A Friend in Need?
Research agenda for electronic Second Language infrastructure

Elena Volodina[1], **Beata Megyesi**[2], **Mats Wirén**[3],
Lena Granstedt[4], **Julia Prentice**[1], **Monica Reichenberg**[1], **Gunlög Sundberg**[3]

[1]University of Gothenburg, [2]Uppsala University, [3]Stockholm University, [4]Umeå University
[1]Box 200 40530 Göteborg, [2]Box 635 75126 Uppsala, [3]SE-10691 Stockholm, [4] 90187 Umeå
elena.volodina@svenska.gu.se

Abstract

In this article, we describe the research and societal needs as well as ongoing efforts to shape Swedish as a Second Language (L2) infrastructure. Our aim is to develop an electronic research infrastructure that would stimulate empiric research into learners' language development by preparing data and developing language technology methods and algorithms that can successfully deal with deviations in the learner language.

1. Introduction

The last decade has seen an uprise of infrastructural initiatives in language technology both at the national and multi-national levels, pointing out the need for coordinating efforts in data collection and tool development for different end-user groups. Two excellent examples of existing and continuously evolving infrastructures are Språkbanken (https://spraakbanken.gu.se/eng/research/infrastructure) and CLARIN (https://www.clarin.eu/).

In general terms, an electronic research infrastructure ideally consists of:

1. (free accessible) data in electronic format
2. technical platform for exploring the data, including tools and algorithms for data analysis, and visualization
3. a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
4. relevant expertise within the area

1.1 Societal needs

In the setting of an escalating refugee crisis in Europe and a growing number of people seeking asylum in Sweden (Migrationsverket, 2016), the need for research within second language acquisition (L2), assessment and teaching, and the evolvement of such a practice is in every way important to the Swedish society. In the recent debates, the Swedish government has been strongly encouraging immigrants to take a "fast path" to learn Swedish so that immigrants can be sooner considered for work in Sweden. However, the fast path is not a solution, according to SLA researchers. Professor Josefsson in her debate article (Josefsson, 2016) argues that the two immediate investments for improving teaching of Swedish as second language (L2) should be:

1. Development of effective IT-based solutions for use anywhere without the requirement of a teacher being present;
2. Education of a larger number of second language teachers that can offer SFI (Swedish For Immigrants) and other types of courses to greater number of immigrants, especially to those planning to take Swedish university courses as a step to validate their education.

With mature language technology tools at hand, the research infrastructure for L2 Swedish could target the first point on Josefsson's agenda and indirectly support the second one.

1.2 Second language research needs

Empirical studies on learner language have been carried out since the late 1960s, but one problem that especially Swedish Second Language Acquisition research is still facing today is the lack of larger annotated L2 corpora; while L2 corpora have been available for e.g. English and Norwegian for the past 2–3 decades (Hawkins and Buttery, 2010; Tenfjord et al., 2004), resources for this kind of studies have been largely lacking for L2 Swedish. However, researchers of Swedish L2 vocabulary and grammar acquisition, and language testing and assessment researchers are in great demand of annotated authentic L2 production data that can help verify hypotheses generated by experimental studies and smaller scaled empirical studies and move research beyond such studies. This is also true for those who pursue research on structures in-between grammar and lexicon, captured by usage-based models of construction grammar (Goldberg, 1995; Goldberg, 2006), which are internationally increasingly applied within the context of L2 learning and L2 pedagogy (Ellis, 2013; Loenheim et al., 2016).

Overtime, Swedish L2 learner essays have been collected in a number of projects and resulted in several learner corpora, e.g. ASU, CrossCheck, Swedish EALA. Previously, ASU (Hammarberg, 2005) and parts of CrossCheck (Lindberg and Eriksson, 2004) were available for researchers through an ITG-system (Saxena and Borin, 2002), but due to the outdated technology ITG is nowadays "retired" and the corpora need to be adjusted to new formats to become searchable through other applications, e.g. Korp (Borin et al., 2012).

2. Challenges for L2 research infrastructure

When it comes to L2 infrastructure, there are three major challenges: availability of data, the need of coordination and availability of methods for processing L2 data. This largely depends upon the following:

(1) L2 learner data, such as essays, is non-trivial to collect since it is not available online for download as is, it requires good contacts with teachers/assessors and via them with learners or their parents who have to be convinced to sign permits for use. This data is essentially sensitive often containing personal details that need to be anonymized.

