Linha D'Água: São Paulo, v. 34, n. 02, p. 83-102, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p83-102

Article / Artigo

Building Digital Humanities on the Linguistic Background: Methodological Basis for Digital Humanities Education in Undergraduate and Graduate Programs

Construir Humanidades digitais num contexto linguístico: bases metodológicas para o ensino de Humanidades Digitais no 1º e 2º ciclo de ensino universitário

Lukáš Zámečník*

lukas.zamecnik@upol.cz
https://orcid.org/0000-0001-8098-4238

Ľudmila Lacková**

ludmila.lackova@upol.cz
https://orcid.org/0000-0001-9852-4280

Abstract

The transformation of society towards digitalization and automatization cannot be ignored by the higher education system. While this has been naturally reflected by the education system regarding technical sciences, humanities are still struggling to catch up with the latest trends in the digitalization of society. The field of Digital Humanities (DH) is very young and lacks a solid methodological basis or ontological principles. This paper aims at proposing some philosophical and methodological grounding for the field of DH and its practical applications in the higher education system. We describe two case studies of the creation of new study programs at the Palacký University Olomouc, Czech Republic.

Keywords: Digital Humanities; Education; Methodology; Philosophy of Science; Quantitative Linguistics; Study Programs.

Resumo

A transformação da sociedade em direção à digitalização e automação não pode ser ignorada no sistema educacional superior. Embora o processo tenha vindo a ser naturalmente refletido pelo sistema educacional de ciências técnicas, as Humanidades continuam a enfrentar dificuldades para se manterem atualizadas com as últimas tendências da

* Department of General Linguistics, Palacký University, Olomouc, Czech Republic

** Department of General Linguistics, Palacký University, Olomouc, Czech Republic.

digitalização da sociedade. A área das Humanidades Digitais (HD) é muito jovem e carece de uma base metodológica sólida ou de princípios ontológicos. Este artigo tem como objetivo propor alguns fundamentos filosóficos e metodológicos para a disciplina das HD, e respectivas aplicações práticas no sistema de Ensino superior. Descrevemos dois estudos de caso sobre a criação de novos programas de programas de estudo na Universidade Palacký em Olomouc, República Tcheca.

Palavras-chave: *Humanidades digitais; Educação; Metodologia; Filosofia da Ciência; Linguística Quantitativa; Programas de Estudo.*

Introduction

The history of the field of DH has its origins in the practical demands of philologies and the humanities based on them (LEE, 2018). It is based on the tradition of digital word processing. The symbolic birth of Digital Humanities is considered to be the decade-long digitization of the work of Thomas Aquinas by the priest Roberto Busa (started in 1946). In the case of Busa's project, we cannot talk yet about the digitalization in the sense of the today's meaning of the word, but it was a project of lemmatization of the data with a semi-automatic process (see BUSA, 1980). Digital Humanities can be classified alongside Digital Social Science in the broader category of Data Science. All these post-disciplines are conceptually based on the theory of dynamical systems (KELLERT, 2008). This theory (built since the 1960s) makes it possible to explain the behavior of complex systems regardless of their ontology, by capturing common features or isomorphic (or analogous) structures across similar systems.

Some mathematical models are created using statistical methods, with the goal to find some common and stabile characteristics of data. The universality of the found statistical distributions of data (in diverse ontologies) led to the definition of some principles of the theory of dynamical systems (especially in the context of the so-called scale-free networks, see CALDARELLI, 2007). Until recently, the main limitation of the development of the theory of dynamical systems for the needs of the systems described in the social sciences and humanities was the lack of data. At present, however, computer systems based on machine and deep learning methods (namely artificial intelligence) provide tools for extracting a growing amount of information from the Internet (social networks, etc., so-called Internet Artificial Intelligence), from the digital archives of government sections, from hierarchical lists of corporate data repositories (so-called Corporate Artificial Intelligence), and from the emerging data packages of the emerging Internet of Things (so-called OMO, online-merge-offline, so-called Sensory Perception Artificial Intelligence). In the near future, data collected by autonomous AIs (smart homes, autonomous vehicles, etc.) will also be added (see mainly the chapter "The Four Waves of AI", in: LEE, 2018, p. 104–139).

Digital Humanities is one of several interesting outcomes of the history of the humanities, starting with the turn to the Geisteswissenschaften and ending with Foucault's efforts to build "human sciences" (FOUCAULT, 1966). In this sense, Digital Humanities are the culmination of the ambitions of the Archaeology of Knowledge (FOUCAULT, 1969). Where historical (and then social) causality exists for Foucault, but is not traceable and is therefore useless as a concept, there is a massive pattern recognition ability in Digital Humanities using AI (finding correlations to an extent that is inaccessible to humans) which overcomes the limitations associated with the search for simple causal relationships. The Archaeology of Knowledge is based on the AI's ability to extract data and with the help of heuristics, which often escape explanations, provide their interpretation, recognize a stable pattern of behavior (consumers), actions (social agents), or decision-making (judges, teachers, etc.).

1 Integrating DH in education

The still growing tendency towards creating new study programs in DH is motivated by the changes in society during the Fourth Industrial Revolution, especially by the expected changes in the labor market (the DH Lab in University Nova of Lisbon, a two-year master's program Digital Humanities and Digital Knowledge at University of Bologna and many others). Expressed by means of the theory of dynamical systems, there is a real risk that the self-regulatory mechanism of the free market will eventually cease to function, because cheap labor will cease to be a competitive advantage. According to some analysts, AI technologies will lead to further growth of corporate structures of the global economy. This is not an ideological statement, but a statement of a probable change in system dynamics. A change in several key regulatory parameters will lead to a change in the attractor of the global economic system. In order to prepare our society for these changes, we should accelerate the adoption of technologies. According to the strategical document Digital Europe Programme, EU member states should try to accelerate the adoption and best use of digital technologies, including the latest digital capacities, across the economy and society. According to this strategic document, "all Member States can identify, analyse and adapt to digital trends, establish the needs and priorities of the public and private sectors, share best practice, and contribute to common specifications and standards"(Digital Europe Programme Draft orientation, p. 29).

Ross (2016) and Lee (2018) show that in the developed world, not only manual-based sectors will be affected (this will be the area in the developing world where the advantage of cheap labor will have dramatically reduced), but also some sub-occupations in the services sector, which are also trained through the faculties of philosophy, will probably be significantly automated (such as some areas of media, interpreting, translation, etc.). In addition, automation will affect the industry of mental activities (e.g., leading positions in company offices, customer services, tax consulting, etc.). As an example, we can mention the eGrants from the European Commission, where the whole process of project submissions, evaluations, etc. has been automatized to a great extent, mostly during the Covid pandemics period (see the document Strategic Plan 2020–2024 - Research and Innovation, p. 43).

AI technologies are undeniably replacing some of the historically existing work positions at the labor market. Nevertheless, at the same time, new positions requiring a human interpretive approach are arising. As a consequence of the great progress in the developing technologies – and as a paradox – there is new room for a person. Not all patterns automatically recognized by AI are equally relevant, and it is the human mind that should be able to assess their relevance and usefulness. As a matter of fact, the field of AI is probably at the beginning of the third-wave AI, but the vision of an artificial agent able to make decisions in the way that a human being (genuine intelligence) does is far from close to be achieved (SMITH, 2019). In the current state of DH, the great optimization ability of AI at the level of quantity is connected with the optimization ability of a person at the level of quality.

This is the main motivation for the need of investments in the new study programs of DH. According to the strategic document of the European Commission Strategic Plan 2020–2024 – Education, Youth, Sport, and Culture,

> technology and the future of work, digitalisation of society and learning, or the transition to a circular economy necessitate that education and training systems across Europe can deliver the knowledge and skills, including digital skills and sustainable education that people need to participate fully in society (EUROPEAN COMMISION, 2020, p. 16-17).

Scholars are still more aware of the fact that the more technology advances, the more we need to stick to what makes us different from machines. This is why not only the education in humanities and social sciences is being radically modified in the form of the common name of DH but also the education in technical sciences and engineering started to be questioned and modified according to the new conditions of the Fourth Industrial Revolution. Some institutions are trying models of a revolutionary change of the education in engineering. For example, Texas Tech University in Lubbock launched a project (it is called DREAM: The Developing Reflective Engineers through Artful Methods) which already shows positive results in the education of engineering programs. The core idea of the change in the education of technical programs is that in order to be a more effective engineer, creative thinking (*Artful Methods*) is an indispensable predisposition. Insertion of artistic and creativity-based courses (such as creative writing) in the curricula of engineering programs proves positive outcomes of the students in the field (CAMPBELL *et al.,* 2020).

Nonetheless, there is also a counterpart: not only programs in the humanities are being digitalized and more *technical*, but also technical study programs are being *humanized*. This fact demonstrates the need of investment of energy and funds into quality education at the intersection of both directions. In this way, we will be ready, and the next generation will be prepared for the symbiosis with AI agents in everyday life. But not only that, the next generation will be prepared to face the changes of the labor market. The transformation of the labor market will largely affect graduates of humanities in general and mostly the graduates of philological disciplines – the small philologies as well as the large ones. Therefore, the overall integration of digitalization in the education of humanities is needed more then ever before.

2 Linguistic Digital Humanities

As was already mentioned in the introduction, the field DH originated in philologies as a necessity to preserve old manuscripts in the digital form (with the first digitized opera being the work by Thomas Aquinas).

Since the digitization of Thomas Aquinas' texts, the development of DH has continued in the direction of digitizing other important philological manuscripts or prints. Besides the transcription of old texts with aid of modern technological tools, the discipline of DH has nothing more in common with philology itself. It can be said that DH in the most general meaning is, strictly speaking, nothing but methodology – the researcher has digitized data, sets of computational methods, statistical tools, etc. and perceives them as tools that can be used in traditional disciplines (e.g., digital competences in medieval studies, in film studies, in gender studies, etc.). It can be said that DH conceived in this way are only a means of modeling data (GIERE, 2006, p. 68-69). Only with a linguistic background can we speak of DH as a new own-standing discipline that puts these data models into a theoretical framework and consequently allows their interpretation. We believe that the above-mentioned understanding of DH as pure methodology can be generalized, and it can be said that some areas of data science, despite their close affiliation with the natural sciences, remain mere clusters of methods. And the extent to which they can fulfill a role analogous to linguistics in another discipline (computational science) determines their future. In the humanities, we can expect greater integration through DH, and perhaps even the completion of the Foucault's project, as indicated above.

Merging the digital tools with linguistic theory, the discipline of DH can get closer to the original philological direction. Despite all the differences between the two authors, both Foucault and Derrida (1976) spoke about the central role of language (or text) in the scheduling of the humanities (or less categorically: *in capturing human destiny*). And even in the critique of the overly descriptive nature of their approaches, of their constructivism and discontinuism, of the relativism they have left room for, the central role of language still remains crucial in their theories. However, we are talking about language not only philosophically assumed, but more importantly, a language precisely defined and described by linguistics.[1]

The concepts of artificial language and natural language are central for DH – methodologically (artificial programming languages in methods), ontologically (DH are in most cases dealing with textualizable objects[2]) and axiologically through linguistic interpretation in a semiotic way.[3] Methodologically, because computational methods that allow data processing are artificial languages whose specific algorithmic set of rules represents the grammar of the code. Ontologically, because the basic elements of the DH ontology are objects in their

1 Both *Archeology of Knowledge* (FOUCAULT, 1966), and *Grammatology* (DERRIDA, 1976) refuse strict formal aspect of structuralism.

2 We have to keep in mind ontological commitments of linguistics, see Quine (1953).

3 Following Larry Laudan in his view of three-fold structure of paradigm (see LAUDAN, 1998).

textualized form. The choice of these objects is pragmatic as only textualized objects are accessible through the methods of artificial grammar processing. Axiologically, since the goal of DH is to map interpersonal interactions in all their complexity. These interactions are understood as the process of establishing, keeping, and transforming characters in the communication interchange.

Central to DH is also the concept of speech. However, not in the traditional structuralistic way of the term *speech* as in contradiction to *language* but in a new form that leaves room for the study of speech diachronicity in its continuous transformation. The original structuralist preference for synchrony over diachrony, as well as the subsequent poststructuralist rejection of the system, is overcome in the equal position of synchronous and diachronic ways of studying phenomena in DH. Following Veyen's Foucault[4], we can say that DH makes it possible to illustrate how the discourse is entrained by the dispositive. It will be possible to model this entrainment for individual disciplines in humanities that work with the historical dimension, similarly to the mood entrainment on social networks. In this sense, DH represents a natural continuation of corpus linguistics and text linguistics going in the direction of the developing of technologies applied to the research of authentic texts. The traditional methods of text linguistics (KOŘENSKÝ, 2003) are being extended and accompanied by various tools for automatic text processing, using corpora both of spoken and written speech.

We comprehend three core elements of Linguistic Digital Humanities (LDH): the basis is the segmentation of textualized objects accompanied by indispensable qualitative linguistic concepts and finally analyzed by digital tools for the analysis which lead to numerous possible applications. We comment accordingly on the three core elements in the following part of this section.

2.1 Segmentation of textualized objects

The exaggerated postmodern statement that "everything is text" has taken on a new message in the context of DH – we must treat textualizable objects as constructs composed of constituents – linguistic planes consequently set the frameworks for segmentations of a given type (morphological plane, syntactical plane, etc.). The determination of units – the basic components of segmentation and their constituents – is governed by some type of compositional principle[5] in which quantitatively traceable universal principles of scale-free networks can manifest themselves.[6]

4 See Veyne (2010, p. 92-110). See Conspiracy Pedagogies: QAnon, Social Media, and the Teaching of Far-Right Extremism. In: ihr.asu.edu/seed-grants/conspiracy-pedagogies

5 For the relation between the register hypothesis and the concept of power law see Köhler (2012, p. 84-92); for the principle of compositeness see Hřebíček (2003).

6 Caldarelli (2007); Ferrer-i-Cancho; Solé (2001).

The linguistic areas that we consider to be very important for DH include computational and quantitative linguistics – universal statistical tendencies in text analysis from Herdan (1966) to Altmann (1978) and Köhler (1986); and corpus linguistics (see JENSEN, 2014) – also works on extracting semantic relations from language corpora, for instance extracting semantic relations from Portuguese corpora (AMARO, 2014).

2.2 Qualitative concepts

As already mentioned, if DH is not to remain a mere group of methods, it is necessary that theoretical tools are involved – concepts, hypothetical principles, but also the whole deductive structure of theories. Theory building in DH inevitably entails the need to work with qualitative concepts of linguistics.[7] The neglecting of qualitative theoretical tools and over-reliance on machinery of technology (statistics in particular) can also cause misleading results. An essential example is the importance of lemmatization – experiments with non-lemmatized data may be valid in some cases (e.g., when looking for a distribution function that expresses the relationship between the length and frequency of a lexical unit), but in others it leads to completely invalid conclusions (e.g., if we want to express the relationship between many meanings of a lexical unit and the length of the lexical unit[8]). These facts have long been known in linguistics (see KÖHLER, 1986) but are sometimes ignored by some current analyses within data science.

2.3 Tools for analysis, comparative tools and aplicability

Merging elements (1) and (2) we can obtain useful tools for analysis in many disciplines. One of the possible applications results in the creation of the new generation of vocabularies or glossaries. Terminology in the traditional sense (CABRÉ, 1999) of the discipline is being developed thanks to the possibility of big data analysis. The digitalization of the discipline of Terminology facilitates creation of electronical vocabularies, dictionaries, or glossaries for specific disciplines or for specific research projects. Besides the creation of new glossaries, the discipline can be also comprehended in the sense of digital editions of already existing vocabularies or dictionaries (SALGADO; COSTA, 2020) or creation of new digital tools for lexicographers (SALGADO; COSTA, 2019). Not only the access to a big amount of data helps the discipine of Terminology to become more „digital", but also the digital outreach of the discipline started to have an important impact on the society. We can mention the ongoing project Glossário Colaborativo COVID-19 by colleagues from the University Nova of Lisbon

[7] Meyer (2002), Grzybek (2006).

[8] There is a long discussion about the appropriate variants of measuring the lenght of linguistic unit, for the new approach see Benešová; Faltýnek; Zámečník (2015). The realtion between the lenght of lexical unit and the number of its meanings (including the role of lemmatization) is explicated in Köhler (2005, p. 767-770).

developed as an orienting tool – a terminological glossary for public use regarding the pandemics. The aim of the project is to help increase the public knowledge about the disease, virus, and pandemics in general. Wordnets (for Portuguese see for instance AMARO, 2006; AMARO *et al.,* 2013) are other tools for working with lexemes or terms, depending on the area in question. Another possible application sphere is automatic sentiment detection, potentially usable in robotics, chatbots, or similar spheres of AI (LESCH, 2015). One more possible example of applicability of Linguitic Digital Humanities is Forensic Linguistics (see FALTÝNEK; MATLACH, forthcomming). In the field of forensic linguistics, the most common application of the digital tools is the automatic recognition of authorship. This application field has begun to grow recently in the Czech Republic, and it became a research topic of many projects in cooperation with students at Palacký University in Olomouc (JANEČKOVÁ *et al*., 2021). Distant Reading represents another possible application, where stylistic, periodization, genre and other categories are studeied through Corpus Linguistics analysis of big data.

Conclusion

DH point a new direction not only for humanities, but also for social sciences, and the consequences of this new direction have started to be incorporated in the educational system (see the following section Case Studies). Since the field of DH is rather young and is still defining its status within the scientific and academic environment, there is no unification so far regarding methodology or the very philosophy of the filed. This fact might be considered as an advantage for everyone who is interested in the field: every scholar contributing to the field of DH can help in defining and shaping the discipline with his/her proper approach. We also take this opportunity and propose a unique, linguistic, understanding of DH. In our understanding, DH is not a mere methodology for already existing disciplines. DH represents a whole specific approach to every textualizable research object. This approach is composed of three main components (segmentation as a quantitative view, qualitative linguistic analysis, and digital tools for specific applications).

We tried to present here the educational effort or even goal of DH as the main and most important part of the initiative across *human sciences* and *digital knowledge* integrated in the project of DH. However, the general effort must be implemented in some concrete activities which can transform the boundary field of humanities as well as digital and data science. Concrete projects should be prepared, like our new study programs in Linguistics and DH, which will incorporate the *linguistic basics* into the educational perspective of DH.

Case studies

At the Department of General Linguistics of Palacký Unviersity Olomouc, we incorporate most of these linguistic areas into teaching strategies and build new curricula in which the DH affinities to linguistic subdisciplines have an important role to play. In the sense of the aforementioned preliminaries of *Linguistic Digital Humanities*, new study programs have been created under the name *Linguistics and Digital Humanities*. Two were accredited in the last two years: a three-year bachelor program and a four-year doctoral program. In the next section, we will shortly present both of the newly created study programs at the Palacký University Olomouc.

Case study 1: Bachelor program Linguistics and Digital Humanities

The bachelor study program Linguistics and Digital Humanities got accreditation in 2020 and we expect the first round of applications in the year 2021. It is a pilot case of a bachelor program in DH in the Czech Republic.[9] The program is largely focused on future applicants to the faculties of arts and humanities in general. It will provide them not only with basic theoretical knowledge, but above all with rich equipment of digital tools and quantitative and qualitative methods that will allow the graduates to regain a competitive advantage. Mastering the methods of work in a digital environment, the ability of software programming and orientation in databases will allow graduates of the study program to successfully enter the transformed labor market (as described above). For example, a translator who is able to integrate software tools into his work and at the same time further develop the very software will still have an advantage over mere automatic translators available to consumers as well as over human translators working with traditional CAT tools.

The latter role of DHin the case of research shows that graduates of the study program *Linguistics and Digital Humanities* can also move towards further study, which will focus on the professional profiling of students with research ambitions. Further research career of the graduates of this program is possible both under the auspices of social sciences and humanities, as well as within the theories and methods of DH themselves.

[9] The study program Linguistics and Digital Humanities fulfills the intention of the Czech Republic in the Olomouc Region to support the employment of university graduates in the areas affected by the Fourth Industrial Revolution. The preparation of the study program was supported by a grant from the European Union from the European Regional Development Fund, in the INTERREG program. The project is entitled "Digital Humanities for the Future", CZ.11.3.119 / 0,0 / 0,0 / 18_031 / 0002217. The project is implemented in cooperation with the University of Wroclaw, its Institute of Information Studies and Librarianship. Polish partners will also implement a part of the teaching in application courses.

The study program *Linguistics and Digital Humanities* is based on three pillars: (1) theory of Digital Humanities, (2) methods available for Digital Humanities, and (3) specific applications of Digital Humanities mainly from the field of general linguistics and individual philologies.

The theory pillar (1) is the subject of basic compulsory courses of the theoretical basis. These courses are: Theory of Humanities 1 and 2, Semiotics, Introduction to General Linguistics, The Past of Database Systems and The Present of Database Systems, Critical Discourse Analysis, and Forensic Linguistics.

Being a post-discipline, DH share most of their theoretical foundations with other disciplines and subdisciplines – their own autonomy is achieved through their specific interconnection / amalgamation. Theory of the Humanities 1 draws mostly from philosophy, both from the specific philosophies of the first half of the 20th century and from the philosophy of science (but also historiography, sociology of science, etc.). The connections between Humanities Theory 1 and 2 represent Michel Foucault's theoretical concepts. The Theory of Humanities 2 also draws from philosophy, namely from the critique of postmodernism and naturalism, and from the theory of dynamical systems, whose theory has become the basis for the creation of a specific theory of Digital Humanities. Introduction to General Linguistics and Semiotics represent the linguistic and general semiotic theoretical foundations of Digital Humanities. Basic theories of general linguistics, from structuralism and generativism to psycholinguistics, systems-theoretical linguistics, or cognitive linguistics, are a necessary source of concepts and hypotheses on which to understand the concepts of DH. Semiotics as a general theory of meaning is the most important qualitative contribution to the interpretation of data that make the methods of Digital Humanities more effective.

The courses Past and Present of Database Systems present various ways of organizing knowledge in the past and present. One of the key theoretical findings of DH is the idea of conceptual framing and organization of knowledge. Here, DH draw mainly from terminology and information science, but also from history and historiography. Critical Discourse Analysis and Forensic Linguistics will present two different conceptualizations of ways of analyzing language data (specific discourse and specific language corpora).

The pillar (2) is the subject of courses: Data processing in DH 1 and 2 and four courses of the profiling basis – Formal language processing 1 and 2, Quantitative language processing 1 and 2.

The Data Processing courses in DH 1 and 2 will introduce students to both the basic and advanced knowledge of statistical analysis, to the issue of data visualization, and also to interpretation of such analyses. Apart from that, the students will be acquainted with natural language processing (NLP), image processing, exploitation of social networks, etc.

The courses Formal Language Processing 1 & 2 and Quantitative Language Processing 1 & 2 will provide students with the knowledge of the basics of formal language analysis and of the creation of text-processing algorithms. The outline of the whole series is designed in such a way that the students will deal with increasingly complex tasks: from formal word processing to algorithmization. The main objective for them is to efficiently exploit prefabricated applications to solve their tasks. The series of courses is completed by the elaboration of a year project in which students apply a set of acquired skills and knowledge. The natural continuation of this project will be the project of the bachelor's thesis itself.

The pillar (3) is the subject of courses: Applied Semiotics, Text Attribution, Creative Visuality, Natural Language Processing, Mathematical Modeling 1–3, and other optional courses in the profiling basis.

The courses, which represent a certain section of possible applications of Digital Humanities, will be continuously supplemented in connection with the development of the discipline, the expansion of the study program portfolio, and the expansion and transformation of the team of lecturers researchers, and interns. Due to the existence of the doctoral study program *Linguistics and Digital Humanities*, the composition of applied courses will also be based on the teaching activity of doctoral students of the mentioned program. Last but not least, the list of applied courses will be based on the offer of our Polish colleagues from the Institute of Information Studies and Librarianship, who will be involved in teaching.

Case study 2: PhD program in Linguistics and Digital Humanities

The DH doctoral program at the Department of General Linguistics of Palacký Unviersity Olomouc reflects a specific approach to this scientific discipline, which is based on methods of quantitative linguistics. Thanks to this methodology, students are able to examine texts on the basis of a wide range of qualitative and quantitative properties and with the requirement to process big data. The students become familiar with various software programs for text analysis, one of them being the QUITA software created at the Department of General Linguistics of Palacký University Olomouc (see KUBÁT; MATLACH; ČECH, 2014). QUITA evaluates a wide range of quantitative text properties, such as entropy, type-token ratio, average word length, etc., allows work with various text transformations, such as n-grams, hapax legomena, bag-of-words model, reduction, randomization, etc., provides visual representation, and serves data mining. At the same time, the fact that it can be used for the analysis of data from genetic banks demonstrates the possibility of extending linguistic methods into new areas within DH. QUITA software is now being used by a wide range of quantitative linguists, and its use has resulted in more than 60 professional studies that have moved the research in quantitative linguistic significantly forward (LIU-LINAG, 2017; POPESCU *et al.,* 2017; CHEN; LIU, 2018; GLOGAROVÁ–KUBÁT, 2020, and many others). It has also been used in dozens of diploma theses (latest ones: ZÁVODNÍ, 2020; VARMUŽOVÁ, 2020).

The program is based on three main research orientations (profiling lines). Each student decides for a specific orientation in relation to the topic of the dissertation: The three orientations are:

a. linguistic description for analysis of digitalized text,

b. analysis of digitized text for use by humanities (philology, history, etc.),

c. linguistic analysis of genetic text.

All of these orientations involve working with data mining methods.

The first of the profiling lines of study is focused on the use of linguistic description of language, text and its properties used in combination with methods of data mining and natural language processing (automatic text attribution for example). The language features that enter the analysis include, among others, grammatical and lexical categories. Such an analysis might further yield tools for a wide variety of linguistic disciplines (text theory, stylistics, pragmatics, etc.). The aim is to associate methods that lack a uniform *tertium comparationis* – are based on a different view of language and text – but as a union can very pregnantly express the specifics of individual texts either under academic research or under assessment in the application sphere. The courses aiming at the application sphere are mostly Forensic Linguistics and Linguistic Applications – those are developed and taught in collaboration with Institute of Formal and Applied Linguistics at Charles University in Prague (taught by doctor Kateřina Lesch). The courses of Programming and Corpus Linguistics (taught by prof. Amaro from University Nova of Lisbon) are also crucial for this line of study.

The second profiling line of study is focused on the use of linguistic analysis of the text within the research of the humanities. The aim is to implement the methods of data mining based on linguistic analysis in the research in their individual disciplines. At the same time, this approach assumes that it can be supplemented by a description of the research topic from the given humanities discipline – it will associate the methodology of the given discipline with an integrated linguistic description and use them together in data mining tools. Semiotics will play a protective role here, which will enable the description of the text, cultural artifacts, social phenomena, etc. to be viewed in a uniform characteristic framework – which a successful analysis of the studied phenomena presupposes. In addition to Semiotics, Terminology represents another unifying approach to the wide range of disciplines. Terminology in the above-mentioned sense (see the previous section) is an important part of DH in the second profiling line of study in that it is a discipline with wide-range applicability across the humanities – but not only. The course of *Terminology and organization of knowledge* is taught by prof. Rute Costa from the University Nova of Lisbon.

The third line of study is focused on the transfer of linguistic methods to the analysis of genetic text. The initial premise is again a unified semiotic framework – the concept of the genetic code, the structure of the genetic text, sign, and its function. The aim is to use linguistic analysis of text together with data mining methods, in this case for the analysis of biopolymer chains – DNA / RNA and proteins. The linguistic methods included in the analysis of the genetic text will be presented in the study as corpus approaches and approaches verifying the manifestations of linguistic laws and quantitative metrics of the text in the genetic strings. Attention will be paid to the possibilities of using n-gram analysis and cluster analysis for taxonomic purposes. This line of study has been developed thanks to the cooperation with the University of Haifa in Israel, where the tradition of DNA Linguistics started decades ago and has been developing up to this day (BEREZOFSKY *et al.,* 2002; BOLSHOY, 2003). Methods from bioinformatics are used accompanied by tools from quantitative linguistics. The courses of DNA linguistics 1 and DNA linguistics 2 are taught by prof. Bolshoy from Haifa University. The Department of General Linguistics at Palacký University Olomouc also has its own tradition in analysis of genetic strings (FALTÝNEK *et al.,* 2019).

References

ALTMANN, G. (1978). Towards a theory of language. In: ALTMANN, G. (Ed.). *Glottometrika* 1. Bochum: Studienverlag Dr. N. Brockmeyer, 1978, p. 1-25.

AMARO, R. WordNet as a base lexicon model for the computation of verbal predicates. *Proceedings of GWC 2006*, Global WordNet Association Conference, 2006.

AMARO, R. Extracting semantic relations from Portuguese corpora using lexical-syntactic patterns. In: Conference LREC 2014 - 9th Language Resources and Evaluation, 2014, Reykjavik.

AMARO, R.; MENDES, S.; PALMIRA, M. Increasing Density through New Relations and PoS Encoding in WordNet.PT. *International Journal of Computational Linguistics and Applications*, v. 4, n. 1, p. 11-27, 2013.

BENEŠOVÁ, M.; FALTÝNEK, D.; ZÁMEČNÍK, L. Menzerath-Altmann Law in Differently Segmented Texts. In: BENEŠOVÁ, M.; MAČUTEK, J.; TUZZI, A. (Eds.) *Recent Contributions in Quantitative Linguistics*. Berlin: De Gruyter Mouton, 2015, p. 27–40.

BEREZOVSKY, I. N.; KIRZHNER, V. M.; KIRZHNER, A.; ROSENFELD, V. R.; TRIFONOV, E. N. Protein sequences yield a proteomic code. *Molecular Biology*, v. 36, n. 2, p. 239–243, 2002.

BOLSHOY, A. DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Applied bioinformatics*, v. 2. p. 103–112, 2003.

BUSA, R. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, v. 14, n. 2, p. 83-90, 1980.

CABRÉ, T. M. *Terminology: Theory, methods and applications*. Amsterdam: John Benjamins Publishing, 1999.

CALDARELLI, G. *Scale-Free Networks:* Complex Webs in Nature and Technology. Oxford University Press, 2007. Available in: https://EconPapers.repec.org/RePEc:oxp:obooks:9780199211517. Last accessed: 27 Apr. 2021.

CAMPBELL, R.; REIBLE, D.; TARABAN, R.; KIM, J. *More than a Dream:* The Developing Reflective Engineers through Artful Methods *(DREAM).* Project Paper presented at Gulf Southwest Section Conference, 2020. Avaible in: https://jee.org/36012. Last accessed: 27 Apr. 2021.

CHEN, R.; LIU, H. Thematic concentration as a discriminating feature of text types. *Journal of Quantitative Linguistics*, v. 25, n. 1, p. 53-76, 2018.

COLLIER, P. *The Future of Capitalism.* Facing the New Anxieties. London: Penguine Random House, 2018.

COSTA, R., SILVA. R. et al. *Glossário Colaborativo COVID-19.* Available in: https://www.lexonomy.eu/ec25mm79/ Last accessed: 27 Apr. 2021.

DERRIDA, J. *Of Grammatology.* London: Johns Hopkins University Press, 1976.

EUROPEAN COMMISSION. *Strategic Plan 2020-2024 – Education, Youth, Sport and Culture.* 2020. Available in: https://ec.europa.eu/info/publications/strategic-plans-2020-2024_en Last accessed: 27 Apr. 2021.

EUROPEAN COMMISSION. *Strategic Plan 2020-2024 – Research and Innovation*. 2020. Available in: https://ec.europa.eu/info/publications/strategic-plans-2020-2024_en Last accessed: 27 Apr. 2021.

EUROPEAN COMMISSION. *The Digital Europe Programme*. Available in: https://ec.europa.eu/digital-single-market/en/europe-investing-digital-digital-europe-programme Last accessed: 27 Apr. 2021.

FALTÝNEK, D.; MATLACH, V.; LACKOVÁ, Ľ. Bases are Not Letters: On the Analogy between the Genetic Code and Natural Language by Sequence Analysis. *Biosemiotics* 12, p. 289–304, 2019. Available in: https://doi.org/10.1007/s12304-019-09353-z Last accessed: 27 Apr. 2021.

FALTÝNEK, D., MATLACH, V. (2021). Forthcoming

FERRER-I-CANCHO, R., SOLÉ, R. V. The small world of human language. *Proceedings of the Royal Society B. London*, 268, p. 2261–2265, 2001.

FOUCAULT, M. *The Archaeology of Knowledge*. New York: Pantheon Books, 1972 [1969].

FOUCAULT, M. *The order of things:* An archaeology of the human sciences. New York: Vintage Books, 1994[1966].

GIERE, R. N. *Scientific Perspectivism*. Chicago: The University of Chicago Press, 2006.

GLOGAROVÁ, J. D., & KUBÁT, M. Srovnávací frekvenční analýza exilových projevů Klementa Gottwalda a Edvarda Beneše z let 1939-1945. *Slovo a Slovesnost*, v. 81, n. 1, p. 65-77, 2020.

GRZYBEK, P. Introductory Remarks: On the Science of Language in Light of the Language of Science. In: Grzybek, Peter (Ed.) *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht: Springer, 2006, p. 1-14.

HERDAN, G. *The Advanced Theory of Language as Choice and Change*. Berlin: Springer-Verlag, 1966.

HŘEBÍČEK, L. Some aspects of Power Law. *Glottometrics*, v. 6, p. 1-8, 2003.

JANEČKOVÁ, B. A.; TICHÁ, A.; FIEDLER, J. *Třikrát o autorství*. Olomouc: Palacký University Press, 2021.

JENSEN, K. E. Linguistics in the digital humanities: (computational) corpus linguistics. MedieKultur: Journal of Media and Communication Research, v. 30, n. 57, p. 115-134, 2014. Available in: https://doi.org/10.7146/mediekultur.v30i57.15968 Last accessed: 27 Apr. 2021.

KELLERT, S. H. *Borrowed Knowledge: Chaos Theory and the Challenge of Learning across Disciplines*. Chicago: The University of Chicago Press, 2008.

KLEIN, N. *No Is Not Enough:* Resisting Trump's Shock Politics and Winning the World We Need. Toronto: Knopf Canada, 2017.

KÖHLER, R. *Quantitative Syntax Analysis*. Berlin: De Gruyter, 2012.

KÖHLER, R. *Synergetic Linguistics*. In: KÖHLER, R.; ALTMANN, G.; PIOTROWSKI, R. G. (Eds.) *Quantitative Linguistics:* An International Handbook. Berlin: Walter de Gruyter, 2005, p. 760–774.

KÖHLER, R. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer, 1986.

KOŘENSKÝ, J. Procesuální gramatika v kontextu současných tendencí lingvistického myšlení. *Slovo a Slovesnost*, v. 64, p. 1-7, 2003.

KUBÁT, M.; MATLACH, V.; ČECH, R. *QUITA. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag, 2014.

LAUDAN, L. Dissecting the Holist Picture of Scientific Change. In: CURD, M., COVER, J. A. (Eds.) *Philosophy of Science:* The Central Issues. New York: W. W. Norton & Company, p. 139-169, 1998.

LEE, K.-F. *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston, Mass: Houghton Mifflin, 2018.

LESCH, K. *On the Linguistic Structure of Emotional Meaning in Czech*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, 2015.

LIU, H.; LIANG, J. (Eds.). Motifs in Language and text. v. 71. Berlin: De Gruyter Mouton, 2017.

MEYER, P. Laws and Theories in Quantitative Linguistics. *Glottometrics*, v. 5, p. 62-80, 2002.

POPESCU, I. I.; MIANGAH, T. M.; GNATCHUK, H.; CECH, R.; BODOC, A; ALTMANN, G. On Rank-Frequency Distributions in Poetry. *Glottometrics*, v. 38, p. 30-54, 2017.

QUINE, W. V. O. *From a Logical Point of View*. New York: Harper, 1953.

ROSS, A. *The industries of the future*. New York: Simon & Schuster, 2016.

SALGADO, A.; COSTA, R. (2019). *LeXmart: a smart tool for lexicographers*. In: Electronic lexicography in the 21st century (eLex 2019): Smart lexicography, 2019, Sintra.

SALGADO, A.; COSTA, R. (2020). O projeto "Edição Digital dos Vocabulários da Academia das Ciências": o VOLP-1940. *Revista Da Associação Portuguesa De Linguística*, v. 7, p. 275-294, 2020. Available in: https://doi.org/10.26334/2183-9077/rapln7ano2020a17 Last accessed: 27 Apr. 2021.

SMITH, B. C. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge: The MIT Press, 2019.

VARMUŽOVÁ, B. Určování autorství Slezských písní. Olomouc, [cit. 2021-01-17]. Dostupné z: https://theses.cz/id/3p64qh

VEYNE, P. *Foucault*: His Thought, His Character. Cambridge: Polity Press, 2010.

ZÁVODNÍ, Š. Proměny verbálního projevu účastnic výchovně vzdělávacího procesu ve vztahu k biorytmům. Hradec Králové, 2020 [cit. 2021-01-17]. Dostupné z: https://theses.cz/id/divq2d/

Appendices

Appendix: Structure of the study plan of bachelor and doctoral programs in Linguistics and Digital Humanities (LDH) at Palacký University Olomouc

1 Bachelor program LDH: structure of the program

LDH can be studied either as a separate study program or in combination with another field of study.

In the case of an independent study program, there is a higher expectancy for the graduates to immediately transition into practice (with possible further study along the employment). In the case of combination with another study field, it is more likely that the student will follow a master's degree.

The independent study program LDH includes, in addition to the basic composition of the courses of the theoretical basis (TB) and profiling basis (PB), an elaboration of an individual project. If a student creates only one Independent Project, he/she must complete at least one Internship (10 credits). Individual projects are selected on the basis of consultation with the guarantor of the study cycle or its authorized representatives. A separate project ideally forms the basis for creation of a diploma thesis. The internship is arranged by the student with regard to their future employment.

In combination with another study field, LDH can be studied either as a *maior* or as a *minor*. For the *maior variant,* the student must complete the compulsory courses of the theoretical basis (TB, 54 credits) and profiling basis (PB, 18 credits). The student must also complete a diploma module (15 credits) and prepare a bachelor thesis in the field of LDH. The *minor* variant differs only in absence of the diploma module.

The three possible study programs can be schematized in terms of ECTS as follows:

Table 1: ECTS differences of bachelor LDH program in three possible varieties

Program	Independent	Maior 60%	Minor 40%
1st year	60	36	24
2nd year	60	36	24
3rd year	60	36	24
Total	180	108	72

The whole structure of the study program is schematized in table 2.

Table 2: Bachelor study program LDH

Course title	**ECTS**	**Year/ semester**	**profiling basis**
Theory of Humanities 1	5	1/WS	TB
Semiotics	6	1/WS	TB
Critical Discourse Analysis	5	1/WS	TB
Theory of Humanities 2	5	1/SS	TB
Forensic Linguistics	6	1/SS	TB
Introduction to General Linguistics	5	2/WS	TB
Data Processing in DH 1	5	2/WS	TB
Past of the Database Systems	5	2/SS	TB
Data Processing in DH 2	6	2/SS	TB
Present of the Database Systems	6	3/SS	TB
Alghoritminc Language Processing 1	4	1/SS	PB
Alghoritminc Language Processing 2	5	1/SS	PB
Alghoritminc Language Processing 3	4	2/WS	PB
Alghoritminc Language Processing 4	5	2/WS	PB
Diploma Thesis Topic	5	2/SS	
Diplomoma Seminar 1	5	3/WS	
Diploma Seminar 2	5	3/SS	
Individual Project 1	20	2/WS	PB
Individual Project 2	20	3/WS	PB
Internship 1	10	2/SS	PB
Internship 2	10	3/WS	PB
Aplied Semiotics	5	3/WS	PB
Text Atribution	5	2/WS	PB
Creative Visuality	5	1/SS	PB

Natural Language Processing	5	2/WS	PB
Mathematical Text Modelling 1	5	2/WS	PB
Mathematical Text Modelling 2	5	2/SS	PB
Mathematical Text Modelling 3	5	3/WS	PB
Science Methodlogy	4	1/WS	PB
Basis for Experimental Analysis of Language	6	1/SS	PB
Text Theory and Pragmatics	6	3/WS	PB
Introduction to Communication Theory	5	1/WS	PB
Psycholinguistics	4	–	PB
Biosemiotics	4	–	PB
Linguistic Applications	3	–	PB
Fiction and Reality Theory in Praxis	4	–	PB
Models of Linguistic Explanations	4	–	PB
Seminars of Invited Speakers	3	–	PB
Philosophy	3	1/WS	PB

2 Doctoral program LDH: structure of the program

The doctoral programs at Palacký University have a particularity of having ECTS system. Thus, every student has to acquire a particular number of credits in order to complete the study cycle. There are several modules (profiling, publications, teaching, foreign languages, etc.), each of the modules is characterized by a minimum of ECTS. The distribution of ECTS in every module is represented in Table 3.

Table 3: Doctoral study program LDH

Profiling Mandatory Courses	ECTS
Introduction to Digital Humanities 1 – Introduction to Quantitative Methods	10
Introduction to Digital Humanities 2 – Bases in DH: Text Processing and Multimedia	10
Philosophy of Science	5

Foreign Language	
To choose from the Faculty database of language courses	10

Profiling Optional Courses	
Linguistic Data Mining 1 – Data Analysis	10
Linguistic Data Mining 2 – Corpus Linguistics	10

Data Mining of Digitalised Text 1 – Introduction to Machine Learning 1	10
Data Mining of Digitalised Text 1 – Introduction to Machine Learning: NLP and Multimedia	10
DNA Linguistics 1	10
DNA Linguistics 2	10
Terminology and Organization of Knowledge	10
Presentation of Data and Access to Data	10
Python Programming	10
Biosemiotics	10
General Linguistics	10
Semiotic Approach to DH	10
Regulation of Cultural Industry and Digital European Market	10
Linguistic Applications	10
Linguistic Analysis of Historical Text: Application in History Studies and Germa Philology	10
Publication Activity	
Publication 1	10
Publication 2	20
Publication 3	30
Conference 1	5
Conference 2	10
Conference 3	20
Stay Abroad	
Stay Abroad over 30 days	20

Pedagogy Module	
Course teaching	5
Supervision of a bachelor thesis	5
Opponent to a bachelor or master thesis	3

Dissertation Module	
Quodlibet 1	5
Quodlibet 2	5
Dissertation Thesis Submission	60

According to the three profiling lines of the doctoral program, the profiling courses are:

In the specialization Linguistic data mining:
Forensic linguistics
Introduction to quantitative methods
Python programming

Data analysis
Data presentation and data access
Corpus linguistics
General Linguistics
Linguistic Applications

In the specialization of data mining of digitized text:
Linguistic analysis of historical texts – possibilities of use in German studies and History
Semiotic approach towards Digital Humanities
Introduction to machine learning
Introduction to machine learning: NLP and multimedia
Data analysis
Data presentation and data access
Python programming
Terminology and organization of knowledge
Regulation of cultural industries and the digital market in Europe

In the specialization of the DNA Linguistics:
DNA linguistics 1
DNA linguistics 2
Biosemiotics
Python programming

Submitted: 02/27/2021.
Accepted: 05/25/2021.