ΓΕ 77 Computational Linguistics

Essays/ Project Topics

1. General Information

- Instead of EXAMS
- 2.000 to 3.000 words +15% (depends on the topic)
- Essay (1-student) OR Project (group of 2-3 students)
- Starting date: 2 Apr 2018
- Topic notification date: until 18 Apr 2018
- (Short) Presentation date: Last lecture
- Deadline: 20 Jun 2018

2. Summary

- (Research | Survey | Literature Review) Papers
- Existing NLP Tools (testing | evaluation)
- Annotation Project
- Corpus data Project
- Phonological, Morphological, Syntactic, Semantic Project
- Language Learning Platform evaluation

Languages: English or Greek or another language

3. Essay Topics

- You select an area and I will recommend some articles/books to get you started. One book chapter or big article is going to be the core part of an essay; several articles will be the additional and supplementary bibliography.
- Research Areas: Machine Translation, Machine Learning, Summarization, Question Answering, Sentiment Analysis, Information Extraction, Reference Resolution, Speech Analysis/ Synthesis, Thesauri and Ontologies, Building Corpora, Corpus annotation, Language Learning Platforms, etc.

You should use examples from English plus another language. Also you can have joint research areas. You can check the book of Speech and Language Processing for additional research areas.

4. Project Topics

4.1. Annotation Project (APHASIA) OR any other spoken data (even video)

- Using a tool for multi-level annotations (ELAN), you will annotate aphasic speech data based on an annotation template.
  - Task 1: Multi-level annotation of a file (2-4 minutes).
  - Task 2: Point out the linguistic errors and speech events.
  - Task 3: Present the aphasic speech and its characteristics/parameters based on annotations.
  - Task 4: Short evaluation of the tool and the annotation template.

Refs

Ide, Nancy, and Keith Suderman. 2014. “The linguistic annotation framework: A standard for annotation interchange and merging.” Language Resources and Evaluation 48:395–418.

MacWhinney, Brian, Fromm, Davida, Forbes, Margie, and Audrey Holland. 2011. “AphasiaBank: Methods for studying discourse.” Aphasiology 25: 1286–1307.

MacWhinney, Brian, Fromm, Davida, Holland, Audrey, and Margie Forbes. 2012. “AphasiaBank: Data and methods.” Στο Nicole Müller, and Martin J. Ball (επιμ.), Methods in Clinical Linguistics, 31–48. New York: Wiley.

Varlokosta S., Stamouli S., Karasimos A., Markopoulos G., Kakavoulia M., Nerantzini M., Pantoula A. & V. Fyndanis (2016). A Greek Corpus of Aphasic Discourse: Collection, Transcription and Annotation Specifications. In Proceedings of LREC 2016 Workshop: RaPID 2016, pp. 14-21.

Wittenburg, Peter., Brugman, Hennie, Russel, Albert, Klassmann, Alex, and Han Sloetjes. 2006. “ELAN: a Professional Framework for Multimodality Research.” Στο Proceedings of the 5 th International Conference on Language Resources and Evaluation, 1556–1559.

Βαρλοκώστα Σ., Σταμούλη Σ., Καρασίμος Α., Μαρκόπουλος Γ., Κακαβούλια Μ., Νεραντζίνη Μ., Φυνδάνης Β., Παντούλα Α., Οικονόμου Α. & Α. Πρωτόπαπας (2017). Ελληνικο σωμα κειμενων αφασικου λογου: μελετη, σχεδιασμος και πολυεπιπεδη επισημειωση. Στο Α. Χριστοφίδου (εκδ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών (Όψεις της Σωματοκειμενικής Γλωσσολογίας: Αρχές, εφαρμογές, προκλήσεις)*, Τεύχος 14^{o}. Αθήνα – Ακαδημία Αθηνών, Κέντρον Ερεύνης Επιστημονικών Όρων και Νεολογισμών.

