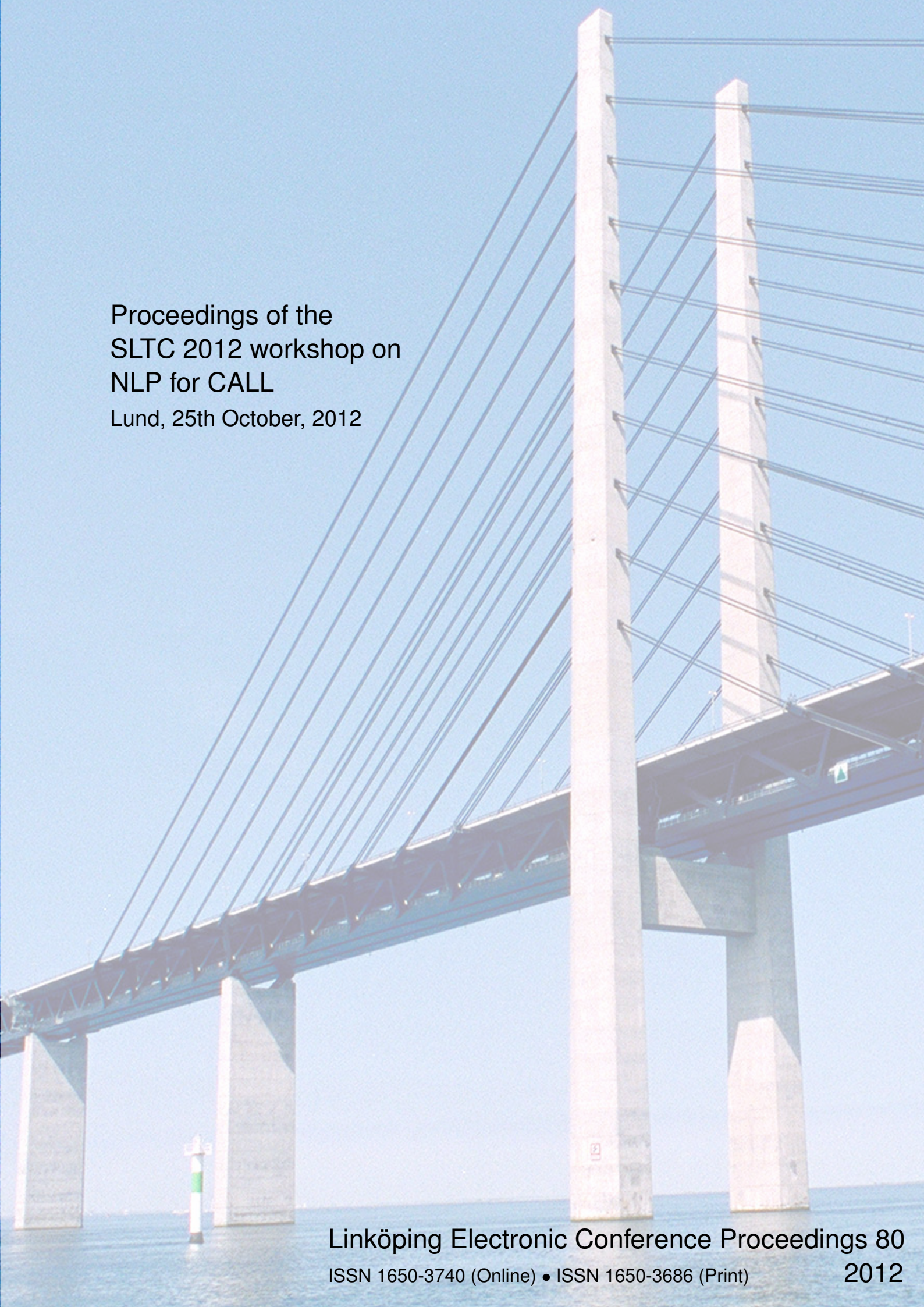

Proceedings of the SLTC 2012 workshop on NLP for CALL

Lund, 25th October, 2012

Linköping Electronic Conference Proceedings 80
ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)
2012

Proceedings of the SLTC 2012 workshop on NLP for CALL

Lund, 25th October, 2012

edited by

Lars Borin and Elena Volodina

Linköping Electronic Conference Proceedings 80

ISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2012

Preface

Learning and teaching languages with the assistance of a computer, i.e. computer-assisted language learning (CALL), has become widespread since the early 1980s. Traditional CALL applications provide limited exercise types, along with limited ability to provide feed-back, because the exercises are static, i.e. pre-programmed, and the answers have to be pre-stored.

To try to overcome this disadvantage, some researchers have started to use techniques from the field of Natural Language Processing (NLP) in CALL systems, i.e. supplying CALL applications with some kind of intelligence. As a result, the interdisciplinary field of Intelligent CALL (ICALL) – combining NLP and CALL – has emerged over the past 20 years or so. The workshop on *NLP for CALL* arranged in conjunction with the 4th Swedish Language Technology Conference in Lund on 25th October, 2012, is a reflection of this development.

In arranging the workshop, we specifically wished to address the following question: There is an array of NLP resources and tools potentially available for re-use in ICALL applications for Swedish as well as for many other languages equipped with NLP tools and resources, but this opportunity has so far remained relatively underdeveloped. Consequently, we invited submissions that would

- describe research directly aimed at ICALL;
- demonstrate actual or discuss potential use of existing NLP tools or resources for language learning;
- describe ongoing development of resources and tools with potential usage in ICALL, either directly in interactive applications, or indirectly in materials, application or curriculum development, e.g. collecting and annotating learner corpora; developing tools and algorithms for readability analysis, selecting optimal corpus examples, etc.;

and especially submissions describing work on Swedish or other Nordic languages.

We received a total of 12 submissions, out of which one was rejected out-of-hand by the workshop organizers for not conforming to the submission format. Each of the remaining 11 papers was reviewed by two (anonymous) members of the program committee (see below). On the basis of the reviews, 8 submissions were accepted for presentation at the workshop and inclusion in the workshop proceedings (subject to revisions required by the reviewers).

The workshop was designed to be a highly interactive event. After an invited oral presentation by Hrafn Loftsson (University of Reykjavik) – *Ongoing development of an NLP toolkit with potential usage in ICALL* – the other contributions to the workshop were presented in the form of two poster sessions. A general discussion concluded the workshop.

Workshop organizers

- Lars Borin, Språkbanken, University of Gothenburg, Sweden
- Elena Volodina, Språkbanken, University of Gothenburg, Sweden

Program committee

- Lars Borin, University of Gothenburg, Sweden
- Sofie Johansson Kokkinakis, University of Gothenburg, Sweden
- Petter Karlström, Stockholm University, Sweden
- Ola Knutsson, Stockholm University, Sweden
- Hrafn Loftsson, Reykjavik University, Iceland
- Detmar Meurers, University of Tübingen, Germany
- Serge Sharoff, University of Leeds, UK
- Elena Volodina, University of Gothenburg, Sweden

Contents

Improving feedback on L2 misspellings – an FST approach

Lene Antonsen
University of Tromsø, Norway
lene.antonsen@uit.no

Abstract

This paper suggests enriching the finite-state transducers (FST) analyser with erroneous forms marked with error tags, as a way of improving feedback on L2 misspellings. This approach can be useful both in isolated word error correction and in detecting real word errors in context-dependent word correction. But most important, it makes it possible to give metalinguistic feedback on the nature of the errors.

1 Introduction

When learning a language with a rich system of inflectional morphology like North Saami, the learner has to focus on form if the goal is to achieve near-native fluency in L2. The learner's awareness of the relevant morphological processes in the language plays a crucial role.

A computer can parse a language on the basis of its standard linguistic forms. We are looking for a way to enable the computer to parse a language even when the forms produced by learners deviate from the target language forms. In other words, we want to find a way for the computer to interpret learners' intentions as represented in their interlanguage forms. This would make it possible for the computer both to recognise forms even when they are misspelled (overlooking the errors) and to heighten the learners' awareness of morphological processes by correctly interpreting their mistakes.

This paper is structured as follows: Section 2 discusses L2 misspellings and looks at different kinds of feedback and how L2 misspellings are recognised by a generic spell checker. Section 3 describes the enriching of the morphological analyser with systematic misspellings and section 4 describes how the analyser functions in an ICALL program with free input. Finally, in sections 5 and 6 I present a conclusion and some assumptions of how an enriched FST can be utilised in automatic writing assistant tools for language learners.

2 Background

A misspelling is a written form that deviates from the conventions in the written language. The misspelling can result in a non-word, an unintended word form of the same lemma, or a new lemma. A human teacher can usually interpret the student's intention behind the misspellings, but the misspelling makes it more difficult for a computer to give the correct syntactic analysis, and in that way it complicates the human-computer interaction.

2.1 North Saami

The Saami languages are morphologically complex suffixing languages with much suprasegmental morphology. Nouns and verbs have about 100 inflected forms, half of the forms for verbs are finite forms. North Saami is the largest Saami language, with only approximately 17 000 speakers, but both schools and universities offer courses for students who want to learn the language.

The orthography conventions differ substantially from the native language of most of the students. North Saami extends the Latin alphabet with seven letters by means of diacritical marks (e.g. *š, č*), where Norwegian and Swedish use letter combina-

Lene Antonsen 2012. Improving feedback on L2 misspellings – an FST approach. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 1–10.

tions (*skj, tsj*). All diphthongs are different from Norwegian and Swedish diphthongs, and some of the graphemes represent other phonemes than in Norwegian and Swedish. Compared to Finnish the differences are smaller, but also the Finnish alphabet does not contain consonant letters with diacritical marks. Especially the Norwegian and Swedish language learners often fail to write the correct form, because of differences between their L1 and Saami, both with regard to morphology and orthographical conventions.

2.2 L2 misspellings

In the learner's production there will be both accidental mistyping and incorrect word forms due to misconceptions of the target language. Corder (1967) makes a distinction between **errors of performance**, which characteristically are unsystematic, and **errors of competence**, which are systematic. From the latter it is possible to reconstruct the learner's knowledge of the language.

The errors of competence can be divided into two groups: morphologically irrelevant, but still systematic ones, like writing *a* instead of *á* in the stem, and morphologically relevant ones, like omitting suprasegmental processes in certain kinds of inflections, e.g. skipping monophtongization when going from the nominative form *viessu* 'house' to the illative form *vissui* 'to the house' (which gives the erroneous form *viessui*), or choosing a wrong inflection for the context. According to the system proposed by James (1998), the former group consists of **substance errors** that violate certain convention for representing phonemes by means of graphemes, and the latter one consists of **text errors**. Also these are systematic errors.

2.3 Feedback

The student usually needs feedback to correct his own errors. The feedback can be a comment saying that something is wrong in the sentence, the erroneous word or phrase can be highlighted, or the student can be provided with the target word or a list of possible target words. Another kind of feedback is a metalinguistic comment saying what is wrong and why, possibly hyperlinked to more information about the phenomenon. Above all the feedback should support and facilitate learning, and the error should be seen as a chance of getting the language learner not only to correct the word or phrase, but also understand the reason for his misconception.

If the misspelling is an error of performance, it is sufficient to make the student aware of it. But if it is an error of competence, the student needs a correction, and if it is a metalinguistic comment, it is crucial to give a feedback according to the student intended writing and at his own level of competence. This is the challenge when coming real word errors. The student will be confused when getting feedback on syntax instead of the misspelling, e.g. feedback on using an infinite form instead of a finite form, when the student believes he has written a finite form.

2.4 L2 and spell checkers

Most spell checkers are generic and made for L1 users, but also language learners use them. The feedback from the spell checker is usually a suggestion for a more appropriate target word, or more often, a list of candidates for the target word. Most spell checkers detect errors and suggest corrections without using context, and therefore only detect non-word errors.

For detecting real word errors it is necessary to use the context. A real word error can lead to a syntactic or a morphosyntactic error, the challenge for the spell checker is to point out which word in the sentence is the incorrect one. There is a work in progress on building a grammar checker for North Saami that also considers real word errors, with L1 users as the target group (Wiechetek, 2012).

Another challenge, independent of whether it is a non-word error or a real word error, is to give the correct suggestion for how to correct the word (Kukich, 1992). The algorithm for suggesting correct candidates in spell checkers for native writers, is based on using as few editing steps as possible, going from the misspelled word to the target word.

A few spell checkers for non-native writers have been developed, most of them are specifically targeting certain error classes. There are spell checkers that incorporate lists of common misspellings in the target language, retrieve suggestions based upon the phonological representation of the misspelling, address morphological triggered misspellings and oth-

ers provide references to alternative spellings, e.g. on the internet. Nevertheless, spell checkers for non-native writers are rare. (Rimrott and Heift, 2008a). There is no such spell checker for North Saami.

Spell checkers are constructed in order to identify errors and give the most relevant suggestion for the correction, but in a language learning context, it can be even better to be able to give metalinguistic feedback to the student, e.g. 'Remember diphthong simplification when adding the suffix -i' for the misspelling *viessui*, which is used as example in section 2.2. Alternatively, one can ignore the misspelling in favour of concentrating upon the syntax of the learner's input.

In order to test the effect of a spell checker, I annotated errors in a corpus consisting of L2 sentences (4633 words, 800 sentences, 739 misspellings). Rimrott and Heift (2008a) present a similar testing for German, but unlike them, I considered also real word errors.

The North Saami spell checker[1] is based on dictionary lookup and dynamic compounding, and is designed for native speakers. The word forms are produced with finite-state transducers, which are explained in section 3.

The error model is based upon edit distance, which is the number of operations applied to the characters of a string: deletion, insertion, substitution, and transposition. In the literature, the edit distance has usually been found to cover more than 80 % of the misspellings at distance one. (Levenstein, 1965; Damerau, 1964). In the algorithm of the North Saami spell checker there are additionally phonetic rules. Errors with the same error distance are ranked based upon phonetic likelihood.

Testing the L2 corpus on a test bench[2] for the spell checker gave a precision of 0.92 and a recall of 0.74. The real word errors constitute 26.0 % of the errors and are therefore not detected by the spell checker. I also looked at the generation of the correct suggestions. In table 1 it appears that for 19.9 % of the misspellings, the program could not generate a correct candidate at all. The average edit distance for these misspellings were 2.74.

Testing shows that for L2 writers, the order in which the words appear in the suggestion list, seems to be an influencing factor for selecting one word over another (Rimrott and Heift, 2008b). This implies that an L2 student is probably not able to choose between a large number of candidates. Table 1 shows that only in 67.7 % of the cases the correct suggestion is among the top 3. This result is poorer than the accuracy level above 90 %, which is usually reported on L1 misspellings, when the first three guesses are considered (Kukich, 1992). For the North Saami spell checker the level is 85 % for L1.

true positives	**correct cand. among top 3**	**no correct cand. among top 3**	**no correct cand.**
99.9 %	67.7 %	12.3 %	19.9 %
aver. edit distance	1.39	1.59	2.74

Table 1: The spell checker's candidates for the true positives. For 32.3 % of the words that were correctly indentified as misspelling, there was no correct candidate among the top three candidates. N=563

This test demonstrates that the spell checker is not sufficient for L2 writers because a relatively big part of their misspellings are real word errors that are not identified, and for the non-word errors the generation and ranking of candidates was not good enough for 32.3 % of the cases. The main reason is that the average edit distance for the L2 misspellings was as high as 1.54. A similar annotated corpus of L1-sentences gave an average edit distance of 1.26. The second reason is probably that the phonetic rules, which rank candidates, do not suit L2 writers, who often are not sure about the word's pronunciation.

3 Enriching the FST with systematic misspellings

3.1 Finite-state transducers

Instead of listing all word forms of a language, one may list all the stems and affixes, and combine them to word forms by means of finite-state automata, see figure 1 for an example.

A finite-state transducer is a finite-state automaton that maps between two strings of characters: the

[1] `http://divvun.no/`

[2] Moshagen (2008) describes the test bench.

Figure 1: This finite-state automaton produces the word forms *lávka, lávkan* ('bag.N') and *girji, girjin* ('book.N').

word form itself and the grammatical word, like in figure 2: *girjin* (lower level) and *girji+N+Ess* (upper level).

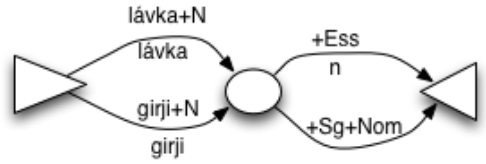

Figure 2: This finite-state transducer produces the same word forms as in figure 1, but it also maps between the word form and the grammatical word.

3.2 Modelling misspellings with FST

The FST models the language in question by producing the correct word forms. But the FST can also model systematic misspellings with specific error tags in the upper level. In that way the analyser identifies the word as an erroneous form of a certain grammatical word. The modelling of misspellings be utilised in several ways:

1. The ranking of suggestion candidates in isolated word correction can be improved by giving priority to systematical L2 errors, some of them with an edit distance bigger than 1.

2. The morphological analysis combined with error tag makes it easier to detect real word errors in context-dependent word detecting.

3. The specific error tag also makes it possible to give metalinguistic comments about the morphological nature of the misspellings, both for non-word and real word errors.

According to the system of errors in section 2.2, two kinds of systematic errors can be added to the FST: substance errors (errors in encoding/decoding), and text errors (usage errors), like omitting suprasegmental processes.

The FST that the North Saami spell checker is based upon, consists of a lexical transducer `lexc` and a phonological transducer `twolc` for the suprasegmental processes (Koskenniemi, 1983). It is compiled with the Xerox compilers (Beesley and Karttunen, 2003), and is available as open source[3] under the terms of the GNU General Public License. I have added systematic misspellings to both the lexical and the phonological transducers. Additionally, certain kinds of misspellings are taken care of by concatenating the final transducer with another transducer containing these misspellings.

3.3 Adding paths to the lexical transducer

Suffixes are added and some vowel and consonant changes are made in the lexical transducer. The ordinary illative suffix *-ii* for nominals with trisyllabic stem, is added in `lexc`. For the same stems I made an extra path with the suffix for nominals with bisyllabic stem, *-i*, marked with the error tag `IllErr` (= incorrect illative suffix) in the upper (here: right) level. In example 1 are the analyses for the misspelling *hivssegi* and the target form *hivssegii*.

Ex. 1

```
"<hivssegi>" "hivsset" N Sg Ill IllErr
"<hivssegii>" "hivsset" N Sg Ill
                          'to the toilet.N'
```

Some suprasegmental processes that are taken care of in the phonological transducer, are triggered by a dummy symbol in the lexical transducer. The erroneous path is made without this dummy, e.g. inflections with strong grade for the consonant centre when there should have been weak grade, as in figure 3. The error tag in upper level is `CGErr` (= lacking consonant gradation), see example 2. The target form is *áhku*.

[3] `https://victorio.uit.no/langtech/trunk/gt/`

Ex. 2

```
"<áhkku>" "áhkku" N Sg Nom 'grandmother.N'
"<áhkku>" "áhkku" N Sg Acc CGErr
"<áhku>" "áhkku" N Sg Acc
```

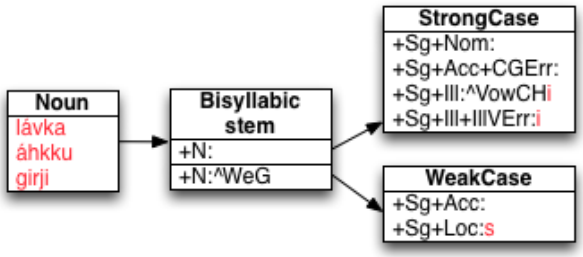

Figure 3: The lexical transducer, `lexc`, is adding both suffixes (*-s, -i*) and dummies for the phonological transducer to the stem. The dummies here are `WeG` for consonant centre in weak grade and `VowCH` for vowel change. The erroneous paths without the dummies are marked with error tags: +CGErr and +IllVErr.

3.4 Generating misspellings with the phonological transducer

The phonological transducer changes letters under specific conditions. In figure 4 the consonant centre is changed from *hkk* to *hk*, or *vk* to *vkk*, if it is followed by one or more vowels and `WeG`, which is a dummy[4].

```
hkk -> hk, vk -> vkk, ... || _ Vow* WeG ;
```

Figure 4: The phonological transducer, `twolc`.

Some misspellings are generated by first adding a path with error tags to both upper and lower level in `lexc`, and then removing the error tag from the lower level under special conditions in `twolc`. The analyses with error tag in both levels are then removed from the output of the FST, by means of regex-rules.

The erroneous path can be a rule that changes letters generally from a letter with a diacritic mark to a letter without, e.g. changing *á* into *a*. The path with the error tag `AErr` remains in the upper level only if the change happens. In example 3, the misspelling *barru* and the target form *bárru* are analysed:

[4]For details, see Trosterud and Uibo 2005.

Ex. 3

```
"<barru>" "bárru" N Sg Nom AErr
"<bárru>" "bárru" N Sg Nom      'wave.N'
```

Other rules change letters under special conditions, such as diphthong simplification, and the erroneous path with error tag `DiphErr` (= omitted monophthongization) will remain only if the diphthong simplification does not happen. The misspelling *viessui* and the target form *vissui* are analysed in example 4:

Ex. 4

```
"<viessui>" "viessu" N Sg Ill DiphErr
"<vissui>" "viessu" N Sg Ill
                              'to the house.N'
```

3.5 Adding paths by concatenating transducers

There is also a special transducer for lowercase initial letters in place names, which is concatenated to the main transducer after the first compilation process. All forms have the tag `LowercaseErr` in the upper level, and in example 5 is the analysis of the misspelling *lundas* and of the correct *Lundas* ('in Lund'):

Ex. 5

```
"<lundas>" "Lund" N Prop Plc Sg Loc
       LowercaseErr
"<Lundas>" "Lund" N Prop Plc Sg Loc
```

3.6 More readings before disambiguation

Table 2 lists the systematic misspellings I added to the FST. Two of them are substance errors, `Lowercase` and `AErr`. The latter one is *a* instead of *á*, the most frequent letter with diacritic mark in the North Saami alphabet.

The other misspellings in table 2 are text errors, products of incorrect inflection. All erroneous forms are marked with an error tag that characterises their nature, like `AiErr` (= a inflection error): *a* is written where it should be vowel change from *i* to *á* caused

of an inflection. Most of the systematic misspellings were added to nouns.[5]

error tag	erroneous form	target form
Lowercase (place names)	londonis	Londonis 'London.SgLoc'
AErr (general rule)	manna	mánná 'child.SgNom'
AiErr (verbs)	boahtan	boahtán 'come.V.PrfPrc'
CGErr (nouns)	skuvlas	skuvllas 'school.SgLoc'
DiphErr (nouns)	viessui	vissui 'house.SgIll'
IllVErr (nouns)	skuvlai	skuvlii 'school.SgIll'
IllErr (nouns)	hivssegi	hivssegii 'toilet.SgIll'

Table 2: Systematic misspellings added to the FST.

By enriching the morphological analyser with erroneous forms, the number of possible readings increases. In figure 5, the morphological analysis of the sentence is done with the regular FST. There are two misspellings, which are unknown to the analyser.

```
"<Ahkku>"
 "Ahkku" ?
"<manná>"
 "mannat" V IV Ind Prs Sg3
"<lundii>"
 "lundii" ?
"<odne>"
 "odne" Adv
```

Figure 5: 'Grandmother goes to Lund today.' analysed with the regular FST.

The same input is then analysed with the error-FST in figure 6, and the misspellings are recognised as an erroneous form of *áhkku* 'grandmother' (*a* instead of *á*), and an erroneous form of the place name *Lundii* (in illative case 'to Lund'), with initial lowercase letter. Also the correctly spelled word *manná* 'goes' gets several possible erroneous readings.

Disambiguation of the multiple readings will be explained in section 4.1.

[5]The makefile and source files can be downloaded: https://victorio.uit.no/langtech/branches/errortag/gt/

```
"<Ahkku>"
 "áhkku"  CGErr Sg Acc AErr
 "áhkku"  CGErr Sg Gen AErr
 "áhkku" N Sg Nom AErr       <- correct
"<manná>"
 "mannat" V IV Ind Prs Sg3   <- correct
 "mánná" Hum N Sg Nom AErr
 "mánná" Hum N CGErr Sg Acc AErr
 "mánná" Hum N CGErr Sg Gen AErr
"<lundii>"
 "Lund" N Prop LowercaseErr Plc Sg Ill
"<odne>"
 "odne" Adv
```

Figure 6: 'Grandmother goes to Lund today.' analysed with error-FST. The error tags are explained in table 2. The correct analyses are marked.

4 Evaluation

The erroneous forms with error tags in the analysis make it possible to recognise the target form. The evaluation will show to which extent the added erroneous forms cover the L2 misspellings, and how the multiple readings influence upon the disambiguation. I will also discuss the consequences for the size of the FST.

4.1 Test bench

I use the syntactic analyser from an existing ICALL-program[6] as a test bench for the error-FST. The ICALL-program accepts free-input, and has L2 learners as its target group.

In the ICALL-program there are three different question-answer drills with free input. For two of them, the questions are generated, for one of them the student can answer freely. The other one presents 2-4 lemmas, which should form part of the answer. The third drill is a tailored dialogue, but the student can answer freely to the questions. All three QA-drills use the same analyser. The tutorial feedback concerning grammatical errors is given in a separate window, and the user is allowed to correct the answer until it is accepted.

The morphological analyser is the one described in section 3. The morphological disambiguator is implemented in the Constraint Grammar (CG) framework (Karlsson et. al, 1995). The rules are

[6]http://oahpa.no/davvi/

compiled with vislcg3[7], and they are manually written, context dependent rules used for selecting and discarding analysis.

The CG-rule set consists of two parts. The first part is a rule set that disambiguates the user's input only to a certain extent. The rule set is relaxed compared to the ordinary disambiguator, in order to be able to detect relevant readings despite a certain degree of grammatical and orthographic errors in the input. The second part of the rule set contains rules for giving feedback to grammatical errors. Question and answer are merged, and given to the analyser as one text string, with only a tag as delimiter between question and answer, so that one can refer to the question and the answer separately in the CG-rules[8].

These ICALL programs are designed with students at introductory level as the target group. Till now feedback to misspellings in the program is handled by pointing to the unrecognised word form, asking the student to check the spelling. Only a couple of systematic real word errors give more specific feedback to the student on the nature of the misspelling. The misspellings constitute the biggest problem for the human-computer interaction (Antonsen et.al., 2009b). Pointing out the misspelled word is not enough help to the student. The system is not like a human reader able to read the answer in a robust way, and detect what the student intended to write.

By using an error-FST as morphological analyser in the ICALL-program, it should be to some extent possible to recognise the student's intended word, and also to make more CG-rules, which trigger metalinguistic feedback as help for the student, e.g. 'X misses diphthong simplification'.

4.2 Test results

I have been testing a part of the programs' student log, consisting of 2705 question-answer pairs. The pairs were parsed with the regular and the error-FST, respectively, and then parsed with the CG-rules.

The erroneous forms in the error-FST cause the number of analyses to increase from 74 517 to 83 582 (+ 12.1 %), from 2.26 to 2.54 per word form before disambiguation. But the disambiguation is quite efficient, as shown in table 3. The erroneous path `CGErr` is the most productive one. It gives only real word errors, and therefore many correct word forms get a possible error analysis in addition to the correct analysis. But the extra readings do not mess up the disambiguation, which removes 93.7 % of the extra readings.

The `IllErr`-path, incorrect illative suffix added to nominals with trisyllabic stem, gives only non-word errors, and the word forms get only this analysis. The other erroneous paths can both give real word and non-word errors.

errortag	**before disamb.**	**after disamb.**
CGErr (nouns)	1786	113
AErr (general rule)	1395	524
Lowercase (place names)	534	65
AiErr (verbs)	214	95
IllVErr (nouns)	74	27
IllErr (nouns)	28	28
DiphErr (nouns)	22	16

Table 3: Parsing 2705 QA-pairs with error-FST. The number of analyses with error tag before and after disambiguation.

The analysis also recognises combinations of the erroneous forms, like in example 6, where the word *fallejohkas* is recognised as a misspelling of the target form *Fállejogas* despite of an edit distance of 4.

Ex. 6

```
"<fallejohkas>" "Fállejohka" N Prop Plc Sg
       Loc LowercaseErr CGErr AErr
"<Fállejogas>" "Fállejohka" N Prop Sg Loc
```

The disambiguation does not need to be complete, because of the special CG-rules deciding whether the student gets an error feedback or not.

Table 4 lists how many misspellings were found in the corpus, and what kind of analysis they got. By parsing the test corpus with the regular FST combined with the CG rule set, the target form was recognised for only 8.1 % of the misspellings. They were recognised by means of special CG-rules for systematic real word errors. By parsing the test

[7] `http://beta.visl.sdu.dk/constraint_grammar.html`

[8] For details, see Antonsen et.al. 2009a.

corpus with the error-FST, the target forms could be recognised regardless of whether they were real word errors or non-word errors. The target form was recognised for 44.0 % of the misspellings.

Errors The target form was	**Reg.FST.**		**Err.FST**	
not recognised	871	91.9 %	563	56.0 %
recognized	77	8.1 %	443	44.0 %
Total	948	100 %	1006	100 %

Table 4: Parsing 2705 QA-pairs. Comparing the regular FST with the error-FST. Some sentences have more than one misspelling.

Table 5 contains a comparison of the error messages, which were given with the two different FST's. In addition to feedback on misspellings, the student also gets feedback on syntactic errors, e.g. 'Remember the agreement between subject and verbal', and semantic comments, e.g. 'You must use the given verb.' The latter message is given if the student does not use the given lemmas in the QA-drill that calls for it. All QA-drills require that the students formulate complete sentences, otherwise they get comments on that (here called comment on semantics).

In table 4, the error-FST diagnoses more errors as misspellings than the regular FST, because more of the real word errors are recognised as misspellings instead of syntactic errors, see also table 5. E.g. the frequent misspelling *vuolggan* in example 7 gets a noun analysis with the regular FST. The error-FST gives an additional analysis as a misspelled finite verb with target form *vuolggán*, and the disambiguation can therefore result in a feedback about a misspelling instead of a syntactic error or a messages about a missing finite verb in a sentence:

Ex. 7

```
"<vuolggan>"
 "vuolgga" N Ess                'departure'
 "vuolgit" V IV Ind Prs AiErr Sg1 'I leave'
```

The number of error feedback tags is bigger than the number of actually given feedbacks, since some sentences get more than one error feedback, but the system presents only one at a time to the student. Sometimes two or more feedback tags are related to the same error. Important is that the precision and recall did not decline when using the error-FST compared to the regular FST.

Feedback	**Reg.FST.**	**Err.FST**
Misspellings	751	804
Syntactic errors	1181	1071
Comments on semantics	599	527
Altogether	2531	2402
Number of sentences giving feedback on errors	1560	1561

Table 5: Parsing 2705 QA-pairs. Some sentences have more than one error feedback. Prec=0.96 Rec=0.99 for both FST's.

Among the unrecognised misspellings there are some frequent systematic groups that could be added to the FST, e.g. omitting vowel change in trisyllabic nominal stems and omitting monophthongization and consonant gradation in verbs.

4.3 The size of the FST

All the extra paths make the FST much bigger. The size of the error-FST is almost ten times as big as the regular FST, as shown in table 6, even if most of the error paths added to the error-FST so far are for nouns only. Paths with missing monophthongization and missing consonant gradation are also relevant for inflection of verbs and adjectives. The compilation time increases with 570 %, e.g. on a MacBook Pro (OS 10.6.8) from 3.5 minutes to 23.5 minutes. The time needed for initiating the analysis is more important, but in the ICALL program in which the error-FST was tested, the lookup process is done in a standby server, and start-up delay is thus not relevant. The size of the FST still has impact on the time for analysis, but not so dramatically. However, it is possible to make the error-FST smaller by removing rare dynamic compounding and derivation paths, which are not likely to occur in the language of L2-students.