(2) Hitherto research on learner data has been carried out in different fields, including linguistics, computational linguistics, and Second language acquisition, in a rather uncoordinated fashion - from different points of view and with different purposes and methods - and so far there has been little dialogue or coordination within or between the fields. Scattered individual efforts to collect L2 learner data such as essays, exercise logs and oral transcripts have been driven by project purposes, which has influenced the type of learner metadata, permits, data formats, databases and search tools. As a result, collected data from one project often cannot be compared to or complemented with data collected in another project. Sometimes permit types may even lead to data being forbidden to be used in new projects.

(3) Automatic annotation of L2 data is problematic due to presence of an excessive amount of deviations from the normative Swedish. The existing computational linguistics methods for text processing are developed with a normative language in mind, and cannot be applied in their current form to L2 texts. However, annotating learner data manually is an extremely time-consuming enterprise. To cater for the grammatical and orthographical infelicities in L2 texts, and to make annotation of L2 data more time-effective, computational linguistics methods need to be adapted to the challenges set by interlanguage (Hawkins and Buttery, 2010; Rosen et al., 2014).

Thus, empirically-based data-driven L2 research is in acute need of coordination and structuring at the national level, with centralized databases, uniform metadata, methods for L2 processing, visualization tools, and cross-disciplinary expertise which interested parties can turn to. L2 infrastructure and research can benefit from computational linguistics methods for tasks such as error detection and correction, automatic essay grading, and proficiency level assignment, but approaches to techniques of these kinds for Swedish have not taken the needs of L2 learning into account (Östling et al., 2013; Grigonytė et al., 2014).

In the longer term, there is a development towards digitization of knowledge assessment in compulsory school (grundskolan) and upper secondary school (gymnasieskolan), including national tests in Swedish as a second language in compulsory school and municipal adult education in Swedish for Immigrants (SFI). Specifically, the Inquiry on National Tests (SOU, 2016) has suggested that all national tests shall be digitized by 2022. This development has the potential to significantly facilitate the construction of Swedish L2 research infrastructure in the future. In the meantime, however, the build-up of L2 research infrastructure depends on independent initiatives for collection and analysis of data.

In an ideal world, all L2-related resources and related technologies, tools and methods should be collected under one and the same national infrastructure which is at the moment largely non-existent for L2 Swedish. However, it would be a natural extension to Språkbanken's infrastructure and, through Språkbanken, to the Swe-Clarin infrastructure (https://sweclarin.se/), both of which aim at the creation of an eResearch infrastructure that makes language resources (e.g., corpora, lexicons), tools (tokenizers, taggers, parsers), methods and expertise available and readily usable to scholars of all disciplines.

3. Initial steps

3.1 Data collection and preparation

During 2013–2016 we have collected and prepared part of data that we intend to use for building the L2 infrastructure prototype. All data is linguistically annotated, though only the core data described below is planned to be normalized and error-annotated. The openness/availability of data varies depending upon the previously signed permits, as well as new permits that we will collect.

Our *core data* consists of L2 essays written during the past 10 years by learners aged 16 or older. We have carried out a pilot project during 2013–2016 aimed at collecting and digitizing essays, an experience that allows us to make estimations of what different steps may cost in time. The pilot SweLL corpus (Volodina et al., 2016) contains at the moment 339 digitized essays, with 144,087 tokens, and approximately 150 essays in a pipeline to be digitized; and the collection is steadily growing. The essays cover several developmental stages, from absolute beginners up to advanced proficiency levels. During the pilot, we established contacts with the Ethics Committee (Etikprövningsnämnden), and drafted permits according to their recommendations. We also developed a simple editor to produce uniform metadata annotation for learner variables, digitized a subset of essays, as well as assessed a subset of essays according to the Common European Framework of Reference scale, CEFR (Council of Europe, 2001).

Reference data: The Uppsala Corpus of Student Writings consists of Swedish texts produced as part of a national test of students ranging in age from nine (in year three of primary school) to nineteen (the last year of upper secondary school) who are studying either Swedish or Swedish as a second language. National tests have been collected since 1996. The corpus currently consists of 2,500 texts containing over 1.5 million tokens. Each token is annotated with lemma, part-of-speech and morphological features as well as syntactic dependency structure. The texts have been annotated automatically using existing state-of-the-art natural language processing tools on several linguistic levels. Since spelling and grammatical errors are common in student writings, the texts are automatically corrected while keeping the original tokens in the corpus. It is a monitor corpus which has a restricted research permit, but cannot be made available to the public (Megyesi et al., 2016).