Βαρλοκώστα, Σπυριδούλα, Καρασίμος, Αθανάσιος, Σταμούλη, Σπυριδούλα, Μαρκόπουλος, Γεώργιος, Κακαβούλια, Μαρία, Γούτσος, Διονύσης, Νεραντζίνη, Μιχαέλα, Φυνδάνης, Βαλάντης, και Αικατερίνη Παντούλα. 2013. Οδηγός Επισημείωσης του Ελληνικού Σώματος Κειμένων Αφασικού Λόγου. ΘΑΛΗΣ «Επίπεδα διαταραχής του λόγου ελληνόφωνων ατόμων με αφασία: σχέσεις με ελλείμματα επεξεργασίας, εγκεφαλική βλάβη και προσεγγίσεις θεραπείας». Αθήνα: ΕΚΠΑ.

Μαρκόπουλος, Γ. & Καρασίμος, Α. (2017). Πολυεπίπεδη επισημείωση του Ελληνικού Σώματος Κειμένων Αφασικού Λόγου. Στο Proceedings of 12th International Conference of Greek Linguistics (ICGL12), σσ. 725-740. Berlin.

4.2. Phonetics/ Phonology Project

Using the PRAAT application, you will test and compare words with secondary stress in Greek.

- Task 1: Record various sentences containing words with secondary stress plus several mock sentences.
- Task 2: Measure some parameters (formats, length, etc.)
- Task 3: Discussion about this phenomenon.

Refs

Arvaniti Amalia (1999), Secondary stress: evidence from Modern Greek.

Botinis A. (1989), Stress and Prosodic Structure in Greek: A Phonologican, acoustic, physiological and perceptual study, Lund University Press.

Malikouti-Drachman and Drachman (1980), Slogan chanting and Speech rhythm in Greek, In Phonologica, 275-283, W. Dressler (publ.).

Nespor Marina – Ralli Angela (1996), Morphology – Phonology interface: Phonological domains in Greek compounds, Linguistic Review 13, England.

Nespor Marina (1999), Φωνολογία, μτφρ. Α. Ράλλη, εκδ. Πατάκης.

Setatos M. (1969), Phonological problems of Modern Greek Koine, Thessalonica.

Σετάτος (1974), Φωνολογία της Κοινής Ελληνικής, Αθήνα, Παπαζήσης.

+ some new and updated articles and researches about Greek stress are necessary!

4.3. A specific NLP/Corpus application tool will be tested and evaluated.

I will provide you with a list of several online tools (syntactic parser, morphological parser/tagger, etc.)

Tasks: Depend on the tool (ambiguity is a good parameter to be tested).

- Present the tool
- Test the tool
- Evaluate the tool
- Discuss about the results.

List of several tools

Online Natural Language Processing Demos

List of corpora

- British National Corpus (BNC): http://www.natcorp.ox.ac.uk/
- BNCWeb: http://bncweb.lancs.ac.uk/cgibinbncXML/BNCquery.pl?theQuery=search&urlTest=yes
- The Brigham-Young version of the BNC, http://corpus.byu.edu/
- Corpus of Contemporary American English (COCA), http://corpus.byu.edu/
- Corpus of Historical American English (COHA), http://corpus.byu.edu/
- Time Magazine Corpus, http://corpus.byu.edu/
- BYU Corpus Portal, https://corpus.byu.edu/
- Google Books for American English, http://corpus.byu.edu/
  Links to various learner corpora around the world (some of them are not freely available): http://www.uclouvain.be/en-cecl-lcworld.html
- Voice: Vienna-Oxford International Corpus of English: http://www.univie.ac.at/voice/page/corpus_availability
- CORPS: A Corpus of tagged Political Speeches http://hlt.fbk.eu/corps
- American National Corpus (comparable across genres to the BNC): http://www.americannationalcorpus.org/OANC/index.html
- Junk email corpus: http://clg.wlv.ac.uk/resources/junk-emails/index.php
- MICASE (Michigan Corpus of Academic Spoken English): http://quod.lib.umich.edu/m/micase/
- Michigan Corpus of Upper-Level Student Papers: ttp://searchmicusp.elicorpora.info/simple/
- British Academic Spoken English corpus (BASE): http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/
- Business Letter Corpus (U.S. and U.K. samples) &Online KWIC Concordancer (letters and novels): http://www.someya-net.com/concordancer/
- The Santa Barbara Corpus of Spoken American English: http://www.linguistics.ucsb.edu/research/sbcorpus.html
- European Parliament Proceedings Parallel Corpus: http://www.statmt.org/europarl/
- Archive of Prime Minister's Questions: March 2011: http://www.parliament.uk/business/news/2011/april/pmqs-revisited-feb-march/
- Parliament Archive: http://www.parliamentlive.tv/Main/Archive.aspx
- Multilingual and Parallel Corpora: The Polylingual Document Collection (collection of newspaper articles from financial newspapers in 6 languages and the Multilingual