5 Conclusion

Enriching the FST-analyser with erroneous forms marked with error tags gives promising results. It

	Regular FST	Error FST
size	41.5 Mb	398.8 Mb
	100 %	959 %
states	497 632	4 739 590
arcs	1 062 995	10 297 121

Table 6: The size of the regular FST and the error-FST.

makes the syntactic analyser able to recognise systematic misspellings, both real word errors and non-word errors, even if the edit distance is big.

Even though the number of analyses per word form increases, it does not destroy the disambiguation in a restricted ICALL program. In fact, by means of the erroneous forms some errors are reclassified from syntactic or semantic errors to misspellings.

The error tags make it possible not only to recognise the target form, but also to give tutorial feedback on the nature of the error to the student. When the analyser identifies the grammatical word despite the misspelling, it is possible to ignore misspellings in favour of giving feedback on syntax.

The size of the error-FST expands exponentially, but it can be trimmed for L2 users.

6 Future work

It will be useful to have a closer look at the nature of L2 misspellings in a larger material, and give more erroneous forms to the FST, combined with restricting of the dynamic derivations and compounding, so the FST will not be too large for implementation in end-user applications.

In a spell checker for isolated non-word errors one may test how useful it is to rank the correction candidates with a combination of edit distance and the erroneous forms from the FST, instead of using phonetic rules as was done for the L1 spell checker.

I will also try out the combination of error-FST and constraint grammar in free-input student tasks that are less restricted than the present ICALL-program. Constraint grammar has been tried out for ruling out correction candidates that are grammatically unacceptable in spell checker programs for English L1 and Danish dyslectics (Agirre et.al., 1998; Bick, 2006).

The combination of erroneous forms with error tags and constraint grammar parsing makes it possible to give metalinguistic feedback. It is important to look more into the human-computer interaction, e.g. by means of looking at the log to see how the students correct their input after getting metalinguistic feedback and making a survey for the students about how useful they find the feedback.

Acknowledgements

I would like to thank my supervisor, professor Trond Trosterud, for discussions and valuable input. Thanks also to Linda Wiechetek for help with proofreading.

References

Eneko Agirre, Koldo Gojenola, Kepa Sarasola and Atro Voutilainen. 1998. Towards a Single Proposal in Spelling Correction. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*: 22–28. Association for Computational Linguistics.

Lene Antonsen, Saara Huhmarniemi and Trond Trosterud. 2009a. Constraint Grammar in Dialogue Systems. *Proceedings of the 17th Nordic Conference of Computational Linguistics.* NEALT Proceeding Series. Volum 8: 13–21. Odense, Denmark.

Lene Antonsen, Saara Huhmarniemi and Trond Trosterud. 2009b. Interactive pedagogical programs based on constraint grammar. *Proceedings of the 17th Nordic Conference of Computational Linguistics.* NEALT Proceeding Series Volum 4: 10–17. Odense, Denmark.

Kenneth V. Beesley and Lauri Karttunen. 2003. *Finite State Morphology.* CSLI publications in Computational Linguistics. USA.

Eckhard Bick. 2006. A Constraint Grammar Based Spellchecker for Danish with a Special Focus on Dyslexics. (ed.) Suominen, Mickael et al. *A Man of Measure – Festschrift in Honour of Fred Karlsson*: 387-396. The Linguistic Association of Finland.

S. P. Corder. 1967. The significance of learner's errors. *International Review of Applied Linguistics* 5: 161–169.

Frederick J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* 7.

Carl James. 1998. *Errors in language learning an use: exploring error analysis.* Longman. USA.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila. 1995. *Constraint grammar: a*

language-independent system for parsing unrestricted text. Mouton de Gruyter.

Kimmo Koskenniemi. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* Publications of the Department of General Linguistics, University of Helsinki, No. 11.

Karen Kukich. 1992. Techniques for Automatically Correcting Words in Text. *ACM Compiling Surveys* 24(4): 377–439.

Vladimir I. Levenstein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10.

Sjur Nørstebø Moshagen. 2008. A language technology test bench – automatized testing in the Divvun project. Proceedings of the Workshop on NLP for Reading and Writing – Resources, Algorithms and Tools. *NEALT Proceeding Series* 3: 19–21.

Anne Rimrott and Trude Heift. 2008a. Evaluating automatic detection of misspellings in German. *Language Learning & Technology* 12(3): 73–92.

Anne Rimrott and Trude Heift. 2008b. Learner responses to corrective feedback for spelling errors in CALL. *System* 36(2): 196–213.

Trond Trosterud and Heli Uibo. 2005. Consonant Gradation in Estonian and Sami: Two-Level Solution. (Eds) Antti Arppe et al. *Inquiries into Words, Constraints and Contexts:* 136–150.

Linda Wiechetek. 2012. Constraint Grammar based Correction of Grammatical Errors for North Sámi. *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMil 8 – AfLaT2012):* 35–40.

Towards fine-grained readability measures for self-directed language learning

Lisa Beinborn, Torsten Zesch and Iryna Gurevych

Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt

Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for International Educational Research

www.ukp.tu-darmstadt.de

Abstract

In this paper, we analyze existing readability measures regarding their applicability to self-directed language learning. We identify a set of dimensions for text complexity and focus on the lexical, syntactic, semantic, and discourse dimensions. We argue that for the purposes of self-directed language learning, the assessment according to the individual dimensions should be preferred over the overall readability prediction. Furthermore, due to the heterogeneity of the learners in such a setting, modeling the background knowledge of the learner becomes a critical step.

1 Introduction

Readability measures have a long history, especially in American education research (DuBay, 2004). The need for these measures is rooted in a very practical task: teachers search for texts that best fit the knowledge level of their students. According to Vygotsky's zone of proximal development (Vygotsky, 1978), the range of suitable texts that a learner can manage without help is very small. Texts that do not challenge the student easily lead to boredom, while overly complex language might lead to frustration when no tutoring is available. In order to prepare useful reading material for students, readability measures assign the most suitable school grade level to each text. Thus, the readability measures provide an approximation of the text complexity.

The existing readability measures have been developed for standard classroom teaching. As an alternative, *self-directed learning* has lately been on the rise. Self-directed learning refers to a learning setting that does not involve a teacher. It's emergence is closely related to the increased availability of educational material on the web. Students use online exercises for additional training, and companies have discovered digital courses as a flexible alternative to educate their employees. The main advantage of self-directed learning, as opposed to standard classroom education, is the focus on the independence and individuality of the learner (see table 1). The learner can work in her own rhythm independent of time slots or opening hours of institutions and can make her own decisions about the learning content and strategy. A readability measure for self-directed learning thus needs to account for the individual user profiles of learners.

A typical application for self-directed learning is *language learning* where exercises are usually coupled with a text that introduces new vocabulary or new grammatical constructions.[1] It is very important for the learning process that the text fits the proficiency level of the learner. This goal can be reached by applying readability measures which provide an objective analysis about the text complexity. It should be noted, though, that most readability measures have been developed for native speakers rather than for foreign language learners. However, the acquisition of the native language and the learning of a second language are very different processes (see table 2). A readability measure for language learning should take these differences into

[1] In this paper, we use the term language acquisition to refer to the process of acquiring the native language and language learning to the process of learning a foreign language.

Lisa Beinborn, Thorsten Zesch and Iryna Gurevych 2012. Towards fine-grained readability measures for self-directed language learning. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 11–19.

	Classroom	Self-directed
Learner	Homogeneous group	Individual
Learning Mode	Teacher	Independent
Background Knowledge	Curriculum	Individual

Table 1: Differences between classroom learning and self-directed learning

	L1 acquisition	L2 learning
Learner	Young child	Unspecified
Learning Mode	Unstructured	Structured
Background Knowledge	No language knowledge	L1

Table 2: Differences between native language acquisition and second language learning

account.

Self-directed language learning gives the learner the opportunity to improve their command of a language on their own. Advanced learning systems can abstract from pre-defined curricula and adapt the content to the specific learner resulting in a very personalized learning setting (Karel and Klema, 2006). Such an advanced learning framework requires flexible technology that is able to react to the user feedback and continuously update the assumptions of the learner's knowledge.

To the best of our knowledge, the requirements for readability measures for self-directed language learning have not yet been studied in detail. Previous surveys give a historical overview of the evolution of readability measures in the classroom setting. DuBay (2004) introduces the most popular traditional approaches to readability in detail and also presents experimental readability studies. Benjamin (2011) evaluates readability measures according to their usability for teachers. Recently, progress in the field of text classification has led to a new perspective on readability measures not yet captured by previous surveys. Text features from various dimensions are taken into account and combined by supervised learning. In this paper, we present the different approaches to measuring readability and group the introduced text features according to their linguistic dimension. In addition, we discuss how the existing approaches can be adapted to other languages, to language learning, and to self-directed learning.

2 Dimensions of text complexity

A text can be difficult in several ways. A reader might for example know each word of the text, but still fail to capture the constructed meaning. In language learning, these differences are even more evident. If a text uses an unknown grammar construction, the learner will fail to comprehend the text regardless of the vocabulary used. In order to understand and predict the difficulties a learner might have with a given text, a system that aims to support language learning first needs some objective assessment of the text's complexity or its corresponding operationalization, the text's readability. Text complexity is characterized by the following dimensions:

Lexical The text contains rare or ambiguous words.

Morphological Rare morphological particles (word formation processes) are used. This factor is particularly important for agglutinative languages (e.g. Japanese, Turkish).[2]

Syntactic Complex grammatical structures are used. Advanced syntactic constructions (e.g. embedded sentences) increase the complexity of the text.

Semantic Infrequent senses of words are used or meanings are composed in an unusual way (e.g. idioms).

Discourse The argument structure of the text is not explicitly mentioned.

Conceptual The text requires domain knowledge. Texts about philosophy or math might be stylistically easy, but require extensive conceptual background knowledge.

Pragmatic The interpretation of the text is twisted by the text genre. The content might be understandable, but the author's intention needs advanced interpretation as is often the case in satire.

Readability measures automatically estimate the complexity of a text based on features from several

[2]Agglutinative languages use a high number of affixes to change the meaning of a word.

of the above dimensions. In the following, we group the features according to their dimension and elaborate on their use in readability measures. The dominant language for readability approaches is English. Therefore, we neglect the morphological dimension in the overview.[3] As readability measures are not yet capable of capturing the conceptual and the pragmatic dimension, we also omit them. In general, newer approaches incorporate most of the features from the previous work. Therefore, we only discuss the new features each approach contributes.

2.1 Lexical dimension

Features in the lexical dimension capture the difficulty of the vocabulary of a text. The choice of words has a strong effect on the comprehensibility of a text; this holds especially for language learners.

Surface-based measures The traditional readability measures rely on two main features: word length and sentence length. They are computed by the average number of characters (or syllables) per word and the average number of words per sentence[4], and are combined with manually determined weights resulting in a grade level as output. Most prominent methods of this type are the Flesch–Kincaid Grade Level (Kincaid et al., 1975), the Automatic Readability Index (Smith and Senter, 1967) and the Coleman–Liau Index (Coleman and Liau, 1975). The Fry Formula (Fry, 1977) plots the word length and the sentence length on a graph and defines areas for each grade level. The corresponding grade level for a text and also the distance to neighboring grade levels can then easily be read from the graph. In addition to the word and sentence length, the SMOG grade (McLaughlin, 1969) and the Gunning–Fog Index (Gunning, 1969) also consider the number of complex words defined as words with three or more syllables. Some of these surface-based approaches are employed in standard word processors. However, they have also been subject to criticism as they only capture structural characteristics of the text and can easily be misleading.[5]

[3] In section 3, we summarize readability measures for other languages

[4] As a common pre-condition, the text should usually contain a minimum of 100 words.

[5] See DuBay (2004) for a very detailed overview of the strengths and weaknesses of the surface-based measures.

For English, word length is a very good approximation of difficulty, as frequently used words tend to be rather short compared to more specific terms (Sigurd et al., 2004). However, there exist of course many exceptions to this.[6] Alternatively, the method described by Dale and Chall (1948) proposes the use of word lists that are based on the frequency of words. If many words of a text do not occur in the list, this serves as an indicator for higher text complexity.

Language models Instead of absolute frequencies as in word lists, language model approaches are based on word probabilities. The use of language models is a common technique in speech recognition and machine translation in order to determine the probability of a term in a given context. Collins-Thompson and Callan (2005) have shown that this notion of the probability of a term can easily be transferred to readability, since it is generally assumed that a sentence is more readable if it uses very common terms and term sequences. In combination with smoothing methods and pre-processing (e.g. stemming), language models can also account for novel combinations of words. Higher n-gram models as used by Schwarm and Ostendorf (2005) can even account for collocation frequencies indicating different usages of content words (e.g. *hit the ball* / *hit rock bottom*). Language models can easily be re-trained for new domains and new languages; they are therefore particularly suitable in self-directed learning. They return a probability distribution of terms over all readability levels.

Lexical variation The lexical difficulty of a text is not only determined by the choice of words, but also by the amount of lexical variation. If the same concept is expressed by different words, the reader has to recognize the similarity relation of the words in order to understand the shared reference. Lexical variation is usually measured by the type-token ratio (Graesser and McNamara, 2004), where type is a word and token refers to the different usages of the word in the text. A low ratio indicates that words are frequently repeated in the text. This characteristic might decrease the stylistic elegance of the text, but it facilitates text comprehension.

[6] Compare, for example, *together* (length 8, ANC frequency 4004) and *sag* (length: 3, ANC frequency: 27)

2.2 Syntactic dimension

Syntactic features measure the grammatical difficulty of the text. Especially for language learners, complex syntactic structures are major text comprehension obstacles. The surface-based measures estimate the syntactic difficulty by considering sentence length (see section 2.1). However, although a longer sentence might indicate a more complex structure, it could also simply contain an enumeration of concepts. In recent approaches, the grammatical structure is thus represented by part-of-speech (POS) patterns and parse trees, as described below.

POS tagging In readability measures, POS tagging is mainly used for the distinction of content and function words. Content words carry lexical meaning, while function words like articles or conjunctions indicate syntactic relations. A high number of content words indicates high lexical density (Vajjala and Meurers, 2012). Feng and Huenerfauth (2010) additionally determine the absolute and relative numbers of the different POS tags in the sentence and found that a high number of nouns and prepositions is an indicator for text complexity. Heilman et al. (2007) highlight the occurrence of different verb tenses as indicators for text complexity, especially for second language learners. Grammatical constructions are usually acquired step by step and complex structures such as the use of the passive voice occur in later stages. Infrequent verb tenses might thus strongly inhibit a learner's comprehension of the text.

Parsing In addition to POS information, parsing features are used for predicting readability. Syntactic parsers analyze the grammatical structure of a sentence and return a formal syntax representation. For readability measures, the number and type of noun and verb phrases are determined (Schwarm and Ostendorf, 2005; Heilman et al., 2007). In addition, Schwarm and Ostendorf (2005) include the depth of the parse tree and the number of subordinated sentences in order to model the sentence complexity. Similarly, Vajjala and Meurers (2012) consider the number of clauses per sentence and the number of subordinations and coordinations. Another parsing feature, used by Tonelli et al. (2012), is the syntactic similarity of sentences. A text is easier to read if it exhibits low syntactic variability. This can be computed by detecting the largest common subtree of two sentences. When accessing user profiles for second language learning, it is possible to determine even more concrete syntactic features that decrease the comprehensibility of a text for a specific learner.

2.3 Semantic dimension

The semantic dimension is related to the meaning of words and sentences. Lexical semantics captures the meaning of words, while compositional semantics describes the sentence meaning.

Lexical semantics Polysemous words complicate the interpretation of a sentence because they have to be disambiguated first. Words denoting abstract concepts, on the other hand, are considered difficult because they do not describe a concrete object. In the CohMetrix readability framework (Graesser and McNamara, 2004), polysemy and abstractness are determined on the basis of WordNet relations (Fellbaum, 1998). Polysemy is measured by the number of synsets of a word and abstractness is determined by the number of hypernym relations.

Compositional semantics The semantics of a sentence can be represented by semantic networks consisting of conceptual nodes linked by semantic relations. Vor der Brück et al. (2008) applied the semantic Wocadi-Parser (Hartrumpf, 2003) for their readability measure on German texts. They considered the number of nodes and relations in the semantic representation as indicators of semantic complexity. These features correlate well with human judgments of readability, but the parser often fails to build a representation, limiting the robustness of their approach. The concepts of polysemy and abstractness can be determined more easily.

2.4 Discourse dimension

In the readability literature, all intersentential relations are perceived as discourse related. Discourse features model the structure of the text as indicated through cohesive markers and the coherence of arguments through reference resolution.

Cohesion An important indicator for text cohesion is the use of discourse connectives. Pitler and

Nenkova (2008) build a discourse language model based on the annotations from the Penn Discourse Bank. This model determines how likely it is for each grade level that the text contains implicit or explicit discourse relations. Tonelli et al. (2012) manually create a list of additive, causal, logical, and temporal connectives for Italian. In addition, they capture the "situation model dimensions of the text" by calculating the ratio between causal or intentional particles and causal or intentional verbs. Causal and intentional verbs are identified manually by exploiting category and gloss information from WordNet.

Coherence The coherence of a text can be measured by the pronoun density. If concepts are not named directly, but referenced by a pronoun, the resolution of the meaning is more difficult. Graesser and McNamara (2004) analyze co-references in more detail by determining the relations between two consecutive sentences. Noun overlap and stem overlap in the sentence pair are both indicators for coherence. Alternatively, Pitler and Nenkova (2008) generate entity grids that capture how the center of attention shifts from one entity in the text to another as postulated in the centering theory (Grosz et al., 1995). Feng and Huenerfauth (2010) keep track of the number of entity mentions. Additionally, they assume that a higher number of active entities poses a higher working memory load on the reader. In order to determine the active entities, they identify lexical chains. A lexical chain is formed by entities that are linked through semantic relations such as synonymy or hyponymy. The length and the sentence span of the chain are interpreted as indicators for text complexity.

2.5 Combining features

From a diachronous view, readability measures have continuously taken more and more features into account. Early measures in the 1960s worked only with surface-based features and manually adjusted the parameters. Later approaches successively added features from the lexical, syntactic, semantic, and discourse dimensions as the respective technologies became available. As the number of features was steadily growing, the need for machine learning methods emerged. Supervised learning methods use training data to determine the significant features for each grade level. Using the learned feature weights then enables the prediction of grade levels for unseen texts. A common training corpus contains news articles for educational use from the WeeklyReader[7] that are labeled according to the US grade levels. Several learning algorithms have been applied for readability measures—e.g. Naïve Bayes (Collins-Thompson and Callan, 2005), k-nearest neighbors (Heilman et al., 2007), support vector machines (Schwarm and Ostendorf, 2005) and linear regression (Pitler and Nenkova, 2008). Tanaka-Ishii et al. (2010) used data annotated with only two different reading classes. This enabled the use of a sorting algorithm that sorts texts according to their readability instead of returning an absolute value. Heilman et al. (2008) compare different machine learning approaches that respectively interpret the readability grades as nominal, ordinal and interval scales of measurements. In their setting, interpreting the readability scores as ordinal data performed best. Thus, the scores are considered to have a natural ordering, but they are not evenly spaced.

The use of feature combination for readability measures has become the common approach, but it has not yet been discussed how these need to be adapted to other languages, to language learning, and to self-directed learning.

3 Adaptation to other languages

The applicability of the explored readability features to other languages is poorly studied because most approaches focus on English. Statistical methods such as language models can easily be adapted to other languages, parsers and POS-taggers are not always available in a comparable quality. Several researchers ported the methods that worked successfully for English to other languages. François and Fairon (2012) implement a readability measure for French, and Aluisio et al. (2010) for Portuguese. Tonelli et al. (2012) rely on the CohMetrix framework and implement an Italian version of the features.

However, features established for English are not necessarily significant for languages with different properties. The particular characteristics of a given language should also be considered in the feature se-

[7] http://www.weeklyreader.com/

lection. Collins-Thompson and Callan (2005), for example, come to the conclusion that their language model–based approach heavily benefits from stemming when applied to the more inflected language French. Similarly, Dell'Orletta et al. (2011) introduce morphological features for Italian. Vor der Brück et al. (2008) present a readability measure for German and also rely on extensive morphological analysis. In addition, they add features specific to German such as the distance between a verb and its separable prefix. Larsson (2006) introduce a new feature for Swedish that identifies subordination by the use of the Swedish conjunction *att*. Sato et al. (2008) present a readability measure for Japanese and introduce new features in order to deal with the different character sets. Another problem for Japanese is the detection of word boundaries as they are not indicated by white space. Al-Khalifa and Al-Ajlan (2010) experiment with readability measures for Arabic and address similar issues related to the different character set.

These examples show that readability can be measured by different text characteristics depending on the specific language. More focused research is necessary in order to determine the most predictive features for each language. However, some major features such as lemma frequency are shared across most languages. They can approximate the readability even for under-resourced languages.

4 Adaptation to language learning

The acquisition of the native language (L1) and the process of learning a second language (L2) evolve in different ways. The three main differences are the age of acquisition, the mode of acquisition and the background knowledge (see table 2). Most of the introduced readability measures have been established for native speakers of English, while aspects of foreign language learning have not yet been studied in detail. Vajjala and Meurers (2012) use features that are motivated in the evaluation of language learners' written production. However, it remains unclear how these features differ from those for native speakers.

L2 learner grades The native language is usually acquired in the first years of childhood, while an L2 is generally learned on top of the L1.[8] This means that a certain level of proficiency in the L1 already exists. As learners are older when learning an L2, they also tend to have a more advanced educational background and have already developed higher intellectual abilities (Cook et al., 1979). On the other hand, L2 learning usually progresses significantly slower than the native language acquisition. Due to these differences, school grade levels indicating the readability of L1 texts cannot be directly mapped to foreign language learning, but rather need to be learned individually from L2 data. The readability for L2 texts should thus not be expressed in school grades, but in L2-specific learner levels.

Fine-grained feedback Language acquisition of the native language is a strongly debated topic in psychology and pedagogy. We will not further elaborate on the cognitive aspects of this process. However, one general difference of the learning setting needs to be considered: the basic L1 knowledge is learned from the unstructured input children receive from the environment, while an L2 is usually learned gradually by instruction (Cook et al., 1979). An L2 can also be learned by simple exposure to the language (informal language learning (Bahrani and Sim, 2012)), but it is usually a more conscious process that also requires more structured input (Schmidt, 1995). School children have already acquired the basic structures of their L1, while L2 students need to actively learn new syntactic regularities. This indicates that the output of readability measures has to be more fine-grained than standard school grades.

The evaluation of supervised learning approaches has shown that syntactic features in isolation perform significantly worse than lexical features in predicting the correct school grade for L1 texts (Heilman et al., 2007; Feng and Huenerfauth, 2010). The syntactic features contribute only slightly to the improvement of the overall readability prediction. However, for L2 learners the extensive use of an unknown verb tense can be a stand-alone criterion for the readability of a text. In the feature combination, this individual information might be lost and cannot fully characterize the text complexity. An ap-

[8] Except for bilingual children who acquire two languages simultaneously

propriate readability measure for L2 learning should thus provide more fine-grained information about the readability. As a result, the language learning system receives information about the lexical, syntactic, semantic, and discourse difficulty of the text and can adapt the learning setting accordingly.

Consideration of L1 The L2 learning is influenced by the background knowledge of the learner. As L1 is already present, basic concepts of languages such as the different behavior of word classes or the syntactic coordination of arguments are already known. In addition, the specific properties of the L1 influence the acquisition of the L2. The phenomena of cross-linguistic transfer have been heavily researched (Odlin, 1989; Zobl, 1980). For example, foreign words that have a similar stem as the translation in the mother tongue are acquired more easily. Similarly, syntactic structures that are comparable across the two languages are less error-prone than idiosyncratic aspects of the L2. Thus, readability measures should account for the native language of the learner and should be adapted to groups of users sharing a common mother tongue. The consideration of a learner-specific feature establishes a focus on user profiles which is even more relevant for self-directed learning.

5 Adaptation to self-directed learning

In the setting of self-directed learning, the user profiles can be more heterogeneous than in school classes. The users differ in age, previous knowledge, intellectual ability, and educational and cultural background, and also might have different learning goals. To account for this, a fine-grained learner model is needed, which captures the learner's knowledge and preferences (Al-Hmouz et al., 2010). A model needs to be instantiated based on the learner's knowledge and updated according to the ongoing performance. The previous knowledge can either be estimated by a pre-test or automatically learned from texts that the learner has already mastered. The update function should dynamically assess the performance in exercises and also consider the learner's usage patterns of the system in order to identify preferences for certain exercises. For example, Virvou and Troussas (2011) maintain an error model in order to keep track of the learner's weaknesses.

The learner model needs to be incorporated into the readability measure in order to determine the readability of a text for one specific learner. The readability measure should model the discrepancy between the characteristics of the text (represented by the extracted features) and the learner's knowledge (represented by the learner model). Thus, the measure models not only the general readability of the text, but also its *suitability* for a specific learner.

A personalized language model that represents the learner's lexical knowledge could be directly compared to the lexical features of the text. However, a one-to-one mapping from knowledge representation to features is not always possible. For example, if the learner has a recorded preference for sports texts, this translates into features from several dimensions (i.e. advanced sports vocabulary, preference for factual style, acquaintance with sports entities, and domain knowledge). As an approximation, the readability measure could assign a degree of difficulty to each dimension. Each dimension can then be looked up in the learner model to verify the competence level of the learner. The suitability of a text for a specific learner could then be expressed by the discrepancy between the learner competence and the text characteristics for each dimension. This allows more fine-grained support for the text elements that cause difficulties for the learner.

6 Application

In an adaptive language learning system, automatic exercise generation plays an important role in accounting for the variability of learners. A precondition for useful automatic exercise generation is a readability measure that gives fine-grained information about the suitability of a text for a certain learner.

Generating suitable exercises for language learning can be approached from two perspectives: it can either be input-driven or determined by a curriculum. The input-driven method utilizes the learner's interests and is embedded into her routines. The learner can select a text in the foreign language that appears particularly interesting or that needs to be read anyway. The system then generates questions on the basis of the text (bottom-up) in order to fa-

cilitate comprehension and to assist with unknown words or constructions.

In the curriculum method, the learning goal is predefined by a learning framework (i.e. realizations of the learner levels as defined by the *Common European Framework of Reference for Languages*[9]). The learner is supposed to learn a new concept (e.g. a grammatical phenomenon, a group of related words) and the exercises are generated in order to reach this goal (top-down). In addition to the learning goal, the exercises should also consider the previous knowledge of the student. A text that meets the learner's interests and knowledge level better stimulates the intrinsic motivation to learn.

For input-driven scenarios, the readability measure can help to extract the dimension of the text that causes comprehension difficulties and trigger exercises to resolve them. The exercise type and the exercise difficulty will thus be determined by the readability outcome—e.g. low readability in the lexical dimension triggers vocabulary exercises. In the case of a given learning goal, the measure helps to acquire the most suitable reading material that best matches the user's profile and fulfills the requirements of the learning goal.

7 Conclusions

In this paper, we gave an overview of readability measures from the perspective of self-directed language learning. We discussed how readability measures need to be adapted in order to consider the requirements of other languages, the different progress levels in L2 acquisition, and the characteristics of user profiles. We suggest the introduction of L2 learner grades and a more fine-grained level of readability feedback. In addition, we propose to assess the suitability of a text with respect to a user model. In the future, we will further develop and implement the proposed measures, and apply them for automatic exercise generation.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Klaus Tschira Foundation under project No. 00.133.2008.

[9] http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

References

Ahmed Al-Hmouz, Jun Shen, Jun Yan, and Rami Al-Hmouz. 2010. Enhanced learner model for adaptive mobile learning. In *12th International Conference on Information Integration and Web-based Applications & Services - iiWAS '10*, pages 783–786, New York, USA. ACM Press.

Hend S. Al-Khalifa and Amany Al-Ajlan. 2010. Automatic Readability Measurements of the Arabic Text: An Exploratory Study. *The Arabian Journal for Science and Engineering*, 35(2C).

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, June.

Taher Bahrani and Tam Shu Sim. 2012. Informal language learning setting: technology or social interaction? *The Turkish Online Journal of Educational Technology*, 11(2):142–149.

Rebekah George Benjamin. 2011. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24(1):63–88, October.

Mery Coleman and TL Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting Reading Difficulty with Statistical Language Models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Vivian J. Cook, John Long, and Steve McDonough. 1979. First and second language learning. *The Mother Tongue and Other Languages in Education*, pages 7–22.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, July.

William H. DuBay. 2004. The Principles of Readability. *Impact Information*, pages 1–76.

Christiane Fellbaum. 1998. *WordNet: An electronic database*. MIT Press, Cambridge, MA.

Lijun Feng and Matt Huenerfauth. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of COLING 2010*, pages 276–284, August.

Thomas François and Cédrick Fairon. 2012. An AI readability formula for French as a foreign language. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, July.

Edgar Fry. 1977. Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*.

Arthur C. Graesser and Danielle McNamara. 2004. CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2).

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Robert Gunning. 1969. The Fog Index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Sven Hartrumpf. 2003. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany.

Michael J Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of NAACL-HLT*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications - EANL '08*, pages 71–79, Morristown, NJ, USA, June.

Filip Karel and Jií Klema. 2006. Adaptivity in e-learning. *Current Developments in Technology-Assisted Education*, 1:260–264.

John P. Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, DTIC Document.

Patrik Larsson. 2006. *Classification into Readability Levels Implementation and Evaluation*. Master's thesis, Uppsala University, Sweden.

G. Harry McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Richard Schmidt. 1995. Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and Awareness in Foreign Language Learning*, pages 1–63.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 523–530, June.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost van Weijer. 2004. Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, 58(1):37–52, April.

E. A. Smith and R.J. Senter. 1967. *Automated readability index.* Cincinnati University Ohio.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting Texts by Readability. *Computational Linguistics*, 36(2):203–227, June.

Sara Tonelli, Ke Tran Manh, and Emanuele Pianta. 2012. Making Readability Indices Readable. In *Proceedings of NAACL-HLT: Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 163–173.

Maria Virvou and Christos Troussas. 2011. Web-based student modeling for learning multiple languages. In *International Conference on Information Society (i-Society)*, pages 423–428.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. In *11th International Multiconference: Information Society-IS*, pages 92–97.

Lev Vygotsky. 1978. Mind in society: The development of higher psychological processes. Harvard University Press.

Helmut Zobl. 1980. Developmental and transfer errors: their common bases and (possibly) differential effects on subsequent learning. *TESOL Quarterly*.

An academic word list for Swedish

- a support for language learners in higher education

Carina Carlund
carina.carlund@svenska.gu.se
Sofie Johansson Kokkinakis
sofie@svenska.gu.se
Judy Ribeck
judy.ribeck@svenska.gu.se

Håkan Jansson
hakan.jansson@svenska.gu.se
Julia Prentice
julia.prentice@svenska.gu.se

All authors at the Department of Swedish
University of Gothenburg, Sweden

Abstract

The paper describes the ongoing development of compiling and introducing a Swedish academic word list (SAWL), inter alia intended to be used as a lexical resource in CALL-applications in relation to higher academic studies. When it comes to language acquisition, resources like these play an important part in instructed language learning. So far, no such resource exists for Swedish. The format of SAWL has been elaborated in collaboration with the Language Support Service at the University of Gothenburg. SAWL is compiled with methods from corpus linguistics inspired by research on English academic words (Coxhead 2002). Our work includes collection and syntactic annotation of learner corpora of Swedish academic texts from a wide range of university subjects within the Faculty of Arts. The corpora are freely accessible through Språkbanken. SAWL are designed with university students and language learners with Swedish or other linguistic backgrounds in mind. The word list and the corpora can be used for studies of one's own or in classroom situations, as well as forming a component of computer computer-based language assessment and CALL-related application platforms.