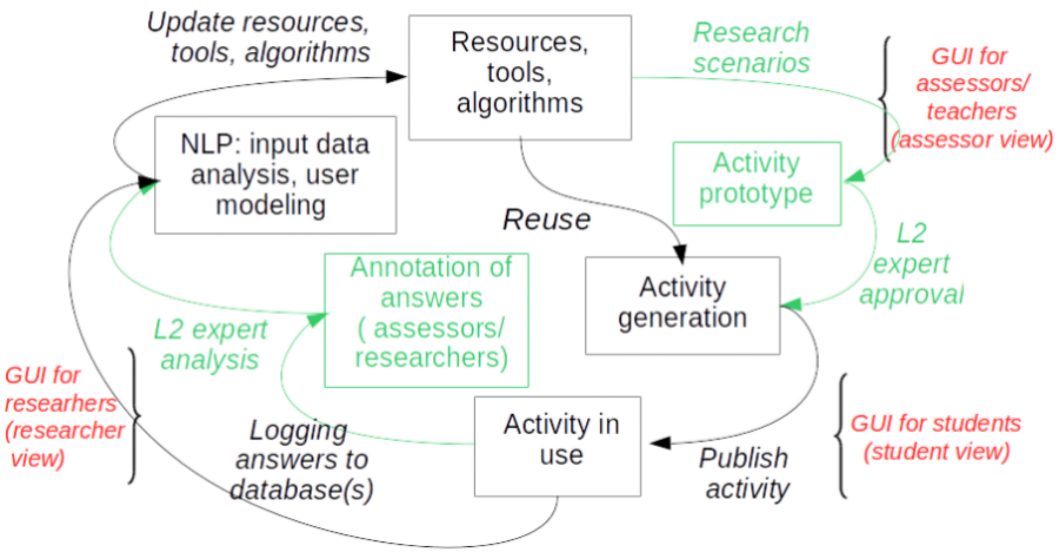
Update resources, tools, algorithms
Resources, tools, algorithms
Research scenarios
GUI for assessors/ teachers (assessor view)
NLP: input data analysis, user modeling
Activity prototype
Reuse
L2 expert approval
Annotation of answers (assessors/ researchers)
Activity generation
L2 expert analysis
GUI for researhers (researcher view)
Logging answers to database(s)
Activity in use
Publish activity
GUI for students (student view)

Figure 1: L2 infrastructure dynamic activity cycle

3.2 Normalization experiments

Standard corpus annotation follows a number of steps, including tokenization, PoS-tagging, lemmatization and syntactic parsing. L2 learner language, however, requires us to handle texts exhibiting a great amount of deviation from standard Swedish. While texts with normative Swedish can be relatively accurately annotated with existing automatic methods, annotating learner language with the same tools is error-prone due to various (and often overlapping) orthographic, morphological, syntactic and other types of errors, e.g.:

- segmentation problems: "jag har två kompisar som hete S och P de är från Afghanistan också jag älskar de för att när jag behöver hjälp de hjälpar gärna mig och jag också hjälpa de."
- misspelling variations: "sommern", "kultor", "frust kyckling lever"
- unexpected morphological forms/agreement errors: "Min drömar"
- word order errors: "Jag bara studera 4 ämne i skolan och på fritiden träna jag på gym"

We have, therefore, started experiments with so-called *normalization*, an extra step which we intend to add to the L2 annotation procedure before applying standard annotation pipeline. During this step deviating learner language is re-written to follow standard language norms. This covers, among others, word-level errors, errors stretching across several words and sentence structure errors. Ideally, part of the corpus should be normalized manually to produce training data. However, so far experiments were run for automatic word-level error normalization, i.e., for words that the lemmatizer failed to identify in a lexicon. Three methods for word-level normalization have been tested:

(1) Using Levenhstein-based normalization, developed for historical texts and adapted to student writings to correct the misspelled words (Pettersson et al., 2014).

(2) Using LanguageTool (Naber, 2003) for a list of suggestions, and picking one variant based on mutual information score with co-occuring words.

(3) Using Levenstein distance for a list of suggestions, picking the first variant with the shortest distance (Llozhi, 2016).

The normalization experiments are ongoing and the results are not yet evaluated.

4. L2 research infrastructure: agenda

In the future, pending funding, our L2 research infrastructure would include the following steps:

- Preparing a gold standard corpus enriching it with manually-added normalization and error-annotation, as well as manually-proofread linguistic annotation
- Developing automatic methods for normalization and error-detection based on annotations in the gold standard corpus
- Technical development of electronic L2 infrastructure with data portal and exercise generator for collecting new data
- Preparing technical solutions for statistic and analytic visualization of L2 data
- Fostering expertise within questions relevant for electronic L2 research infrastructure, e.g., legal issues (copyright, privacy), agreements and permits, standardized metadata for learner data, etc.