Parallel Corpus consisting of translated data in nine European languages: http://www.elda.org/catalogue/en/text/W0023.html

- ELFA: English as a Lingua Franca in Academic Settings. Corpus of spoken academic English as a Lingua Franca (not free, more information available online) http://www.helsinki.fi/englanti/elfa/elfacorpus.html#licence

Learner Corpora (data produced by foreign language learners)

- ICLE, the International Corpus of Learner English contains argumentative essays written by higher intermediate to advanced learners of English from several mother tongue backgrounds (not available online) http://www.uclouvain.be/en-cecl-icle.html
- FRIDA, The French Interlanguage Database contains texts written by learners of French as a foreign language (not available online) http://www.uclouvain.be/en-cecl-frida.html
- LINDSEI, Louvain International Database of Spoken English Interlanguage. Contains oral data produced by advanced learners of English from several mother tongue backgrounds. (not available online) http://www.uclouvain.be/en-cecl-lindsei.html
- LONGDALE, Longitudinal Database of Learner English. a large longitudinal database of learner English containing data from learners from a wide range of mother tongue backgrounds (not available online) http://www.uclouvain.be/en-cecl-longdale.html
- VESPA, The Varieties of English for Specific Purposes dAtabase (VESPA) learner corpus. a large corpus of English for Specific Purposes texts written by L2 writers from various mother tongue backgrounds. (not available online) http://www.uclouvain.be/encecl-vespa.html

Pedagogical Corpora

- TeMa; pedagogical corpus, which contains pedagogical materials, for instance textbook materials http://www.uclouvain.be/en-cecl-tema.html

Multilingual Corpora

- PLECI. The Poitiers-Louvain Échange de Corpus Informatisés. a large bidirectional English-French translation corpus that includes literary prose and newspaper articles. (not available online) http://www.uclouvain.be/en-cecl-pleci.html
- MULT-ED. The Multilingual Editorial Corpus is a multilingual comparable corpus of newspaper editorials written in English, Dutch, French and Swedish. (not available online) http://www.uclouvain.be/en-cecl-multed.html

TV Scripts

- Scripts of different American TV shows: http://www.script-o-rama.com/tvscript.shtml
- Scripts of the American TV show Seinfeld: http://www.seinfeldscripts.com/
- Scripts of the TVshow Roswell, German translation: http://www.tvscripte.de/roswell/index.html

List of several corpora: http://martinweisser.org/corpora_site/online_corpora.html

Refs

Based on the chosen NLP/Corpus application tool.

4.4. Corpus data project

- Choose a specific Corpus (English or Greek or any other language) and concordance the data.
- I will provide you with a list of several corpora
    - Task 1: Check a specific word group (phrasal verbs, NPs, etc)
    - Task 2: Extract some result
    - Task 3: Present the corpus and discuss its potential (based on your results)

List of (English) Corpora

Check the list from 4.3 project proposal

Refs

Aijmer, K. (2009). Corpora and language teaching. Amsterdam: John Benjamins.

Aston, G. (2001). Learning with corpora. Houston: Athelstan.

Aston, G., S. Bernandini, & D. Steard (eds) (2004). Corpora and language learners. Amsterdam: John Benjamins.

Boulton, A. (2008). DDL: reaching the parts other teaching can't reach? In A. Frankenberg Garcia (ed.) Proceedings of the 8th Teaching and Language Corpora Conference. Lisbon, Portugal: Associacao de Estudos e de Investigacao Cientifica do ISLA_Lisboa, pp. 3844.

Burnard, L. & T. McEnery (eds) (2000). Rethinking language pedagogy from a corpus perspective. Frankfurt: Peter Lang.

Gabrielatos, C. (2005). Corpora and language teaching: just a fling or wedding bells? Teaching English as a Second Language – Electronic Journal, 8/4, p. 135. http://teslej.org/ej32/a1.html

Granger, S., J. Hung & S. Petch Tyson (eds) (2002). Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: John Benjamins.