1 Introduction

The language in academic studies and in teaching is often a challenge for both L1 students without an academic background and L2 students. In order to meet the language demands, university students must not only master a subject's specific vocabulary, but also be able to understand and use a more general academic vocabulary, which is common within a range of study areas. To meet the students' need for knowledge of this type of vocabulary a number of English academic word lists have been developed. Our aim is to compile and offer a similar resource in the Swedish academic context.

An academic word and phrase list would serve as a valuable resource for L2 students in particular, but also for L1 students during their first year of university studies, a period during which many students struggle to meet the demands set by their academic studies, not least linguistically. In order to master both written and spoken academic language use, one has to be able to understand and use conventionalized formulaic expressions that are typical for academic discourse. Hence, in addition to a list of individual academic words, L2 students and students who are lacking experience of academic studies can be expected to have use for a resource that lists and describes multi-word expressions that are relevant for Swedish academic language (c.f. Ellis et al. 2008:379).

The Language Support Service at the University of Gothenburg had conducted a small user study of words in academic text and further user studies are planned. It is thought that the word and phrase list will be used in the language tutors' work, by other course teachers and by the students themselves. Today the development is towards computer based applications in the teaching of language and an academic word and phrase list is a resource that is suitable for CALL.

Carina Carlund, Håkan Jansson, Sofie Johansson Kokkinakis, Julia Prentice and Judy Ribeck 2012. An academic word list for Swedish – a support for language learners in higher education. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 21–28.

The project of compiling a Swedish academic word and phrase list, which is also part of a wider Nordic collaboration, must also been seen from a language perspective. New documents from many Nordic universities have expressed concern about the increased use of English within academia to the detriment of the national language. For example, a study from the language council in Sweden demonstrated that 20% of all Swedish theses are now written in English (Salö 2010).

Increased internationalization in the academic world has the positive effect of increasing dissemination of research results and has increased academic mobility, but the fact that teaching and research more and more are conducted in English can lead to domain loss of the native language in certain areas. In addition, studies have pointed out a number of negative effects on study results when lectures and the interaction between Swedish students and teachers are mainly conducted in English (see Salö 2010: 8, 14-19).

2 Previous research

There has been a few attempts on the creation of academic vocabulary resources, so far mainly for learners of English but also for Portuguese (Baptista et al. 2010) and French (Cobb and Horst 2004). In this paper we describe the English one, being the best documented. The Academic Word List (Coxhead 2000), contains words believed crucial to higher education independent of study orientation, for instance *analyze*, *distribution* and *indicate*.

Also, academic vocabulary is highlighted in some general learners' dictionaries of English. However for students of the Swedish language, similar support is not yet available (see Johansson Kokkinakis et al. 2012).

2.1 The Academic Word List for English

[In the late 1990s, Coxhead presented her Academic Word List (AWL) for English. She believed in her approach that the content of an academic word list should be based on relevant principles within corpus linguistics. Therefore Coxhead compiled a corpus of academic texts to be able to extract the word list from.

The Academic Corpus consists of 3.5 million tokens. It contains 414 texts (mainly articles and text books) by more than 400 different authors. The data is spread equally across four disciplines: the arts, commerce, law and science. Each discipline is divided into seven subject areas (see table 1).

Arts	Education, history, linguistics, philosophy, politics, psychology, sociology
Commerce	Accounting, economics, finance, industrial relations, management, marketing, public policy
Law	Constitutional, criminal, family and medicolegal, international, pure commercial, quasi-commercial, rights and remedies
Science	Biology, chemistry, computer science, geography, geology, mathematics, physics

Table 1. Subject areas in the four AWL disciplines (Coxhead 2000:220).

The arts discipline contains subject areas such as education, history and psychology. To be included in the AWL, the members of a word family (West 1953; Bauer and Nation 1993) cumulatively had to occur at least 100 times in the entire corpus, ten times in each of the four disciplines and in 15 of the subject areas. The entries in the AWL are word families, each of which is a stem plus all closely related affixed forms (Coxhead 2000). An example of a word family is: contribute - contributed, contributes, contributing, contribution, contributions, contributor, contributors.

The AWL contains 570 word families frequently found in Coxhead's Academic Corpus. The word families are not among the 2,000 most frequently occurring English words, as described in The General Service List (West 1953). By using the concept of word families Coxhead concur in the tradition of previous creators of vocabulary lists for language learners (cf. West 1953, Xue and Nation 1984). Her motivation for this choice is that the use of word families "is supported by evidence suggesting that word families are an important unit in the mental lexicon" (Coxhead 2000:217f.).

As the name indicates, the AWL is a plain word list. It consists of word families, graphically indicated with an initial head word followed by family members – in the case there are any. There is however no information on the head words' or the family members' pronunciation, grammatical paradigms, meaning or collocational properties. The fact that there is so little information included in the list limits its use in academic settings as well as its use for lexicographic purposes. Advice for language learners on how to use the list is described at:

<http://www.victoria.ac.nz/lals/resources/academicwordlist/>.

Criticism

Since its release, the AWL has hugely influenced the curricula of English for academic purposes and English as a second/foreign language (Hyland and Tse 2007, Granger and Paquot 2009). Nevertheless, Coxhead's selection methods and presentation have been criticised.

Like Hyland and Tse (2007), one can certainly question Coxhead's division into disciplines and subject areas. As Nesi (2002) points out, it would be favorable if the division were transferable across institutions to enable comparison of different academic corpora. We believe that the difference in the word list's coverage within different disciplines and the dominance of commerce words, reported by Coxhead (2000), have to do with the fact that commerce is more homogenous than for instance science.

Eldridge (2007) and Hyland and Tse (2007) also question the usability of the actual list – for reception and production, as well as the benefit of word families for learners at different proficiency levels. They call for sense descriptions in general and subject-specific senses in particular, as well as combinatorial properties in relation to the words. They argue that the members of a word family should rather be taught separately, since their collocational patterns tend to differ.

3 Resources and Method

Building on previous work on academic word lists, as presented above, there would be two main routes for this project to pursue: One could either simply translate Coxheads English list into Swedish or one could compile a corpus of Swedish Academic texts.

3.1 Translation of Coxheads AWL?

The translation path has been followed by a similar Portuguese project (P-AWL, Baptista et al, 2010), and also by other similar projects. Thus a Finnish WordNet has been produced, applying translations techniques to Princton WordNet (Lindén and Carlsson 2010) and a Norwegian LEXIN learners dictionary has been made based on the translation of the corresponding Swedish dictionary, (Bjørneset 2001). There are however some limitations connected to the translation method.

Martola (2011) lists some of the shortcomings of the Finnish WordNet, which are tied to its translation from English. Apart from the culturally specific semantic problems pointed out by Martola, there are also issues that are of a more lexical/morphological nature, which are partly connected to Coxheads notion of word families. These problems came to light when 60 headwords from sublist 1 of the AWL were compared to their Swedish translation equivalents in the dictionary *Norstedts stora engelsk-svenska ordbok* (2000).

Only a few of the words, e.g. *percent*, are easy to translate. More than a third of the words e.g. *contact* and *issue* are homographs and most words are polysemous. The English word families will inevitably be split up in a translation. For a further discussion of the translation method and some of its issues, e.g. the problems with the implications of the notion of word families, c.f. Sköldberg and Johansson Kokkinakis (2012).

3.2 Corpus collection

The translation option was subsequently discarded. Instead a decision to aim for a Swedish corpus of academic texts was taken. After finishing some pilot studies, designed to evaluate the effectiveness of different corpus compilation methods, reported on in Jansson et al. (2012), it was decided to compile a corpus from documents published in the Swedish national academic online database, SwePub <http://swepub.kb.se/>, kept by the National Library of Sweden.

An advantage with the use of that particular source is that all the documents have been catalogued in compliance with the guidelines set by the Swedish National Agency for Higher Education, which in turn are based on the OECD classification Field of Science and Technology (OECD. Organisation for Economic Co-operation and Development 2007). This foundation of our corpus in an official typology of Academic subjects provides an unbiased text subjects division and facilitates an easy comparison between countries, since it falls back on an OECD standard. As noted above, Nesi (2002) stresses that more uniform corpus subdivisions across different languages and groups would enable comparison of different academic corpora.

It should be noted that the subject of one entire subcorpus of Coxheads e.g. *commerce*, compares to OECDs *business and management* which is a secondary subdivision of the field *social sciences* in OECD typology, Coxhead (2002:75), OECD (2007).

3.3 The Arts corpus

Since the use of the Swedish language is not evenly spread over the different fields of science, we decided to start with a corpus using theses and other academic publications from the arts, which is the most widely represented field in Swedish (see Salö 2010). The subjects chosen were ethnology, history, linguistics, literature, philosophy and religious studies.

The corpus comprises approximately 220 documents by more than 140 authors and contains roughly 11 million tokens (punctuation marks excluded). It has been divided into subcorpora with regard to the already mentioned subjects, as well as the document types Ph.D. theses, Articles, and Other. The SwePub database allows searches with the above specifications, so the corpus compilation was uncomplicated, although each document had to be downloaded manually.

	Ph.D. theses	*Articles*	*Other*	Total
Ethnology	1,210,735	69,047	168,712	**1,448,494**
History	2,119,048	93,721	95,312	**2,308,081**
Literature	1,753,839	205,482	26,616	**1,985,937**
Linguistics	1,544,166	156,921	228,058	**1,929,145**
Philosophy	454,266	48,157	140,892	**643,315**
Religious studies	2,282,125	48,794	288,615	**2 619,534**
Total	**9,364,179**	**622,122**	**948,205**	**10,934,506**

Table 2. Subjects and text types in the Arts corpus

Table 2 shows the distribution of words in the corpus. As can be seen, the subcorpora vary in size. More specifically, philosophy is considerably smaller and ethnology somewhat smaller than the other subjects, but this reflects the total amounts of documents in the SwePub-database.

The texts were first cleaned from markup and code by uploading them into the Sketch Engine (for ref. see Kilgarriff et al., 2004). Then they were downloaded and subsequently tokenised, lemmatised and pos-tagged at Språkbanken.

3.4 Word selection

The principle for word selection for the list is based on the aim of finding an academic-specific vocabulary that is common for all subjects at the university, but not part of the everyday language.

As pointed out by Savický and Hlaváčová (2002), there is no formal definition of the intuitive notion of "commonness" when trying to rank words of the language. Most often, absolute or relative frequency of words in a corpus has come to denote commonness. This however is far from an optimal measure.

To obtain a more objective measure of word commonness, one has to look not only at frequency, but also at the distribution of that frequency. This is what is done by means of different types of *corrected frequencies* (Savický and Hlaváčová 2002).

Reduced frequency

The sort of corrected frequency we applied is called *reduced frequency*, RF[1] (Hlaváčová 2000; Savický and Hlaváčová 2002) and is calculated as follows:

Let *f(x)* be the frequency of word x in a corpus consisting of *N* tokens. Then divide positions of the whole corpus into *f(x)* intervals $< i , j >$. For $n = 1 \ldots f(x)$, the *n*:th interval is:

$$< \ [(n-1)N/f(x) +1] \ , \ [nN/f(x)] \ >$$

Let *Fx* be the partial frequency of *x* as:

Fx(n) = 1, if x occurs in the *n*:th interval
Fx(n) = 0, otherwise

RF(x) is then simply the sum of all partial frequencies for *x*:

$$RF(x) = \sum_{n=1}^{n=f(x)} Fx(n)$$

RF ensures the frequencies to be spread across the corpus without requiring the corpus to be divided into sub corpora according to for example genres or text types. This is a great advantage to other measures of dispersion, since "any trial of text annotation brings plenty of problems, which are difficult, if not even impossible to resolve… Moreover there is no strict border between genres…" (Savický and Hlaváčová (2002:216f.).

The RF for evenly distributed words is closer to their absolute frequency, and the RF for unevenly distributed words is smaller than their absolute frequency.

Keywords

To automatically identify domain-specific vocabulary, we ranked the lemmas according to keywordness (Scott 1997). The reference corpus was set to a 2.5-millon token collection of novels from Nordstedts, available through Korp at Språkbanken. The first selection criterion we

[1] After conducting some tests on our material, we decided not to use the Average Reduced Frequency described in Savický and Hlaváčová (2002) and Hlaváčová (2006). The results showed that RF was sufficient, since the values of RF and ARF hardly differed.

applied was for a lemma to score above 1.1 in keywordness.

Range
The second selection criterion was a requirement for the lemmas to have a relative RF of at least 15 per million tokens in each of the university subjects. By applying this demand of range, we increased the remedy for the "burstiness" problem (Kilgarriff 2009), which still was salient in our preliminary list. Moreover, we wanted to be sure that the words really were common to all subjects included.

Some examples of lemmas ruled out at this stage were: *präst* 'priest', *världskrig* 'world war', *sexualitet* 'sexuality', *kung* 'king', *författarskap* 'authorship', *medeltid* 'Middle Ages', *lagstiftning* 'legislation', *ordbok* 'dictionary' and *syntaktisk* 'syntactic'.

Filtering out non-everyday words
The third selection criterion was that the lemmas should not be part of the most frequent words of everyday Swedish. The filtering was done by removing all lemmas that belonged to the 1000 most frequent words of the 1.1-million token corpus LäSBarT available through Korp at Språkbanken. This corpus contains children's books and other easily read texts.

Words ruled out at this stage were for instance: *svensk* 'Swedish', *exempel* 'example', *språk* 'language' and *istället* 'instead'.

Manual processing
The final step was to manually clean the list from unwanted noise, such as abbreviations like *s.* 'p.', *t.ex.* 'e.g.', *jfr* 'cf./cp.' and *eds.*, numerals and text-structuring tokens as *ii.*.

We also brought some entries together that were tagged as different parts-of-speech [2], although according to modern lexicographic tradition belong to the same entry. As an example, words tagged as both adjectives and adverbs, e.g. *speciell* 'special', only appears as an adjective in the final list.

4 The resulting list

Our methodology for identification of academic words has resulted in a word list of 750 entries.

4.1 Entries

The 10 topmost entries of the list according to keywordness are: *dock* 'however', *relation* 'relation', *samt* 'and', *studie* 'study', s*ocial* 'social, public', *begrepp* 'concept', *form* 'form', *betydelse* 'meaning, importance', *analys* 'analysis' and *utifrån* 'on the basis of'.

We regard the lack of information about the words in the AWL to be a drawback. The entries in our list are annotated with: [3]

1. part of speech
2. inflectional forms
3. meaning
4. one (or more) editorial examples based on instances in the corpus
5. English translations.

To exemplify what the entries look like, we can look at the word *innebära* (imply, mean).

> **innebära** (verb) *innebar*, *inneburit*; *innebär* • betyder, medför. *Vårdnadsansvaret innebär både rättigheter och skyldigheter för dig som förälder; Romerskt medborgarskap innebar en mängd friheter och privilegier*. 'imply, mean'.

As far as the meanings are concerned, all the meanings given in *Lexins svenska lexikon* (2011) are included, even the ones that may not be that common in the academic texts. This approach was chosen since not all instances of this dilemma were entirely intuitively obvious.

The examples should function as an aid to the information about meaning. The intention is that they should be illustrative of one of the given meanings – preferably the one most common in the corpus. To facilitate for the users, the examples are editorial, which means that they are based on authentic occurrences in the corpus, but depicted with less or simplified context when needed. In the online version of the list, the user can easily follow a link to the corpus and look at actual concordances.

[2] Pos-tagging was made by means of the open source hunpos-tagger, which implements the TnT-tagger. The tagger is trained on data from SUC 2.0 from which the pos-tags derive.

[3] So far, this work has been carried out for the first 100 entries by a lexicographer. The information about part of speech, inflection and meaning are drawn from the recently revised 4th edition of *Lexins svenska lexikon* (2011) supplied by Språkrådet. The English translations are taken from *Lexins svensk-engelska lexikon* supplied by Språkbanken.

4.2 Coverage

With regard to a previous categorization of word types, Nation (2001) concludes that the vocabulary of academic texts consists of 80% of the most common and frequent words, 8-10% general academic words and 5% subject specific and technical words.

The 750 words of our list cover on average 8.7% of our Arts corpus (10.1% linguistics, 7.9% history, 8.1% ethnology, 7.9% literature, 10.4% philosophy and 9.1% religious studies). This can be compared with the 10.0% coverage of the AWL reported by Coxhead (2000). We believe the smaller coverage of our list can be explained by at least three factors.

First and foremost, we apply much more rigorous selection criteria. The words of the AWL are chosen as a consequence of frequency and range alone, while we also require certain keywordness in relation to a reference corpus, as well as considering the distribution of the frequencies (dispersion). We strongly believe that this approach will assure a high precision of academic vocabulary. Besides that, total recall was never our goal. Most important for us was to identify a crucial vocabulary for academic achievements, in that knowledge of the words would help students in their academic studies.

Second, the entries of the AWL are word families (see 2.1), while we have lemmas. Word families may contain lemmas from different parts of speech as well as affixated word forms, e.g. [*available, availability, unavailable*]. Since the selection procedure for the items of the AWL adds the frequencies of all the members of a word family, not all members alone need to fulfill the requirement for inclusion. Still, these "additional" members contribute to the overall coverage of the AWL.

Third, academic texts written in Swedish contain a non negligible amount of non Swedish language, for example in citations and summaries. Since we only included Swedish words in the list, foreign language in the corpus was never going to be covered.

5 Conclusions and accessibility

This paper describes the use for and the creation of an academic word list for Swedish. The method describes an approach where a list of 750 lexical items is extracted from a compiled corpus of Swedish academic texts publically available through Språkbanken. The overall coverage of the word list is 8.7% of the corpus.

The word list is available, both from Språkbanken and as a freely downloadable lexical recource – *En svensk akademisk ordlista*, version 1.0, <http://spraakbanken.gu.se/ao/>.

The list is shown online and is downloadable in two formats. On the one hand, there is a listing of all the 750 headwords, which can be viewed in alphabetical order or according to keywordness. On the other hand, there is the fully annotated top-100 list of words according to keywordness.

6 Future research and applications

The described lexical resource, SAWL, is intended to be used in language learning both individually and in academic class room settings.

6.1 Research

The next immediate step will be an evaluation process of the usefulness of the extracted lexical items, in collaboration with the University of Gothenburg language support service.

As an extension to SAWL, inspired by the research carried out by Ellis et al (2008), and Simpson-Vlach and Ellis (2010) that resulted in an Academic Formula List (AFL) for English, we are also aiming to use various methods within the fields of i.e. language technology, corpus linguistics and psycholinguistics, to develop a list of conventionalized multi-word expressions for Swedish academic language. As Ellis et al. point out, it has been established in relatively recent research "that highly frequent formulaic expressions are not only salient but also functionally significant: Cognitive science demonstrates that knowledge of these formulas is crucial for fluent processing" (Ellis et al. 2008:379).

In addition to the research questions mentioned above, the next step in extending the corpus in the subject areas social sciences and natural sciences. The latter being more difficult since English is used more often in those subjects.

6.2 Applications

Regarding how to implement SAWL in computer-based applications, the aim is twofold; one goal is to use it as a validated and reliable lexical resource in language assessment platforms similar to those implemented in the testing part of the "Complete Lexical Tutor" for assessment of English general and language specific vocabulary tests <http://www.lextutor.ca> and in the testing of Swedish vocabulary for secondary and upper secondary school in the OrdiL-project (Lindberg

and Johansson Kokkinakis, 2007). Another Swedish project modeling different aspects of the lexical knowledge of a language learner in vocabulary assessment is the MOA-project (Lindberg and Johansson Kokkinakis, 2011). In these two projects language pedagogical aspects are emphasized and benefits from focusing on everyday vs. scientific language. Research has so far shown that students with a different language background encounter difficulties with polysemous words, in particular those with subject-specific senses which sometimes also have a more general everyday sense.

Another goal is to incorporate SAWL as a lexical resource in CALL-based platforms cf. the Swedish Lärka <http://spraakbanken.gu.se/larka> which is under development in Språkbanken at the University of Gothenburg.

References

Jorge Baptista, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria Cabral and Nuno Mamede. 2010. P-AWL: Academic Word List for Portuguese. Computational Processing of the Portuguese Language, In: Lecture Notes in Computer Science, 6001/2010:120–123.

Laurie Bauer and Paul Nation. 1993. Word families. International Journal of Lexicography, 6:253-279.

Tove Bjørneset. 2001. Introduksjon til ordboksprosjektet NORDLEXIN-N. In: Martin Gellerstam, Kristinn Jóhannesson, Bo Ralph and Lena Rogström (Eds.) Rapport från Konferens om lexikografi i Norden Göteborg 26-29 maj 1999. (Nordiska studier i lexikografi 5.): 44–53. Göteborg.

Tom Cobb and Marlise Horst. 2004. Is there room for an academic word list in French? In: Paul Bogaards and Batia Laufer (Eds.) Vocabulary in a second language: Selection, acquisition, and testing, 15-38. John Benjamins, Amsterdam.

Averil Coxhead. 2000. A new academic word list. TESOL Quarterly, 34(2):213–238.

Averil Coxhead. 2002. The Academic Word List: A corpus-based word list for academic purposes. In: Bernard Ketteman and Georg Marko (Eds.) Teaching and Language Corpora (TALC) 2000 Conference Proceedings. Rodopi, Atlanta.

John Eldridge. 2007. "No, There Isn't an 'Academic Vocabulary,' But…": A Reader Responds to K. Hyland and P. Tse's "Is There an 'Academic Vocabulary'?". TESOL Quarterly, 42(1):109-113.

Nick. C. Ellis, Rita Simpson-Vlach and Carson Maynard, (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpuslinguistiscs and TESOL. TESOL Quarterly 42(3):375–396.

Jaroslava Hlaváčová. 2000. Rarity of words in a language and in a corpus. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000):1595-1598.

Jaroslava Hlaváčová. 2006. New Approach to Frequency Dictionaries – Czech Example. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006):373-378.

Sylviane Granger and Magali Paquot. 2009. Lexical Verbs in Academic Discourse: A Corpus-driven Study of Learner Use. In: Maggie Charles, Diane Pecorari and Susan Hunston (Eds.) Academic Writing: At the interface of corpus and discourse. Continuum, London/New York.

Ken Hyland and Polly Tse. 2007. Is There an "Academic Vocabulary"? TESOL Quarterly, 41(2):235-253.

Håkan Jansson, Sofie Johansson Kokkinakis, Judy Ribeck and Emma Sköldberg. 2012. A Swedish Academic Word List: Methods and Data. In: Ruth Vatvedt Fjeld and Julie Matilde Torjusen (Eds.) Proceedings of the 15th EURALEX International Congress 7–11 August, 2012. Oslo University, Oslo.

Sofie Johansson Kokkinakis, Emma Sköldberg, Birgit Henriksen, Kari Kinn and Janne Bondi Johannessen. 2012. Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project. In: Ruth Vatvedt Fjeld and Julie Matilde Torjusen (Eds.) Proceedings of the 15th EURALEX International Congress 7–11 August, 2012. Oslo University, Oslo.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. In: Geoffery Williams and Sandra Vessier (Eds.) Proceedings of the Eleventh EURALEX International Congress, EURALEX2004: 123–131, Lorient, France, May 6-10, 2004. Université de Bretagne Sud, Lorient. [online: http://www.euralex.org/elx_proceedings/Euralex2004/011_2004_V1_Adam%20KILGARRIFF,%20Pavel%20RYCHLY,%20Pavel%20SMRZ,%20David%20TUGWELL_The%20%20Sketch%20Engine.pdf].

Lexins svenska lexikon. 2011. 4:th ed. [www] http://lexin.nada.kth.se/lexin/.

Inger Lindberg and Sofie Johansson Kokkinakis. 2007. OrdiL. En korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år. (ROSA, Rapporter om svenska som andraspråk 8.) Institutet för svenska som andraspråk, Göteborgs universitet, Göteborg. [on-line: http://gupea.ub.gu.se/dspace/handle/2077/20503]

Inger Lindberg and Sofie Johansson Kokkinakis. 2011. Identification of lexical cohesive ties in secondary school text books. AILA 2011. The 16th World Congress of Applied Linguistics, Beijing, China.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. LexicoNordica, 17:119–140.

Nina Martola. 2011. FinnWordNet och kulturbundna ord. LexicoNordica, 18:111–133.

Hilary Nesi. 2002. An English Spoken Academic Word List. In: Anna Braasch and Claus Povlsen (Eds.) Proceedings of the Tenth Euralex International Congress 2002 (vol. 1): 351–357. Center for Sprogteknologi, Copenhagen.. [on-line: http://www.euralex.org/elx_proceedings/Euralex2002/036_2002_V1_Hilary%20Nesi_An%20English%20Spoken%20Academic%20Wordlist.pdf].

OECD. Organisation for Economic Co-operation and Development. 2007. Working Party of National Experts on Science and Technology Indicators: Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. [on-line: http://www.oecd.org/dataoecd/36/44/38235147.pdf].

Linus Salö. 2010. Engelska eller svenska? En kartläggning av språksituationen inom högre utbildning och forskning. (Rapporter från Språkrådet 1). Språkrådet, Stockholm.

Petr Savický and Jaroslava Hlaváčová. 2002. Measure of word commonness. Journal of Quantitative Linguistics 9,:215−231.

Mike Scott. 1997. PC analysis of key words – and key key words. System 25/2, p. 233−245.

Rita Simpson-Vlach and Nick C. Ellis (2010). An Academic Formulas List: New Methods in Phraseology Research. Applied Linguistics 31(4):487-512.

Emma Sköldberg and Sofie Johansson Kokkinakis. 2012. A och O om akademiska ord. Om framtagning av en svensk akademisk ordlista. In: Birgit Eaker, Lennart Larsson and Anki Mattisson (Eds.). Nordiska studier i lexikografi 11. Rapport från Konferensen om lexikografi i Norden: 575–585. Lund 24−27 maj 2011.

Michael West. 1953. A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longman, London.

Guoyi Xue and Paul Nation. 1984: A university word list. Language Learning and Communication 3(2):215–229.

SweVoc - A Swedish vocabulary resource for CALL

Katarina Heimann Mühlenbock and **Sofie Johansson Kokkinakis**

Dept of Swedish, University of Gothenburg, Gothenburg, Sweden
katarina.heimann.muhlenbock@gu.se, sofie@svenska.gu.se

Abstract

The core in language teaching and learning is vocabulary, and access to a delimited set of words for basic communication is central for most CALL applications. Vocabulary characteristics also play a fundamental role for matching texts to specific readers. For English, the task of grading texts into different levels of difficulty has long been facilitated by the existence of word lists serving as guides for vocabulary selection. For Swedish, the situation is with a few exceptions less fortunate, in that no base vocabulary organized according to aspects of usage has existed. The Swedish base vocabulary – SweVoc – is an attempt to remediate this. It is a comprehensive resource, aimed at differentiating vocabulary items into categories of usage and frequency. As we are of the opinion that no corpus of written text can do fully justice of general language use, we have utilized materials from a second language as reference for delimiting the category of core words. Another belief is that the task of defining a base vocabulary can not be fully automatic, and that a considerable amount of manual, traditional lexicographic work has to be invested. Hence, the present approach is not an innovative, but a methodological approach to word list generation for a specific purpose, much like LSP. We anticipate SweVoc to be integrated in CALL applications for vocabulary assessment, language teaching and students' practice.

1 Background

Vocabulary knowledge plays a central role in a person's ability to communicate, as well as reading and understanding written text. It is therefore a central issue in many readability assessment approaches. Prominent researchers within readability and language assessment, such as (Thorndike, 1921; Vogel and Washburne, 1928; Patty and Painter, 1931; Thorndike and Lorge, 1944; Dale and Chall, 1948; Spache, 1953), and more recently (Nation, 1990; Nation, 2001), all included specific word lists as a criterion to measure text difficulty for English. In quantitative associative studies of readability, some scheme for measuring the vocabulary difficulty is set up, compared to a predefined criterion, and expressed by a coefficient of correlation. In this way, the word lists may be constructed in order to mirror vocabulary difficulty corresponding to school grade levels. Thorndike's (1921) word list of 10,000 words, later on revised into a list of 30,000 words (Thorndike and Lorge, 1944) and Spache's revised word list (Spache, 1974) of 1,040 entries, were mainly constructed by judgment and common sense. West published in 1953 the General Service List – a list of 2,000 words selected to represent the most frequent words in an English corpus.

Vocabulary is also an important issue when producing language-supportive aids for persons with deficient communication capability. Insufficient vocabulary knowledge implies a decrease in expressive power of an utterance or written text, and the receptive language skills are also heavily dependent upon the individual vocabulary range. In order to obtain

Katarina Heimann Mühlenbock and Sofie Johansson Kokkinakis 2012. SweVoc – A Swedish vocabulary resource for CALL. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 29–35.

maximum benefit from language supportive tools, the resources provided as word lists ought to be chosen with care in order to conform to individual and situational needs. Also in generating LSP (language for specific purposes) and particular domain vocabulary lists, a list of general base vocabulary is needed in order to exclude the most common and general words.

In the following we are making a distinction between *base vocabulary* and *core vocabulary*. A language teaching situation might involve a more extensive base vocabulary, while assistive technology applications such as symbol boards for communication would benefit from a restricted core vocabulary, expandible with complementary vocabulary items from different domains. The present approach is an attempt to combine both models, i.e. it is a Swedish core vocabulary word list, supplied with words belonging to a broader base vocabulary.

Defining a core vocabulary is a task associated with several methodological challenges. Lee (2001) has enumerated some of them. First of all, the concept of *core vocabulary* has to be settled. Several working definitions exist, out of which the most contested point seems to be whether the list is based on, and intended for, applications within written or spoken language, or both. If one decides to adopt the view that a core vocabulary is by definition that which is central to the language as a whole, it rules out for instance approaches based on frequency countings of words in written language. Furthermore, it should be untarnished from any stains of genre, style, register or lect association.

In addition to the theoretically founded issues, also problems of more practical nature arise. Although a major part of verbal communication is said to take place with the use of 1,500 - 2,000 words (West, 1953), this figure must be considered in the light of language-specific properties, of the type of communication, and above all, as a function of the *word* concept. Counting lexemes, lemmas, baseform orthographic words or multiwords render different figures. For English, the notion of *word family* plays a central role when defining word list for educational purposes. Lee (2001), citing Schmitt (2000) maintained that

> people in the field seem to agree that the "word family" is the most meaningful unit to work with and pedagogically most useful.

The word family concept was put forth by Bauer and Nation (1993), from a reader's perspective defined to comprise

> a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately.

If all the lemmas belonging to a specific word family are considered as one member of the word list, Hirsh and Nation (1992) found that a vocabulary size of at least 5,000 entries were needed in order to read unsimplified fiction texts. The same study also showed that graded readers beginning at a level of 2,600 word families would be of great benefit in language teaching.

An attempt to construct a levelled base vocabulary for another language than English was made by De Mauro (1980) when he published a list of 7,400 Italian words, categorized into three different groups according to use. The only attempt in this direction for Swedish was made by Forsbom (2006), who derived a base vocabulary pool from a corpus of 1 million words – the Stockholm-Umeå Corpus (SUC) (Källgren, 1992). This was achieved by ranking base word forms according to adjusted frequency over the entire corpus, and then adopting a subsequent filtering technique that sorted out entries which did not occur in more than three out of nine genres in the corpus. The result was a Swedish base vocabulary pool (henceforward referred to as SBVP), with a total amount of $\approx$ 8,200 word base forms, mirroring the use of written Swedish in the early nineties.

SBVP alone neither be considered to reflect modern language use, nor to be enough informative to independently serve as a source of words pertaining to a restricted core vocabulary, since it is based solely on written language. As already mentioned, the base word forms in SBVP are ranked according to adjusted frequency (AF: see equation 1), i.e. relative frequency weighted with dispersion over the 9 categories (genres) in SUC. It implies that the vocabulary are those words that are not genre dependent, given the subdivisions of a small-size text corpus. Furthermore, it lacks information at the lexeme

level, which reduces its feasability for purposes demanding a semantic disambiguation between words. A base form word like the Swedish noun *gång* has for instance four lexeme representations, belonging to different base vocabulary categories. The first refers to 'time' and is considered to be a core vocabulary item, while the sense 'path' is not. The second Issues regarding a distinction between lemma and lexeme concepts are discussed in Gardner (2007). Another flaw in SBVP is the absence of internal levelling, which would be required in order to serve as a list of core vocabulary words. In the present approach, it was hence enriched with labels indicating levels of general use from three additional sources; (1) a translated base vocabulary, (2) a list of words from modern vocabulary, and (3) a dictionary of words denoting domestic life activities and participation in community activities. The final product is SweVoc, a base vocabulary word list, consisting of ≈ 8,500 words, mainly lemma forms, divided into five different categories.

$$AF = (\sum_{i=1}^{n} \sqrt{d_i x_i})^2$$

where
AF = adjusted frequency
d_i = relative size of category i
x_i = frequency in category i
n = number of categories

(1)

2 Material

SweVoc is a comprehensive resource, based on lists of lexical items and texts from four different sources:

1. The backbone was the monolingual Swedish base vocabulary pool (SBVP) (Forsbom, 2006), derived from the SUC corpus (Källgren, 1992), containing 8,213 base form entries. Personal nouns, numbers and punctuation marks were omitted, which reduced the number of entries to ≈ 7,400.

2. The second major resource is a translation of the earlier mentioned work by (De Mauro, 1980), *Guida all'uso delle parole* henceforward referred to as *GUP*. It consists of a vocabulary of 7,400 words, mainly lemma forms, divided into three categories:

   - 2,100 basic words, regarded as fundamental for communication, representing a core vocabulary (C)
   - 2,400 words used in every-day communication (D)
   - 2,900 words highly frequent in written text (H)

3. The *Kelly* modern vocabulary list (Johansson Kokkinakis and Volodina, 2011) was used in order to ensure that frequent words used in modern settings were included. The Swedish version of Kelly is derived from a large modern corpus of web texts, and a subset of ≈ 500 words translated between Swedish and Italian was employed.

4. The *ICF* (Socialstyrelsen, 2003) is a classification of health and health-related domains, ranging from body structure to individual and societal issues. It was used as a reference word list in order to ensure coverage of words related to every-day matters.

3 Preprocessing

The Italian list of basic words was translated into Swedish by a second-language-speaker of Italian. Localisms and archaisms in the source language were ignored. The reason for using a foreign resource was two-fold; First, a base vocabulary should be selected in order to cover both universal concepts and essential phenomena and situations in the main local environment. Secondly, the manual translation task revealed ambiguities due to different usage of words and word senses among two syntactically and lexically distant languages, which contributed to a more fine-grained levelling of words into different subcategories.

The Kelly modern word list was a result of the EC-financed project Kelly <http://kellyproject.eu>. The aim of the project was the generation of monolingual word lists of nine languages, Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish. The lists were generated from

many sources including web corpora in order to reflect a modern vocabulary. The lists were then all translated into the eight other languages, generating 72 language pairs. The Italian-Swedish is one of them. The lists were then finally merged to 36 lists. These lists are used in the Keewords language learning tool <http://Keewords.com>.

Several structural differences between the two main sources – SBVP and GUP – caused problems already at the preprocessing stage of SweVoc. As is shown in table 1, the tag set used in SBVP is in PAROLE-format with morphosyntactic information, while GUP was based simply on part-of-speech. In addition to automatic conversion into SUC-format part-of-speech labelling, a considerable amount of manual work was required to make the lists comparable. However, as mentioned already in the introduction, we are of the firm view that no wordlist aimed at specifying a base vocabulary can be produced without a considerable degree of human intervention. We hence regard the present approach to be a pragmatic and feasible way to perform a restricted task.

4 Word list compilation

Entries in SBVP were checked against GUP in order to find candidates for inclusion into SweVoc. As already mentioned, the lists were comparable in size (≈ 7,400 words), but differed largely as regards to compilation methods and contents. As was expected, many words in the each of the two lists corresponded to multiple entries in the other. Multiword expressions and structural differences between the languages also required particular consideration.

One such example is the Swedish verb *be* 'ask, pray', present among the 1,000 words with highest adjusted frequency in SBVP. GUP provides three different lexemes for this verb, either *chiedere*, 'ask', *pregare* 'pray' and *supplicare* 'beseech'. All the words fall into category (C) in Italian, which would not necessarily be true for Swedish. In the opposite direction, the Italian polysemous noun *rapporto*, also among the words in category (C), is covered by three different entries in SBVP, either *förhållande* or *relation* 'relationship', both among the top 1,000 entries, but also 'rapport' 'report', with a lower adjusted frequency.

The degree of coverage of GUP lemmas in SBVP was also measured. It turned out that on overall 37.5% of the translated lemmas were present also in SBVP, but that the (C) group had a significantly higher coverage. Of the total 2,143 candidate lemmas considered as fundamental for communication, 81.4% were also present in the SBVP. Entries in the daily vocabulary (D) group were covered to 20.6%, while 28.6% of the high-frequency lemmas (H) in GUP were present also in SBVP. Of the 483 entries in the Kelly word list which did not occur in GUP, 288 were present in SBVP, i.e. 59.6%.

4.1 The final SweVoc

The GUP and Kelly word list entries that were present in SBVP were used to populate the first four categories in SweVoc, i.e. the core vocabulary items (C), words belonging to every-day language (D), high frequency words (H), and words from modern vocabulary (K). Additionally, items lacking in SBVP but present in both ICF and GUP, denoting daily activities or phenomena, were included. An example of such a word is *andning* 'breathing'. Words present only in ICF, denoting every-day situations and objects, were also added. The Swedish verb *möblera* 'furnish', exemplifies such a word. Finally, a supplementary group of words present only in SBVP were preserved, denoted by the category label (S). The word *samband* 'connection' serves as example from this category. An entry in SweVoc consists of information regarding rank in SBVP, the lemma form, the part-of-speech, and one or more category belongings. The entry *form* is given as example, illustrated below. It is a polysemous noun, found among the 223 most frequent base forms in SBVP, and different senses of the lemma belong to different SweVoc categories.

Rank	**Lemma**	**POS**	**Categories**
223	form	NCU	C, D

In conclusion: the present version of SweVoc contains 7,572 lemmas pertaining to one or more of five different categories. A lemma that is present in more than one category has discriminatory lexical senses, which implies that the number of lexemes amounts to 8,468, see table 2. Category (C) is dominated

Rank	Lemma	Adj.Freq.	Contr.	(WF.PoS.Freq)
5	en.DI	25958.046833	9	ett.DI@NS@S.7952 en.DI@US@S.18050
140	en.MC	726.135618	9	en-.MC0000C.2 ett.MCNSNIS.276 en.MCUSNIS.463
167	en.PI	606.653923	9	ett.PI@NS0@S.147 en.PI@US0@S.467 enom.PIUS0S.1
5708	en.RG	5.661842	4	en.RG0S.9

Table 1: Four different entries of the word *en* in SBVP

by nouns (38%), verbs (23%) and adjectives (13%). The category of words related to every-day matters (D), is mainly composed of nouns (66% of the total amount of lexemes), while verbs and adjectives only occur in 18 and 12% of the totality. In the group of high-frequency words (H), nouns were found to cover 55%, verbs 21% and adjectives 18% of the lexemes. From the perspective of core vocabulary alone, category (C) include 21% of the total nouns in SweVoc, 31% of all verbs, and 23% of all the adjectives. All pronouns and determiners were included in (C). All prepositions except one were found in category (C), except the word *tills* 'until', which was referred to the (D) category. One instance of all conjunctions (*visserligen* 'certainly') was found in the (H) category, while 58% appeared in category (C) and the remaining 40% in category (S). Figures regarding ratios of participles and adverbs are generally somewhat unreliable since different principles were used for corpus part-of-speech tagging in SUC and word list creation of GUP. Specifics regarding the part-of-speech distributions in each category are given in table 3.

Label	Category	Ex	Lexemes
C	Core vocabulary	säga	2,201
D	Words for every-day communication	soffa	1,019
H	High frequency words	sorg	1,518
K	Words in Kelly modern vocabulary	debatt	288
S	Supplementary words from SBVP	ting	3,442
Total			8,468

Table 2: SweVoc entries per category

POS	C	D	H	K	S
Nouns	844	670	831	139	1,436
Verbs	502	181	323	24	575
Adj	295	123	277	52	510
Adv	176	12	8	66	427
Part	168	29	58	7	194
Prep	42	1	0	0	0
Conj	29	0	1	0	20
Pron	65	0	0	0	0
Det	16	0	0	0	0
Other	64	3	20	0	280
Total	2,201	1,019	1,518	288	3,442

Table 3: Part-of-speech distributions in each SweVoc category

5 Evaluation

In order to validate the reliability of the SweVoc, evaluation was performed by coverage tests. It was assumed that the coverage of SweVoc would vary between texts of different types and from various genres. If the core vocabulary items were correctly chosen, the degree of words from this category would correspond to textual complexity, i.e. easier texts would contain more words from category (C). Another assumption was that the ratio of words from category (D) would vary depending on genre, that it would be much smaller, and that the words from the Kelly list (K) would appear more frequently in recent texts. In order to test these hypotheses, evaluation was performed on texts from three different sources:

1. The corpus LäsBarT (LB), which is a corpus of 1.4 million words, containing children's fiction for ages 6-12, and four easy-to-read text varieties:

- Easy-to-read news texts
- Easy-to-read community information texts
- Easy-to-read children's fiction
- Easy-to-read adults' fiction

2. The corpus SUC

3. News text from the daily newspaper Göteborgs-Posten (GP) published in 2007

It was found that, on overall, 91.4% of the tokens in LB, 82.7% of the tokens in SUC, and 83.0% of the tokens in GP were represented at the lemma level in SweVoc, while tokens belonging to the core vocabulary (C) amounted to 80.3% in LB, 68.4% in SUC, and 69.9% in GP texts. The ratios of words related to daily matters (D) were about the same in all texts ($\approx$ 1.3%), but the ratios of high-frequency words were significantly higher in SUC and GP than in LB ($p < 0.001$). Supplementary words (S) were found to be more frequent in SUC than in both GP and LB, which was expected since the original SBVP was retrieved from SUC. By studying the figures in table 4 we can see that the the degree of words in category (C) differ substantially between the ordinary and the easy texts, and also that the percentage of core vocabulary items is higher in fiction than in news and informative texts.

Type/ genre	SweVoc	C	D	H	K	S
ECF	92.5	82.4	0.8	2.1	0.7	6.5
OCF	90.6	80.4	1.0	1.9	0.7	6.6
EAF	93.4	83.1	0.9	2.2	0.6	6.6
OAF	86.3	75.8	1.0	2.4	0.8	6.3
EN	91.5	78.8	1.8	3.9	0.6	6.5
ON	82.2	67.6	1.7	3.8	1.1	8.0
EI	90.6	79.2	1.3	3.3	0.5	6.4

Table 4: SweVoc lemmas, percentage of tokens in different subcorpora
ECF = Children's easy-to-read fiction
OCF = Children's ordinary fiction
EAF = Adults' easy-to-read fiction
OAF = Adults' ordinary fiction (SUC K)
EN = Easy-to-read news
ON = Ordinary news (SUC A and GP)
EI = Easy-to-read community information

6 Foreseen improvements

Entries in the present version of SweVoc preserve information "inherited" from the translated word list GUP, in that a lemma might be categorized with several labels depending on which lexeme it refers to. The Swedish polysemous word *panna* ('front', 'pan', 'oven') is for instance labelled both as a core word (C) and as a word referring to every-day issues (D). One valuable resource for disambiguation is the Swedish word association lexicon Saldo (Borin and Forsberg, 2009), which is a modern Swedish semantic and morphological lexical resource. It is superficially similar to Princeton WordNet (Fellbaum, 1998), but different in the principles by which it is structured. The organizational principles of Saldo consist of two primitive semantic relations, or descriptors, one of which is obligatory and the other optional. When looking up *panna* in Saldo, we find three competing lexemes:

- *panna..1* ansikte..1 ('face') PRIM..1
- *panna..2* laga..2 ('cook') PRIM..2
- *panna..3* elda..1 ('make fire') PRIM..1

The semantic paths in Saldo for each of the three senses are illustrated below, each of the length of 6.

panna..1→ansikte→huvud→kropp→varelse→vem
('face'→'head'→'body'→'being'→'who')
panna..2→laga→mat→äta→leva→vara
('cook'→'food'→'eat'→'live'→'be')
panna..3→elda→eld→brinna→het→varm
('make fire'→'fire'→'burn'→'hot'→'warm')

Frequency counts in SUC reveal that 77% of the instances referred to *panna..1*, 15% to *panna..3*, and 8% to *panna..2*. From these figures, it seems plausible that *panna..1* would be referred to category (C), and either *panna..2* or *panna..3* or possibly both referred to category (D).

Regarding the CALL perspective of this lexical resource, we foresee it as an asset for vocabulary instruction and also as a resource in various CALL-oriented learning platforms and applications, as

for instance the Lextutor, <http://www.lextutor.ca/>. It is also relevant for integration into a Swedish CALL platform under development, cf. Lärka <http://spraakbanken.gu.se/larka/>.

7 Results and conclusion

We found that 81% of the GUP lemmas translated and selected as candidates for inclusion into category (C) were actually to be regarded as pertaining to a core vocabulary for Swedish. Additionally, 21% of the lemmas in category (D) and 29% in category (H) were appropriate for inclusion as complementary vocabulary words.

The resulting word list – SweVoc – of ≈ 7,600 Swedish lemmas is expected to be an asset in language learning and teaching and in readability checkers. The performance of other NLP applications, such as classification tools and morphological analyzers, would also improve with the access of a restricted set of base vocabulary words.

References

Laurie Bauer and Paul Nation. 1993. Word families. *International Journal of Lexicography*, 6(4):253–279.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27:37–54.

Tullio De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Eva Forsbom. 2006. A Swedish Base Vocabulary Pool. In *Swedish Language Technology conference*, Gothenburg.

Dee Gardner. 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, 28(2):241–265.

David Hirsh and Paul Nation. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2):689–696.

Sofie Johansson Kokkinakis and Elena Volodina. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project KELLY issues on reliability, validity and coverage. In *eLex Conference*, Slovenia.

Gunnel Källgren. 1992. SUC - the Stockholm - Umeå Corpus Project: Corpus-based research on models for processing unrestricted swedish text. Technical report, Stockholm.

D. Y. W. Lee. 2001. Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics*, 29:250–278.

Paul Nation. 1990. *Teaching and learning vocabulary*. Heinle & Heinle, New York.

Paul Nation. 2001. *Learning vocabulary in another language*. Cambridge University Press, Cambridge.

W.W. Patty and W.I. Painter. 1931. Improving our method of selection high-school textbooks. *Journal of Educational Research*, XXIV:23–32, June.

Norbert Schmitt. 2001. *Vocabulary in language teaching*. Cambridge University Press, Cambridge, UK.

Socialstyrelsen. 2003. Klassifikation av funktionstillstånd, funktionshinder och hälsa.

George D. Spache. 1953. A new readability formula for primary-grade reading materials. *Elementary School Journal*, LIII:410–413.

George D. Spache. 1974. *Good reading for poor readers*. Garrard Publishing, Champaign, IL.

Edward L. Thorndike and I. Lorge. 1944. *The teacher's word book of 30,000 words*. Columbia University Press, New York.

Edward L. Thorndike. 1921. *The teacher's word book*. Teacher's College, Columbia University, New York.

M. Vogel and C. Washburne. 1928. An objective method of determining grade placement of children's reading material. *Elementary School Journal*, 28:373–381.

Michael West. 1953. *A General Service List of English Words*. Longman, London.

A web-deployed Swedish spoken CALL system based on a large shared English/Swedish feature grammar

Manny Rayner, Johanna Gerlach, Marianne Starlander, Nikos Tsourakis
University of Geneva, FTI/TIM, Switzerland
{Emmanuel.Rayner,Johanna.Gerlach,
Marianne.Starlander,Nikolaos.Tsourakis}@unige.ch

Anita Kruckenberg
Royal College of Technology, Stockholm, Sweden
anita.kruckenberg@comhem.se

Robert Eklund, Arne Jönsson, Anita McAllister
Linköping University, Sweden
{robert.eklund,arne.jonsson,anita.mcallister}@liu.se

Cathy Chua
Swinburne University of Technology, Melbourne, Australia
cathychua@swin.edu.au

Abstract

We describe a Swedish version of CALL-SLT, a web-deployed CALL system that allows beginner/intermediate students to practise generative spoken language skills. Speech recognition is grammar-based, with language models derived, using the Regulus platform, from substantial domain-independent feature grammars. The paper focusses on the Swedish grammar resources, which were developed by generalising the existing English feature grammar into a shared grammar for English and Swedish. It turns out that this can be done very economically: all but a handful of rules and features are shared, and English grammar essentially ends up being treated as a reduced form of Swedish. We conclude by presenting a simple evaluation which compares the Swedish and French versions of CALL-SLT.

1 Introduction and background

People studying a foreign language need to practise four main skills: reading, writing, listening and speaking. All of these, especially the fourth, are challenging to do well. The increased emphasis on spoken language in education means that the issues involved have been brought more sharply into focus. In Europe, for example, the influential "Common European Framework of Reference for Language" (CEFR; http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf) has led to substantial changes in language teaching methods. Human teachers cannot easily cope with the increased demand for time spent helping students develop productive speaking skills, and the case for developing mechanical aids has become correspondingly stronger. For these reasons, the CEFR document suggests that CALL technology and the Web should be harnessed to try and offload some of the teaching burden on to machines.

There are many applications designed to help improve pronunciation: an impressive and well-documented example is the EduSpeak® system (Franco et al., 2010), and some commercial offerings, like RosettaStone and TellMeMore, have become very popular. These systems, however, generally limit themselves to teaching the student how to imitate: the student listens to a recorded sound file, imitates it to the best of their ability, and is given informative feedback. This does indeed help with pronunciation, but it is less clear that it helps improve spontaneous speaking skills.

A more ambitious approach is to design an application where the student can respond flexibly to the system's prompts. The system we will describe in this paper, CALL-SLT (Rayner et al., 2010), is based on an idea originating with Wang and Seneff (2007); a related application due to Johnson and Va-

Manny Rayner, Johanna Gerlach, Marianne Starlander, Nikos Tsourakis, Anita Kruckenberg, Robert Eklund, Arne Jönsson, Anita McAllister and Cathy Chua 2012. A web-deployed Swedish spoken CALL system based on a large shared English/ Swedish feature grammar. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 37–46.

lente (2009) is TLTCS. The system prompts the user in some version of the L1, indicating in an abstract or indirect fashion what they are supposed to say; the student speaks in the L2, and the system provides a response based on speech recognition and language processing.

The system is accessed via a client running on a web browser; most processing, in particular speech recognition and linguistic analysis, is carried out on the server side, with speech recorded locally and passed to the server in file form. The current version, available at `http://callslt.org`, supports French, English, Japanese, German, Greek and Swedish as L2s and English, French, Japanese, German, Arabic and Chinese as L1s.

The system is based on two main components: a grammar-based speech recogniser and an interlingua-based machine translation (MT) system, both developed using the Regulus platform (Rayner et al., 2006). Each turn begins with the system giving the student a prompt, formulated in a telegraphic version of the L1, to which the student gives a spoken response; it is in general possible to respond to the prompt in more than one way. Thus, for example, in the version of the system used to teach English to French-speaking students, a simple prompt might be: `DEMANDER DE_MANIERE_POLIE BIÈRE` ("`ASK POLITELY BEER`"). The responses "I would like a beer", "could I have a beer", "please give me a beer", or "a beer please" would all be regarded as potentially valid.

The system decides whether to accept or reject the response by first performing speech recognition, then translating to language-neutral (interlingua) representation, and finally matching against the language-neutral representation of the prompt. A "help" button allows the student, at any time, to access a correct sentence in both written and spoken form. The text forms come from the initial corpus of sentences or can be created by the MT system to allow automatic generation of variant syntactic forms. The associated audio files are collected by logging examples where users registered as native speakers got correct matches while using the system. Prompts are grouped together in "lessons" unified by a defined syntactic or semantic theme. A response which is correct but which does not match the theme of the lesson produces a warning.

The student thus spends most of their time in a loop where they are given a prompt, optionally listen to a spoken help example, and attempt to respond to the prompt. If the system accepts, they move on to a new prompt; if it rejects, they will typically listen to the help example and repeat, trying to imitate it more exactly. If they are still unable to get an accept after several repetitions, they usually give up and move to the next example anyway. On reaching the end of the lesson, the student either exits or selects a new lesson from a menu.

The architecture presents several advantages in the context of the web-based CALL task. The system is not related to a particular language or domain, as in (Wang and Seneff, 2007). The Regulus platform offers many tools to support addition of new languages and new coverage (vocabulary, grammar) for existing languages: the recogniser's language model is extracted by specialisation from a general resource grammar in order to get an effective grammar for a specific domain, with the specialisation process driven by a small corpus of sentences. The general grammar can thus easily be extended or specialised for new exercises by changing the corpus, enabling rapid development of new content.

In this paper, we will describe a Swedish-language version of CALL-SLT. The main focus is the Swedish resource grammar, which we constructed by generalising the English grammar into a shared English/Swedish grammar. It turned out that this could be done very economically, creating a grammar in which English is essentially treated as a reduced form of Swedish. The rest of the paper is organised as follows. Sections 2 and 3 give a brief overview of multilingual grammars, Regulus and the original Regulus resource grammar for English. Section 4 describes how the English grammar was extended to cover Swedish as well. Sections 5 and 6 describe the Swedish version of the CALL-SLT system, and presents results from a simple evaluation. The last section concludes.

2 Shared grammars

Large computational grammars were unfashionable for a while, but are attracting more interest again. One high-profile example is PARC's XLE (Maxwell

Figure 1: The version of CALL-SLT (Swedish for English-speakers) used in the main study.

and Kaplan, 1993; Crouch et al., 2007), which has formed the basis of the PARGRAM parallel LFG grammar consortium (Butt et al., 2002); a second is the Open Source Grammar Matrix project (Bender et al., 2002). Other substantial grammar-based programs include Gothenburg University's GF (Ranta, 2004; Ranta, 2007) and the Open Source Regulus platform (Rayner et al., 2006).

Multilingual efforts like these highlight the fact that languages are related. When a grammar for a related language already exists, it is unusual to attempt to develop a new grammar from scratch. The typical strategy is, rather, copy-and-edit; the related grammar is adapted to the new language by making suitable changes. A less common idea is grammar-sharing: a single, parametrized grammar is written which covers two or more languages simultaneously. When languages are closely enough related, the advantages of this approach are obvious. The grammar-sharing strategy has, for example, been successfully applied within the PARGRAM/LFG framework for Japanese and Korean (Kim et al., 2003), within the Regulus framework for Romance languages (Bouillon et al., 2007), and within the GF framework for both Romance and Scandinavian languages (Ranta, 2009). It is possible to construct shared grammars for groups of languages that are less closely related. This is the basic idea of the Grammar Matrix project; another example is (Santaholma, 2007). Nonetheless, the grammars produced by these projects are small, and the general belief is that the shared grammar approach most obviously makes sense when languages have similar structures.

Here, we have developed a substantial shared grammar that covers the greater part of English and Swedish. Considered as Germanic languages, it is not generally acknowledged that English and Swedish are especially close. As already noted, the GF project makes extensive use of grammar-sharing, but does not merge English with its Scandinavian grammar; similarly, the Spoken Language Translator project (Rayner et al., 2000) based on the SRI Core Language Engine, had separate grammars for English and Swedish. In fact, the only previous example of a shared English/Swedish grammar known to us is BiTSE (Stymne and Ahrenberg, 2006), constructed inside the DELPH-IN framework (Bond et al., 2005). The BiTSE grammar, however, appears to be small in scale, only covering core constructions, and, as far as we are aware, has not been tested in any real applications; the description in the paper also suggests that only about two-thirds of the grammatical structure is shared between the

two languages. We were thus surprised to discover that an extremely efficient shared grammar could be constructed, in which English structure, to a good approximation, turns out to be included within Swedish structure.

3 Regulus and the Regulus grammar for English

Regulus is an Open Source platform for building grammar-based speech-enabled applications. A distinguishing feature is that all language processing is based on the use of large, domain-independent feature grammars. These are compiled into grammar-based language models in two main steps. The first uses a small domain corpus, typically of a few hundred examples, to extract a specialised version of the feature grammar. The second compilation step converts the specialised feature grammar into a CFG approximation, which is then compiled into a recognition package using a third-party recognition engine. The current version of Regulus employs the Nuance 8.5 and Nuance 9 engines for this purpose. It is also possible to compile grammars into generator form, for example for use in translation applications.

The Regulus grammar formalism permits definition of feature grammars with finite-valued features (this restriction is motivated by the requirement that the grammars should be capable of compilation to CFG form). The notation is Prolog-based, and is similar to that used in the earlier Core Language Engine and Gemini projects (Alshawi, 1992; Dowding et al., 1993). Grammar rules are associated with a compositional semantics defined in the Almost Flat Functional semantics framework (AFF; (Rayner et al., 2008)), an intelligent compromise between nested predicate/argument structures and flat lists of feature-value pairs. For example, "Does coffee give you headaches?" is represented in AFF as

```
[null=[utterance_type,ynq],
 null=[action,give],
 agent=[cause,coffee],
 indobj=[pronoun,you],
 obj=[symptom,headache],
 null=[tense,present],
 null=[voice,active]]
```

Structure-sharing in Regulus grammars is primarily implemented using macros, which perform a function similar to that of *templates* in the XLE.[1] Macros are, for example, typically used in the lexicon to define classes of words with similar syntactic properties, and in grammar rules to define groups of features shared between the mother of a rule and one of its daughters.

The Regulus English grammar, described in Chapter 9 of (Rayner et al., 2006), is largely modelled on the earlier Core Language Engine grammar (Pulman, 1992). It contains about 220 feature-grammar rules, and covers most of the core constructions of English, including declarative clauses, YN- and WH-questions, most common types of verbs, nominal and verbal PPs, adverbs, negation, prenominal and predicative adjectives, compound nominals, partitives, pronouns (including expletive pronouns), relative clauses, embedded questions and verbs taking embedded question complements, subordinating conjunctions, constituent conjunction of NPs, PPs, ADJPs and clauses, dates and times. There is also a function-word lexicon containing about 450 words, and a set of macros for defining regular content-words (nouns, various types of verbs, adjectives, etc).

The English grammar has been used to construct over a dozen different speech-enabled applications, some very substantial. We have already mentioned the CALL-SLT system. Other prominent examples are NASA's Clarissa procedure navigator (Rayner et al., 2005), the Ford Research/UCSC SDS in-car information system and Geneva University's MedSLT (Bouillon et al., 2008), a multilingual interlingua-based medical speech translator.[2]

As described by Bouillon et al. (2007), shared grammars in Regulus can readily be constructed using the macro mechanism. The language-dependent portion of a lexicon-entry or rule is encoded using a suitable macro; this macro's expansion is then defined in two or more ways, one for each of the languages involved. Each language is associated with a different file of language-dependent macro definitions.

[1] http://www2.parc.com/isl/groups/nltt/xle/doc/walkthrough.html#W.templates

[2] The MedSLT application has also been ported to Swedish, using the grammar described here. This work will be described elsewhere. Some examples below refer to the MedSLT domain.

4 A shared English/Swedish grammar

We started with the English Regulus grammar described in Chapter 9 of Rayner et al. (2006) and broadened it to cover both English and Swedish, using a macro-based parameterization scheme. In this section, we present a complete list of the changes made, and the resulting differences between English and Swedish inherent in the shared grammar. We organise the material under the following headings: question-formation, verb-second word-order and periphrastic "do"; gender, definiteness and agreement; verb inflections; inherent reflexives and lexical passives; adverbs; the lexicon; and other issues.

The fact that the grammar is intended for speech applications allows us to simplify it in several places, and ignore issues which are primarily orthographic in nature. For example, English writes the possessive as the suffix "'s", while Swedish uses a plain "s". As far as speech recognition is concerned, both alternatives can equally well be considered as a separate word, "s". Speech recognisers also have no ability to recognise orthographical conventions such as punctuation or capitalization. Thus the grammar represents both English "Anna's" and Swedish *Annas* (possessive form of "Anna") as the same string

```
anna s
```

In a similar way, we can finesse the fact that Swedish compound nominals are conventionally written as single words (*busshållplats*, *morgonkaffe*), while English orthography adds intervening spaces ("bus stop", "morning coffee").

4.1 Question-formation and related issues

As explained in Chapter 9 of Rayner et al. (2006), the rules in the English grammar relevant to inverted (V2) word-order are implemented in a slightly unusual way, primarily motivated by the requirement of efficient compilation to CFG form for purposes of generating language models for speech recognition. Following the earlier Core Language Engine grammar, the binary feature `inv` is set on V constituents, and percolated up to their projections; it encodes whether the V is the main verb in a clause with uninverted (`inv=n`) or inverted (`inv=y`) word-order. Non-main verbs are always `inv=n`. In clauses with inverted word-order, the main V is combined with the inverted subject to form a constitutent called, for want of a better term, a VBAR. Figure 2 shows a minimal Swedish example illustrating use of the VBAR constituent.

```
.MAIN
 /  utterance_intro null
 |  utterance
 |     s
 |        s
 |           vp
 |           /  vp
 |           |  /  vbar
 |           |  |  /  v lex(har)
 |           |  |  |  np
 |           |  |  \     pron lex(du)
 |           |  |  np
 |           |  |  /  np
 |           |  |  |     nbar
 |           |  |  |        n lex(bröd)
 |           |  \  \  post_mods null
 |           \  post_mods null
 \  utterance_coda null
```

Figure 2: Analysis tree (slightly simplified) for the Swedish sentence *Har du bröd* ("have you bread" = "do you have bread")

The most important differences in word-order between English and Swedish derive from the fact that only periphrastic "do", auxiliaries, "have" and "be" can invert in English, while all verbs can invert in Swedish. This is captured in different values for the `inv` feature defined in the lexicon.

In Swedish, `inv` is always unset in the lexicon, since it can take either value. In English, the default value for `inv` is `n` (most verbs cannot invert). Periphrastic "do" has `inv=y` (it *must* be used inverted), while auxiliaries, "have" and "be" have `inv` unset (they can be used both inverted and uninverted). The semantics for periphrastic "do" are similar to those for other auxiliaries, with the verb contributing only tense information.

The only divergences in grammar rules related to these issues are in the rules for fronting of `wh-` constituents, where a language-specific macro specifies that English requires the uninverted word-order ("him she likes") while Swedish requires the inverted one (*honom gillar hon*).

4.2 Gender, definiteness and agreement

The `agr` feature mediates agreement, and is one of the two features whose spaces of possible values are language-dependent. In English, `agr` has six possible values, constituting the cross-product of `[1, 2, 3]` with `[sing, plur]`. In Swedish, it takes 12 possible values, since it is also necessary to include the component `[common, neuter]` to encode gender (Swedish has two grammatical genders). The marking for person is almost not required, since Swedish verbs do not inflect by person; all forms in the present and imperfect tenses are the same. It is, however, needed in order to enforce agreement between subjects and reflexive pronouns (cf. §4.4).

The `agr` feature was added to the grammar in many places, to enforce agreements which do not exist in English. In particular, possessive pronouns and ADJ projections carry the `agr` feature, so that these constituents agree with the nouns they modify, and `agr` is passed down though VPs, so that past participles agree with subjects. Thus for example *är din huvudvärk associerad med stress* ("is your headache associated with stress") but *är din***a** *huvudvärkar associerad***e** *med stress* ("are your-PLUR headaches associated-PLUR with stress").

D, N and ADJ projections carry the extra `def` feature, which marks for definiteness. In Swedish, these constituents agree in definiteness, e.g. *en stor kopp* ("a large cup") but *den stor***a** *kopp***en** ("the large-DEF cup-DEF").

The feature `def` exists in the English grammar, but is always unset.

4.3 Verb inflections

Swedish verbs have more inflectional forms than their English counterparts. We have already mentioned the fact that past participles are marked for gender and number; these forms are also distinct from the supine, which is used to form the perfect tense. For example, "I have written" is *jag har skriv***it** but "The book was written" is *boken blev skriv***en**. In addition, the imperative, considered as the base form, is in general distinct from the infinitive; to continue the example, *skriv* is the imperative, but *skriv***a** is the infinitive.

This motivates the other instance in the grammar of a feature where the range of possible values is different in the two languages. The feature in question is `vform`, like `inv` set on the V and percolated up to its projections. In English `vform` takes the range of values:

```
[base, finite,
en, en_passive,
ing, to, null]
```

(this is again closely based on the English Core Language Engine grammar). `ing` is for the present particle, `en` for the past participle, `en_passive` for past participle used as a passive, and `to` for VPs preceded by a 'to' complementizer. The Swedish `vform` feature's range is slightly different:

```
[imperative, infinitive, finite,
supine, en, en_passive,
ing, to, null]
```

The fact that Swedish makes strictly more fine-grained distinctions than English renders it straightforward to parameterize the grammar cleanly. Rules are written in such a way that they refer to notional infinitive and imperative forms, using macros to specify the concrete values of `vform` that correspond to these notional forms. Thus, in Swedish, the macros `notional_infinitive` and `notional_imperative` respectively expand to `infinitive` and `imperative`. In English, both expand to `base`.

4.4 Inherent reflexives and lexical passives

Like most modern European languages, but unlike English, Swedish has inherently reflexive verbs; thus, for example, "move" is *röra sig* (literally "move oneself") and "decide" is *bestämma sig* (literally "decide oneself"). The reflexive pronoun agrees with the subject, thus *jag rör mig* but **jag rör sig*.

To accommodate inherent reflexives (what Stymne and Ahrenberg (2006) call "fake reflexives"), we added the extra feature `takes_refl` to V and VBAR, marking Swedish verbs that require a reflexive pronoun, together with a rule of the schematic form

```
vbar:[takes_refl=n, agr=Agr] -->
   vbar:[takes_refl=y, agr=Agr],
   refl:[agr=Agr].
```

Swedish and the other Scandinavian languages also have lexically passive inflections of verbs; these

are finite, passive forms, which consist of an active form followed by a terminal 's'. The passive present is formed from the imperative, and the passive supine, imperfect, and infinitive from the corresponding active forms. Thus for example *skrivs* ("write-INF-PASSIVE") means "is-written", *har skrivits* ("has write-SUPINE-PASSIVE") means "has been-written", and so on. (There are subtle semantic differences between the lexical passive and the passive formed using the auxiliary, which we will not discuss here for lack of space).

To cover lexical passives, we added the extra feature `lex_passivisable` to V, marking verbs that may be combined with the passivising affix 's', together with a rule-schema which expands out into four rules for each subcategorisation class of verb which can be passivised. Somewhat to our surprise, no other changes were required in the grammar; a VP whose main verb is lexically passivised behaves exactly like one whose main verb is a form of the passive auxiliary *bli*. The features `takes_refl` and `lex_passivisable` exist in the English grammar, but are always unset.

4.5 Negation and adverbs

The Swedish negation particle *inte* is syntactically an adverb, which appears after the main verb in a main clause and before it in a subordinate clause. Thus *jag skriver inte*, "I write not" but *därför att jag inte skriver*, "because I not write". Several other common adverbs — so-called "mobile adverbs" — have the same distribution.

In order to capture this alternation, S carries the extra binary feature `main_clause`. This distinguishes main from subordinate clauses, and is passed to adverbial modifiers. Again, the feature exists in the English grammar, but has no function there.

4.6 The lexicon

Although it is possible to suggest correspondences between English and Swedish words (especially function-words), it seemed dangerous to us to use this strategy systematically. For example, although it is certainly the case that a connection exists between the Swedish modal verbs *ska* and *vill* and their English counterparts "shall" and "will", the meanings of these words in modern English and Swedish are substantially different.

With regard to parametrization of the lexicon, we have consequently adopted a more conservative approach; we write macros that define classes of lexical items with the same syntactic properties, and as far as possible share these macros between the two languages. In this way, we can talk about *syntactic classes* of words which can be identified between English and Swedish, and do not attempt to address the question of whether individual words can be put in correspondence. Lexical macros are defined hierarchically (this is the way the Regulus framework encodes inheritance in the lexicon); we will thus often identify a class of English words with a corresponding class of Swedish ones, leaving the proviso that a macro lower down in the hierarchy is language-dependent. To take a simple example, the macro defining an intransitive verb entry is common to the two languages, but depends on the language-dependent macro which expands out the different inflected forms of the verb from its base entry. As previously mentioned (§4.3), Swedish verbs have different inflectional forms from English ones, and there is a language-dependent macro, `verb`, which encodes this fact. All of the macros for specific syntactic classes of verb invoke `verb` in some way.

Divergences between the English and Swedish lexica are thus best studied at the level of lexical macros: the question is which macros, and thus which pieces of lexical structure, turn out to be language-specific. It turns out that only a few language-specific macros are required. We have just mentioned `verb`. Similar macros deal with the divergent inflectional morphology of nouns and adjectives. English requires an extra macro, `be_verb`, to cover the special case of "be", which has multiple suppletive forms ("am", "are", etc).

Higher up in the hierarchy, there are language-specific macros for syntactic types of verb. English has macros for verbs which subcategorise for verbs in the "-ing" form ("start running"), and Swedish for verbs which subcategorise for inherent reflexives (§4.4) and plain infinitives (*jag tänker gå* = "I intend go"). The macro for particle verbs is language-dependent, encoding the fact that Swedish particle verbs are separable: for example, the past participle of *ta bort* ("remove") is *borttagen*.

Unsurprisingly, the largest differences in the func-

tion word lexica arise from the fact that Swedish marks for number, gender and definiteness. The Swedish lexicon macros for determiners and possessives adds some of this structure to the corresponding English ones; for example, English "my" is unmarked, while Swedish has the three forms *min* (common, singular), *mitt* (neuter, singular) and *mina* (plural). Similarly, English "the" is unmarked, while the Swedish forms are both marked for gender and number, and are also `def=y`, agreeing in definiteness with nouns and adjectives.

The other differences in function-word macros are surprisingly few in number. Swedish, as already noted several times, has inherent reflexive pronouns, and it also has infinitive modal verbs (*jag skulle* **kunna** *komma* = "I would **can** come"). English has periphrastic "do"; reduced negated modals (Swedish lacks words like "won't" or "can't"), auxiliary "be" taking "-ing"; frequency adverbials like "once" and "twice"; distinguished subject and non-subject versions of the `wh+` personal pronoun (Swedish does not distinguish "who" from "whom"); and "please".

4.7 Other issues

Finally, we list a few other divergences which do not fit into any particular category. The object following the particle in a particle verb needs to be `pron` in English ("*I picked up it") but not in Swedish (*jag tog emot det*); the possessive marker attaches to the head noun in Swedish, but to the NP in English; the partitive marker is "of" in English, and null in Swedish; and the syntax of date and time expressions is slightly different in the two languages.

5 The Swedish CALL-SLT system

The initial Swedish version of the CALL-SLT system contains seven lessons, divided into two separate domains; content was largely derived from corresponding material in the existing English and French versions of the system. The first two lessons are for basic introductory Swedish. One covers greeting and politeness expressions, and the other simple questions and answers for talking about oneself: where do you come from, what language do you speak, what are you studying, and so on.

Lessons 3 to 7 are in a tourist restaurant domain, and repectively cover asking for something; asking for something using a question; numbers; payment expressions; and time expressions. The grammatical topics covered include simple noun phrases, declarative sentences in the present tense, some modal verbs, basic Y-N and WH-questions, measure phrases and numbers.

The total vocabulary included consists of 500 surface forms. The development effort, excluding work on the shared grammar described earlier, was about two to three person-weeks. The system is freely available at `http://callslt.org`.

Subject	Level	WER	SER
CC	Beginner	38.5	55.2
MR	Interm.	7.0	20.0
SG	Fluent	6.6	23.1
SR	Fluent	6.6	26.0
SC	Fluent	4.5	15.1
JG	Native	2.2	7.3
VB	Native	0.7	2.8
PB	Native	0.2	1.3

Table 1: Gross speech recognition measures for French.

Subject	Level	WER	SER
CC	Beginner	44.3	55.6
NT	Beginner	31.8	42.6
JG	Beginner	20.5	27.0
SS	Native	14.6	23.1
HH	Fluent	14.4	27.7
RS	Fluent	14.3	24.6
AX	Native	12.1	19.6
RE	Native	12.1	18.5
MS	Native	11.5	17.2
AB	Fluent	11.2	20.0
JM	Fluent	6.6	10.8
LS	Beginner	3.3	6.2
MR	Fluent	3.3	6.2
CS	Native	0.5	1.5

Table 2: Gross speech recognition measures for Swedish.

6 A simple evaluation

In previously reported work, we have carried out various kinds of evaluation of different versions of

the CALL-SLT system. In (Bouillon et al., 2011) and (Rayner et al., 2011), we presented evidence suggesting the students could improve their linguistic competence by interacting with the system; in (Rayner et al., 2012), we showed that judges, presented with randomly ordered pairs of responses made by the same subject to the same prompt, strongly preferred ones that had been accepted by the recogniser. In the present paper, we use a very simple strategy that we had not previously tried. We asked 25 subjects, with different levels of ability in Swedish and French, to log into the two versions of the system for about half an hour to an hour and practise the content of a few of the easier lessons; since the French lessons contained fewer examples than the Swedish ones, we used five lessons for French (73 examples) and only two for Swedish (65 examples).

Subjects were asked to begin by familiarising themselves with the system until they were comfortable with headset placement, use of the interface, appropriate speaking rate, and so on. They were then asked to attempt the contents of the selected lessons, using the help examples and trying each example once, and achieve as good a score as possible. The results were recorded and transcribed to enable calculation of Word Error Rate and Sentence Error Rate. Since many subjects did not follow the instructions carefully and attempted examples multiple times even after the "familiarisation" part of the session, results were normalised by including only the first response to each prompt. We also removed the data from four subjects (three Swedish and one French) who were having clear problems with the audio connection, resulting in very low recording volume. Tables 1 and 2 present the figures.

The French version of CALL-SLT is a mature system, which represents perhaps six to twelve person-months of effort and has gone through multiple design iterations; as already noted, the Swedish version is very new. Unsurprisingly, the French version performs rather better. The higher error rates in Swedish, compared to French, can reasonably be ascribed to two main causes. First, the current Swedish system has just one language model for all the lessons. The French one, in contrast, is set up so that there are multiple language models, with a specialised model for each group of lessons, giving lower perplexity and correspondingly lower error rates. It is easy to add similar declarations to the Swedish system and support multiple language models there too. A second issue is missing vocabulary. Looking at the results of the Swedish tests, it is clear that some important items should be added; for example, subjects often try to use *jobba* as a synonym for *arbeta*, *hur har du det* as a synonym for *hur mår du*, *läsa till att bli* as a synonym for *läsa till*, and so on. Two or three iterations of tuning would plug the important holes, after which our guess, based on previous experience, is that performance of the two versions would be fairly similar.

We had expected to find a correlation between system recognition accuracy and speaking ability. For the French system, the results are roughly as we thought they would be. The native speakers get low WER scores averaging under 2%; the intermediate/fluent speakers averaged around 6%; and the beginner was much higher. The pattern in Swedish, however, was not as clear. Native and fluent non-native speakers did about equally well, and we were startled to find that subject LS, who had no previous experience in Swedish, had made the third best score. Although this at first seemed so anomalous that we assumed it had to represent some kind of bug, human examination of the recordings suggested, to our surprise, that the machine had made a reasonable evaluation. LS, a Dutch native speaker, is a gifted linguist, speaking several languages to near native-speaker level, and had picked up a credible Swedish accent with astonishing rapidity.

We find these preliminary results interesting, but are not yet sure how to interpret them. More data is clearly needed; we hope to perform another data collection when the next version of the system is ready, hopefully before the end of 2012.

7 Summary and conclusions

We have described a preliminary Swedish version of the Web-enabled CALL-SLT spoken CALL system. Although very new, it already performs quite well, with at least some native and fluent speakers getting near-perfect recognition scores. Some simple tuning, along the lines of that performed on the French version, would probably improve performance considerably.

The limited-domain Swedish speech understanding technology used is generic, and has already been used to port another non-trivial application, the MedSLT medical speech translator, to Swedish.

References

Alshawi, H., editor. 1992. *The Core Language Engine*. MIT Press, Cambridge, Massachusetts.

Bender, E.M., D. Flickinger, and S. Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of COLING 2002 workshop on Grammar Engineering and Evaluation*.

Bond, F., S. Oepen, M. Siegel, A. Copestake, and D. Flickinger. 2005. Open source machine translation with DELPH-IN. In *Open-Source Machine Translation Workshop at MT Summit X*.

Bouillon, P., M. Rayner, B. Novellas, M. Starlander, M. Santaholma, Y. Nakao, and N. Chatzichrisafis. 2007. Une grammaire partagée multi-tâche pour le traitement de la parole: application aux langues romanes. *TAL*.

Bouillon, P., G. Flores, M. Georgescul, S. Halimi, B.A. Hockey, H. Isahara, K. Kanzaki, Y. Nakao, M. Rayner, M. Santaholma, M. Starlander, and N. Tsourakis. 2008. Many-to-many multilingual medical speech translation on a PDA. In *Proc. AMTA*, Waikiki, Hawaii.

Bouillon, P., M. Rayner, N. Tsourakis, and Q. Zhang. 2011. A student-centered evaluation of a web-based spoken translation game. In *Proceedings of the SLaTE Workshop*, Venice, Italy.

Butt, M., H. Dyvik, T.H. King, H. Masuichi, and C. Rohrer. 2002. The parallel grammar project. In *Proceedings of COLING 2002 workshop on Grammar Engineering and Evaluation*.

Crouch, D., M. Dalrymple, R. Kaplan, T. King, J. Maxwell, and P. Newman. 2007. XLE documentation. `http://www2.parc.com/isl/groups/nltt/xle/doc`.

Dowding, J., M. Gawron, D. Appelt, L. Cherny, R. Moore, and D. Moran. 1993. Gemini: A natural language system for spoken language understanding. In *Proc ACL*.

Franco, H., H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda. 2010. Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401.

Johnson, W.L. and A. Valente. 2009. Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures. *AI Magazine*, 30(2):72.

Kim, R., M. Dalrymple, R.M. Kaplan, T.H. King, H. Masuichi, and T. Ohkuma. 2003. Multilingual grammar development via grammar porting.

Maxwell, J.T. and R.M. Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–590.

Pulman, S.G.˙ 1992. Syntactic and semantic processing. In Alshawi (Alshawi, 1992), pages 129–148.

Ranta, A. 2004. Grammatical framework. *Journal of Functional Programming*, 14(02):145–189.

Ranta, A. 2007. Modular grammar engineering in GF. *Research on Language & Computation*, 5(2):133–158.

Ranta, A. 2009. GF: A Multilingual Grammar Formalism. *Language and Linguistics Compass*, 3(5):1242–1265.

Rayner, M., D. Carter, P. Bouillon, V. Digalakis, and M. Wirén, editors. 2000. *The Spoken Language Translator*. Cambridge University Press.

Rayner, M., B.A. Hockey, J.M. Renders, N. Chatzichrisafis, and K. Farrell. 2005. A voice enabled procedure browser for the International Space Station. In *Proc. ACL*, Ann Arbor, MI.

Rayner, M., B.A. Hockey, and P. Bouillon. 2006. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. CSLI Press, Chicago.

Rayner, M., P. Bouillon, B.A. Hockey, and Y. Nakao. 2008. Almost flat functional semantics for speech translation. In *Proceedings of COLING-2008*, Manchester, England.

Rayner, M., P. Bouillon, N. Tsourakis, J. Gerlach, M. Georgescul, Y. Nakao, and C. Baur. 2010. A multilingual CALL game based on speech translation. In *Proceedings of LREC 2010*, Valetta, Malta.

Rayner, M., I. Frank, C. Chua, N. Tsourakis, and P. Bouillon. 2011. For a fistful of dollars: Using crowdsourcing to evaluate a spoken language CALL application. In *Proceedings of the SLaTE Workshop*, Venice, Italy.

Rayner, M., P. Bouillon, and J. Gerlach. 2012. Evaluating appropriateness of system responses in a spoken call game. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Santaholma, M. 2007. Grammar sharing techniques for rule-based multilingual NLP systems. In *Proceedings of NODALIDA*, pages 253–260.

Stymne, S. and L. Ahrenberg. 2006. A bilingual grammar for translation of English-Swedish verb frame divergences. In *Proc. EAMT*, pages 9–18.

Wang, C. and S. Seneff. 2007. Automatic assessment of student translations for foreign language tutoring. In *Proceedings of NAACL/HLT 2007*, Rochester, NY.

Waste not, want not: Towards a system architecture for ICALL based on NLP component re-use

Elena Volodina, Lars Borin
Språkbanken (Swedish Language Bank)
University of Gothenburg, Sweden
first.last@svenska.gu.se

Hrafn Loftsson
School of Computer Science
Reykjavík University, Iceland
hrafn@ru.is

Birna Arnbjörnsdóttir
School of Humanities
University of Iceland, Iceland
birnaarn@hi.is

Guðmundur Örn Leifsson
School of Engineering and Natural Sciences
University of Iceland, Iceland
gol1@hi.is

Abstract

It is a surprising fact that, despite the existence of various mature Natural Language Processing (NLP) tools and resources that can potentially benefit language learning, very few projects are devoted to development of Intelligent Computer-Assisted Language Learning (ICALL) applications. This paper presents an on-going collaborative project whose overall aim is to develop an open-source system architecture for supporting ICALL systems that will facilitate re-use of existing NLP tools and resources on a plug-and-play basis. The two language teams – Icelandic and Swedish – have tested the architecture design by implementing two ICALL applications which convincingly show how principles defined by Service-Oriented Architecture (SOA), with web services as implementation technology, can benefit re-use of existing NLP components in ICALL applications. This paper introduces the project, provides the theoretical and practical background, describes the different paths adopted within the two language teams, and presents the first results.

1 Introduction

The project described in this paper was prompted by the surprising fact that existing NLP tools and resources do not tend to find their way into the language learning classroom, despite their obvious potential uses in language learning. The reasons may be twofold. On the one hand, there is a lack of interested sponsors. On the other hand, there is a general lack of interest in the NLP community in CALL applications. Borin (2002), for example observed that "[...] while certainly not part of the core of NLP, CALL seems not to have a place even in its periphery", and "[...] most NLP work on Nordic languages has nothing to do with CALL". While this might have changed for English, and a small number of other languages in the past ten years,[1] it still holds true for the Nordic languages.

We are aware of only three ICALL[2] systems that are an integral part of a real-life foreign language program in universities today: TAGARELA for Portuguese (Amaral and Meurers, 2011; Amaral et al., 2011), E-tutor for German (Heift, 2003), and Robo-Sensei for Japanese (Nagata, 2009). It seems that the few systems that have been developed are either copyrighted and restricted by high licensing fees – and hence too expensive for universities and schools – or fall short of the required quality in linguistic or pedagogical functionality.

This situation calls for a change. Since ICALL is a truly interdisciplinary field, it is important that researchers from several areas, like linguistics, pedagogy, NLP, and human-computer interaction (HCI) cooperate for the purpose of making ICALL projects successful. In view of that, we have joined

[1] Major NLP conferences tend to organize workshops on the use of NLP technologies in language learning, e.g. NAACL and COLING. The same holds true for the main conferences within computer-assisted language learning where AI and NLP approaches are studied within the area of pedagogy, e.g. CALICO and ICCE.

[2] Intelligence in CALL systems can be understood differently by different researchers. In this paper, we define ICALL as NLP-based CALL, i.e. intelligence in CALL is ensured through the use of NLP tools and resources like parsers, taggers, corpora, lexicons, etc.

Elena Volodina, Lars Borin, Hrafn Loftsson, Birna Arnbjörnsdóttir and Guðmundur Örn Leifsson 2012. Waste not, want not: Towards a system architecture for ICALL based on NLP component re-use. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 47–58.

forces in order to design and develop open-source system architecture for supporting ICALL systems. The architecture is open-source in order to encourage participation from other researchers and developers, and to facilitate re-usability of existing NLP tools and resources in the area of CALL. This is an ongoing collaboration, and some preliminary results and earlier versions of the implementations described below have been presented (in much less detail and without evaluation) in other contexts (Volodina and Borin, 2012; Volodina et al., 2012).

Our main argument is that the use of NLP tools and annotated resources can ensure linguistic analysis of input data, thus adding generative power. This is accomplished by applying the same analysis model to different (authentic) language samples, e.g. for generating exercises or detecting errors in learner text production. This, in our view, will not only relieve teachers of monotonous tasks that can be performed by computers, but can also support autonomous learning by students. And last, but not least, we hope it will increase the applications of NLP tools among CALL end-users.

For this purpose, we need access to existing NLP tools (e.g. sentence segmenters, tokenisers, part-of-speech [PoS] taggers, lemmatisers, syntactic parsers, error parsers, spell-checkers, etc.), as well as to existing (available and reliable) annotated resources (e.g. corpora, lexicons, learner-oriented word lists, etc.). We intend to re-use existing NLP tools and resources as much as possible (as opposed to developing new ones).

However, one problem is that most available resources and tools are difficult to deploy in CALL applications since (1) they are monolithic and inflexible and need to be individually adapted to each new application; (2) they are not readily available as the rights to their use are held by individuals or institutions all over the world and they are physically located in different places; and (3) they are not interoperable via standardised interfaces.

In order to achieve more flexibility, we need to cooperate with the owners of tools and resources. We need a standardisation effort within the ICALL community. One of the goals of this project is to design an architecture for deploying NLP tools and resources that will have well-defined principles and requirements, as well as provide easy-to-follow guidelines. We hope it will generate an interest in ICALL standardisation, and at best, if we are fortunate – encourage owners to provide a wrapper layer to their tools and resources making them re-usable in ICALL (and other) applications via web services. One overarching goal of our project is to test web services as a possible approach to making tools and resources available for re-use.

To avoid being too abstract, we are also implementing two end-user applications that will help us (1) test and refine the architecture; (2) produce guidelines for making a service wrapper layer to the tools and resources; (3) define relevant input/output formats and documentation standards; as well as (4) demonstrate the architecture design in practice for potential end-users and web service providers.

The rest of this paper is structured as follows. In section 2, we present the technical framework which we have adopted for the development of our architecture. Sections 3 and 4 are descriptions of the two examples where web services are used in development of ICALL applications, one for Icelandic (section 3) and one for Swedish (section 4). Section 5 concludes the paper with some general considerations about the effectiveness of the adopted approach and its future.

2 Technical framework

2.1 Background

The idea of re-usability as a paradigm for software development is not original. It is well-known that programmers often make chunks of their code available to each other in order to save time on implementation of something similar. With the appearance of the Free Software Foundation[3] in 1984, developers could have access to each other's code, copy it, modify and built upon it, which speeded up development times and reduced costs. Initiatives like that are very popular, but they have some limitations: first, the code comes in various different programming languages and it is not certain that it will be available in the language you need; second, they often lack documentation with explanation of their design or how the program works; and third, they are often centered around one problem specific for the current project, which is most probably not the one that is relevant for your needs (Wood, 2008).

Standardisation is a key notion in such initiatives. In addition to work carried out on standardisation of e-learning (IMS Global Learning Consortium,[4] ADL,[5] etc.) and of text

[3]http://www.fsf.org/
[4]http://www.imsglobal.org/
[5]http://www.adlnet.gov

corpus and lexicon resource formats (TEI,[6] EAGLES,[7] etc.), some successful standardisation efforts have been initiated for NLP components as well, e.g. GATE,[8] NLTK,[9] UIMA,[10] which are frameworks for integrating NLP tools and resources. However, the NLP components are still bound to particular programming languages: Java (GATE and UIMA) and Python (NLTK).

2.2 NLP component re-use through web services

The original initiative of re-using different existing programming functionalities in applications without re-writing the code is known as *Service-Oriented Architecture* (SOA).[11] SOA is an architectural style based on a set of global principles and requirements defined first by Erl (2005) and later by the SOA Manifesto Working Group.[12] SOA emphasises implementation of components as modular services that can be re-used by other clients. The main idea is that, despite different programming languages or platforms, the existing functions have a common communication layer consisting of a well-defined interface, where the user can formulate a request and get a response which can be re-used in other applications. The data is passed in standardised formats between the service and a client or between several services through coordinated calls. The key requirements are interoperability, re-use, standards-compliance, and well-documented metadata. Services can be made accessible to a closed group, e.g. within a company's intranet, or be open to anybody concerned via internet, for a fee or for free. Services are loosely coupled, and can be combined and re-combined for different purposes in production of other applications.

If SOA is an architectural style, then web services[13] are an implementation technology (one of many) for SOA. Web services make programs accessible through Internet protocols independent of platforms or programming languages. They can represent new applications or wrap around existing tools, becoming a port of access to them. Each service in the SOA architecture has, in turn, its own architecture. It includes all the resources used by a service, e.g. databases, software components, other services, and the physical design of their communication.

The basic principles and ideas behind SOA, particularly with web service technology as its implementation form, seem to be the answer to the question of accessibility of existing NLP tools and resources over the internet, and not only for ICALL applications. The software can still be residing on the original server and in the original programming language. It is the wrapper layer (web service) that makes it available to the users world-wide.

2.3 A platform for supporting ICALL

From an end-user perspective, Learning Platforms (LP), virtual learning environments (VLE) and learning/content management systems (LMS/CMS) serve different pedagogical purposes. They are different types of online services facilitating communication between teachers and students, e.g. for delivery of course-related information, resources and tools; as well as for synchronous (e.g. web chats, video conferences) and asynchronous (e.g. forums) meetings between students and teachers where course-related questions can be discussed. Such systems model a real-life communication between all involved parties, and may be used either in e-learning/ distance learning, i.e. without any class meetings, or as enhancement of face-to-face courses. Examples of such platforms are Moodle (Martín-Blas & Serrano-Fernández, 2009) and Fronter.[14]

Viewed from a developer's perspective, LPs can be compared to operating systems since they share some common characteristics, e.g. they are composed of a number of web-based applications that can be run within some environment.

ICALL is a specific area of learning, and thus a platform aimed at language learning requires a more specific design. Further, a platform offering intelligent analysis of language input needs to be designed for re-use of the components that can perform such analysis.

We therefore define an ICALL platform in technical terms as a structured backend, i.e. a "machinery" for deploying different NLP tools

[6] http://www.tei-c.org
[7] http://www.ilc.cnr.it/EAGLES96/home.html
[8] http://gate.ac.uk
[9] http://nltk.org/
[10] http://uima.apache.org/
[11] Architecture is a description of a system, defining its purpose, functions, externally visible properties and interfaces; including the description of its internal components and interoperability along with the principles governing its design, operation and evolution. It is thus a design of a system, not its implementation (Srinivasan and Treadwell, 2005).
[12] www.soa-manifesto.org, 2009
[13] A web service is an implemented software component that can be accessed via a network to provide functionality to a service requester/client (Srinivasan and Treadwell, 2005).
[14] http://com.fronter.info/product/

and lexical resources for supporting language learning activities, as well as specifically tailored algorithms for various language learning tasks (e.g. exercise generators). We neglect most of the administrative and content management functions that pedagogical platforms described above usually imply.

In particular, we build two ICALL platforms on SOA principles where the collection of web-services are the basis of the platforms.[15] The user interface,[16] on the other hand, is a top layer that is used for delivering the results of existing web services and should not necessarily be viewed as an integral part of the platform. It is rather an environment for presenting the output of web services and may be developed by different users according to their tastes and needs.

The advantage of separating ICALL modules into a frontend (user interface) and a backend (web services) parts is that the algorithms for required language learning task can be made language independent, i.e. they will rely only on the availability of corresponding NLP tools and lexical resources for other languages with the same type of annotation.

Another advantage is that in case we optimise or change the backend algorithm, the user interface remains unaffected; it is just a container for collecting user input and for showing the results of the web service.

One more advantage is that the web services are made re-usable for any other applications/user interfaces.

That is our starting point, and we are currently testing this approach by building two ICALL applications based on NLP components accessed through web services. The two ICALL applications are aimed at different language learning tasks: error analysis and feedback on L2[17] learner written input for the Icelandic partner; and corpus-based exercise generation for the Swedish partner, as described in the sections that follow.

3 ICALL through web services – an Icelandic example

3.1 NLP and ICALL for Icelandic

A decade ago, Icelandic could have been categorised as a less-resourced language, i.e. a language for which only a few, if any, NLP resources exist. Ten years later, the situation has changed dramatically (Rögnvaldsson, 2008). A number of BLARK[18] (Krauwer, 2003) components have now been developed, e.g. the open-source *IceNLP* toolkit,[19] a collection of tools for processing and analysing the Icelandic language (Loftsson and Rögnvaldsson, 2007b).

Among other tools, IceNLP contains a tokeniser, the PoS tagger IceTagger (Loftsson, 2008), and the shallow parser IceParser (Loftsson and Rögnvaldsson, 2007a). IceTagger, which performs morphosyntactic disambiguation, is the current state-of-the-art tagger for Icelandic (Loftsson et al., 2009). IceParser, which receives disambiguated input from a PoS tagger and whose task is to label constituents and syntactic functions, is the only publicly available parser for the language.

Two lexical resources are important parts of the Icelandic BLARK. First, the Icelandic Frequency Dictionary (Pind et al., 1991), a PoS-tagged corpus, and, secondly, the morphological database BÍN[20] (Bjarnadóttir, 2005). Both resources are available for research purposes, while the data of the latter can be used for developing language technology applications.

Currently, no ICALL application exists for the Icelandic language. On the other hand, the development of the web course (CALL application) *Icelandic Online* (IOL)[21] began in 2000. The sequential course is pedagogically driven in that instructional goals were served by the available pre-web 2.0 technology (the opposite was true for most CALL courses at the time). The technology used by IOL was only limited by the Digital Divide. This meant that, at the time, students in countries other than the most technologically advanced did not have the bandwidth to download websites heavily based on videos and interactive learning objects with many images.