Multiple linguistic and pedagogical exploitation scenarios can be envisaged given that L2 corpora with rich linguistic and error annotation become available, such as to search for all (mis)spelling variants of some lemma, e.g. “mycket” (“much”) and get hits with all variations “mycekt*”, “miket*”, “micke*”. Another example is to trace (in)correct use of possessive constructions in essays written by the same student over time, or students sharing the same mother tongue, and get results showing types and percentage of erroneous/correct use at the beginner level (e.g. min familjen*, min livet*, gick hennes hemma*) compared to more advanced levels.

In the long run, we envisage a larger dynamic electronic platform for collection of new L2 learner data, annotation of data as well as searches and visualization, see Fig.1. On top of that, L2 learners will be given a possibility to engage in learning activities and through those add more data to the databases. Exactly ”what the doctor prescribed”.

Acknowledgements

The pilot project on SweLL corpus has been financed through Språkbanken (https://spraakbanken.gu.se/), Center for Language Technology (http://clt.gu.se/) and Department of Swedish (http://svenska.gu.se/), all the three at the University of Gothenburg. The Uppsala Corpus of Student Writings and the web-based tools for annotation and analysis was financed by SWE-CLARIN (https://sweclarin.se) and the Department of Scandinavian Languages, Uppsala University (http://www.nordiska.uu.se).

References

L. Borin, M. Forsberg, and J. Roxendal. 2012. Korp - the corpus infrastructure of Språkbanken. In *Proceedings of LREC*, pages 474–478.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

N. Ellis. 2013. Constructions Grammar and Second Language Acquisition. In Th Hoffmann and G Trousdale, editors, *The Oxford handbook of construction grammar.*, pages 348–378. Oxford and New York: Oxford University Press.

A. E. Goldberg. 1995. *A Construction Grammar Approach to Argument Structure.* Chicago: The University of Chicago Press.

A. E. Goldberg. 2006. *Constructions at work: the nature of generalization in language.* Oxford: Oxford University Press.

G. Grigonytė, M. Kvist, S. Velupillai, and M. Wirén. 2014. Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). Association for Computational Linguistics.*

B. Hammarberg. 2005. *Introduktion till ASU-korpusen, en longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska med en motsvarande del från infödda svenskar*. Institutionen för lingvistik, Stockholms Universitet, Sweden.

J.A. Hawkins and P. Buttery. 2010. Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, 1(1):1–23.

G. Josefsson. 2016. *http://www.svd.se/professor-snabbspar-till-svenska-fungerar-inte/om/debatt*. Svenska Dagbladet.

J. Lindberg and G. Eriksson. 2004. CrossCheck-korpusen - en elektronisk svensk inlärarkorpus. In *Proceedings of the ASLA Conference 2004*.

L. Llozhi. 2016. *SweLL List. A list of productive vocabulary generated from second language learners’ essays.* Master Thesis in Language Technologies. Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden.

L. Loenheim, B. Lyngfelt, J. Olofsson, J. Prentice, and S. Tingsell. 2016. Constructicography meets (second) language education: On constructions in teaching aids and the usefulness of a Swedish constructicon. In S de Knop and G Gilquin, editors, *Applied Construction Grammar*. Berlin: De Gruyter Mouton.

B. Megyesi, J. Näsman, and A. Palmér. 2016. The Uppsala Corpus of Student Writings: Corpus Creation, Annotation, and Analysis.). In *Proceedings of Language Resources and Evaluation (LREC)*.

Migrationsverket. 2016. *http://www.migrationsverket.se/Om-Migrationsverket/Statistik/Asylsokande—de-storsta-landerna.html*. Migrationsverket.

D. Naber. 2003. A rule-based style and grammar checker. Master’s thesis, Bielefeld University, Germany.

E. Pettersson, B. Megyesi, and Nivre J. 2014. A Multilingual Evaluation of Three Spelling Normalization Methods for Historical Text. In *In Proceedings for Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities, European Association for Computational Linguistics, LaTeCH 2014, EACL 2014.*

A. Rosen, J. Hana, B. Štindlová, and A. Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resource and Evaluation Journal*, 48:65–92.

A. Saxena and L. Borin. 2002. Locating and reusing sundry NLP flotsam in an e-learning application. In *LREC 2002. Workshop Proceedings. Customizing knowledge in NLP applications: strategies, issues, and evaluation.*

SOU. 2016. *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning.* Statens offentliga utredningar 2016:25.

K. Tenfjord, P. Meurer, and Hofland K. 2004. The ASK-corpus - a language learner corpus of Norwegian as a second language. In *Proceedings from 5th International Conference of Language Resources and Evaluation (LREC).*

E. Volodina, I. Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, and M. Sandell. 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In *LREC Proceedings 2016.*

R. Östling, A. Smolentzov, B. Tyrefors Hinnerich, and E. Höglin. 2013. Automated Essay Scoring for Swedish. In *Association for Computational Linguistics.*, volume Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications.