Co-funded by the
Erasmus+ Programme
of the European Union

i. Name of the course
Applications of Linguistics: From computational linguistics via clinical linguistics to forensic linguistics

ii. Level of the course
MA (can also be taught to advanced BA students)

iii. Workload
5 ECTS

iv. Institution
University of Graz

v. Course instructor(s)
Stefan Milosavljević

vi. Brief course description

In this semester (winter, 2021/2022), the interdisciplinary course *Applications of Linguistics: From computational linguistics via clinical linguistics to forensic linguistics* focuses on the applications of the software-based Social Network Analysis (SNA) in a variety of fields linking linguistics to both natural and social sciences.

The course is organized in six main units.

In the first part of the course, network theory and SNA are introduced and their application in different areas are discussed: in exploring social relationships (e.g. online social networks such as *Facebook* or *Twitter*), sports (e.g., networks of players based on mutual interactions), psychology (e.g., the mental lexicon as a network); epidemiology (e.g., how viruses spread through a network); movies (networks of actors), etc. This unit is meant to introduce the main concepts and methodology by using familiar examples.

The second part of the course is devoted to the application of SNA in exploring narratives. The students learn how to use the software-based SNA to analyze the basic properties of (literary) texts, such as identifying the main characters, extracting the most important segments of a given text, comparing 'fictional' and historical ('realistic') texts, and identifying the authorship. SNA is conducted in the programming language *R*. For concreteness, the SNA of the *Dictionary of the Khazars* by Milorad Pavić is used as a case-study.

The identification of authorship, which is introduced in this part of the course, anticipates the role of Forensic Linguistics, which is a topic of the fifth part of the course.
In the third part of the course, the focus is on the software-based SNA of language networks: the mental lexicon, semantic networks of the creative mind, and/or networks based on grammatical and/or derivational relations. A practical part consists in analysing a network of Serbo-Croatian prefixes, which is extracted from the database of Western South Slavic languages (WeSoSlaV), being developed within the project *Hyperspacing the verb* at the University of Graz and the University of Nova Gorica.
The fourth part of the course brings together Network Science, Computational Linguistics and Clinical Linguistics, by focusing on the software-based analysis of networks extracted from the clinical language data (autism, schizophrenia, aphasia, late-talking children, children with Down syndrome). In this part of the course, the focus is on discussing the published papers on the topic. The students learn how computational and network science tools and metrics (familiar from the previous classes) can be applied in analysing clinical cognitive networks and how these sets of tools can be used for the diagnosis and classification of language-related diseases, the quantification of condition severity, as well as for preparing treatment protocols.
In the fifth part of the course, Forensic Linguistics is introduced from the interdisciplinary perspective: the focus is on using the computational network tools and metrics in the field of author identification and plagiarism, as prominent topics in Forensic Linguistics.
Each of the above described five parts of the course consists of three main building blocks: i) the instructor's brief introduction of the phenomenon, ii) reading an article related to a given topic, and iii) practical exercises (modelling the relevant networks, preparing the material using online corpora, data analysis in the programming language *R*).
At the transition point from the first to the second unit, the students are introduced with a form of a final report that they handle over at the end of the course, in which they describe a computational analysis of a network based on some language/textual data. They choose (in consultation with the instructor) some narrative and/or language phenomenon that can be analyzed from the SNA perspective. During the entire course, a part of each class is devoted to discussing their projects.
In the final, sixth part of the course, the students work exclusively on finalising their projects and writing up the final report, which they submit at the end of the course (or after the course, depending on the established deadline).

vii. Research related subject

Language and narrative networks

viii. Data the students work with

Data obtained from corpora: Stanford Large Network Dataset Collection (http://snap.stanford.edu/data/index.html); Project Gutenberg (www.gutenberg.org); BCMS corpora srWaC, hrWac, bsWac, meWac (https://www.clarin.si/noske/index.html); the databases of the project Hyperspacing the verb (in preparation).

ix. Topics

A: Research design

A1: General research design

[Teaching materials]
quanteda: Quantitative Analysis of Textual Data (tutorial in English): http://quanteda.io/
Network Analysis and Visualization with *R* and *igraph* (tutorial in English): https://kateto.net/networks-r-igraph
Network science (an interactive online book in English): http://networksciencebook.com/

A2: Adapting the general research design to the specific topic of interest

Identifying the key properties of textual/narrative networks: the main characters, the most important segments of texts, authorship, etc.; analyzing the structure of mental lexicon by extracting the most prominent types of words and their relations; learning about the possibilities of diagnosing different types of language impairments based on network properties; detecting plagiarism and identifying authors based on network properties

A2.1: Formulation of questions and hypotheses in terms of variables

A2.2: Selection of appropriate research techniques, selection and creation of corresponding data sources

- Computational tools for textual analysis (the programming language *R*, especially the library *quanteda*)
- Computational tools for SNA (especially the *igraph* package in the programming language *R*)
- Developing and exploiting databases and corpora (e.g. extraction of language networks from the BCMS corpora *srWac* and *hrWac*, as well as from the databases within the project *Hyperspacing the verb*)

A2.3: Identifying the optimal data analysis method

A2.4: Inferring theoretical consequences from the specific data analysis results

A3: Adapting the research design to the available research infrastructures

Familiarising with the available computational tools relevant for SNA of narratives and language
Familiarising with the type of data extractable from the available datasets

A3.1 Selection of optimal research techniques, selection and creation of corresponding data sources (see also A2.3)

- computational analysis of textual data, Social Network Analysis (SNA);
- modelling textual and language networks, selecting optimal computational tools (packages, libraries) for extracting the relevant material; choosing appropriate SNA metrics for performing the analysis and visualising the results

A4: Research reporting

Identifying optimal formats for presenting different types of research outcomes and making them available to a wider audience: using the repositories for the public storing of the relevant code (e.g., github.com), using appropriate visualisation techniques (graphs, charts, tables, etc.) depending on the type of a written form (reports, presentations, posters, articles)

A4.1 Presentation modes for research reporting (short oral presentation, poster, squib, report, article etc.)

A4.2 Established procedures and conventions in research reporting, such as:

- the ordering of thematic units in an article/squib/report,
- organization of the presentation,
- amount of text and graphical items on a poster (including text size),
- amount of text and graphical items on a slide/handout,
- terminology,
- citing conventions

- B: Infrastructures & techniques

B1: For obtaining literature

[GENERAL-PURPOSE REPOSITORY] ResearchGate, Google Scholar, Academia.edu, slideshare.net
[DISCIPLINARY REPOSITORY] lingbuzz, Stanford Large Network Dataset Collection (http://snap.stanford.edu/data/index.html), Project Gutenberg (www.gutenberg.org)

B2: For obtaining, sharing and managing data

Use of the repositories for sharing the relevant code (e.g. guthub.com), advanced use of the available corpora for extracting narrative and language networks
B2.1: Definition of research infrastructures and the main concepts around **data interoperability**, such as **data**, **metadata** and **standards**
B2.2: Platforms and repositories ■ **General-purpose repositories** and **disciplinary repositories** ● [GENERAL-PURPOSE REPOSITORY] FigShare, github.com, pixabay.com; ● [DISCIPLINARY REPOSITORY] CLARIN; Stanford Large Network Dataset Collection; Project Gutenberg
B2.3:Identifying, collecting, creating and/or using relevant data for research projects ● Searching, identifying and selecting relevant corpora from language resources platforms and repositories hosting them ● Citing linguistic data sets as appropriate. ● Depositing research data in a **certified repository** and selecting an appropriate licence for sharing their data ● The **versioning** policy of repositories ● Familiarity with online survey tools
B2.4: **Data management plan** ● Understanding the **data lifecycle** ● Understanding how to generate data, analyse and handle it ● Understanding the **legal and ethical issues** around data generation and use (e.g. licensing, GDPR compliance, anonymisation, the importance of FAIR principles and Open Access). ● Secure storage and backup of research data ● Documenting workflows and what metadata to use to describe the nature of the data based on existing standards. ● What data needs to be destroyed, preserved in a data repository and made available for reuse
B3: For analysing data
B3.1: Software for computational text analysis (the programming language *R*, library *quanteda*)
B3.2: Software for SNA (the programming language *R*, package *igraph*)
C: Subject-specific topics
C1: Basic concepts, terms and methods in Computational Linguistics and their application

C2: Basic concepts, terms and methods in Clinical Linguistics and their application
C3: Basic concepts, terms and methods in Forensic Linguistics and their application
C4: Basic concepts, terms and methods in Network Science and their application
C5: Basic concepts, terms and methods in Social Network Analysis and their application

x. Learning outcomes

A: Research design
A1: Students will be able to make an overview of the general research design.
[Teaching materials] *quanteda*: Quantitative Analysis of Textual Data (tutorial in English): http://quanteda.io/ Network Analysis and Visualization with *R* and *igraph* (tutorial in English): https://kateto.net/networks-r-igraph Network science (an interactive online book in English): http://networksciencebook.com/
A2: Students will be able to create a suitable research design for the specific topic of interest.
Students will be able to identify the key properties of textual networks (the main characters, the important segments of texts, autorship), to analyze the structure of mental lexicon by extracting the most prominent types of words and their relations, to recognize the network properties of textual/language data that may be relevant in diagnosing different types of language impairments, in detecting plagiarism, and in identifying the authorship.
A2.1: Students will be able to formulate questions and hypothesis in terms of variables.
A2.2: Students will be able to select optimal research techniques, and create corresponding data sources • Computational tools for textual analysis (the programming language *R*, especially the library *quanteda*) • Computational tools for SNA (especially the *igraph* package in the programming language *R*) • Developing and exploiting databases and corpora (e.g. extraction of language networks from the BCMS corpora *srWaC* and *hrWaC*, and from the databases within the project Hyperspacing the verb).

A2.3: Students will be able to select and implement the optimal data analysis method.
A2.4: Students will be able to infer theoretical consequences from the specific data analysis results.
A3: Students will be able to adapt a research design to the available research infrastructures.
Students will be familiar with the available computational tools relevant for the language and texts analysis from the perspective of network theory; Students will be familiar with the type of data extractable from the available datasets.
A3.1 Students will be able to select optimal research techniques and data sources, in particular: • computational analysis of textual/language data, Social Network Analysis (SNA); • modelling textual and language networks, selecting optimal computational tools (packages, libraries) for extracting the relevant material; choosing appropriate SNA metrics for performing the analysis and visualising the results.
A4: Students will be able to report on their performed research in accordance with standards and conventions in the field.
Students will be able to identify optimal formats for presenting different types of research outcomes and make them available to a wider audience: to use the repositories for the public storing of the relevant code (e.g. github.com), and to apply appropriate visualisation techniques (graphs, charts, tables, etc.) depending on the type of a written form (reports, presentations, posters).
A4.1 Students will be able to select and implement different presentation modes for research reporting (short oral presentation, poster, squib, report, article etc.).
A4.2 Students will be able to implement established procedures and conventions in research reporting, such as: • the ordering of thematic units in an article/squib/report, • organization of the presentation, • amount of text and graphical items on a poster (including text size), • amount of text and graphical items on a slide/handout, • terminology, • citing conventions.
B: Infrastructures & techniques