Hadley, G. (2002). Sensing the winds of change: an introduction to datadriven learning. RELC Journal, 33/2, 99124 http://www.nuis.ac.jp/~hadley/publication/windofchange/windsofchange.htm

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Krieger, D. (2003). Corpus linguistics: what it is and how it can be applied to teaching. Internet TESL Journal, 9/3. http://iteslj.org/Articles/KriegerCorpus.htm.

McCarthy, M. (2004). Touchstone: From corpus to coursebook. Cambridge: CUP. http://www.cambridge.org/us/esl/Touchstone/teacher/images/pdf/CorpusBookletfinal.pdf

McEnery, T., R. Xiao & Y.Tono (eds). (2006) Corpus-based language studies. An advanced resource book. Oxon: Routledge.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins.

O'Keefe, A., M. McCarthy, & R. Carter (2007). From corpus to classroom: Language use and language teaching. Cambridge: CUP.

Partingon, A. (1998). Patterns and meanings. Amsterdam: John Benjamins.

Sinclair, J. McH. (ed). (2004) How to use corpora in language teaching. Amsterdam: John Benjamins.

West, M. (1953). A General Service List of English Words. London: Longman. Available at John Bauman's website: http://jbauman.com/aboutgsl.html

4.5. Language Learning Platforms

From a list of language learning platforms, you will test its sufficiency.

- Task I: Present and test the platform.
- Task II: Analyse your results
- Task III: Discuss about the effectiveness of the platform and your results.

List of LLPs

- Duolingo (since 2009 | iPhone App of the Year 2013 and Google's Best of the Best 2014 | 22 languages)
- Busuu (since 2008 | 12 languages)
- LiveMocha (since 2007 |it gives you access to natives speakers, teachers, language enthusiasts, and language experts around the world from over 190 countries. The community encourages language learning through interaction)
- LingQ (since 2007 |it has cartoonish, children's-type interface, it can be used by any language learner of any age — and at any stage of language learner (beginner to advanced)
- Byki (since 2008 | 76 languages | Also a pro version | focuses on teaching adults vocabulary)
- Lang 8 (since 2006 | a community of native speakers who connect with you and correct what you write; this is a platform for those speakers who are already proficient in writing and reading a foreign language)
- Lingualia (since |AI "Lingu" treats you like a student. Lingu will make sure that you improve on areas that are challenging for you and stay motivated to learning your language)
- Papora (since 2009 | rather than focusing purely on vocabulary, Papora will also incorporate grammar so that you can form proper sentences; the "bite-sized" lessons are very easy with excellent audio)
- Digital Dialects (since 2016| it is focused on providing users with games to learn a new language with the help of phrases, vocabulary, numbers, spelling, verb conjugations, and alphabets)
- Memrise (since 2010 |it has over 300,000 courses; if you're also interested in learning more about a specific country's history, culture, or geography, it is offered)

Refs

Holmberg, B. (1986). Growth and Structure of Distance Education. London: Croom Helm.

Peters, O. (1988). Distance Teaching and Industrial Production: A Comparative Interpretation in Outline. In D. Sewart, D. Keegan, & B. Holmberg, Distance Education: International Perspectives, 95-113. New York: Routledge.

Psychogyiou, A. (2016). The effectiveness of learning a foreign language via the Duolingo application. MEd Dissertation. Hellenic Open University.
Vesselinov, R. and Grego, J. (2012). Duolingo Effectiveness Study. Final Report. Duolingo.

Wedemeyer, C. A. (1981). Learning at the Back Door: Reflections on Non-Traditional Learning in the Lifespan. Madison, WI: University of Wisconsin Press.

Ye, F. (2014). Validity, Reliability, and Concordance of the Duolingo Test. Duolingo.

4.6. FRAMENET/ WORDNET Project

Presenting and Evaluating the FrameNet or WordNet Framework.

- Task 1: Present the properties of FrameNet
- Task 2: Analysing the extracted data from FrameNet
- Task 3: Compare the errors from the data, explain them and provide the correct annotation.
- Task 4: Discuss the results from the data
- OR TASK 2-4: Present and Compare it with Related Projects (FrameNets In Other Languages, Automatic semantic role labeling (ASRL), etc.)

