companies creates a certain search option the other offers its equivalent. For information on FREESTYLE SEARCHING see <http://www.lexisnexis.com/custserv/freestyle>.

²⁶ Where a certain word occurs in one document only, its discriminatory ability is the highest of all possible. If it occurs in each of the documents present in the database, it is the lowest of all possible. This is why terms that appear frequently in many documents in the database belong to the group of the so called non-informative words and are often removed from the computer search files such as i.e. inverted file. See G. Salton, *Experiments in Automatic Information Organization and Retrieval*, New York–St. Louis–San Francisco–Toronto–London–Sydney 1968.

search instructions, i.e., some of the other terms that appear in a document considered to be relevant. The tool to use is the *More-Like-This* option. On the next step, to broaden the list of search terms the statistical thesaurus is used²⁷. New words are added to the instruction on the basis of statistical analysis of the proximity of words in the documents in the database. It is assumed, however it is an idealistic assumption, that terms that frequently appear in texts next to the terms already used in the initial instruction are semantically close, and if they are added to the instruction documents with a close semantic meaning will be found²⁸. To achieve this result specially created association coefficients are used. Finally the instruction is supplemented by the user with the so-called mandatory terms, i.e. terms that the user believes should appear in the document in order for it to be relevant. It is important that the user has the possibility to create the final version of the instruction; because if the instruction is broadened automatically only by the computer which uses, for this purpose, information taken from statistical thesaurus, the instruction would probably include some terms that are in fact irrelevant for the user's information needs and, as a consequence, irrelevant documents would be retrieved. It is caused by a fact that high value of association coefficient does not always reflect the semantic proximity of the words.

The above review of search options based on natural language in the WIN and the FREESTYLE SEARCHING systems shows that the latter system is more advanced, mainly due to the way it uses the statistical thesaurus, although both systems are built on similar assumptions. However, it is hard to say if the method it follows leads to good results²⁹.

3.3. Five years ago, the natural language-based methods for the search of legal information through the WIN and the FREESTYLE SEARCHING systems described above were modified by West Publishing Company in its new product called WESTLAW NEXT³⁰, and more recently in LEXIS ADVANCED created by Lexis Nexis. The new concepts involving the use of natural language in the search for legal information depart from a traditional search method and implement methods that can be termed as *Google for lawyers*. As only general

²⁷ For details about a statistical thesaurus see J. Petzel, *Statistical systems for computerized information retrieval*, (in:) *Mélanges offerts à la mémoire de Jason Hadjadinis*, Piraeus University 1989. See also op. cit. J. Petzel, *Informatyka prawnicza...*, s. 157–170.

²⁸ This assumption is typical for all statistical thesauri. It should be added that, in the opinion of the creators of FREESTYLE SEARCHING, the use of a statistical thesaurus brings much better results than using any other type of thesauri.

²⁹ About FREESTYLE SEARCHING see C. Griffith, *FREESTYLE:LEXIS/NEXIS Goes Natural*, Information Today, 1994, , at <https://www.questia.com/magazine/1G1-14777062/freestyle-lexis-nexis-goes-natural> (visited January 17, 2017). No legal writings discuss the results of tests of the system.

³⁰ For details about WESTLAWNEXT see R. E. Wheeler, Jr., *Does WestlawNext Really Change Everything? The Implications of WestlawNext on Legal Research*, "Law Library Journal" 2011, Vol. 103, No. 3.

information on LEXIS ADVANCED is available³¹, the presentation of the functions of this type of systems will refer to WESTLAW NEXT.

Three features distinguish WESTLAW NEXT from the classic WESTLAW. It has a new search programme called WestlawSearch; at the beginning of the search operation the user need not, and in fact, cannot, determine in which part of the database he or she wants to search for information; finally, there is a change, as compared with the classic Westlaw, in how user charges are collected³².