IOL I and II were launched in 2004 and 2005. The goal of those courses is to introduce the structure and lexicon of Icelandic in a meaningful context using 40 pre-programmed learning objects, the contents of which can be altered and geared to the particular pedagogical goals of the lesson. The first courses were also heavily dependent on

[15]The terms platform and backend are used interchangeably in the text.
[16]The terms GUI, user interface, and frontend are used interchangeably in this text.
[17]L2 covers both foreign and second language learning.
[18]BLARK – Basic LAnguage Resource Kit, a joint initiative for European countries which has been extended to many other than European languages, see http://www.blark.org/.
[19] http://icenlp.sourceforge.net
[20] http://bin.arnastofnun.is
[21]http://icelandiconline.is

individually programmed interactive Flash lessons that introduced new vocabulary and grammar. The limitations of the courses were that they taught perceptive language with limited activities for students to practice productive skills other than form focused discrete vocabulary and grammar exercises (Arnbjörnsdóttir, 2004).

In 2010, Icelandic Online 3 and 4 and IOL for Immigrants were launched. These courses use the 40 learning objects but also introduce lesson content through authentic videos, texts and interactive websites, chosen and sequenced to advance the lesson goals. This was post web 2.0 which made available different social networks and functionalities that allow learners to interact with each other and practice their target language and negotiate meaning in social situations. This has been made full use of in Icelandic Online 3 and 4 (Arnbjörnsdóttir, 2008).

Currently, Icelandic Online has almost 90,000 registered users and has received universally positive feedback. IOL has revolutionised accessibility to Icelandic language and culture for teachers and students at the University of Iceland and worldwide. IOL is free and open to all.

To date, technology has not been able to provide CALL projects, like IOL, with meaningful intelligent feedback on second language writing. Despite the availability of spelling and grammar checkers in some languages, these tend to correct, rather than instruct, which is not always optimal for language learning.

3.2 The Icelandic platform

In the Icelandic part of the project, the platform connects various pre-existing NLP tools. Internally, the platform uses a particular XML format, the Text Corpus Format (TCF), proposed in the WebLicht SOA project (Hinrichs, 2010), for communication of information between the various components. Each annotation (e.g., at the level of tokens, PoS tags, or constituents) is stored in a separate layer, but all annotations for a particular text is stored in a single XML file. In addition to using the layers proposed in the WebLicht project, we have added our own layer for information about grammatical errors.

Using a web service, a user asks the platform to carry out a given task. Thus, the platform does not need to be set up on the user's machine. Moreover, the server running the web service and the platform do not have to be located on the same machine.

3.3 Writing support for second-language learners

In IOL, second-language learners of Icelandic can receive feedback from a teacher on short written texts. Currently, teachers use special codes for hand-marking specific types of errors, i.e. spelling errors, feature agreement errors, case errors in objects of verbs, etc.

In order to automate part of the hand-marking, and to test our platform, we are currently in the process of developing a web service which allows students of IOL to send texts to the service for the purpose of detecting particular types of grammatical errors. This will allow the students to correct potential errors and re-submit the texts for error detection again, and so forth, before finally submitting the text to the teacher. The web service merely identifies error candidates, but does not attempt to correct errors. At this stage, the goal is to help students correct second language grammar issues, and free instructors to focus on content.

The web service uses the platform, which, in turn, uses tools from the IceNLP toolkit, to detect the following types of grammatical errors, chosen for this first version: (1) feature agreement errors in noun phrases, i.e. errors in gender, number and case; (2) feature agreement errors between subjects and verb complements; (3) feature agreement errors between subject and verbs, i.e. errors in person and number; and (4) incorrect case selection of verb objects.

In using the feedback feature, the student inputs Icelandic text through a web application. The application submits the text to a web service, requesting it to analyse the text and carry out error detection. In turn, the platform calls components from IceNLP for carrying out the given tasks. IceNLP outputs XML in TCF, which the platform forwards to the web service, which in turn sends it back to the client application. The TCF contains all information from the analysis, i.e. information about the individual tokens, their PoS tags, individual constituents and error candidates. The client application converts the TCF to HTML and displays the resulting page to the student, where the original text submitted is shown with error candidates highlighted. In addition, by clicking on a word in a given sentence, the student can see morphological information for each word of the sentence.

Figure 1 shows the feedback given to a student for the sentence *Hann er góð kennari* 'He is (a) good teacher', in which the adjective *góð* 'good (feminine)' does not agree in gender with the

following noun *kennari* 'teacher (masculine)' in the noun phrase *góð kennari*. The phrase containing the disagreement is displayed, as well as morphological information for each word.

Figure 1. Feedback given to a student for a sentence containing a disagreement in a noun phrase.

Preliminary tests of the application have been carried out with two groups of students – a group of 11 advanced and 12 intermediate students in a summer course in Icelandic as a foreign language. The purpose of the test was twofold: First to elicit feedback from students about their experiences using the application and, second, to test the functionality of the application itself – the accuracy of the error detection.

In general students found the system helpful for error detection and that it aided them in their writing. Most found the directions for use clear. Two respondents wanted clearer suggestions for corrections or even declension tables to be attached to the system. The latter could be accomplished using the morphological database BÍN (see above).

The accuracy of the error detection was evaluated using the first texts submitted by the second group (12 intermediate students). The results are shown in table 1. In total, the system pointed to 25 grammatical errors, out of which 19 were true positives. This is equivalent to 76% accuracy, which is too low for practical use. Note, however, that the third error type, feature agreement errors between subject and verbs, is mainly to blame. Out of seven error candidates signalled by the system for this error type, only three were true positives. All the four false positives are due to the same error made by the error detector when analysing a sentence like: *Konan og drengurinn voru að þvo* ... 'The woman and the boy were washing … '. For this sentence, the error detector signals a disagreement in number between the singular noun phrases 'the boy' and the verb form 'were', not taking into account that the two singular noun phrases 'The woman' and 'the boy' indeed constitute a plural subject! When we account for this, both the precision and the recall will presumably increase.

Error type	Precision	Recall
agreement errors in noun phrases	80%	100%
agreement errors between subjects and verb complements	100%	87.5%
agreement errors between subjects and verbs	42.9%	42.9%
incorrect case selection of verb objects.	100%	50%
All error types	76%	76%

Table 1. Accuracy of the error detection.

Overall, we feel that the system has shown its value as a first step in the development of a semi-automatic writing feedback feature for Icelandic as a second language.

4 ICALL through web services – a Swedish example

4.1 NLP and ICALL for Swedish

Language technology research has a long history in Sweden, going back to the 1960s, and is conducted in a number of groups at the main Swedish universities and in some groups in industry. Consequently, most of the basic BLARK components exist for Swedish in quite stable and mature forms. For example, there are several PoS taggers and parsers, annotated reference corpora, and large lexical databases with morphological analysers available for Swedish, many (but not all) under open-source licenses.

Swedish ICALL has a shorter history. In recent years, there have been four main, partly overlapping, strands of research (ignoring speech-based ICALL, which is also being pursued at the Royal Institute of Technology in Stockholm, but which is out of scope for this paper) (see also Borin, 2006):

(1) Supporting reading of authentic texts by automatic selection of texts containing vocabulary and linguistic constructions at a suitable level for a particular language learner proficiency level (Nilsson and Borin, 2002).

(2) Automatic generation of focus-on-form exercises from annotated corpora, for PoS and syntactic functions such as subject and object (the ITG project;[22] Saxena and Borin, 2002; Borin and Saxena, 2004), and for vocabulary (Volodina, 2010).

(3) Writing support for second-language learners using online (bilingual and monolingual) lexicon access, and spelling and grammar checkers (the Grim project; Knutsson, 2005).

(4) Research on the characteristics of learner language and text complexity with an explicit aim of informing the research described under the previous three points (Magnusson and Johansson Kokkinakis, 2008; Johansson Kokkinakis, 2009).

Both the ITG project (2) and the Grim project (3) have resulted in concrete ICALL applications. The ITG application is open-source and is maintained by University of Gothenburg. It has been used extensively in university-level linguistics courses at the universities in Uppsala and Stockholm, and also in a high school in Uppsala. Its point of departure is what Second Language Acquisition (SLA) researchers have dubbed "focus-on-form" (FoF; contrasted to more traditional form-based drills, referred to as "focus-on-formS" in the SLA literature):

> Whereas learners are able to acquire linguistic forms without any instructional intervention, they typically do not achieve very high levels of linguistic competence from entirely meaning-centered instruction. For example, students in immersion programs in Canada fail to acquire such features as verb tense markings even after many years of study. This had led second language acquisition researchers [...] to propose that learners need to do more than simply engage in communicative language use; they also need to attend to form. (Ellis et al., 2002: 401)

In the ITG application, annotated Swedish text corpora are the basis for guided form exercises as well as curiosity-driven corpus exploration of particular linguistic features (the application includes a general corpus search interface), in both cases using authentic language material directly from the corpora, rather than made-up exercises and examples.

The Grim writing support application is not open-source (although the language tools used in it are) and it can be accessed only via a web page. Both ITG and Grim use a technology for the user interaction with the tool *Java Web Start* which was state of the art at the time, but which practical experience shows is not the optimal solution today, when web technology has developed to a point where pure web solutions will provide equivalent or better functionality in a much more transparent way to the user. Important for our purpose is that the language tools used in both these applications are to a large degree open-source and independent of the technology for realising the user interaction part.

4.2 Lärka and its architecture

In designing the new architecture for the Swedish application, we first ported the existing Swedish FoF exercises developed for the ITG application and started adding the Swedish vocabulary exercises developed by Volodina (2010). Having the existing ITG exercises allows us to quickly assess the viability of the architecture for this kind of application. Together with the new modules to be developed in this project, they make up a broad and varied spectrum of ICALL applications which will allow us to test the flexibility of the architecture. The ITG exercises use manually annotated corpora and although the text material is authentic, it is also now slightly dated and becoming more so all the time. One goal of this project is thus to adapt the language tools at our disposal with the aim of achieving the same kind of functionality using arbitrary text, e.g. from the internet. Another goal is to extend the range of FoF exercises offered and to explore how these exercises should be connected to other language learning activity types.

The application developed as a part of this project is web-based and has been given the name *Lärka*[23] *(*LÄR språket via KorpusAnalys 'learn language by corpus analysis'), with the English equivalent *Lark* (Language Acquisition Re-using Korp). The two main guiding principles for the implementation of Lärka have been modularity and re-use. The main components of Lärka are, as shown in figure 2):

- *frontend* – the graphical user interface that handles user interaction, sends requests to the backend, prettifies its output and assigns behaviour to the buttons and fields;
- *backend* – a number of web services for creating language training exercises, selecting

[22]ITG stands for *IT-based collaborative learning in grammar*; see http://spraakbanken.gu.se/swe/itg

[23]The Swedish word *lärka* means 'lark' (the bird), hence the logo; see http://spraakbanken.gu.se/larka/

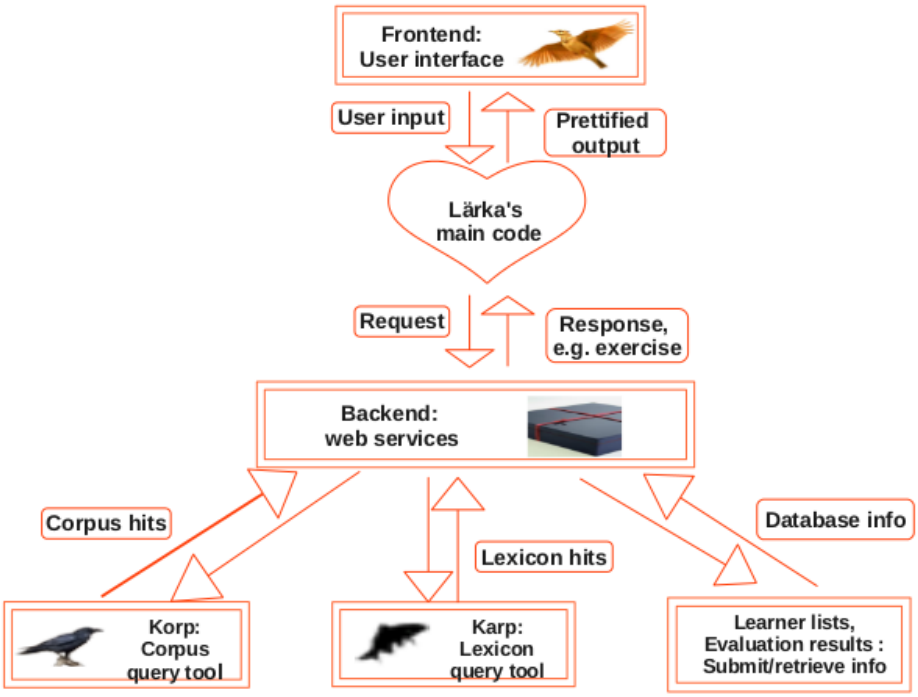

Figure 2. The architecture of Lärka

distractors, generating syntactic trees and rating corpus hits according to their appropriateness for particular exercise types;

- *Korp*[24] is Språkbanken's web-service based infrastructure for maintaining and searching a constantly growing corpus collection, at the moment amounting to over one billion words of Swedish text (Borin et al. 2012a). The corpora available through Korp contain multiple annotations, e.g. lemmatisation, compound analysis, PoS tagging, and syntactic dependency trees, which can form the basis for versatile exercises;

- *Karp*[25] is the corresponding web-service based infrastructure for maintaining and retrieving information from Språkbanken's collection of computational lexical resources (Borin et al. 2012b);

These four components together constitute Lärka's *architecture*. Below, we describe the backend and the frontend, discuss the functionality that Lärka can provide at the moment, and outline future work.

Lärka's frontend (figure 2, top) is the graphical user interface that collects user input and sends requests to the backend. The design has been inherited from the two other applications mentioned above – Korp and Karp. Similarly to these, Lärka will have the functionality to encode the exercise type in a URL (defining the exercise type, training mode, corpus, learner level, etc), so that exercise configurations can be referenced directly as URLs – i.e., bookmarked and passed around – saving users the extra effort of always going through the menus on the main webpage.

Each exercise (or any other future learner activity) is added as a separate module with minimal additions to the user interface code and as a web service. Exercises and other learning objects can thus be developed separately and get integrated with minimal efforts.

At the moment of writing, Lärka offers three exercise types: (1) training PoS; (2) training syntactic relations; and (3) multiple-choice vocabulary exercise items for language learners (re-implemented from Volodina 2010). The first two types are intended for linguistics students and ported from ITG. Each of the exercise types can be run in test mode or in self-study mode, see figure 3. As soon as one item is answered, the next one is generated. The result tracker shows the learner progress.

Lärka's backend is the heart of the architecture; see figure 2. Lärka depends heavily on the corpora and their annotation, and therefore uses Korp's web service for sentence selection. The rich

[24]http://spraakbanken.gu.se/korp/ (*korp* means 'raven').
[25]http://spraakbanken.gu.se/karp/ (*karp* means 'carp' [the fish]).

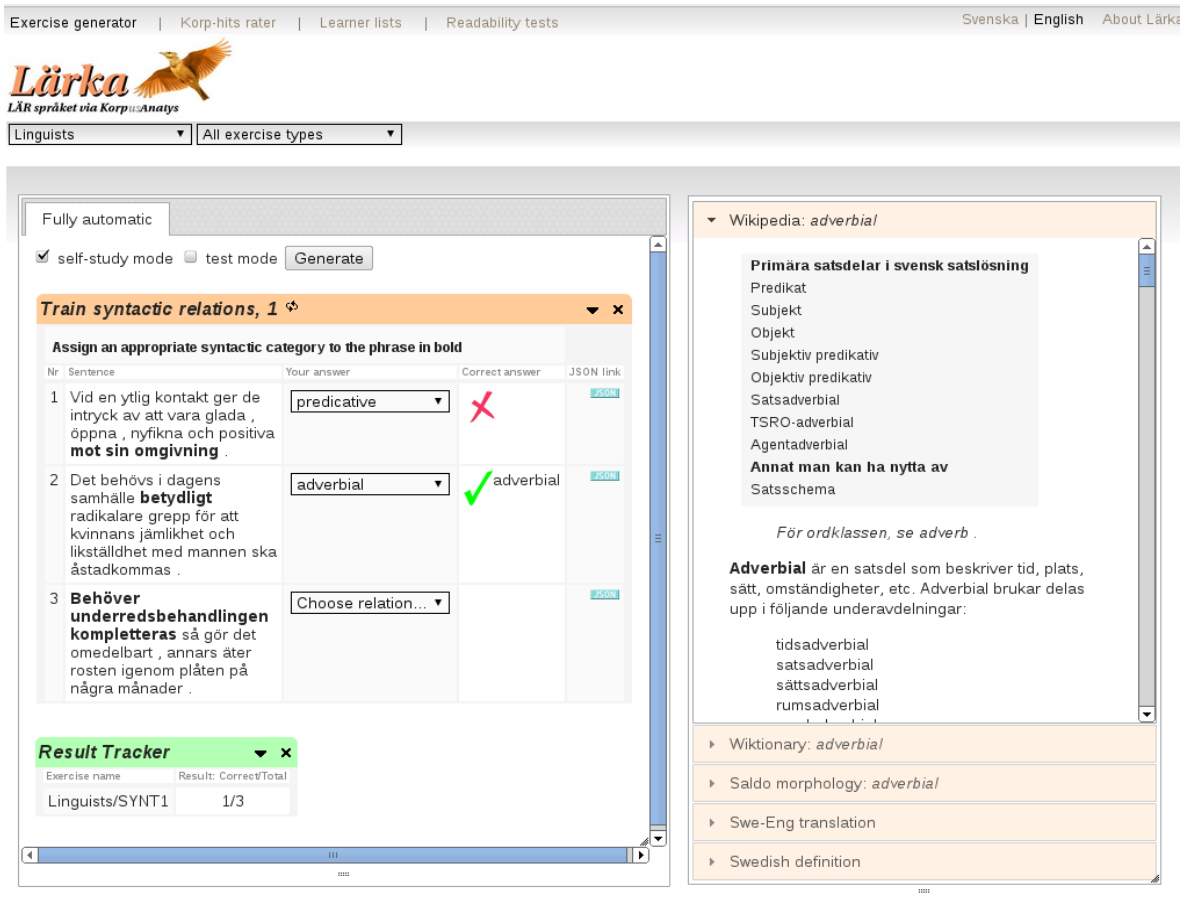

Figure 3. Lärka's frontend (GUI)

annotations available in Korp facilitate generation of exercise types other than the ones that have already been implemented; these are planned for future implementation.


```
{"corpus": "TALBANKEN",
 "distractors":["FV","IO","IV","OO","SP",
                "SS","VA"],
 "distractors_en_sv": {
  "FV":{"en":"finite verb","sv": "finit verb" },
  "IO": {"en":"indirect object","sv":
         "indirekt objekt"},
  "IV": {"en":"nonfinite verb","sv": "infinit
         verb"},
  "OO": {"en":"object","sv": "objekt"},
  "SP": {"en":"predicative","sv":
         "predikativ"},
  "SS": {"en":"subject","sv": "subjekt"},
  "VA": {"en":"adverbial","sv": "adverbial"} },
 "exetype": "synt1",
 "sent_index": 5648,
 "sentence_left": "De hanliga
         vinkarkrabborna bygger ",
 "sentence_right": "i stranden . ",
 "target": "hålor ",
 "target_deprel": "OO",
 "target_index": 4}
```


Figure 4. Example output (in JSON[26] format) from Lärka's backend.

The output from Lärka's backend (i.e. web services) can be used by any program, e.g. in mobile apps. An example of the output from the web service is shown in figure 4. Here you can see all the necessary information for the syntactic training exercise (in JSON, which currently is the common data communication format used by all Språkbanken's web services):

- sentence_left, target and sentence_right make up a complete sentence;
- the target is the part of the sentence that needs to be matched with a syntactic relation;
- the target's syntactic relation (correct answer) is provided as a tag in target_deprel;
- the list of distractors is provided together with the Swedish and English terms for each tag;
- the extra information, like corpus, sent_index (sentence index), target_index (position of the target item in the sentence), etc. are provided in case the user would want to replicate exactly the same item once again through a call to the backend.

In the user interface a JSON link is provided for every single exercise item for those who want to see the web service output.

The web service algorithms for exercise

[26]JSON is an acronym for JavaScript Object Notation.

generation are language independent since they rely on the annotation only. The exercise generation can therefore be made language independent provided there are resources (corpora and underlying word lists) for other languages using the same annotation.

At the moment, the web service output is provided in one format only – JSON. Eventually, other formats will be added, e.g. QTI (Question and Test Interoperability; IMS 2006) and TCF (Hinrichs 2010).

Next on our to-do list is to add syntactic tree visualisations, show relevant encyclopedia entries as an accompanying feature for exercises, design morphological and semantic exercise items based on Karp's web-services (Borin et al. 2012b), add gap cloze and wordbank items as well as diagnostic tests for vocabulary knowledge training. In the more distant future we are planning to:

- add an option of editing existing exercises by providing word lists, texts or selecting other distractors;
- extend the Lärka with Hit-ex – a web service and frontend for showing results from an algorithm for rating corpus searches according to different combinations of linguistic parameters. Tests with Hit-ex are ongoing;
- add the possibility to measure text readability using several readability indices;
- and of course add more exercise types, for grammar, word-building, etc.

5 Concluding remarks

The main idea of our project is to stimulate the re-use of existing accurate NLP tools and resources in language learning by designing and implementing a system architecture for ICALL, at the moment on a more abstract level – where our two subprojects share the general philosophy of making NLP components available via web services – and in the next phase of the project on the concrete level of having a common data exchange format (e.g. TCF). ICALL researchers and developers clearly stand to benefit from our project. In addition, language learners will also be affected because the system architecture and the two test applications will benefit language learners in the form of a more versatile and open-ended CALL experience, thanks to the NLP components.

Our experiences so far indicate that web services are a promising approach to re-use of existing NLP components: they are easy to develop and they preserve their independent stable form despite the changes introduced to the user interface. However, web service providers – including ourselves – should keep in mind, that (1) the services need to be stable and predictable over time, i.e. not undergo sudden changes in their output formats or any other unwelcome changes that can influence the performance of the application(s) based on them; (2) they should deliver as much information as possible to allow the end-user some variation in using their output, e.g. in the case of Lärka's syntactic exercises, the output from the web service could contain not only strings of left and right contexts, but also all associated annotation information for each token coming from the corpus web service.

Practical experience also shows that the web services as far as possible should be split into one separate component that reads information in the request and makes calls to separate request-specific components. In other words, the service-based architecture should be consistently applied all through the application. In the long run, this makes maintenance of the components easier.

It is at the moment undecided which formats we will adopt as standards in the final versions of our web-services. The two formats – TCF and JSON – adopted at the moment by the two language teams work well for us at this testing stage. We should, however, consider the end user interests; for example there is one format we know is used for exercises – QTI (see above) – that we consider important for inclusion as an output format for the exercise generator; there might be other relevant formats that need to be considered.

However, we believe that once our web-service based philosophy is adopted by other owners of NLP components, the two applications described in this paper may become a potential portal for delivering results gained by researchers in CALL, NLP and HCI to the general user and therefore fulfil a very important aim: to make NLP and ICALL research results available outside academia in the form of hands-on applications, thus making technology benefit language learning.

Acknowledgments

The work in this paper has been partially funded with support from NordPlus Sprog, grant LA-2011_1a-25339. The Swedish group has also been partially funded by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken.

References

Luiz A. Amaral and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL* 23(1): 4–24.

Luiz A. Amaral, Detmar Meurers, and Ramon Ziai. 2011. Analyzing learner language: towards a flexible natural language processing architecture for intelligent language tutors. *Computer Assisted Language Learning* 24(1): 1–16.

Birna Arnbjörnsdóttir. 2008. Kennsla tungumála á netinu: Hugmyndafræði og þróun Icelandic Online [The teaching of languages through the Net: The ideology and development of Icelandic Online]. *Hrafnaþing* 5: 7–31.

Birna Arnbjörnsdóttir. 2004. Teaching morphologically complex languages online: Theoretical questions and practical answers. *CALL for the Nordic languages*, ed. by Peter Juel Henrichsen. (Copenhagen Studies in Language 30.) Copenhagen: Samfundslitteratur.

Kristín Bjarnadóttir. 2005. Modern Icelandic inflections. In H. Holmboe, editor, *Nordisk Sprogteknologi* 2005, 49–50. Museum Tusculanums Forlag, Copenhagen.

Lars Borin. 2002. What have you done for me lately? The fickle alignment of NLP and CALL. In Proceedings of the EUROCALL 2002 pre-conference workshop "NLP in CALL". Jyväskylä, Finland.

Lars Borin. 2006. Sparv i tranedansen eller fisken i vattnet? *Språkteknologi och språklärande. Från vision till praktik: Språkutbildning och informationsteknik*, ed. by Patrik Svensson. Rapport 1:2006, Nätuniversitetet. 25-49.

Lars Borin and Anju Saxena, A. 2004. Grammar, incorporated. *CALL for the Nordic languages*, ed. by Peter Juel Henrichsen. (Copenhagen Studies in Language 30.) Copenhagen: Samfundslitteratur. 125–145.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012a. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA. 474–478.

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012b. The open lexical infrastructure of Språkbanken *Proceedings of LREC 2012*. Istanbul: ELRA. 3598–3602.

Rod Ellis, Helen Basturkmen, and Shawn Loewen. 2002. Doing focus-on-form. *System* 30: 419–432.

Thomas Erl. (2005) *Service-Oriented Architecture: Concepts, Technology, and Design*, Prentice-Hall, USA

Trude Heift. 2003. Multiple learner errors and meaningful feedback: A challenge for ICALL systems. *CALICO Journal*, 20(3), 533–548.

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. WebLicht: Web-based LRT services in a distributed eScience infrastructure. *Proceedings of LREC 2010*. Valletta, Malta: ELRA.

IMS (2006). IMS Question and Test Interoperability overview. Version 2.1 Public Draft (revision 2) Specification. IMS Global Learning Consortium. http://www.imsglobal.org/question/qtiv2p1pd2/imsqti_oviewv2p1pd2.html (Retrieved on 29th Juny, 2012).

Sofie Johansson Kokkinakis,. 2009, Readability and multilingualism. *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics,*. Studia Linguistica Upsaliensia 8 (Acta Universitatis Upsaliensis). 323–324

Ola Knutsson. 2005. Developing and evaluating language tools for writers and learners of Swedish. Doctoral thesis in human-computer interaction. KTH, Stockholm.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*. Moscow.

Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1), 47–72.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007a. IceParser: An incremental finite-state parser for Icelandic. In J. Nivre, H-J. Kaalep, K. Muischnek and M. Koit (eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*. Tartu, Estonia.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007b. IceNLP: A natural language processing toolkit for Icelandic. In *Proceedings of InterSpeech 2007, Special session:* "*Speech and language technology for less-resourced languages*". Antwerp, Belgium.

Ulrika Magnusson and Sofie Johansson Kokkinakis. 2008. Quantitative measures on student texts. *Papers from the ASLA Symposium in Stockholm, 7-8 November, 2008*, Association suédoise de linguistique appliquée (ASLA), Language and Learning (ASLA:s skriftserie nr 22). 43–56.

Teresa Martín-Blas and Ana Serrano-Fernández. 2009. The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education 52* (2009), p.35–44

Noriko Nagata. 2009. Robo-Sensei's NLP-based error detection and feed-back generation. *CALICO Journal*, 26(3), 562–579.

Kristina Nilsson and Lars Borin. 2002. Living off the land: The Web as a source of practice texts for *learners* of less prevalent languages. *Proceedings of LREC 2002*. Las Palmas: ELRA. 2002. 411–418.

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]. The Institute of Lexicography, University of Iceland, Reykjavik.

Eiríkur Rögnvaldsson. 2008. Icelandic language technology ten years later. In *Proceedings of "Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages", SALTMIL workshop, LREC 2008*. Marrakech: ELRA.

Anju Saxena and Lars Borin. 2002. Locating and reusing sundry NLP flotsam in an e-learning *application*. *LREC 2002. Workshop Proceedings. Customizing knowledge in NLP applications: strategies, issues, and evaluation.* Las Palmas: ELRA. 45-51.

Latha Srinivasan and Jem Treadwell. 2005. An overview of service-oriented architecture, web services and grid computing. Hewlett-Packard Development Company, V02, 11/2005.

Elena Volodina. 2010. *Corpora in Language Classroom: Reusing Stockholm Umeå Corpus in a vocabulary exercise generator.* Saarbrücken: Lambert Academic Publishing.

Elena Volodina and Lars Borin. 2012. Developing a freely available web-based exercise generator for Swedish. *EuroCALL 2012 Proceedings*, Gothenburg.

Elena Volodina, Hrafn Loftsson, Birna Arnbjörnsdóttir, Lars Borin, and Guðmundur Örn Leifsson. 2012. Towards a system architecture for ICALL. In G. Biswas et al. (eds), *Proceedings of the 20th International Conference on Computers in Education*. Singapore: Asia-Pacific Society for Computers in Education.

Peter Wood. 2008. Developing ICALL tools using GATE. *Computer-Assisted Language Learning*, 21:4, 383-392.

Semi-automatic selection of best corpus examples for Swedish: Initial algorithm evaluation

Elena Volodina, Richard Johansson, Sofie Johansson Kokkinakis
elena.volodina@svenska.gu.se
richard.johansson@svenska.gu.se
sofie@svenska.gu.se
Department of Swedish & Språkbanken, University of Gothenburg, Sweden

Abstract

The study presented here describes the results of the initial evaluation of two sorting approaches to automatic ranking of corpus examples for Swedish. Representatives from two potential target user groups have been asked to rate top three hits per approach for sixty search items from the point of view of the needs of their professional target groups, namely second/foreign language (L2) teachers and lexicographers. This evaluation has shown, on the one hand, which of the two approaches to example rating (called in the text below algorithms #1 and #2) performs better in terms of finding better examples for each target user group; and on the other hand, which features evaluators associate with good examples. It has also facilitated statistic analysis of the "good" versus "bad" examples with reference to the measurable features, such as sentence length, word length, lexical frequency profiles, PoS constitution, dependency structure, etc. with a potential to find out new reliable classifiers.

1 Introduction

This evaluation has been carried out as a part of a pre-study partly financed by the Centre for Language Technology (CLT) at the University of Gothenburg.

In this study we have evaluated two different approaches, namely algorithm #1 and #2, to the selection of examples. Both algorithms perform in such a way that, given a number of corpus hits for a search item, examples are sorted withdrawing or awarding points for presence or absence of formalized linguistic features, so called constraints. This brings to the top examples that correspond best to the constraints.

Using a specifically designed user interface and database, we performed the first evaluation. This step has provided us with a body of linguistic evidence for further refinement and tuning of the algorithm in *general* terms for Swedish.

Our hypothesis is, though, that users of different target groups would value presence (or absence) of different linguistic features; and that the same set of parameters cannot satisfy all potential target groups. Moreover, even within different target groups, the definition of a "good example" would change depending upon the practical aim at hand, e.g. examples for learners of different levels will need to take into account different language characteristics.

Thus, during the second iteration planned for near future our intention is to implement a user interface for working with different configurations of extended set of parameters according to the results of the first evaluation. We intend to evaluate parameter configurations again, this time concentrating on whether requirements set on examples differ between different target groups, and different tasks at hand. As a result we hope to suggest optimal parameter configurations for each individual target group, and eventually for different practical tasks at hand.

2 Background

Selection of authentic examples that can appropriately demonstrate vocabulary items of interest is a vital question for lexicographers and L2 teachers. At present it is often unknown for

Elena Volodina, Richard Johansson and Sofie Johansson Kokkinakis 2012. Semi-automatic selection of best corpus examples for Swedish: Initial algorithm evaluation. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 59–70.

instance, on what principles dictionary examples are selected or where examples for illustrating new vocabulary for L2 learners come from. One way of providing examples is to make them up – they are then as typical as the person that comes up with them thinks they should be, but they lack authenticity. Another way is to use some source of authentic texts, e.g. a linguistic corpus, and select examples using concordance software. The only constraint set on the corpus hits is then the occurrence of the target word in the text span (as opposed to sentence) which makes the number of hits often innumerable. In this case examples are authentic, but the selection process can be very tedious and the quality of "candidate" examples can be very different. One more option is to pre-select sentences automatically using a number of constraints downgrading inappropriate samples. The user is then offered top candidate samples he or she can choose from. The resulting list of ranked candidate sentences can be used for further manual or automatic selection (or editing) of top high-quality sentences, reducing the costs and time spent on manual pre-selection of those. The candidate examples can be used: for dictionary entries; to illustrate language features for students of linguistics; to exemplify vocabulary for language learners; to create test items for L2 learners; to accompany electronic texts (e.g. via clicking on the unknown word the user can see another example of the usage of this word), and eventually for a number of other tasks.

The ranking algorithm can eventually be used to test web texts for appropriateness for inclusion into a corpus. The target user groups are therefore lexicographers, L2 teachers, teachers of linguistics, test item creators, designers of electronic course materials and corpus linguists.

The question arising in this connection is whether we can comprehensively describe and model "good examples". This question has been addressed in different studies (Kilgariff et al. 2008, Husák 2008, Kosem et al. 2011, Segler 2007, etc.), though up to date never for Swedish as a target language. Our starting point is that parameters of good examples are language dependent and need to be tested for each language separately.

Algorithms for ranking corpus hits for Swedish have been designed with two practical applications in mind: *Swedish FrameNet* (SweFN, Friberg Heppin and Toporowska Gronostaj 2012) and *Lärka* (Volodina and Borin 2012).

SweFN is a lexical resource under development based on frame semantics, put forward by Charles J. Fillmore. The central idea is that word meanings are described in relation to semantic frames which are schematic representations of the conceptual structures of the language. Work on each frame consists in identifying relevant lexical items and providing authentic corpus (sentence-long) examples for each frame-related meaning. At the moment the work on finding examples involves a tedious look through several hundreds of examples in search of one that is good enough for the task. An algorithm that would be able to sort inappropriate examples away can considerably accelerate work on each frame.

Lärka (Eng. Lark) is an ICALL platform for deploying different language learning activities, at the moment consisting of an exercise generator for linguists and language learners. The language learner part contains a preliminary version of multiple-choice exercise items for vocabulary training. Training context for exercises is at the moment limited to sentences due to copyright restrictions set on most of the corpora available through the Swedish Language Bank. We need, therefore, a reliable automatic approach to selection of appropriate example sentences for language learners, which means, that we need to take into account learner proficiency levels and relevance for different types of vocabulary aspects.

In this study we have evaluated two different approaches to the selection of examples.

In the first algorithm, each example is scored independently of all other examples using a manually defined set of heuristic rules, each of which has an associated weight:

– sentence length: sentences shorter that 10 words or longer than 15 words have 5 points withdrawn for each item not in the range;

– rare words: two relevance points are subtracted for each infrequent word, defined as words above the frequency threshold of 200 based on a frequency list over word forms in the Swedish Wikipedia Corpus;

– keyword position: five points are withdrawn if the keyword item appears after the tenth position in the sentence;

– finite verb: sentences without finite verbs get 100 points withdrawn.

This is in principle similar to the well-known GDEX algorithm often used in lexicography (Kilgariff et al. 2008, Husák 2008).

In the second algorithm (Borin et al. 2012), we additionally took into account the intuition that in order to get a good overview of the usages of a word, e.g. to represent different senses of a lexical item in SweFN context, the examples should not only be typical but also different.

This notion of difference is formalized as a similarity metric. The joint optimization of the sum of goodness score according to the heuristic rules and the dissimilarity scores is a computationally intractable problem in general, but can be approximately solved using diversification methods developed in the information retrieval community (Minack et al. 2011). We used a similarity measure based on the Euclidean distance between feature vectors; these vectors represented words in the context of the search terms, as well as a number of syntactic features derived from dependency trees.

The critical question for the present study is whether the two approaches target the parameters that ensure acceptable example ranking; which of the two approaches performs better; what other parameters might be necessary to consider to improve algorithm performance as predictors of good examples. The goal of the study is, in other words, to evaluate the two above-mentioned algorithms; and as a side effect – to identify other potential parameters for Swedish that need to be considered.

3 Related research

Of all the research aimed at selecting authentic examples, the main bulk of studies have been dealing with text readability as opposed to sentence readability. Text readability measures have been explored in a number of studies (Flesh 1948; Björnsson 1968; Huckin 1983; Cedergren 1992; Fulcher 1997; Collins-Thompson and Callan 2004; Mühlenbock and Johansson Kokkinakis 2009; Volodina 2010, etc.); some of them describe CALL and ICALL applications that make use of the measures for automatic selection of texts of appropriate language learner proficiency levels (REAP,[1] Read-X,[2] Ott & Meurers 2010).

Even though larger contexts, like text, are usually preferred in language learning setting, sentence, nevertheless, cannot be neglected in this discussion. It is a popular linguistic unit when it comes to demonstrating use of vocabulary items for students, e.g. to provide an extra example to usage of an item. In our case it is a necessary limitation imposed by copyright restrictions set on many corpora. Therefore the issue of sentence readability needs to be addressed separately.

When it comes to the source of examples, there have been lively discussions about their nature – should they be authentic, invented or should there be a compromise between the two in the form of simplified corpus examples. Authentic examples, though of course praised by many, are criticized for being rather long and containing too many infrequent words; and that "authenticity" as it is plays greater role for native speakers than for language learners or lexicon users. On the other hand, it is time-consuming to invent examples. Automatic selection of examples from authentic corpora speeds up the process, but it is known to be controversial since the notion of "good examples" is subjective and often conflicts with the notion of "authentic examples". However, it is argued that with semi-automatic approaches using so-called "curation", i.e. applying human proofreading and editing where necessary, authentic materials can acquire the necessary precision, accuracy and appropriateness (Hubbard, 2012).

Good examples change their characteristics depending upon who is defining them. Most of research within automatic example rating has been done within the domain of lexicography (Kilgariff et al. 2008, Husák 2008, Kosem et al. 2011, Didakowski et al. 2012); only a few studies exploring characteristics of good sentence-long examples within L2 learning (Segler 2007) or aimed at people with special needs (Heimann Mühlenbock, forthcoming).

Regardless of the target group, it has been proven that sentence length is one of the most reliable predictors of sentence readability. Other classifiers vary within different projects and for

[1] http://reap.cs.cmu.edu/

[2] https://sites.google.com/site/elenimi2/read-xpublications

different languages. For example, linguistic features such as sentence length, word frequencies, pronouns, main clauses have been found useful as main predictors of sentence readability for English; punctuation and proper names being used as additional indicators of how well-formed and easy-to-understand a sentence is (Kilgariff et al. 2008; Husak 2008). The Slovenian team (Kosem et al. 2011) tested different configurations of linguistic classifiers and compared them in several iterations, having naturalness, typicality and intelligibility as primary criteria for human evaluators. Even sentences showing potential to be turned into good dictionary examples have been considered as good ones. The classifiers that have shown the best predicting ability for Slovene have turned out to be: preferred sentence length, relative keyword position, penalty for keyword repetition, optimal word length.

Different approaches treat linguistic constraints differently. For instance, unlike the English and Slovenian GDEX approaches described above, where all the features are non-obligatory, i.e. none needs to be necessarily met, an approach adopted by the German team (Didakowski et al. 2012) applies harsher selection. They define a set of parameters with some of them being "hard", i.e. examples are not considered at all if the constraint is not met.

4 Method

Starting from the previous practical and theoretical findings, we designed our evaluation set-up:

Given the two existing algorithms for Swedish, we needed to evaluate their prediction performance on authentic examples and compare them with human judgment. To do that, we selected 60 test items (keywords) from the Swedish Kelly-list, an L2 learner frequency list of modern Swedish (Volodina & Johansson Kokkinakis 2012), taking ten items from each learner proficiency level as defined by Common European Framework of References, CEFR (Council of Europe 2001). Only lexical word classes have been considered, i.e. nouns, verbs, adjectives, adverbs. The number of selected items per word class reflects part-of-speech distribution per CEFR level in the Kelly-list. By having items from a learner-oriented list we tried to address both lexicographers, linguists and L2 teachers as potential user groups.

The 60 items have been sent to the algorithms that made corpus searches in Korp (Borin et al. 2012a) and ranked the hits. Three top results per algorithm and keyword have been saved in a specially designed database. We kept all the annotations coming from corpora for later statistic analysis of linguistic parameters.

Search for examples was made in several corpora: *SUC* (Stockholm Umeå Corpus), which is often used as the "gold standard" of POS annotation since it has been manually proofread; it amounts to 1.2 mln tokens (Källgren et al., 2006); *Talbanken*, which is a manually constructed treebank from the 1970s, that is considered to be the "gold standard" of syntactic annotation; the professional prose part used in this project contains 86,000 words (Teleman 1974; Einarsson 1976; Nivre et al. 2006); and *LäsBarT*, a collection of easy-to-read texts from the 2000s amounting to 1 mln. words (Heimann Mühlenbock, forthcoming).

We initially planned to use only the 3 above-mentioned corpora since they can boast reliability in PoS and syntactic annotations. However, the number of hits for some of the keywords on the list (for CEFR levels B2–C2) proved to be not extensive enough. Therefore to ensure variability of hits per keyword, we added some other corpora, namely; 1) four corpora of fiction prose: *Bonniersromaner I and II* from 1976–1981, *Nordstedtsromaner* from 1999 and *SUC romaner* from 1990s, totaling at about 18 mln words; 2) *PAROLE*, a corpus of mixed texts (novels, newspapers, journals and web text) to balance down the amount of novels (about 24.5 mln words).

Once the database was populated with corpus examples, the user interface was set up with an option for "voting" for appropriateness of examples: *acceptable* ("thumbs up"), *unacceptable* ("thumbs down"), *doubtful* ("question mark").

We provided a possibility to leave a comment about each example, but it wasn't obligatory. The user was given an opportunity to go back to the previous answers and change them. To avoid any bias in their answers, users were not given information about which of the two algorithms has suggested this or that example sentence. The JSON[3] button, however, (placed in the same cell as examples) reveals all corpus- and user-related information about each example.

[3] JavaScript Object Notation

All users had to evaluate the same population of example sentences. In the result set we had each particular sentence associated with five human votes and optional comments. In addition, sentences contained linked information about which of the algorithms has suggested them, whereas user votes had information about the user target group.

We have asked 5 people to perform evaluation. They come from two different professional backgrounds, some of them working across several subjects, namely: one lexicographer, one lexicographer/computational linguist, and three L2 teachers/computational linguists. Three of them have Swedish as their mother tongue; two others are non-native proficient users of Swedish; all of the participants have doctoral degrees; two of them are men, three are women.

In selection of evaluators the most important factor was that they all are actively involved in the development of the two resources that the algorithms have been developed for – SweFN and Lärka. They are well-trained and qualified to make judgments about example appropriateness and therefore their answers are relevant in terms of requirements set on the example selection.

The users have been instructed to look at every example and assign it a vote ("acceptable", "unacceptable", "doubtful") following the same judgment they would use selecting examples when working with one of the two projects.

This way we collected information about how often human graders agreed with algorithm judgments and could make conclusions about appropriateness of different rating approaches to example selection. Moreover, the optional comments provided us with insights about the linguistic features that we need to take into account in the future versions of algorithms.

A word about bias and limits of this research: we would like to note that four of five participants are computational linguists which supposedly has influenced the type of comments they provided. We presume that their answers are more reasonable in terms of what technology can perform. This might also have influenced their ratings in favor of the algorithms. Users without technical background tend to set higher requirements on technology. We have been aware of that and in fact very interested in their responses since they could help us pinpoint technically reasonable classifiers and predictors which we overlooked from the start.

5 Results and discussion

5.1 Quantitative data

We have looked into how the approach represented by algorithm #1 has performed compared to the approach in algorithm #2 – first in general and then for each target user group, for individual parts-of-speech and finally for learner proficiency levels.

As shown in table 1, algorithm #1 has "won" over algorithm #2 by 6.3% (56.6% to 50.3%). Reasons could be different, one of them being that #2 presents top examples with dispersion built in, i.e. it presents versatility of a lexical item demonstrating it in a group of examples with different realization of meanings and in various syntactic patterns; and thus should be evaluated as a group of examples, rather than individual examples in isolation.

	acc	*unacc*	*doubtful*	*total*
alg# 1	509 56.6%	177 19.7%	213 23.7%	899 100%
alg #2	453 50.3%	242 27%	204 22.7%	899 100%
Total (#1+#2)	962 53.5%	419 23.3%	417 23.1%	1798 100%

Table 1. Distribution of acceptances between the two algorithms.

Algorithm #2 has also suggested more examples that, evaluated individually, were more often found *unacceptable* than the ones suggested by algorithm #1 (27% to 19.7%). The number of *doubtful* examples, however, is almost equal between the two algorithms (23.7% to 22.7%).

Distribution of acceptances between the two user groups looks as illustrated in table 2.

user groups	*acc*	*unacc*	*doubtful*	*total*
lexico-graphers	458 63.6%	144 20%	118 16.4%	720 100%
alg #1	238 66.1%	67 18.6%	55 15.3%	360 100%
alg #2	220 61.1%	77 21.4%	63 17.5%	360 100%

user groups	*acc*	*unacc*	*doubtful*	*total*
L2 teachers	504	275	299	1078
	46.7%	25.5%	27.7%	100%
alg #1	271	110	158	539
	50.2%	20.4%	29.3%	100%
alg #2	233	165	141	539
	43.2%	30.6%	26.1%	100%

Table 2. Distribution of votes per user group

Table 2 indicates that lexicographers slightly favored algorithm #1 compared to algorithm #2 (66.1% to 61.1%, *acceptable* examples); the *unacceptable* votes do not either have a clear tendency to distinguish algorithm #1 as a better one (18.6% to 21.4%). Numbers for L2 teachers, however, show an obvious tendency to favor algorithm #1: 50.2% to 43.2% of votes given to *acceptable* examples versus 20.4% to 30.6% to *unacceptable* ones. Here, too, individually well-formed examples from #1 seem to play a more important role for L2 teachers than versatility of a lexical item presented in a group of examples which seems to be important for lexicographers.

An interesting tendency has been shown in ratings viewed from the point of learner proficiency levels.

CEFR levels	*acc*	*unacc*	*doubtful*	*total*
A1	153	81	64	298
	51.3%	27.2%	21.5%	100%
alg #1	73	41	35	149
	49%	27.5%	23.5%	100%
alg #2	80	40	29	149
	53.7%	26.8%	19.5%	100%
A2	146	62	92	300
	48.7%	20.7%	30.7%	100%
alg #1	86	18	46	150
	57.3%	12%	30.7%	100%
alg #2	60	44	46	150
	40%	29.3%	30.7%	100%
B1	143	94	63	300
	47.7%	31.3%	21%	100%
alg #1	84	34	32	150
	56%	22.7%	21.3%	100%
alg #2	59	60	31	150
	39.3%	40%	20.7%	100%
B2	175	56	69	300
	58.3%	18.7%	23%	100%
alg #1	91	25	34	150
	60.7%	16.7%	22.7%	100%
alg #2	84	31	35	150
	56%	20.7%	23.3%	100%
C1	161	63	76	300
	53.7%	21%	25.3%	100%
alg #1	83	29	38	150
	55.3%	19.3%	25.3%	100%
alg #2	78	34	38	150
	52%	22.7%	25.3%	100%
C2	184	63	53	300
	61.3%	21%	17.7%	100%
alg #1	92	30	28	150
	61.3%	20%	18.7%	100%
alg #2	92	33	25	150
	61.3%	22%	16.7%	100%

Table 3. Distribution of votes per learner proficiency level (CEFR-based)

In table 3 we can see a clear tendency of algorithm #1 performing better than #2 for items coming from intermediate proficiency levels B1 and B2, both in terms of higher acceptance and lower rejection rates. This tendency is less clear at levels A2 and C1. Performance per algorithm is strikingly equal for items at levels A1 and C2. Hypothetically, this might indicate that the lower the learner level is, the stricter constraints might need to be applied to example well-formedness to make them appropriate. At intermediate levels (B1, B2) “normally” well-formed examples are much more easily accepted; and the requirement for examples to be well-formed decreases by level C2 (= “proficient language user”); so that both algorithms are performing equally well.

Viewed as a whole, the total number of *acceptable* examples (from both algorithms) is nearly equal to the sum of *unacceptable* and *doubtful* examples: 53.5% versus 46.4%. It means that algorithms suggest 54% of examples that users accept as good ones. This leaves us with the task of improving the rating strategies to offer a higher rate of *acceptable* examples.

5.2 Qualitative data

Analysis of the comments left by the evaluators reveals a range of positive and negative arguments, with critical ones prevailing, which can be summarized as follows: structural, lexical, related to annotation and heterogeneous comments.

1. A large group of comments mention structural features of the sentence, among them:

- *Use of ellipsis*. Elliptic sentences can be found both among the approved examples and among the discarded ones, e.g. *Dämpar inflationen.* 'Decreases inflation' or *Sannolikt, sade van Delden*. 'Most likely, said van Delden'. In both cases, ellipsis has been criticized, e.g. one of the comments says: "elliptical construction; it can function as a possible usage example; but it is not a typical use of this word". A possible approach to this problem could be to check each example for finite verb and subject, and especially check for completeness the clause where the keyword is used.
- *Use of passive*. A recurrent criticism has been aimed at sentences containing passive, even in cases where examples have been marked as *acceptable* ones, e.g. *Midsommarens ritualer genomgicks.* 'The rituals of Midsummer was explained.' The evaluator has written: "passive should rather be avoided". Other comments criticized use of passive in combination with other complicating parameters, e.g. "compounds; plus domain-specific vocabulary; plus passive" for the rejected example: *Uppgången dämpades i avvaktan på fredagens sysselsättningssiffror för maj*. 'The increase was dimmed awaiting employment figures for May on Friday.'
- *Limited context*. Some of the examples have been rejected on the basis of being difficult to understand in the provided context. One of the comments says: "too short to be illustrative" for the rejected example *Påstår Gunnar alltså.* 'States Gunnar that is.'
- *Non-typical word order*. Example of the rejected *sentence: Efter semifinalförlusten känns därför behovet av en föryngring akut.* 'After the loss in the semi finals the need for a rejuvenation seemed acute.'
- *Use of anaphora/pronouns*: An example of such is *I åratal hade det sparats till den.* 'One had been saving up for it for years.' Use of anaphoric expressions inhibit understanding, therefore presumingly it would be reasonable to avoid sentences where both subject and object are expressed by pronouns.
- *Not appropriate for the learner level*. This type of comment has often been provided for sentences containing a combination of complicated factors, such as unusual (non-frequent) vocabulary, compound words; structurally difficult sentences with inverted word order or long phrase structure, e.g. *Som kandidat till utrikesministerposten utpekades EU-parlamentarikerm Elisabeth Guigou.* 'Elisabeth Guigou singled out as candidate for the job as minister of foreign affairs.'

Structural parameters seem to have influenced a lot of decisions against the acceptance of suggested examples. Technically viewed, several of the listed parameters can be easily incorporated into the future algorithms, e.g. restriction on elliptic sentences, on use of passive and pronouns; others, e.g. non-typical word order, might require some brainstorming in terms of which structures to classify towards typical versus non-typical word order; and more importantly in which contexts to classify them as unusual (e.g. for language learners at beginner levels). Limited context is another such parameter. It seems that sentences of the same length can sometimes be sufficiently informative, and at other times highly unrevealing of the word meaning.

A more radical way of treating syntactic characteristics could be a *discriminative approach to target word classes*, e.g. building specific parameter configurations for each word class, e.g. for verbs – check semantic and syntactic valency in a GLDB (The Göteborg Lexical DataBase) (Järborg 1989, http://www.ilc.cnr.it/EAGLES96/rep2/node19.html) and look for identified patterns; for adjectives – look for typical patterns, e.g. keyword in pre-modifier, post-modifier or in attributive positions, etc. Checking statistic results for most frequent structural and lexical patterns for the keyword, so-called word pictures, mutual information, Z-score and other measures of degree of collocality between the keyword and its neighboring words is another possible approach.

2. Second group of comments focuses on lexical features of example sentences:

Stricter word frequency filtering. In many cases examples have been rejected because of the difficult word choice, i.e. containing *domain-*

specific or *advanced* vocabulary. One example of such sentence is *Avskrapade smulor av yxorna blandades i bly vid hagelstöpning för att uppnå bättre träffsäkerhet.* 'Scraped crumbs from the axes were mixed with led at the hail steeping in order to achieve better accuracy.' A more detailed recommendation has been provided about frequency range of finite verbs (on more than one occasion), for example: "a finite verb should be more frequent one" as a comment for sentence *Tyvärr snuddar också "Studio Ett" en smula vid den sortens generalisering.* 'Unfortunately, the radio program "Studio Ett" also touches upon that sort of generalization.'

- *Use of proper names*: general recommendation provided by the evaluators is that proper names should be avoided. Examples criticized for (unusual) proper names can, however, be viewed as good potential examples if some human editing is applied. e.g. *Improjekteatern ger 'Ritualer', en improviserad föreställning i Observatorielunden kl.19* . 'The "Improjektteatern" gives 'Rituals', an improvised show in the "Observatorielunden" at 19 o'clock.'
- *Use of acronyms and abbreviations in* example sentences has been criticized on several occasions.
- *Use of compounds*: a repetitive criticism. Swedish is famous for its compounding as a productive word-building pattern. Words can therefore become very long and difficult to interpret, e.g. *Och fredagens relativa marknadslugn kan avläsas i kursdiagrammet för lågräntan* 'And the calm market on Friday could be read in the stock chart of low interest rate.'
- *Semantic definition through antonyms/synonyms:* marked as a positive feature in examples like *Sammanbrottet är roligare än bygget*, and *Flanera blir till promenera* 'The collapse is more fun than the construction' and 'Strolling becomes walking'.
- *Keyword repetition:* avoid sentences where target item is used more than once since examples becomes non-explanatory.

Lexical features have proven to be crucial, especially for L2 teachers. Most of the listed parameters would be trivial to implement. When it comes to compounds, available methods for identification of compounds, e.g. via Saldo morphology, need to be checked and tested for reliability. To impose a stricter word frequency filtering we need to consider the type of vocabulary, and therefore underlying word lists, relevant for different purposes and target groups.

3. Third group of comments directs critics at annotation.

Problems with *errors in PoS annotation* result from the fact that we have been using corpora that were not manually proofread, and therefore certain percent of annotation errors can be expected.

However, some of the frustration has been caused by the fact that keywords have been more or less systematically provided as a different part of speech than the one specified, e.g. participles where verbs have been targets; adverbs instead of adjectives; and proper names for nouns. This depends upon *search strategies used in Korp* (Borin et al. 2012) web service that we are using for primary example selection.

4. The last group of comments is heterogeneous and takes up more general aspects of sentences, such as typicality, metaphoric use, etc.:

- *Prototypical*. Approval of the typical meaning and typical context for the target item, e.g. for the approved sentence *Tidigare verk, "Brödrosten" och "Warszawapakt" var två kortoperor*. 'The previous works "Brödrosten" and "Warszawapakt" were two short operas.'
- *Not demonstrative of structural or semantic patterns of the target word*, e.g. for the approved sentence *Ordet "möjligen" skrämde mig.* 'The word "möjligen" scared me.'
- *Metaphoric use*; e.g. for the approved sentence *Ljuskänglorna dansar i mörkret.* 'The light cones were dancing in the dark.'
- *Strange* (as a variant: *not clear*, etc.), for example for the rejected sentence *Den skällde skräck och lydnad*. 'It was barking of fear and obedience.'
- *Abstract use*, e.g. for the example marked as doubtful: *Avståndet från 'ätbart' till 'jätteäckligt' är mikroskopisk*. 'The distance from 'ätbart' to 'jätteäcklingt' is microscopical.'
- *Innovative modern use*, e.g. for the approved sentence *Öken, tycker Peter om banan i Lierop.* 'Desert, Peter thought of the course in Lierop.'

Categories like "strange", "metaphoric", "abstract" are difficult to account for automatically. Hypothetically, strange and abstract examples will be reduced among the top results,

once we have improved structural and lexical filtering. Techniques derived from word sense discrimination (Purandare and Pedersen, 2004) can also help us reduce such examples among the top results.

5.3 Statistic data over linguistic features

The rich annotation accompanying each sentence token has become an important source of statistical analysis of *acceptable* versus *unacceptable* examples. Below we are looking into whether *acceptable* examples (for both algorithms) share any common features and how these contrast with the *unacceptable* examples.

Linguistic feature	*acc*	*unacc*
Sentence length, range (tokens)	3–27	3–27
Sentence length, average (tokens)	8	9
Sentence length, mean (tokens)	7	7
Word length, range (characters)	1–23	1–23
Word length, average	5	5

Table 4. Surface features in *acceptable* versus *unacceptable* examples

Values for surface features, such as sentence length and word length presented in table 4 do not seem to be discriminating for example acceptability. The optimal sentence length of 7 tokens suggests that sentences do not have complex phrase structure and do not tend to contain subordinate clauses.

As far as the presence of different word classes is concerned, a summary of indications of the examples is found in table 5.

Linguistic feature, % of sentences	*acc*	*unacc*
Absence of nouns	9%	1
Presence of proper names	29%	29%
Presence of pronouns	27%	64%
Presence of adverbs	36%	44%
Presence of numerals	10%	8%
Presence of conjunctions	12%	20%
Presence of subjunctions	2%	3%

Table 5. Presence of selected word classes in *acceptable* versus *unacceptable* examples

Only 9% of *acceptable* examples contain no nouns at all, which means either use of proper names/pronouns or imperative/elliptical sentences. 73% of the *acceptable* examples do not contain any pronouns at all which presumably depends on the fact that pronouns often make anaphoric references which may be difficult to interpret in one-sentence context. The latter fact might have become the reason for rejection of some examples: we can see that 64% of rejected examples contain pronouns.

In 64% cases of *acceptable* examples, they do not contain any adverbs. This might indicate the fact that sentence structure without adverbials is easier to interpret and is therefore to prefer.

Function words indicating more complex sentence structure, like conjunctions and subjunctions, tend to be absent in the *acceptable* examples, e.g. only in 12% of *acceptable* examples conjunctions are used (versus 20% in *unacceptable*); and only in 2% of accepted sentences subjunctions are used.

Some numbers have been obtained for clause level, such as presence of subjects, finite verbs, subordinate clauses, complex phrase structures (table 6).

Linguistic feature	*acc, nr per sentence, in % of sentences*	*unacc., nr per sentence, in % of sentences*
Subject (S)	0 S: 7.2% 1 S: 86% 2 S: 5.8% 3 S: 0.7% 4 S: 0.1%	0 S: 11% 1 S: 80% 2 S: 7.5% 3 S: 0.5% 4 S: 1.4%
Finite verb(FV)	1 FV: 91.2% 2 FV: 8.1% 3 FV: 0.5% 4 FV: 0.1%	1 FV: 87% 2 FV: 11.4% 3 FV: 0.7% 4 FV: 1%
Subordinate clause (SC)	0 SC: 96% 1 SC: 4%	0 SC: 93% 1 SC: 6% 2 SC: 1%
S-passive (SP)	0 SP: 96% 1 SP: 4%	0 SP: 95% 1 SP: 5%
Complex phrases (CP)	0 CP: 11.2% 1 CP: 55% 2 CP: 29% 3 CP: 4.4%	0 CP: 8.8% 1 CP: 60% 2 CP: 28.5% 3 CP: 1.9%

Table 6. Statistics on the clause level

Though showing only a slight difference between the groups of *acceptable* versus *unacceptable* examples, the group of *unacceptable* examples contains more sentences without subjects (11% vs 7.2%); more examples with multiple subjects (9.4% vs 6.6%); they more often contain several finite verbs (13.1%) compared to the group of acceptable examples (8.7%).

Finally, we calculated the lexical frequency profile for each sentence in the evaluation set, see table 8.

LFP, % of sentence tokens	acc, range	unacc, range	acc, average	unacc, average
Voc, CEFR A1	20–100	20–91	60	58
Voc, CEFR A2	0–50	0–40	6	6
Voc, CEFR B1	0–40	0–33	5.3	5.4
Voc, CEFR B2	0–29	0–33	3.8	3.8
Voc, CEFR C1	0–40	0–40	3.2	3.15
Voc, CEFR C2	0–33	0–33	3	2
Voc, C2+	0–75	0–75	18.8	21.4

Table 8. Lexical frequency information

Lexical frequency information has been collected per lemma using Kelly word list; punctuation has been counted towards A1 items assuming that all language users are familiar with it. Words calculated towards C2+ are the ones not appearing among A1–C2 words in the Kelly list, and are thus assumed to be rare and presumingly more difficult to understand.

Looking at the numbers we have received,we can see that lexical complexity of the *unacceptable* sentences only a few percent higher than of the *acceptable* ones: A1 words, i.e. most frequent ones ones (60% vs 58%); and C2+ vocabulary, i.e. less frequent words (18.8% vs 21.4%). We would need to investigate these numbers further to arrive at any relevant measures for sentence lexical complexity measures.

Therefore, we can summarize that lexical frequency statistics and statistics on clause and phrase levels collected for each example do not straightforwardly explain why *unacceptable* examples have not been approved. It can be said, however, that though numbers concerning vocabulary frequency, phrase structure and clause structure differ only slightly between the groups of *acceptable* and *unacceptable* examples, the tendency for difficulty is more consistent in the group of *unacceptable* examples. Taken in isolation, each parameter differs only slightly between the two groups; however in combination these parameters intensify the "complexity" effect making it unattractive for the the end-users.

6 Concluding remarks

We have presented a series of user evaluations of two automatic algorithms for the selection of illustrative examples from corpora. The first algorithm scored the examples independently of each other based on a few manually defined heuristics, while the second one additionally tried to use a distance function to ensure that the selected set was diverse. Contrary to our intuitions, the simpler algorithm with independent scoring consistently outperformed the complex algorithm taking selection diversity into account. There are several possible reasons for this result: our diversity scoring metric may be too simple, and we may need to make use of techniques derived from word sense discrimination (Purandare and Pedersen, 2004); diversification may be of more interest if the target word is highly polysemous, which we did not take into account when selecting lexical items for our evaluations; we selected fairly small output sets, while diversification may be more necessary for large sets.

In addition to the evaluation of the two algorithms, the user study has given us valuable feedback that can lead to the extension and improvement of the heuristic scoring rules. Several new criteria have been proposed by the users: voice and valency features for verbs, word order, the presence or absence of proper names or acronyms, and the strength of collocation with contextual words. The addition of new scoring rules would make the evaluation function more complex and sensitive, but would also allow us to fine-tune it for particular user groups, such as lexicographers or foreign language teachers.