OR

- Task 1: Present the structure and theory behind Wordnet
- Task 2: Provided some examples and present the online database
- Task 3: Compare WordNet with other semantic networks (BabelNet, Mimida Project, MultiWordNet, etc.)

Links

https://wordnet.princeton.edu/

https://framenet.icsi.berkeley.edu/fndrupal/

Refs

Fellbaum, Chr. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

(2012). WordNet 3.0 Reference Manual. Available at: https://wordnet.princeton.edu/wordnet/documentation/

Framenet. Introduction to FrameNet (ppt). Available at: https://framenet.icsi.berkeley.edu/fndrupal/CJFFNintroPPT

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Baker, C.F. & J. Scheffczyk (2016).*FrameNet II: Extended Theory and Practice*. Available at: https://framenet.icsi.berkeley.edu/fndrupal/the_book

4.7. Conversational Agent Project

Interact with a CA or present a research paper about a CA or test a computer game CA.

I will provide you with some CAs or Computer Games.

- Tasks: depend on the CA.

List of conversational agents

- NADIA
- Chatbots
- Chat bot 2 (Game-like app)
- ReBot – Create your chatbot
- Evie
- Elbot
- L. I. C. E. Artificial Intelligence Foundation
- Cleverbot
- Mitsuku
- A list of various serious and funny chatbots (http://ai.wikia.com/wiki/List_Of_Chat_Bots)

List of Games with conversational agents

- Marvel's Guardians of the Galaxy: The Telltale Series
- Batman - The Telltale Series
- Game of Thrones
- Tales from the Borderlands
- The Wolf Among Us
- Jurassic Park
- The Waking Dead: Season one
- Mass Effect

Refs

Jurafsky & Martin (2017). *Speech and Language Processing* (Chapter 28-30).

Mendonça, V., Melo, F.S., Coheur, L. & A. Sardinha (2017).A Conversational Agent Powered by Online Learning. In S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017).

Forbell, E., Kalisch, N., Morbini, F., Christoffersen, K., Sagae, K., Traum, D. & A. Rizzo (2013). Roundtable: An Online Framework for Building Web-based Conversational Agents. *Proceedings of the SIGDIAL 2013 Conference*, pp. 372–374.

Radziwill, N. & M.Benton (2012). Evaluating Quality of Chatbots and Intelligent Conversational Agents. https://arxiv.org/ftp/arxiv/papers/1704/1704.04579.pdf

4.8. Thesaurus and Ontology

Creating and presenting a Thesaurus or Ontology.

- Task 1: Present a Thesaurus/ Ontology approach
- Task 2: Create a Thesaurus or Ontology
- Task 3: Discuss the decisions and the challenges of your thesaurus/ ontology.

List of Thesauri and Ontologies

- THEMAS (FORTH/ ΔΥΑΣ)
- OntologyOnline
- WebVOWL: Web-based Visualization of Ontologies
- OWLGrEd
- NeMO
- Wikilist of ontology editors: https://en.wikipedia.org/wiki/Ontology_(information_science)

Refs

Oberle, D., Guarino, N., & Staab, S. (2009) What is an ontology?. In: "Handbook on Ontologies". Springer(2nd edition).

Gangemi A., Presutti V. (2009). Ontology Design Patterns.[dead link] In Staab S. et al. (eds.): Handbook on Ontologies (2nd edition), Springer.

Maria Golemati, Akrivi Katifori, Costas Vassilakis, George Lepouras, Constantin Halatsis (2007). "Creating an Ontology for the User Profile: Method and Applications". In: Proceedings of the First IEEE International Conference on Research Challenges in Information Science (RCIS).

Maedche, A. & Staab, S. (2001). "Ontology learning for the Semantic Web". In: Intelligent Systems. IEEE, 16(2): 72–79.

Razmerita, L., Angehrn, A., & Maedche, A. 2003. "Ontology-Based User Modeling for Knowledge Management Systems". In: Lecture Notes in Computer Science: 213–17.

Smith, B. Ontology (Science), in C. Eschenbach and M. Gruninger (eds.), Formal Ontology in Information Systems. Proceedings of FOIS 2008, Amsterdam/New York: ISO Press, 21–35.

Staab, S. & Studer, R. (2009). Handbook on Ontologies. 2nd edition. Springer-Verlag, Heidelberg.

Uschold, Mike & Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, 11(2).