www.glrjournal.com
GLR
GLOBAL LANGUAGE REVIEW

Mark Perkins[*]

Vol. IV, No. I (Spring 2019) | Page: 1 – 7
p-ISSN: 2663-3299 | e-ISSN: 2663-3841
L-ISSN: 2663-3299
DOI: 10.31703/glr.2019(IV-I).01
URL: http://dx.doi.org/10.31703/glr.2019(IV-I).01

Approaches to Text Analysis

Abstract:
Text Analysis is a broad term that covers many approaches and technologies. Those initially stemming from the academic sphere have come to enter the commercial, and today there is a wide interplay between the two. A further dichotomy is that between natural language and computational approaches. Over time, approaches have come to draw upon each other, although there are still clear dividing lines and practitioners tend to rely mainly on one approach or the other. This paper seeks to draw out these approaches and give an account of them over time. It also points to future developments where artificial intelligence is increasingly used.

Key Words:
Text Analysis, Lemmas and Concepts, Text Analysis Technologies, Enterprise Search

Introduction

Since the early 2000s, the advent of the internet and digital storage has led to a proliferation of text sources, analytical technologies, natural language and computational approaches to the analysis of the text. Academic and commercial users are actively involved in the quest for meaning and its representation. In the academic sphere, two lines of inquiry have emerged. The first, corpus linguistics, concerns the analysis of large corpora, with a corpus being a systematically collated body of the text of typically more than 100m words, annotated according to linguistic principles. (McEnery, & Hardie 2013). The second, computational linguistics, employs statistical and artificial intelligence (AI) techniques in order to detect patterns in text (Augustyniak et al., 2016, Briscoe et al., 2006). In the commercial world, a large number of companies have sprung up over nearly 20 years. In the beginning, these were almost all small start-ups, although some large companies such as Oracle (2019) and IBM (2019) have had an interest in text analysis for a long time. Indeed, IBM offer IBM SPSS Modeler with a text analytics add-on which is especially useful for the analysis of open-ended survey comments (ibid., 2019). Many small companies have disappeared or have been acquired by larger groups. Others have developed and grown, such as Meltwater (2019). However, it is interesting that even Meltwater, considered to be a large player, only employs 1,500 people at the time of writing, which is much smaller than large companies in many other sectors. On the other hand, although the field now has a track record, market entrants with new approaches are still appearing. Factbase, for example, is a new Swedish company that claims to be able to collect every piece of digital text publicly available and attributable to a person of interest (Factbase 2019). As regards applications in both the academic and commercial areas (and, of course, the two often overlap), there is a focus on how real-world objects, events and processes are represented by lower-level linguistic items such as words and higher-level entities such as concepts.

It is also clear that there has been a gradual move from simply "buzz tracking" based on a combination of behavioral click metrics and keyword (single term or word) trending towards more sophisticated linguistic methodologies based on notions of concept. The activity as a whole has only recently begun to consider wider social contexts as well as point of view and role (Niederhoffer, K., & Smith, M. 2009). Social contexts, indeed, have been well mapped by philosophers (Austin 1979) applied linguists (Fowler et al. 1979; Halliday 1994), as well as socio-linguists (Hudson 1996; Hymes 1972) and have only relatively recently begun to be perceived as relevant in the web analytics community (Mason 2007).

The location of texts has also had a major impact on analytical approaches, with the great move from print to digital. Today, the digital text is found in two main spheres: web and non-web. Within the web, the sphere includes websites (extranet and intranet) and all accompanying documents (such as downloadable pdf files) as well as chat feeds, blogs, and forums. The years from 2004 (the advent of Facebook) and in particular around 2013 were a period of great change when approaches became more sophisticated and companies began to focus on the new social media, such as Facebook and Twitter (Murphy 2013; Leung 2017).

Within the non-web sphere are included all documents digitally held on machines in programs such as Microsoft Office Suite. Some technologies are designed to target web text, others non-web. Enterprise search, for example, is an activity where intelligent search technologies are deployed to search for information held by the enterprise in many locations and forms, from collections of Microsoft documents of all types to text held in databases (such as free text opinion data). The most advanced systems use a concept-based approach to searching and categorizing data, as well as making links between various types of databases and texts (Micro Focus, 2018).

[*] Chief Executive Officer (CEO), Repindex, Cambridge, UK. Email: mark.perkins@repindex.com
(The author has worked in Text Analysis for over 20 years)

The Basic Units of Text Analysis: Terms (Words), Lemmas and Concepts

It is possible to put forward a three-stage model of the main categories of text analysis. In the first stage individual linguistic terms, (words or particles), are the focus of attention. A simple analysis might provide the word count frequency in a document. The second stage involves looking for collocations (groups of two or more words that typically occur together in direct proximity) and concordances (the combination of one or more words with one or more other words in a context, called a context horizon). In the third stage, concepts are identified. A concept is an abstraction that is realized by concrete pieces of text occurring in written or spoken language. One simple type of concept is a lemma. A lemma is a root form and is traditionally denoted in capitals in the field of corpus linguistics. Thus, the lemma LOOK stands for the term look and all forms which may follow it, such as looks, looked, looking, looker, lookers, looker's, and lookers'. Queries based on all these techniques are possible in the well-known corpus linguistics program WordSmith Tools, which was launched in 1996 and is now in its 7th edition (Scott 2016).

Lemmas can be characterized as simple types of concepts. Higher types of concepts might vary from an overall idea exemplified by a number of terms (micro-level) to a topic of a larger text, ranging from a paragraph to a book (macro-level). The topic thus means what a larger text is about (as in "the topic of the article is how we define concepts"). There is a great deal of work to be done within the third stage since there is no universal agreement on the definition of the concept, and there are many possible definitions of the size and types of texts to which it may apply.

These three stages of text analysis, however, should not necessarily be seen as historical in their development. Indeed, although there is a broad historical trend – from term to concept – the various levels of analysis can be found at different stages, such that even when some actors have moved to the concept level, others have remained on the term level or employ a mixed approach. In academic corpus linguistics, the main query types in WordSmith Tools, the term (word) and lemma, have remained the same from its launch in 1996 to the latest version in 2016 (Scott 2016). On the commercial side, Micro Focus, an industry leader, includes "conceptual search, keyword search, field text search, phrase search" in its toolkit (Micro Focus 2018). In addition to the identification of terms and their frequency, a further level of analysis concerns the keyword. In corpus linguistics, this means the statistical significance of a word in a text compared with that in a reference text (ibid 2016). However, in the commercial sector keyword often means little more than an important word or word as the focus of a search (Squarespace 2019).

As I have already indicated, the concept level may cover micro (word and sentence) and macro (paragraph and above) stretches of text, and perhaps the macro-level constitutes the fourth stage of analysis.

Beyond these three stages, it is possible to discern a fourth stage in which larger meaning units are dealt with, and even a fifth where the human interacts with the machine in a natural way.

Linguistic Complications

The text analysis industry often operates on a fairly simple level of semantics, that concerned with word (or collection of word) meaning, which can also be called lexical content, where word in English can be defined as one autonomous linguistic unit with content, and a small collection of two or three words with a semantic unity such as a phrasal verb. However, semantics and other dimensions of linguistics offer plenty of rich, often neglected material that can be used in text analysis. There is, of course, a long tradition of syntactico-semantic research (Katz & Fodor 1963; Katz 1972; Ferris 1983, 1993) with detailed and complex analyses which can be of great use to text analysis (Palek & Perkins, 1985; Perkins, 1985; Perkins 2019, p.335). For example, we might talk about micro-semantics (where one category might be opposites; good and bad, or scales; a rather good phone, a fairly bad laptop) or macro-semantics in terms of discourse (Perkins 2019, p.25 ff). Then again there are semantic dimensions of time present in texts (indications of how likely something is going to happen).

Linguists have also long been engaged in the activity of parsing and tagging. Parsing refers to the process whereby the grammatical parts of speech are identified (such as the noun, verb, and adjective). When a text is tagged these markers are assigned to terms. Here is a simple example.

a. He(pronoun-masculine singular) had(verb-simple past) a(indefinite article) complaint(noun) about(preposition) service(noun).One practical application of this is the detection of gender bias in a text: the recording of the number of masculine personal pronouns as opposed to feminine. Syntax concerns the way in which terms are connected together. For example, word order in English is significant syntactically because a change in word order can determine a change in meaning, as is clear in this example.

b. John loves Jane

c. Jane loves John

Word order is also used in English to indicate emphasis, as here.

d. I never buy Fiat cars

e. Never do I buy Fiat cars

The first example shows the unmarked (default) word order, while the second a marked order, together with a change of grammar.
Morphology is all about the shape of words. English has a poor or reduced morphology. For example, the plural of nouns is most often indicated by the addition of s.

f. Car

g. Cars

Those are the only forms that occur. Slavic languages, on the other hand, have a complex morphology. In Czech, for example, there are four declensions (noun types) and seven cases. Forms are determined by a role in the sentence (such as subject or object) and relationship to other items (such as preposition plus noun combinations), as in these examples.

h. Hrad je starý (The castle is old)

i. Mluvím o hradu (I am talking about the castle) These two sentences can be analyzed in the following way and the complexity can clearly be seen.

Figure 2. morphological analysis of Czech sentences

Hrad	je	starý
The castle	is	old
Noun-nominative singular masculine inanimate	Verb-third person singular	Adjective-nominative singular
Mluvím	o	hradu
I am talking	about	the castle
Verb-present first person singular	Preposition	Noun-dative singular masculine inanimate

Pragmatics is concerned with broader aspects of meaning than just word or sentence meaning. It is about larger units of meaning (such as found in a paragraph or article) and it is also about how dimensions of meaning cut across texts and outside texts. Sociolinguists, for example, study how ways of speaking or writing reflect groups of people in different social and / or geographical locations. Discourse analysts add even more dimensions such as economic and power relations.

Finally, there are several aspects of English which a probability-information model would find hard to cope with. In Systemic Functional Linguistics (SFL), for example, the ideational component includes a function centered on the notion of transitivity, the mechanism whereby participants (people, things and thoughts) are represented by linguistic form (Halliday 1994). Pioneering critical linguists such as Fowler took up this mechanism and showed how, for example, by changing verb form from active to passive it is possible to represent the same event (from the camera's point of view) in different ways (Fowler et al. 1979).

Consider these pairs of examples.

j. Police break up the mob

k. Mob controlled by police

l. Central bank tames inflation

m. Inflation brought within targets

Both the choice of words and choice of grammar shape the representation of events.

Text Analysis Technologies: Starting Points

The type of technology deployed and developed to analyze text depends on starting points.

Those starting from a natural language and applied linguistic perspective have tended towards an analysis using tools ultimately based on the notion of a spreadsheet, such as Excel. In their early work in the mid-2000s, Lexalytics offered simple dashboards with term frequency counts. Later, their software could be deployed as an add-in to Excel, and currently, they use machine learning to detect sentiment, for example (Lexalytics 2019). Since the 1960s, indeed, academic linguists have developed natural language-based desktop (and later cloud-based) programs to analyze large text databases known as corpora (Baker et al., 1994, McEnery & Hardie 2013). The basic unit of corpus linguistics is the context line, typically a search term bounded by a number of words left and right (known as the context horizon).

In parallel with these developments, computer scientists have developed algorithmic methods to identify patterns in texts, without relying on any natural language analytical approach. These two sides have not enjoyed the extensive dialogue. Applied linguists have tended to start from functional and social analyses of language. Computer scientists (and later computational linguists) have tended to start from the question of what a technical system can do with the material it is presented with. Again, the advent of the web has also stimulated search technologies that are developing rapidly.

Text Analysis Technologies: Corpus Linguistics and Natural Language

Linguists have always analyzed text, whether occurring naturally or whether created by the linguists themselves. Indeed, there are two distinct strands of linguistics represented by an approach which requires naturally occurring stretches of language as data on the one hand, and one which relies on examples constructed by the linguist for analysis. The former can be traced back at least and far as John Firth (1957), and the latter was extensively used in the Chomskyan tradition (Chomsky 1965; Akmajian & Heny 1975).

At first, companies drew on early academic work conducted under the umbrella of content analysis (Budd 1967), along with co-word analysis (mentioned in Myers 1996). The Glasgow Media Group (1976; 1980; 1982) conducted a ground-breaking diachronic analysis of terms found in the TV media over time.

As soon as advances in computing permitted, linguists began to collect and analyze text on a large scale, and the field of corpus linguistics emerged (Biber & Reppen 2012, McEnery & Hardie 2013). For example, an early and significant corpus, the Brown Corpus of 1m words (American English), was built in the 1960s and analyzed by Kučera and Francis (1967). A corpus is defined as an amount of text, often large, collected according to systematic principles (such as a representative selection of media sources) and notated or tagged grammatically, for example. Since the 1960s large corpora have been constructed and put to use, such as the Collins COBUILD Corpus which was a pioneer in the construction of English dictionaries based on corpora, and which, at 4.5bn words, is one of the largest corpora of the English language (Collins 2019). There are also many corpora in other languages, such as the National Corpus of Polish (2019) with 1.5bn words, and The Balanced Corpus of Contemporary Written Japanese, with 104m words (NINJAL 2019).

Commercial Approaches: From Keywords to Computational Linguistics

In the commercial world, companies have been using linguistic techniques, ranging from simple manually generated classifications to advanced computational approaches for many years. At first, companies drew on early academic work conducted under the umbrella of content analysis (Budd 1967), along with co-word analysis (mentioned in Myers 1996). The Glasgow Media Group (1976; 1980; 1982) conducted a ground-breaking diachronic analysis of terms found in the TV media over time. The classification of linguistic items in the content analysis came to be known as coding. Analysts followed categories in a coding frame, manually allocating items to categories.

In the early part of the 2000s, many companies sprang up with the primary mission to analyze talk about brands on the internet. Companies came and went in rapid succession. Some went bankrupt or disappeared. Others were bought by larger players such as media or marketing agencies. Of ten text analytics companies I surveyed in 2009 (Perkins 2009), four are untraceable (Strategy Eye, Overtone, Scoutlabs, Market Sentinel), three were acquired (Nielsenbuzzmetrics, Clara, and Autonomy), and two still trade under their own name (Clarabridge 2019a, 2019b; Brandwatch 2019).

A great deal of analysis has been undertaken and still is being undertaken under the umbrella of sentiment analysis (SA). For commercial actors, sentiment usually means positive of negative orientation towards a brand as expressed in free text. That free text may occur on the internet in various locations or in free text opinion surveys, usually known as the open-ended part. Typical questions may vary from "tell us your thoughts" (completely open-ended) to "please tell us what you like or dislike about our brand" (guided open-ended). Positive of negative orientation may be expressed in terms of likes and dislikes, and good and bad. In many cases, this very simple approach is all that companies need to have. So long as the data are extensive, it gives a good idea of attitudes towards a brand. Keywords, loosely defined as prominent or important words, may be identified and simply counted (as opposed to keywords defined in corpus linguistics as those statistically key when a reference and target corpus are compared (Scott 2016)). However, it is possible to do much more than this. For example, a brand may have many aspects. The use of a mobile phone involves screen size, appearance, and arrangement of icons, battery life, and available apps, to name some of them.

Interestingly, it is only now that computational approaches to the evaluation of user opinion concerning aspects (such as comments on aspects of a mobile phone in reviews) are being developed in computational linguistics (Augustyniak, Kajdanowicz, & Kazienko 2019). This could be linked to Halliday's Systemic Functional Linguistic framework (Halliday, 1994) where again, only recently are attempts being made to apply computational techniques (Ito et al., 2005).

Enterprise Search

Search technologies show similarities to those of text analysis when the objective is not only to find information but to classify it in addition. Enterprise search refers to the activity of conducting searches in a wide variety of documents and text sources held on the servers of a large organization, private or public. The process includes identifying concepts in a text, classifying them and creating and displaying links between documents and information.

Micro Focus broadly employs two approaches to text analysis. In the one they use linguistic and natural language derived techniques. These include tokenization (breaking down text into unit instances, where the unit most often is a word), stemming (where a word is reduced to its semantic root), and stoplists (lists of form words such as of and to which are redundant when conducting a frequency count, for example) (Micro Focus 2019). On the other, probability and information theory are used.

Micro Focus IDOL (Intelligent Data Operating Layer) is one of the most advanced systems in the field and offers unified text analytics, speech analytics, and video analytics

Its approach to search and text analysis relies mainly on theories of probability and information. IDOL relies, at least partly, on Bayesian inference (Micro Focus 2018, p.1), whereby it is possible to calculate the probability of terms co-occurring or occurring in sequence and Shannon's theory of information. Here is a brief statement of the latter.

IDOL's approach to concept modeling begins from the tenet of Shannon's theory, which says the less frequently a unit of communication occurs, the more information it conveys. As a result, concepts and ideas that are unusual or distinctive within the context of communication tend to be more indicative of its meaning. IDOL applies this theory to determine the most important (or informative) concepts within a document.

(ibid., 2018, p.4).

However, rarity is only one measure of meaning and thus value. There are other dimensions of meaning, such as social meaning, which may be expressed by a high frequency of term occurrences. A one-time occurrence of one term in a text may not be more indicative of the text's meaning where that term is rare compared to the contents of that text or of a larger text. The text may contain other terms and combinations of terms which, when repeated and combined in different ways may express a social meaning in addition to other types of meaning, such as propositional.

Machine Learning, Artificial Intelligence, Big Data, and Automation

The deployment of artificial intelligence (AI), is now becoming established in many areas, and text analysis is no exception (Evelson 2018). Defined as the application of statistical techniques to big data (large quantities of data of many types), AI is increasingly used to identify patterns and make forecasts, often in real-time. Such systems may employ Supervised Learning (SL) which operates after a training data set has been applied, or Unsupervised Learning (UL) which operates directly on the data it is presented with. Indeed, AI consists of the various types of machine learning, but artificial intelligence is a more attractive term in the commercial world, at least, and it has begun to enjoy widespread use.

AI is currently used in text analysis in certain ways, and Lexalytics identifies three parts (Lexalytics, 2019 p.3). The first is to take training data. These data could comprise a large number of customer reviews which a manual coder could code as to positive or negative. The next step is to apply algorithms to the coded data such that when more data enter the system they can be coded appropriately. This is an example of supervised learning. When the algorithms are working on a satisfactory level, training data may not be needed. New data may be entered from a different domain and the algorithms may produce satisfactory results (unsupervised learning). As mentioned above, state of the art research by Augustyniak et al. (Augustyniak, Kajdanowicz, & Kazienko 2019) now focuses on aspect extraction and levels of accuracy of the above 70% are possible (compared with manual coding samples). The third part identified by Lexalytics is called hyper-parameters. These include non-linguistic data such as timestamps (Lexalytics, 2019 p.4). Hyper-parameters, or metadata, are catered extensively by Micro Focus whose system is capable of making links to many different types of data, such as audio and video (Micro Focus 2019).

Conclusions

From the 1960s to the present day many significant advances have been made in text analysis, and the last 20 years, in particular, have seen rapid progress. This is due to two main factors: technology and commercial factors. Technology has both made available large quantities of text held not only in online platforms but also on the servers of large organizations. Secondly, commercial actors have seen extensive business opportunities in the analysis of the text, ranging from free text opinion data to enterprise search. The most recent advances concern the development and use of AI. It seems certain that this will lead to a generation of predictive text analytics with exciting and useful results.

References

Akmajian, A. & Heny, F. (1975) An introduction to the principles of transformational syntax. Cambridge, Mass.: MIT Press.

Allan, Keith (2013) The Oxford Handbook of the History of Linguistics. Oxford Handbooks in Linguistics.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L., & Sadler, L. (1994) Machine Translation: an Introductory Guide. London: Blackwell.

Augustyniak, Ł., Kajdanowicz, T., Tuligłowicz, W., & Szymański, P., (2015) Comprehensive Study on Lexicon-based Ensemble ClassificationSentimentAnalysis. Retrievedfrom:https://www.researchgate.net/publication/288488744_Comprehensive_Study_onLexiconbased_Ensemble_Classification_Sentiment_Analysis

Augustyniak, Ł., Bartusiak, R., Kajdanowicz, T., & Kazienko, P., & Piasecki, M. (2016) WordNet2Vec: Corpora Agnostic Word Vectorization Method. Retrieved from:https://www.researchgate.net/publication/ 303921745 _WordNet2Vec_Corpora_Agnostic_Word_Vectorization_Method

Augustyniak, Ł., Kajdanowicz, T., & Kazienko, P. (2019) Comprehensive Analysis of Aspect Term Extraction Methods using Various Text Embeddings. Retrieved from https://www.researchgate.net /publication/335755175

Austin, J.L. (1979) Philosophical papers. Oxford University Press.

Baker, M., Francis, G., & Tognini-Bonelli, E. eds (1994) Text and Technology: In Honour of John Sinclair. Amsterdam: John Benjamins.

Biber, D., & Reppen, R. (2012) Corpus Linguistics. London: Sage.

Brandwatch (2019). Retrieved from https://www.brandwatch.com/

Briscoe, T., Korhonen, A., & Krymolowski, Y. (2006) A Large Subcategorization Lexicon for Natural Language Processing. Cambridge University: Computer Laboratory.

Budd, R., Thorpe, R. & Donohew, L. (1967) Content analysis of communications. New York: Macmillan.

Chomsky, N. (1965) Aspects of the theory of syntax. Cambridge Mass.: MIT Press.

Clarabridge(2019a)SentimentAnalysis. Retrieved from: https://www.clarabridge.com/customer-experience-dictionary/sentiment-analysis/

Clarabridge (2019b) Text Analytics. Retrieved from https://www.clarabridge.com/customer-experience-dicti onary/ text-analytics

Collins (2019) The History of COBUILD. Retrieved from https://collins.co.uk/pages/elt-cobuild-reference-the-history-of-cobuild

Evelson, B. (2018) The Forrester Wave™: AI-Based Text Analytics Platforms, Q2 2018: The Eight Providers That Matter Most and How They StacUpRetrieved from https://www.clarabridge.com/wp-content/uploads/ 2019/04/ The-Forrester-Wave%E2%84%A2_-AI-Based-Text-Analytics-Platforms-Q2-2018.pdf

Factbase (2019). Retrieved from https://factba.se/

Ferris, D.C. (1983) Understanding Semantics. University of Exeter.

Ferris, D.C. (1993) The Meaning of Syntax: A Study in the Adjectives of English. London: Longman.

Firth, J.R. (1957) Papers in linguistics. Oxford University Press.

Fowler, R., Hodge, B., Kress, G., & Trew, T. (1979) Language and control. Routledge & Kegan Paul.

Glasgow Media Group (1976) Bad news. London: Routledge.

Glasgow Media Group (1980) More bad news. London: Routledge.

Glasgow Media Group (1982) Really bad news. London: Writers and Readers.

Halliday, M.A.K. (1994) An introduction to functional grammar. 2nd ed., London: Edward Arnold.

Hudson, R.A. (1996) Sociolinguistics. 2nd ed., Cambridge University Press.

Hymes, D. (1977) Foundations in sociolinguistics: an ethnographic approach. London: Tavistock Publications.

IBM (2019) IBM SPSS Modeler. Retrieved from https://www.ibm.com/uk-en/products/spss-modeler

Ito, N., Iwashita, S., Sugeno, M., & Sugimoto, T. (2005). A Computational Framework for Text Processing Based on Systemic Functional Linguistics. Retrieved from http://www.brain.riken.jp/labs/mns/sugimoto/csfgc05.pdf

Katz, J.J. & Fodor, J.A., (1963) The structure of a semantic theory. Language, XXXIX, 2, PP. 170-212.

Katz, J.J. (1972) Semantic theory. New York: Harper & Row.

Kučera, H. and Francis, W. N. (1967) A Computational Analysis of Present-Day American English. Providence: BrownUP.

Leung, J. (2017) A Guide to Text Analytics from a Leader in the Industry. Retrieved from https://community.microfocus.com/t5/Information-Management-and/A-Guide-to-Text-Analytics-From-a-Leader-in-the-Industry/ba-p/2702174/jump-to/first-unread-message

Lexalytics (2019) Machine Learning for Natural Language Processing and Text Analytics. Retrieved from https://www.lexalytics.com/resources/wpcontent/uploads/sites/3/2019/02/Lexalytics_Machine_Learning_Natural_Language_Processing_Whitepaper.pdf

McEnery, T. & Hardie, A. (2013) The History of Corpus Linguistics, in Allan, (ed) The Oxford Handbook of the History of Linguistics. Oxford Handbooks in Linguistics, chapter 34.

Meltwater (2019). Retrieved fromhttps://www.meltwater.com/uk/

Microfocus (2018) Seven Tips for Developing an Effective Unstructured Data Analytics Program. Retrieved from https://www.microfocus.com/media/whitepaper/seven_tips_for_developing_an_effective_unstructured_data_analytics_program_wp.pdf

Micro Focus White Paper (2019): Augmented Intelligence: Helping Humans Make Smarter Decisions, 2018, p.4. Retrieved fromhttps://www.semanticscholar.org/paper/Augmented-Intelligence-%3A-Helping-Humans-Make/8f5dd061171ccff9a12622f2c895f75831292be2#related-papers

Micro Focus (2019) Micro Focus IDOL Expert. Retrieved from https://www.microfocus.com/documentation/idol/IDOL_12_4/IDOLServer_12.4_Documentation/Guides/html/English/expert/index.html

Murphy, L (2013) The Fall of Buzzmetrics & Rise of The New Social Media Analytics Market Research Firm. Retrieved from https://greenbookblog.org/2013/04/22/the-fall-of-buzzmetrics-rise-of-the-new-social-media-analytics-mr-firms/

Myers, G. (1996) Out of the laboratory and down to the bay: writing in science and technology studies. Written Communication, 13 (1), pp.5-43.

National Corpus of Polish (2019). Retrieved from http://nkjp.pl/index.php?page=0&lang=1 NINJAL (2019) The Balanced Corpus of Contemporary Written Japanese. Retrieved from https://pj.ninjal.ac.jp/corpus_center /bccwj/en/

Oracle (2019) Social Cloud. Retrieved from https://www.oracle.com/uk/applications/customer-experience/social/

Palek, B. & Perkins, M.C. (1985) Základy jazykovědy (Foundations of language). Prague: SPN.

Perkins, M. C. (1985) Subject Properties and Predicate Referential Structure. In Linguistica Generalia, Volume 4 pp. 23-61, Acta Universitatis Carolinae Philologica. Prague: Charles University.

Perkins, M.C. (2009) White Paper: The Repindex Approach to Semantic Text Analysis. Montreal: iPerceptions.

Perkins, M.C. (2019) Discourse, evolution and power. Cambridge: Repindex.

Scott, M (2016) WordSmith Tools. Oxford University Press.

Squarespace (2019) Adding keywords for SEO. Retrieved from https://support. squarespace.com/hc/en-us/articles/360001997648-Adding-keywords-for-SEO