The new WestlawSearch programme is the main distinguishing feature. The programme is based on a retrieval algorithm, the firm's commercial secret, that allows the ranking of the retrieved documents in a different manner than before. The algorithm analyses the searches made by the users so far. It is a 'learning' algorithm and it makes it possible to assign a greater weight to those documents that have been marked in previous searches as those well found which is demonstrated by the fact that the users use commands *PRINT*, *SAVE*, *FOLDER* or *VIEW*³³. Thus, the programme uses user knowledge; the programme creators believe this is an example of the crowdsourcing idea³⁴. It is to be noted, however, that the crowdsourcing in WestlawSearch is somewhat unique, as user knowledge is used without the users' consent. The system creators stress that the more the algorithm is used the better the search results it generates will be. However, an algorithm based on information about user conduct may miss a new case, statute or another legal document and fail to include it in the list of most relevant results, as its weight that relates to the number of searches involving it, will be low.

³¹ Information about LEXIS ADVANCED is available only in advertisements placed by LexisNexis on the Internet and is very general. It seems the system is based on the same approach as WESTLAWNEXT.

³² The charging structure, including a charge of 60 Dollars for introducing a search instruction and, which is a novelty, a charge for opening documents, ranging from 13 Dollars for opening one case to 25 Dollars for opening a document from the Federal State Use, Court Rules and Regulations databases, demonstrates that West Publishing Company's price policy is exceptionally unfriendly. It limits the number of users of the system. However, since law firms are the majority of WESTLAW clients and they always negotiate terms of use, maybe the real costs are lower. This, however, must have an impact on the clients of the law firms who ultimately bear the costs of information retrieval.

³³ Interestingly, student searches are not considered for this purpose, which, it seems, means that it is assumed that this class of users is not capable of correctly assessing the relevancy of documents.

³⁴ The idea of crowdsourcing has many manifestations, e.g., the drafting of Wiki (Wikipedia). The first instance of crowdsourcing happened in 1714. The British government offered rewards to every citizen who offered a method for determining a precise geographical length of the position of ships. Crowdsourcing is sometimes used as a tool for engaging citizens in the law-making process. One of the early examples was the drafting of traffic regulations in Finland in 2013. The idea that crowdsourcing is useful is not universally accepted, though, see D. Woods, *The Myth of Crowdsourcing*, September 29, 2009, at <http://www.forbes.com/2009/09/28/crowdsourcing-enterprise-innovation-technology-cio-network-jargonspy.html> (visited January 17, 2017).

Secondly, the user need not select a database for his or her search, which entirely departs from the traditional approach to searching information in WESTLAW. In WESTLAWNEXT it is sufficient to type the search instruction, and the system will carry out the search in all available databases. The classic WESTLAW is different. The classic WESTLAW is based on a structure that includes a large number of databases and at the beginning of the search process, the user must identify the database he or she wants to search. In the opinion of WESTLAWNEXT creators, the system better serves user needs where the user need not anticipate which database includes the requested document. WESTLAWNEXT displays the search results as an orderly list, irrespective of the type of the document. The user can select a category by using an option that groups documents into cases, legislation and many other categories. Limiting the search to a certain category limits further searches in that category, and the search results display only those documents that have already been identified.

The owner of WESTLAWNEXT (Thomson Reuters) describes the system as revolutionary, and one that entirely changes the approach to searching legal information, but the results of tests analysing this search option, as in the case of WIN, show that the system does not adequately fulfil user needs. The owner's enthusiasm for its product is not shared by legal search professionals who point to major deficiencies of the system. The main deficiencies are, first, the limited number of identified documents. Sometimes WESTLAWNEXT identifies half or less as many documents than the classic WESTLAW³⁵. Second, the list of search results appears in a particular order, which is to facilitate identification of the desired information, but the ranking systems are not perfect³⁶, which means that some rare precedents and doctrinal views, expressed in academic writings that are not mainstream, are not identified as relevant. Thirdly, WESTLAWNEXT does not allow one to employ a strategy often used in search operations, namely to begin with a wide search and then to narrow it down. Finally, no mechanisms are available to the user to expand the range of identified documents, and therefore to achieve completeness of the search³⁷. This is a major drawback of the system.

³⁵ A test search for texts concerned with constitutionality of abortion has shown that WESTLAWNEXT returned 317 cases, and that same search in the classic WESTLAW returned 804 cases, i.e. 2.5 times more.

³⁶ Tests made by R. E. Wheeler prove this. They have shown that the applicable system of ranking documents works as follows: the most relevant documents do not appear in the primary search results list or are placed far down the list. The statute he searched for, despite being the most relevant document, did not appear in the primary results list, and in the course of a more detailed search it was to be found on the 9th place. For details see R. E. Wheeler, Jr., *Does WestlawNext Really Change Everything?...*, pp. 366–367.

³⁷ In the theory of legal informatics the view that the search systems should ensure greater completeness of the searches, even at the cost of their precision, is well grounded. The first author to point to this specificity of legal system was J. M. Myers, *Progress in Documentation. Computer and Searching Law Texts in England and North America. A Review of the State of the Art*,

4. The natural language-based search methods described above have several advantages over the Boolean operator-type ones. The advantages are, first of all, that the search instructions can be easily formulated and that the search results are weighted. The tests show, however, that the quality of results leaves much to be desired, due to a number of factors.

First, such a system requires that the users verbalise their information needs in a proper manner, which is true for all legal information retrieval systems, but this aspect is particularly important in the case of retrieval systems based on natural language³⁸. Full text systems, and even more so index systems, require the user to put greater intellectual effort in formulating search instructions than the natural language-based systems. As a result, a user's needs are more precisely formulated in the former type of the systems which follows from the fact that a user can use logical operators as well as positional operators³⁹. Formal criteria can be used to a larger extent, too. Second, a proper functioning of such systems depends on correct design of the weighting methods. The weighting methods are based on algorithms that involve statistical data relating to texts. These algorithms are not perfect⁴⁰. It is sufficient to say that there is no way to proceed if the documents in the database are short or if there are major differences between them as to their length. Also, oftentimes the effect of automated algorithms is that the result is obtained without a user's control and that after a user types a search instruction, he or she cannot control the result of the search, which is a significant shortcoming. It is true that the WIN and the FRERESTYLE systems allow the user to modify the search instruction, but some other systems do not. It is so in the case of the systems using Personal Librarian. Also the fact that some systems use a statistical thesaurus to broaden the search instruction creates additional problems⁴¹. To use

"Journal of Documentation" 1973, Vol. 29, No. 2, p. 217. On completeness and precision see also J. Petzel, *Informatyka prawnicza...*, pp. 193–194.

³⁸ A closer analysis of the relationship between the information needs and their verbalisation shows that a proper verbalisation of needs is very complicated. On the subject see F. Studnicki, *Wprowadzenie do informatyki prawniczej*, Warszawa 1978, pp. 158–162. See also J. Petzel, *Komputerowe systemy wyszukiwania...*, pp. 192–197.

³⁹ Precision in formulating search instructions for the systems is enhanced by the fact that they allow the use of positional operators, as well as formal characteristics of documents, such as i.e. date of entry into force. Anyhow some natural language-based systems allow also search by dates. This option is available in WIN.

⁴⁰ Automatic algorithms do not always lead to correct results, as shown by R. Berring. He stresses the fact that in an automatic environment the search results are accepted by the users as those providing best possible information, which is not always correct. For details see R. C. Berring, *Chaos, Cyberspace and Tradition: Legal Information Transmogrified*, 12 Berkeley Tech. L.J. 189, 190 (1997). See also S. N. Mart, *The Relevance of Results Generated by Human Indexing and Computer Algorithms: A Study of West's Headnotes and Key Numbers and LexisNexis's Headnotes and Topics*, "Law Library Journal" 2010, Vol. 102, No. 2. The author says: **no one algorithm will give you all relevant results**.

⁴¹ On statistical thesauri see J. Petzel, *Informatyka prawnicza...*, pp. 157–170.

such thesauri one should choose a proper correlation coefficient so that counted value of proximity reflects the actual semantic relationships among the words in a text; also, an appropriate strategy for broadening the instruction must be followed. Should these be wrong, the instruction will be supplemented with words that have no relevance for the user information needs stated in the instruction⁴². It causes a big problem, especially in the situation where a statistical thesaurus is used fully automatically. It is to be noted that even where the user can correct an instruction created by the system based on that thesaurus (which is the case with WIN), there is no guarantee that the instruction will adequately express his or her information needs. Should there be a mistake in the structure of the thesaurus, some of the words that could adequately address these needs may not be included in the instruction that is displayed for user review. As a result, the search instruction that is processed by the system will not adequately reflect a user's needs, and, consequently, irrelevant documents will be retrieved.

The most important, anyhow is that the very idea that each of the words that are stated in the instruction is considered to be a search criterion is, it seems, not entirely correct. This method does not allow for a search based on criteria that precisely reflect user needs. As a result, a number of documents are retrieved that are, in fact, not relevant. Although systems of this type can recognize phrases and weight documents, which mitigates the problem, this is not enough to achieve satisfactory results.

In the view of specialists, the systems that use natural language cannot replace Boolean operator-based retrieval systems. In their view, such systems can only be supplementary. It is stressed that they cannot be a single tool to use when one searches databases, especially when one is a professional⁴³. In this situation the question arises as to whether it makes sense to devise such systems, as it is the professionals who are the user target group⁴⁴. No doubt the firms that market such systems are cognizant of this, but they are driven by a desire to increase their profits and assume, especially in relation to WESTLAWNEXT and the new addition on the market, LEXIS ADVANCED, that the general trend to use Google will prompt many non-professionals to search for legal informa-

⁴² In the LexPolonica system that used the standard Personal Librarian model, and whose one search option was based on natural language, this happened quite often, probably because of a faulty structure of a statistical thesaurus and incorrect methods for broadening instructions.

⁴³ Among others, S. E. Deserts support this opinion. In her view, WESTLAW WIN cannot be considered to be capable of replacing a search conducted with use of Boolean operators.

⁴⁴ It may be considered to be useful for professionals to use these systems only if, having carried out a search with the classic methods, they conclude that it would be useful to retrieve more relevant documents. It sometimes happens that a system based on a natural language retrieves a document that is relevant but has not been found with the classic methods.

tion with these tools⁴⁵. One should recognize, however, that at present such systems, in particular those that follow the *Google for lawyers* concept, are not able to provide information of adequate quality. This does not mean that this will not happen in the future. Search algorithms must be created, though, to achieve better results in ranking documents than those available at present. In the case of systems based on the Google idea, the area with legal information must be identified within the information resources. Also, if non-professionals were to use the systems, thesauri with general language terms, legal terms used in the statutes and terms used in legal writing must be created so that (which is very difficult to achieve), search instructions formulated in general language can be transformed into instructions using specialized legal terms. These are complicated problems which have to be solved.

SEARCHING LEGAL INFORMATION USING A NATURAL LANGUAGE

Summary

The article deals with a problem of constructing computer retrieval systems based on the use of natural language. Those kind of systems differ from the classical ones and relay on the idea that the user should have the possibility to introduce to the system search instructions in the same manner in which she or he asks a question. Searches in such systems don't use any retrieval languages. They don't use Boolean operators and every single word from the instruction is used as a separate search criterion. To the retrieved documents weights are assigned on the basis of statistical analysis of the text of the documents. This allows the ranking of documents and presenting the results of the search in an ordered manner. The first attempts of using such a system took place in 1960s, but in the field of legal information from the beginning of 1990s. The article presents the methods of searching used in WESTLAW IS NATURAL (WIN) created by West Publishing Company, FREESTYLE SEARCHING developed by Lexis Nexis and WESTLAW NEXT based on the idea *Google for lawyers*. The critical analysis of functioning and estimation of retrieval results shows that those kind of systems cannot replace traditional ones based on Boolean searching but can sometimes play a supplementary role.

⁴⁵ Searching legal information through Google cannot result in obtaining relevant information, in particular due to the fact that there are lexical differences between the general language and the language of the law.

BIBLIOGRAPHY

- Berring R. C., *Chaos, Cyberspace and Tradition: Legal Information Transmogrified*, 12 Berkeley Tech. L.J. 189, 190 (1997)
- Brodziak K., *O lingwistycznym statusie języka prawnego*, "Studia Prawnicze" 2004, No. 1
- Desert S. E., *WESTLAW Is Natural v. Boolean Searching: A Performance Study*, "Law Library Journal" 1993, Vol. 84, issue 4
- Griffith C., *FREESTYLE:LEXIS/NEXIS Goes Natural*, Information Today, 1994, at <https://www.questia.com/magazine/1G1-14777062/freestyle-lexis-nexis-goes-natural> (visited January 17, 2017)
- Justiss L. K., *A Survey of Electronic Alternatives to LexisNexis and Westlaw in Law Firms*, "Law Library Journal" 2011, Vol. 103, No. 1
- Mart S. N., *The Relevance of Results Generated by Human Indexing and Computer Algorithms: A Study of West's Headnotes and Key Numbers and LexisNexis's Headnotes and Topics*, "Law Library Journal" 2010, Vol. 102, No. 2
- McKenzie E. M., *Natural Language searching: How WIN Works in Westlaw*, "Legal Reference Services Quarterly" 2001, Vol. 18, No. 4
- Myers J. M., *Progress in Documentation. Computer and Searching Law Texts in England and North America. A Review of the State of the Art*, "Journal of Documentation" 1973, Vol. 29, No. 2
- Petzel J., *Informatyka prawnicza. Zagadnienia teorii i praktyki*, Warszawa 1999
- Petzel J., *Język prawny w świetle lingwistycznej teorii rejestru językowego*, "Studia Iuridica" 2006, Vol. XLV
- Petzel J., *Komputerowe systemy wyszukiwania informacji prawnej wspólnot europejskich*, Bazy Europejskie, Biuletyn, Centrum Europejskie Uniwersytetu Warszawskiego. Ośrodek Informacji i Dokumentacji Rady Europy, Warszawa 1993, No. 9
- Petzel J., *Statistical systems for computerized information retrieval*, (in:) *Mélanges offerts à la mémoire de Jason Hadjirinias*, Piraeus University 1989
- Petzel J., *Tezaurusy systemów wyszukiwania informacji prawnej*, Materiały konferencji "Prawo i język", Warszawa 2009
- Salton G., *Experiments in Automatic Information Organization and Retrieval*, New York–St. Louis–San Francisco–Toronto–London–Sydney 1968
- Schweighofer E., *Legal Knowledge Representation. Automatic Text Analysis in Public International and European Law*, The Hague–London–Boston 1999
- Simitis S., *Informationskriese des Rechts und Datenverarbeitung*, Karlsruhe 1970
- Studnicki F., *Wprowadzenie do informatyki prawniczej*, Warszawa 1978
- Tapper C., *Computers and Law*, London 1973
- Weyer S. A., *Questing for "DAO": DowQuest and Intelligent Text Retrieval*, "Online" 1989, Vol. 13, No. 5
- Wheeler R. E., Jr., *Does WestlawNext Really Change Everything? The Implications of WestlawNext on Legal Research*, "Law Library Journal" 2011, Vol. 103, No. 3
- Woods D., *The Myth of Crowdsourcing*, September 29, 2009, at <http://www.forbes.com/2009/09/28/crowdsourcing-enterprise-innovation-technology-cio-network-jargonspy.html> (visited January 17, 2017)

KEYWORDS

natural language in retrieval, ranking algorithms, statistical analysis of the text, learning programmes, Google for lawyers, WESTLAW IS NATURAL, FREESTYLE SEARCHING, WESTLAWNEXT

SŁOWA KLUCZOWE

język naturalny w wyszukiwaniu, algorytmy porządkujące, analiza statystyczna tekstu, programy uczące się, Google dla prawników, WESTLAW IS NATURAL, FREESTYLE SEARCHING, WESTLAWNEXT