



Stephan Oepen (oe@ifi.uio.no)
Leon Derczynski (leod@itu.dk)
Filip Ginter (figint@utu.fi)
Joakim Nivre (joakim.nivre@lingfil.uu.se)
Anders Søgaard (soegaard@di.ku.dk)
Jörg Tiedemann (jorg.tiedemann@helsinki.fi)

To the
NeIC Director, and
NeIC Provider Forum

UNDERSTANDING LANGUAGE
IS THE HOLY GRAIL
OF MACHINE LEARNING.
John Giannandrea
Senior Vice President at Google

October 30, 2019

Dear colleagues:

we are writing on behalf of the Nordic research community in Natural Language Processing (NLP), the sub-discipline of Artificial Intelligence (AI) that enables computational systems to ‘make sense’ (to some degree) of human language. Large enterprises and start-ups alike are investing frantically in NLP and other AI technologies, and there is an acute danger of university and other public-sector research falling behind the sprawling corporate AI laboratories. NLP research today is data- and computing-intensive, and the competition for innovation and talent is, of course, intricately tied to the availability of advanced e-infrastructures for NLP research.

Our community has received support from NeIC (and CSC and Sigma2) over the past three years through the *Nordic Language Processing Laboratory* project (NLPL). In addition to project management and operating expenses, NeIC has contributed three person years of development effort to the project, which have been matched by our universities collectively as another three person years of in-kind contributions, as well as by CSC and Sigma2 (as additional project partners) through ‘earmarked’ computing and storage allocations. We are tremendously grateful for this support; it has been instrumental in the formation of our research community and developing an infrastructure for cross-border resource sharing—which we have dubbed the NLPL virtual laboratory (see below).

As the contract duration of the initial NeIC project is drawing towards completion, we are eager to establish a framework for sustained operation of the laboratory and our research collaboration across borders. In a nutshell, our community is very satisfied with the NLPL development: Its shared research infrastructure and annual training events are by now recognized as a ‘clearing house’ for large-scale NLP research in Northern Europe, and the initiative has already attracted additional associate members from Estonia, Iceland, and Sweden. Therefore, we are writing to ask for financial support from NeIC towards future training and networking events, and for allocations of computing and storage resources by the national providers—for continued cross-border operation of the virtual laboratory. We kindly suggest that the NeIC Director and members of the Provider Forum relay this letter appropriately within their respective organizations.

To help you consider our needs, the following paragraphs offer some background on the status of NLPL to date and our plans for the next few years. More detailed information is available on the project wiki, including on several of the sub-aspects (software, usage, community) that we summarize below:

<http://www.nlpl.eu>

Over the past three years, the project has established a largely uniform data and software environment on two superclusters (originally Abel and Taito, now migrating to Puhti and Saga). The bulk of the data are large samples of natural language texts and derived statistical models; the software comprises mostly discipline-specific tools (e.g. add-ons to common frameworks like TensorFlow and PyTorch for Natural Language Processing), of which some are developed within the consortium and used across sites.

This environment, together with our ability to share access and allocations with NLPL researchers from Denmark, Finland, Norway, and Sweden, is what we call the virtual laboratory. Being able to 'meet' on the same systems has been very productive, e.g. has enabled new collaborations and, more generally, technical knowledge exchange among NLPL users. In our experience, this kind of community-internal self-help substantially reduces the demand on system administrators, for example for custom software installations and general support.

Current NLPL resource demands are non-trivial but not overwhelming. Over the past six months, NLPL users in Norway have consumed about two million core hours (combined on Abel and Saga, including the Saga burn-in phase), and our estimate for combined billing units by NLPL users on Taito and Puhti in 2019 so far are around seven million. Storage usage to date is around ten terabytes 'on-line' (mounted on the systems) and roughly fifty terabytes on the separate NIRD storage facility.

There are around fifty active users associated with NLPL on these systems, of which some regularly compute both in Finland and Norway (while we have contained large groups of users, notably MSc students, to only one system). In April this year, we obtained statistics for software usage on Abel: Close to five percent of all `module load` operations on the system that month were for NLPL-installed and -maintained packages. We are aware of four graduate-level courses that have been taught (at Helsinki, Oslo, and Uppsala) using the NLPL laboratory, and we estimate that at last two dozens of Master and Doctoral projects have been pursued in this environment. Dozens of scientific publications over the past three years reflect large-scale experimentation in the NLPL virtual laboratory.

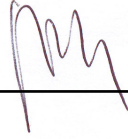
In addition to the NLPL virtual laboratory, the project has greatly contributed to community formation: We have held winter training schools (at Skeikampen in the Norwegian mountains) for the past two years, with close to fifty mostly self-funded participants in 2019. Subsets of NLPL partners have started multiple joint new research projects and submitted project proposals together. With Oslo and Helsinki universities representing our community, NLPL participates as one use case in the recent EOSC-Nordic collaboration. In early November this year, the project hosted a one-day scientific workshop at the University of Turku with about seventy participants. All in all, in our view the project has been very successful already, and to us it is highly desirable for our community to amortize the joint investments in this initiative and secure continued discipline-specific support across borders.

Regarding our plans for the future, we hope to continue the annual winter school, bringing high-profile international experts to Norway each year and providing a focused meeting place in particular for doctoral and other early-career researchers. As a minimum budget for this activity, to cover travel and accommodation of presenters as well as offer bus transfer from and to the Oslo airport, we estimate an annual cost of NOK 50,000. We would like to kindly ask NeIC to consider providing these funds, for example through an 'alumni' programme for past projects.

Regarding future collaborative research, we anticipate computing and storage needs of around five million core hours each in Norway and Finland per year and up to 100 terabytes combined over the various systems. We hope that these estimates are at a scale that the national providers could (once again) 'pledge' to our community for the next three years. To match such earmarking, our consortium will commit to maintaining and further developing the virtual laboratory, so as to continue to provide a first-line level of community-organized 'self-help' and, thus, lessen the demand for discipline-specific support by national providers.

Beyond the current NeIC project, We have formalized our initiative as a Special Interest Group (SIG) within our professional organization, the *Northern European Association for Language Technology* (NEALT). This SIG currently is chaired by Stephan Oepen (who in the past has helped coordinate the NLPL collaboration), and we therefore would like to nominate him as our representative towards NeIC and the national providers. Please do not hesitate to contact us about any additional information you might require in assessing our proposal!

UNIVERSITY OF OSLO
Department of Informatics
Boks 1080 Blindern, 0317 Oslo (Norway)
Professor Stephan Oepen
Head of Division: Machine Learning



IT UNIVERSITY COPENHAGEN
Computer Science Department
Rued Langgaards Vej 7, 2300 Copenhagen S
Assistant Professor Leon Derczynski



UNIVERSITY OF COPENHAGEN
Department of Computer Science
Universitetsparken 5, 2100 København Ø
Professor Anders Søgaard



UNIVERSITY OF TURKU
Department of Future Technologies
Vesilinnantie 5, 20500 Turku (Finland)
Associate Professor Filip Ginter

UNIVERSITY OF HELSINKI
Department of Digital Humanities
Unioninkatu 40, 00014 Helsinki
Professor Jörg Tiedemann

UPPSALA UNIVERSITY
Department of Linguistics and Philology
Box 635, 75126 Uppsala (Sweden)
Professor Joakim Nivre
Acting Head of Department