B1: Students will be able to identify and apply suitable infrastructures & techniques for obtaining literature.
• [GENERAL-PURPOSE REPOSITORY] ResearchGate, Googlescholar, Academia.edu, • [DISCIPLINARY REPOSITORY] lingbuzz, Stanford Large Network Dataset Collection (http://snap.stanford.edu/data/index.html), Project Gutenberg (www.gutenberg.org).
B2: Students will be able to identify and apply suitable infrastructures & techniques for obtaining, sharing and managing data.
[GENERAL-PURPOSE REPOSITORY] ResearchGate, Google Scholar, Academia.edu, slideshare.net, github.com [DISCIPLINARY REPOSITORY] lingbuzz.
B2.1: Students will understand what research infrastructures are, and the main concepts around **data interoperability**, such as **data**, **metadata** and **standards**.
B2.2: Students will be able to identify suitable platforms and repositories, and to understand the difference between **general-purpose repositories** and **disciplinary repositories**.
B2.3: Students will be able to identify, collect, create and/or use relevant data for their research projects, and • to cite linguistic data sets as appropriate, • to deposit their research data in a **certified repository** of their choice and select an appropriate licence for sharing their data, • to understand the **versioning** policy of the repository, • to use online survey tools.
B2.4: Students will be able to create a **data management plan** • Understand the **data lifecycle** • Understand how to generate data, analyse and handle it • Understand the **legal and ethical issues** around data generation and use (e.g. licensing, GDPR compliance, anonymisation, the importance of FAIR principles and Open Access). • Know how to securely store and backup their research data • Know how to document their workflows and what metadata to use to describe the nature of the data based on existing standards. • Know what data needs to be destroyed, preserved in a data repository and made available for reuse.
B3: Students will be able to identify and apply suitable infrastructures & techniques for analysing data.

B3.1: Software for computational text analysis (the programming language *R*, library *quanteda*).
B3.2: Software for SNA (the programming language *R*, package *igraph*).
C: Subject-specific learning outcomes
C1: Students are familiar with the basic concepts, terms and methods in Computational Linguistics and are able to apply them in specific problem-solving.
C2: Students are familiar with the basic concepts, terms and methods in Clinical Linguistics and are able to apply them in specific problem-solving.
C3: Students are familiar with the basic concepts, terms and methods in Forensic Linguistics and are able to apply them in specific problem-solving.
C4: Students are familiar with the basic concepts, terms and methods in Network Science and are able to apply them in specific problem-solving.
C5: Students are familiar with the basic concepts, terms and methods in Social Network Analysis and are able to apply them in specific problem-solving.

xi. Overview of evaluation

Rubric	Weighing
Participation in classes incl. homework (mini-projects related to each of the topics, mini-presentations, discussing the reading material)	40 %
Final written report	60 %

xii. Reading materials

Beckage, N. M., & Colunga, E. (2016). Language Networks as Models of Cognition: Understanding Cognition through Language. In A. Mehler, L. Andy, B. Sven, P. Blanchard, & B. Job (Eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks* (pp. 3–28). Springer.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Castro, N., Stella, M., & Siew, C. S. Q. (2020). Quantifying the Interplay of Semantics and Phonology During Failures of Word Retrieval by People With Aphasia Using a Multiplex Lexical Network. *Cognitive Science*, *44*(9).

Cummings, L. (2013). Clinical linguistics: state of the art. *International Journal of Language Studies*, *7*(3), 1–32.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature*. Springer.

Kenett, Yoed N., and Miriam Faust (2019). Clinical Cognitive Networks: A Graph Theory Approach. In M. S. Vitevitch (ed.), *Network Science in Cognitive Psychology*, 136–165. Routledge.

Olsson, J. (2008). *Forensic Linguistics: Second Edition*. Continuum.

Raj S., Kannan B., Jagathy Raj V. P. (2021) Significance of Network Properties of Function Words in Author Attribution. In: Satapathy S., Zhang YD., Bhateja V., Majhi R. (eds.) Intelligent Data Engineering and Analytics. Advances in Intelligent Systems and Computing, vol 1177. Springer, Singapore.

Stewart, L. L. (2006). Computational Stylistics. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics*. Elsevier.