The algorithms in their final improved form promise to be a useful instrument in applications designed for computer-assisted language learning, for teaching of linguistics, and in lexicographic and linguistic projects. We have plans for embedding the web service for example ranking into Korp,[4]

[4]http://spraakbanken.gu.se/korp/

Karp[5] and Lärka[6] – all of them applications developed and maintained at the Swedish Language Bank.

References

Carl Hugo Björnsson. 1968. Läsbarhet. Liber Stockholm.

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012a. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA. 474–478.

Lars Borin, Markus Forsberg, Karin Friberg Heppin, Richard Johansson, Annika Kjellandsson. 2012b. Search Result Diversification Methods to Assist Lexicographers. P*roceedings of the 6th Linguistic Annotation Workshop*.

Magnus Cedergren. 1992. Kvantitativa läsbarhetsanalyser som metod för datorstödd granskning. <http://iplab.nada.kth.se/pub_all.jsp> (Retrieved 2007-02-08) Stockholm: Inst.för Numerisk analys och datalogi, Kungl. Tekniska högskolan, NADA.

Kevyn Collins-Thompson and James P. Callan. 2004. A Language Modelling Approach to Predicting Reading Difficulty. *Proceedings of the HLT/NAACL Annual Conference*. Boston, MA, USA.

Council of Europe 2001. *The Common European Framework of Reference for Languages*. Cambridge University Press.

Jörg Didakowski, Lothar Lemnitzer & Alexander Geyken. 2012. Automatic example sentence extraction for a contemporary German dictionary. *Proceedings of EuraLex 2012.*

Jan Einarsson. 1976. Talbanken: Talbankens skriftspråkskonkordans/Talbankens talspråkskonkordans. Lund University.

Rudolf Flesch. 1948 A new readability yardstick. *Journal of Applied Psychology*, Vol. 32, pp. 221–233.

Karin Friberg Heppin, Maria Toporowska Gronostaj. 2012. The Rocky Road towards a Swedish FrameNet – Creating SweFN. *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC 2012); Istanbul, Turkey*. p. 256–261

Glenn Fulcher. 1997. Text Difficulty and Accessibility: Reading Formulae and Expert Judgement. *System* vol.25, 497–513.

Jerker Järborg. 1989. *Betydelseanalys och betydelsebeskrivning i lexikalisk databas*. Göteborg: Inst. f. sv. Spr., Göteborgs universitet.

Katarina Heimann Mühlenbock. Forthcoming. *I see what you mean – Assessing readability for specific target groups.* PhD Thesis, Gothenburg University.

Philip Hubbard. 2012. Curation for systematization of authentic content for autonomous learning. *EuroCALL 2012 Proceedings*, Gothenburg.

Thomas N. Huckin. 1983. A Cognitive Approach to Readability. In: Paul V. Anderson, R. John Brockmann and Carolyn R. Miller, Editors, *New Essays in Technical and Scientific Communication: Research, Theory, Practice*, Baywood, Farmington, NY, pp. 71–90.

Milos Husák. 2008. *Automatic Retrieval of Good Dictionary Examples*. Bachelor Thesis, Brno. Retrieved on 2010-09-22 from http://is.muni.cz/th/172590/fi_b/bachelor_thesis.pdf

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. *Proc EURALEX*, Barcelona, Spain.

Sofie Johansson Kokkinakis and Elena Volodina. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project KELLY – issues on reliability, validity and coverage. *Proceedings of eLex 2011*, Slovenia.

Iztok Kosem, Milos Husák and McCarthy Diana. 2011. GDEX for Slovene. *Proceedings of eLex 2011*, Slovenia, pp.151–159.

Gunnel Källgren, Sofia Gustafson-Capková and Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University.

Enrico Minack, Wolf Siberski, and Wolfgang Nejdl. 2011. Incremental diversification for very large sets: a streaming-based approach. In P*roceedings of the 34th International ACM SIGIR Conference on Research and Development of Information Retrieval*, SIGIR'11, pp. 585–594. New York, United States.

Katarina Mühlenbock and Sofie Johansson Kokkinakis. 2009. LIX 68 revisited - An extended readability measure. *Proceedings of Corpus Linguistics 2009*.

Joakim Nivre, Jens Nilsson & Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In

[5]http://spraakbanken.gu.se/karp/

[6]http://spraakbanken.gu.se/larka/

Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006) Genoa: ELRA. 1392–1395.

Niels Ott and Detmar Meurers. 2010. Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education*. Vol 3, No 1-2. Special issue on "Computer-aided language analysis, teaching and learning: approaches, perspectives and applications" edited by George Weir and Shin'ichiro Ishikawa, 2010.

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning* (CoNLL), pp. 41–48. Boston, United States.

Thomas M. Segler. 2007. *Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German*. Doctoral Thesis. University of Edinburgh. Retrieved on 2010-09-22 from http://www.era.lib.ed.ac.uk/bitstream/1842/1750/3/Segler TM thesis 2007.pdf

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund.

Elena Volodina. 2010. Corpora in Language Classroom: Reusing Stockholm Umeå Corpus in a Vocabulary Exercise Generator. *LAP Lambert Academic Publishing*, Colne, Germany.

Elena Volodina and Lars Borin. 2012. Developing an Open-Source Web-Based Exercise Generator for Swedish. *EuroCALL 2012 Proceedings*, Gothenburg.

Elena Volodina & Sofie Johansson Kokkinakis. 2012. Introducing Swedish Kelly-list, a new lexical e-resource for Swedish. *LREC 2012*, Turkey.

Automatic question generation for Swedish: The current state

Kenneth Wilhelmsson
Språkbanken
Department of Swedish
University of Gothenburg, Sweden
kenneth.wilhelmsson@gu.se

Abstract

The research area of question generation (QG), in its current form, has a relatively brief history within NLP. A description of the current question generation implementation for Swedish text built on schema parsing is here presented and exemplified. Underlying the current approach is the view of 'all textual information as answers to questions.' This paper discusses strategies for enhanced functionality for arbitrary Swedish text through extended question generation. It also brings up some theoretical issues regarding the nature of the task, and concerns practical considerations in an area such as Intelligent CALL (*ICALL*) where this type of application has been considered for English.

1 Introduction

The field of question generation can be said to have old roots in AI research, but a young community involved in research on QG, particularly regarding the English language, has now appeared and brought a reoccurring international workshop series into being.[1]

A definition of the QG task in general is to exhaustively produce all questions that a text can be said to provide the answers to. (This definition is taken from proceedings edited by Rus and Graesser, 2009.) Whether or not such a definite *exhaustive* set for any text is possible to determine is of course debatable.

QG for arbitrary text is a type of NLP application that puts special demands on contributing basic NLP techniques. QG for Swedish with the current approach (see figure 2) relies on a parsing format where the identification of spans of functional constituents, such as adverbials, must be exact and include post-modifiers. For Swedish text, there exist a few parser implementations with such suitable capabilities for free text. QG for arbitrary Swedish text has, to the best of the knowledge of the author, only been undertaken using schema parsing (Wilhelmsson, 2010),[2] which currently only gives a parse of the main clause level, see example 1. (QG for Swedish was introduced as one of the suitable applications of schema parsing independently of ongoing research on English, then unknown to the author, *ibid*, chap. 5). In the recent QG research period, a number of plausible areas of application for QG have been identified. Swedish QG has to this point been seen especially in the light of information extraction *(IE)*, see figure 1. In this approach, the text database of Swedish *Wikipedia* was used frequently as a text source, and the role of QG has been as a generic usability resource aimed at enhancing quick access of specific portions of information. Why should then explicit generation of questions be used in a

[1] See http://questiongeneration.org/

[2] Tasks similar to or equivalent of QG for Swedish has however been discussed by researchers considering the inductive dependency parser *MaltParser* (Nivre 2006).

Kenneth Wilhelmsson 2012. Automatic question generation for Swedish: The current state. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 71–79.

general information extraction setting? The idea about QG in natural language query interfaces has been that a UI allowing arbitrary natural language questions as input too often results in only 'best string matches'; there is no guarantee that a freely formulated question actually is answered at all in the current text database. Fully functional QG, on the other hand, ideally only allows a user to choose among the explicit questions produced, answered by the text by grammatical definition (see figure 1). In the GUI of the QG program has an auto-completing dropdown-menu allowing only generated question to be posed, see figure 1. However, the QG program for Swedish described here may be used for other purposes.

The pedagogic situation, such as tutorial dialog, is often mentioned in the literature regarding QG. The international research describes uses of QG in ICALL applications such as automatic generation of exercises (eg. Lefevre Jean-Daubias, and Guin, 2009, Wyse and Piwek, 2009). QG in an ICALL setting has a potential use in automatically created tests of reading comprehension. In an ICALL setting, new subtasks such as selecting a useful subset of all questions generated appear.

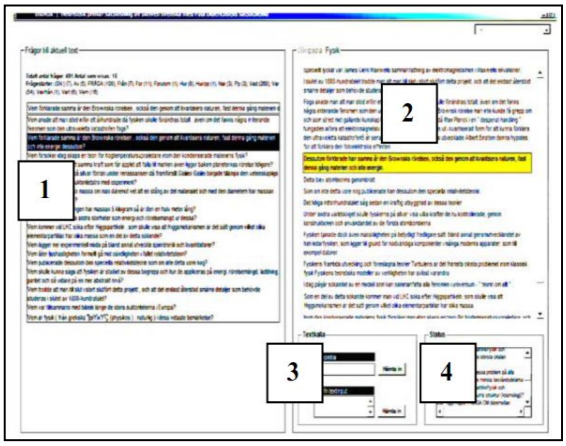

Figure 1: The GUI of the QG program for more general information extraction (taken from Wilhelmsson 2011)
1) Autocompleting input form for choice of question
2) The text source in which the suggested questions will mark and scroll to the corresponding section with answers
3) Forms for choice of Wikipedia article or text input
4) Status box displaying various information during a run.

1.1 Outline

This paper begins by describing experiences with development of Swedish QG so far; it also discusses possible improvements for potential applicability to areas such as Intelligent CALL.

Section 2 starts by describing the existing implementation, where questions concerning explicit full main clause functional constituents are treated. It also brings up additional question types stemming from partial and multiple functional constituents.

Section 3 deals with questions that could be produced using minor deduction techniques, notably pronoun resolution.

Section 4 brings up how a more advanced QG system could use 'shallow reasoning', logical conclusions and possibly exploit semantic lexica to draw conclusions resembling more human-like reasoning.

Section 5 brings up reformulation of information as a strategy, e.g. in ICALL. While not resulting in true change of content, altering syntax and vocabulary of facts can be motivated for testing the reading comprehension of a student.

Section 6 deals with a theoretical aspect of QG and explains why certain texts can appear self-contradictory from this QG viewpoint.

Section 7 discusses weaknesses in the current QG implementation. Achieving high correctness in QG can be regarded as a form of the common situation in NLP where both precision and recall rates must be raised simultaneously.

2 Generation of questions answered by explicitly stated information

Question generation regarding explicit information from functional grammatical constituents in Swedish declarative sentences has been the initial focus for Swedish QG. Generation of questions concerning (i.e., *that are answered by*) the unbounded constituents (subjects, objects/predicatives and adverbials) can in general be carried out as in Wilhelmsson (2011) by a three-step procedure, as in figure 2. Identifying full spans of the constituents considered is a prerequisite. The method has declarative main clauses, or coordinated finite VPs (that inherit subjects from a previous structure) as input.

1. Place the currently fronted constituent in its canonical position, thus creating a yes/no-question (V1 question).
2. Front each unbounded functional constituent of the main clause level. Fronting of the currently available subjects (formal and logical), objects/predicatives and adverbials, thus creates grammatical variations of the same propositions.
3. Substitute each of the adverbial and the nominal constituents with corresponding interrogative phrases, eg. *Wh*-words. This is not always possible or useful for all constituents.

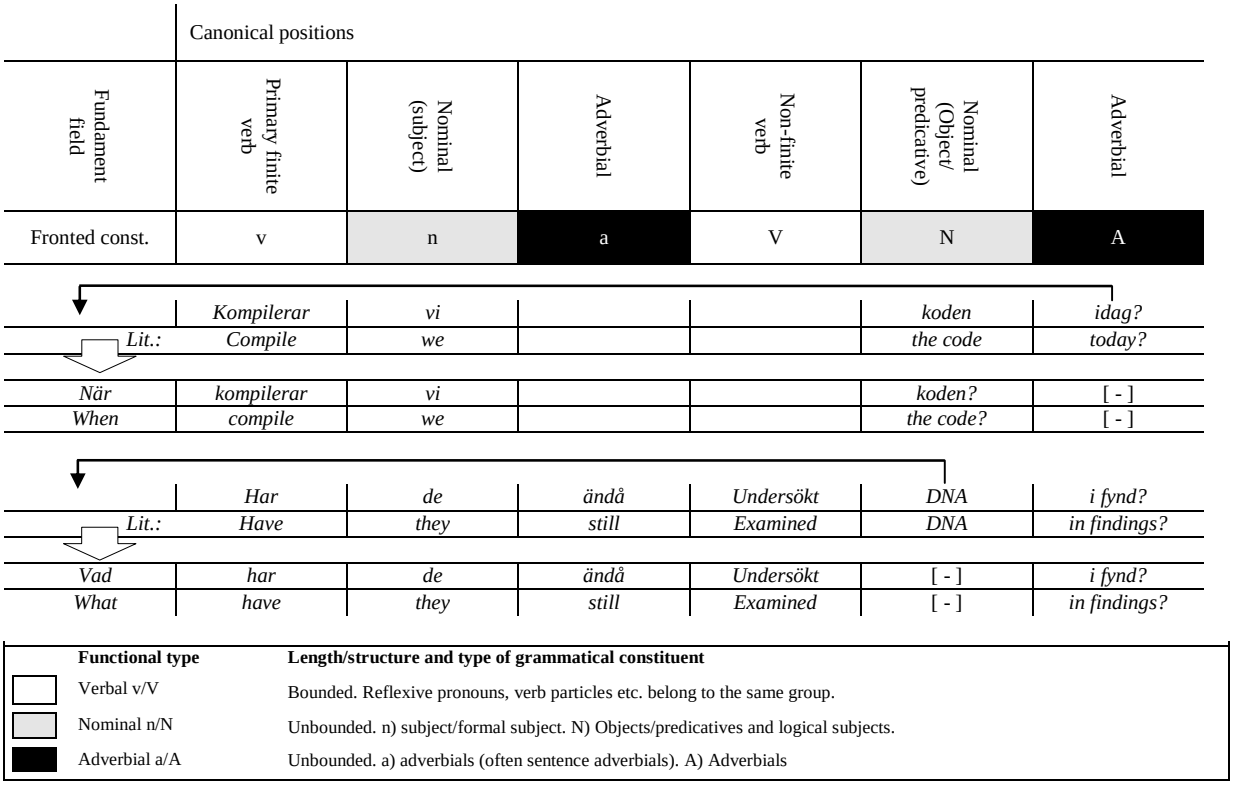

Figure 2. The procedure of fronting the unbounded main clause constituents and substitution with corresponding question segments – e.g. a single *wh*-word (with literal English translations). The sign *[-]* marks *traces* (gaps). Only one adverbial (*idag/today*) and an object (*DNA*) are shown here. However, it also applies (if suitable) to the other candidates *vi/we → who, koden/the code → what, de/they → who, (ändå/still → ε)* and *i fynd/in findings → where*.

The Nordic sentence schema, introduced for Danish by Diderichsen (1946) in figure 2, can be used to describe the general QG process regarding full syntactic constituents dealt with here, theoretically working also for other Germanic languages, except for English, which is not a *V2* language. (In this paper dealing with Swedish, examples will sometimes be given in English when analogous to Swedish.)

The result of the schema parser used here comes in an XML format. The output of the schema parser (example 1) is the input of the QG procedure above (figure 2).

```
<subjekt>Johan</subjekt>
<pfv>skulle</pfv>
<adverbial>alltid</adverbial>
<piv>försöka</piv>
<piv>fundera</piv>
<adverbial>på vad pappa hade sagt att
man skulle göra</adverbial>
<tom>.</tom>
```

Example 1. Tag names: pfv – primary finite verb, piv – primary non-finite verb, tom – 'empty'. So-called prepositional objects are seen as adverbials. 'Primary' here means 'on main clause level'.

The text parsed in example 1 (*Johan skulle alltid försöka fundera på vad pappa hade sagt att man skulle göra./Johan would always try to think about what dad had said that one should do)* originally comes from Stockholm Umeå Corpus 2.0 (Ejerhed, Källgren and Brodda, 2006), unit: kk70-010.

The approach for Swedish QG described here bears some resemblance to an approach for English: *[…] the derived declarative sentence is turned into a question by executing a set of well-defined syntactic transformations (wh-movement, subject-auxiliary inversion, etc.). The system explicitly encodes well-studied linguistic constraints on wh-movement such as noun phrase island constraints […]. The transformation rules were implemented by automatically parsing the input into phrase structure trees with the Stanford Parser […] and using hand-written rules in the Tregex and Tsurgeon tree searching and tree manipulation languages.* (Heilmann & Smith, 2009)

2.1 A brief authentic example

An account of actual question generation for Swedish from an early test with a short text is shown in example 3. The excerpt comes from the article *Brödrost* (*Toaster*) from Swedish *Wikipedia*), five sentences. Each main clause is processed and coordinated main clause level finite VPs are turned into main clauses by inheriting the most recent main clause subject in the text sentence. The example shows questions derived from *explicit full main clause constituents* (subjects, objects/predicatives, adverbials).

1) Brödrost är en elektrisk apparat som värmer upp bröd.
   [A] toaster is an electrical device that heats up bread.
2) Oftast används vitt bröd till detta.
   Usually, white bread is used for this.
3) Brödet får en lite mörkare yta som dessutom blir lite spröd efter att brödet är rostat.
   The bread gets a somewhat darker surface that furthermore becomes a bit crisp after being toasted.
4) En brödrost har elektriska värmeelement bestående av glödande trådar som värmer upp brödet.
   A toaster has electrical heat elements consisting of glowing wires that heat the bread.
5) Den elektriska effekten hos brödrostar är oftast 500-1000 watt.
   The electrical power is most often 500–1000 W.

Eleven unique generated 'candidate questions' was the actual result. As shown, all of these are not grammatically correct.

a) Vad är den elektriska effekten hos brödrostar oftast?
   What is the electrical effect for toasters mostly?
b) Vad är en elektrisk apparat som värmer upp bröd?
   What is an electrical device that heats up bread?
c) ?Vad är brödrost?
   ?What is toaster?[lit., sic]
d) Vad används till detta oftast?
   What is used for this most often?
e) Vad får en lite mörkare yta som dessutom blir lite spröd efter att brödet är rostat?
   What gets a somewhat darker surface that furthermore becomes a bit brittle after being toasted?
f) Vad får brödet?
   What does the bread get?
g) Vad har elektriska värmeelement bestående av glödande trådar som värmer upp brödet?
   What has electrical heat elements consisting of glowing wires that heat the bread.
h) Vad är oftast 500 - 1000 watt?
   What is most often 500–1000 W.
i) ?Vart används vitt bröd oftast?
   ?Whereto is white bread used most often?
j) ?Vilka har en brödrost bestående av glödande trådar som värmer upp brödet?
   ?Which do a toaster have, that consists of glowing threads that heat up the bread?
k) *Vilka har en brödrost elektriska värmeelement?
   **Which has a toaster electrical elements?*

Example 3. A sample text and actual generated questions.

Adverbial structure	Example	Adv./Nom. (as a group)	Possibilities for mapping
AdvP	*Bra/good*	Adv.	Usually pure mapping or none
PartP	*Förvånande/ surprising*	Adv./Nom.	Usually pure vague mapping *(hur/how)* or none
AdjP (or derivations)	*Lyckligt/happily*	Nom./Adv.	Usually pure mapping or none
V1 conditional	*Funkar det/ (lit.: Works it)*	Adv.	(No obvious *wh*-question mapping)
Sub clause	*Som det verkar/ as it seems*	Nom./Adv.	Usually pure mapping or none
NP	*En dag/one day*	Nom./Adv.	Usually pure mapping or none
As-phrase	*Som målvakt/ As a goalkeeper*	Adv.	Usually pure mapping or none
PP	*På bordet/ on the table*	Adv.	Complex situation; many exceptions

Table 1. The various adverbial structures considered with short examples. An in-dept account is given in Wilhelmsson (2012).

Table 1 describes the starting point for finding *wh* correspondences for adverbials from a technical perspective. Adverbials form a large, diverse group. Many types, however, have pure mappings, and question word correspondences can be determined by the phrasal head or similar.

Whereas Swedish adverbials have a large number of potential question counterparts, the situation for nominal constituents (subjects, objects/predicatives) appears to be much simpler. Full nominal constituents mostly correspond *to vad (what), vem (who/whom)* or *vilket/vilken/vilka (which* SING-UTR, SING-NEU, PLU). The choice of the correct counterpart is dependant of the semantics of the head words. Animate references will correspond to *who*, whereas *what* is the default. '*Which'* (Swe: *vilket/vilken/vilka)* is primarily used for full constituents when the set of referents is presumed to be of a fixed size. When a nominal constituent is a named entity (e.g. *Volvo*), there appears to be a need for correct semantic classification, as Volvo will correspond better to *what company* than *what* or *who*.

Only generating these various types of questions stemming from explicit full main clause constituents already means a large set of questions. A rough estimate was around four questions per sentence in some text types with normal settings.

The type of questions answered which have explicit information as answers exemplified above, i.e. those corresponding to full main clause constituents, is the sole question type that has been investigated carefullyand implemented to this point. Particularly, the multi-facetted *wh*-question counterparts of Swedish adverbials have been the focus of a recent research project. [3] Swedish adverbials considered come in roughly eight different structural forms, using the phrase categories naturally discerned when using the de facto standard tagset for Swedish from Stockholm Umeå Corpus 2.0 (Ejerhed et al., 2006), see table 1.

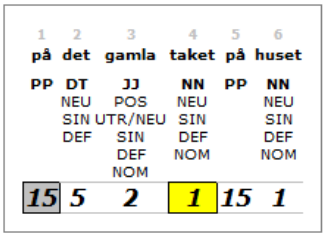

Figure 3. A web GUI of the implementation of identification of head words and heads of prepositional complements *(on the old roof of the house.)*for adverbials. In this case på ('on') and taket ('the roof') are identified and used to decide the *wh* mapping – *var/where*.[4] The rank numbers below determine these words and come from rank-based chunking (see e.g. Wilhelmsson, 2010).

[3] A working paper report in Swedish is available at: http://gup.ub.gu.se/publication/160440-adverbialkarakteristik-for-praktisk-informationsextraktion-i-svensk-text-projektrapport Wilhelmsson (2012).

[4] An online implementation of this is available, currently at: www.ling.gu.se/~kw/applications/adverbialkarakteristik/index.htm It may later be available from: spraakbanken.gu.se

In many cases the head word alone decides appropriate *wh* correspondence for adverbials. In the common PP adverbials, particularly for the common *in/i*, *to/till* etc. the head of the prepositional complement must also be examined, see figure 3. More precisely, the base form of the head of the prepositional complement marked in yellow in figure 3 is preferable. *SALDO* (Borin, Forsberg and Lönngren, 2008) was used for this purpose.

2.2 Other potential questions answered by explicitly stated information

Other explicitly stated information in natural text that could be made to yield new questions includes the following.

2.2.1 Questions answered by full clauses: *Yes/No* questions

Questions answered by full propositions*: It rains today* can easily be used to produce the corresponding *yes/no*-question *(Does it rain?).* In Swedish this means a *V1*-question. As described in Wilhelmsson (2011), this type of question may be less useful – at least in an IE setting: the answers to yes/no questions just confirm the fact (The mere existence of the question 'Does it rain' means that the information 'It rains' is present in the text).

Some sentence adverbials and the like such as *kanske/maybe* may also be seen as answers to *yes/no*-questions. In general, they do not correspond to any particular *wh* question word, like other adverbials.

2.2.2 Questions answered by parts of full constituents and subordinate constituents

Questions regarding smaller information portions than full main clause functional constituents, like modifiers, clearly constitute a large amount of all realistic questions.[5] *The oldest student closed the door → Which student closed the door?* If the schema parser is enhanced to yield a similar detail of analysis for subordinate clause levels, this clearly will lead to a much larger set of information portions and corresponding questions, by allowing additional questions regarding full and partial subordinate functional constituents: *I think that they will buy the car → What do you think that they will buy [-]?*A rough estimate is that half of the grammatical sentences in published Swedish text (SUC 2.0 was examined) include at least one subordinate clause, i.e. sub-clauses (including relative clauses) with finite verb forms.

2.2.3 Questions answered by clause segments spanning more than one full functional constituent

The above description points to several ways of extending the number of questions produced for a text. In actual discourse, however, many question types such as *varför/why* are answered not by one or a few functional constituents (like the *eftersom/because* sub-clause for *why*) but by a series of sentences with an enormous potential syntactic variation. The aim here of course is to investigate QG systematically, and to do so by going from an expression to the corresponding question, rather than the other way around.

Questions of the type *Vad gjorde de/What did they (What did they do)* deal with full VPs including objects etc. These questions appear to be possible to generate, although not all VPs correspond to *do*. See section 6 below.

3 Generation of questions from explicit information with a minor degree of deduction

The previous section has shown that the explicit information can yield a large number of questions. Still, these correspond only partly to the full set of questions that a text provides answers to. Another type of information requiring some deduction comes from treating certain subordinate clauses like main clauses. E.g. from the main clause; *He knew that they were wrong*, the proposition *they were wrong* might be deduced (provided sub-clause analysis). Depending on the nature of the type of sub-clause and verb, such conclusions may or may not be drawn.

Other types of less obvious questions that can be generated include those stemming from anaphoric references, which clearly is a large class of

[5] In a sense, questions regarding smaller parts of information than full constituents is partly present in the current version of the Swedish QG implementation: A pied piping question, *He gave it to me → To whom did he give it,* is a question regarding the prepositional object *to me* but focusses on a part of it (i.e. its complement, *me*). This becomes clearer in the version with a fronted complement and stranded preposition: *Whom did he give it to?*

(particularly nominal) information in text. This may turn out to be particularly useful in ICALL applications. In the QG implementation aimed at information extraction the GUI 'answered' a question by showing the sentence from where the question was extracted, together with its context. The answer is often an anaphoric expression such as *he*, and this has not been problematic since the user herself solves the referent. On the other hand, an application built for practical ICALL purposes might require resolved anaphoric references, so that answers can be given explicitly.

For Swedish, at least two different main strategies for anaphora resolution can be distinguished. The first comes from a heuristic, very economical rule set proposed and tested by Fraurud (1988). The second type is more influenced by English algorithms, especially of one by Mitkov (1998). In the second approach, pronoun resolution takes many surface aspects into account, weighted for optimal performance. Swedish elaborations have incorporated mixes of the two. Recent attempts for Swedish include hybrid methodology (Nilsson, 2010) combing data-driven and rule-based techniques. Other carefully adapted approaches for pronoun resolution have been presented, eg. by Hassel (2000) and by Algotsson (2007).

4 Logically deduced questions

The kind of information content that humans perceive from text widely exceeds the obvious manifestations covered this far.

All persons born in the US are American citizens.
[…] Barack Obama was born in the US.
→ Barack Obama is an American citizen.

Example 4. An example of a deduction with *universal generalization* in English text.

Deduced questions through rules of logic such as *universal generalization* and other techniques, involving e.g. lexica of semantic information such as *Swedish WordNet* (Viberg, Lindmark and Lindvall, 2002) or *Swedish FrameNet* (Borin et al, 2010) is a class of questions whose size becomes extremely hard to estimate. That will make the concept of 'all questions that are answered by a text', used in a definition of QG vague. In fact, the exact set of questions that each text ought to 'produce' remains unknown, as discussed above. A direct consequence is that it will not be possible to assess relative coverage of a question set generated, see Wilhelmsson (2011).

5 Reformulations of questions

In the information extraction setting mentioned, the idea of QG was to let a user only ask questions which were generated, ensuring that there would be answers in the text. An obvious difficulty was that the user had to *find* the question – more precisely: a formulation of a question in the usually very large set of questions produced. A slightly counter-intuitive method for helping the user finding a question was discussed in Wilhelmsson (2010, 2011) – extending the 'set of questions' even more by adding reformulations. The early tests furthermore showed that substituting words by Swedish near-synonyms from *Folkets synonymordlista* (Kann and Rosell, 2005) and Swedish WordNet (Viberg, Lindmark and Lindvall, 2002) to add alternative question formulations, without word-sense disambiguation, produced many erroneous questions. In an ICALL setting, it may however be a well-founded idea to use the slightly altered correct formulation of a question to test the reading comprehension and vocabulary of a student.

In Wilhelmsson (2010), different *syntactic* changes to Swedish text preserving meaning was also discussed. Whether any of these are relevant to pedagogic situations is not clear, although most of these should be accomplishable in QG, and could similarly produce a not too obvious variant of the information in a teaching situation.

6 Identical questions and the time aspect

Consider a text article about a particular person. Many sentences may involve this one person as a subject, perhaps in an anaphoric form. The result for QG will be something that has already been noted in current implementations: there will often be several identical questions produced. Those will have different origins in the text and therefore different answers. Especially the VP type of question sketched above (*What did he do?*) might be generated repeatedly in that context. Clearly, a human teacher or similar would ideally choose not to pose that type of question at all.

Another aspect of this 'question ambiguity' is that many text types (e.g. the '*story'* genre) do not capture a fixed point in time with a world in a static state, but rather a time span involving new events and changes to the states of objects in the world, throughout the text. Consequently, such a text may, by different text sections, indicate both that 'the weather was sunny *and* stormy', that 'two persons have never met *and* that they met' etc. A text written with this ongoing flow of events will generally have this effect on QG.

7 Weaknesses in the current QG implementation

The QG implementation for Swedish was developed independently of the English approach. QG for Swedish was originally an idea about putting schema parsing to optimal use. From the beginning, an idea was to produce *as many questions as possible* – whether useful or not, this 'total' approach was thought to bring forward some interesting aspects.

As mentioned, this strategy turned out to be similar to one of the approaches used for English; *overgenerate and rank* (Heilmann and Smith, 2009). Producing all questions or near-questions will generally lead to many irrelevant or less useful questions. In that approach, the act of total question generation by syntactic means is fairly termed overgeneration. The second step; ranking or selecting what may be useful questions is then the real challenge. It seems likely that linguistic theory of information structure can be helpful here. A *rhematic* portion of text (according to theory of information structure) is likely to produce a more relevant question, in some rather general sense. Identifying those portions automatically would clearly be an interesting task.

The situation described here, that there is currently much 'overgeneration', together with the previously stated fact; that many of the plausible and useful questions are not among the generated ones, gives the picture that current implementations suffer from two 'opposite' weaknesses: too many useless questions generated (weak 'precision') and too few of the truly relevant are generated (weak 'recall').

Acknowledgments

The project mentioned called *Adverbialkarakteristik för praktisk informationsextraktion i svensk text* was funded by Centre for Language Technology (CLT) and by the Department of Swedish at the University of Gothenburg.

References

Algotsson (2007), Gustav, Automatic Pronoun Resolution for Swedish, Master thesis, Department of Computer and Systems Sciences, KTH

Borin, Lars, Dannélls, Dana, Forsberg, Markus., Toporowska Gronostaj, Maria, & Kokkinakis, Dimitrios (2010). The Past Meets the Present in the Swedish FrameNet++. 14th EURALEX International Congress, (ss. 269-281).

Borin, Lars, Forsberg, Markus, & Lönngren, Lennart (2008). SALDO 1.0 (Svenskt associationslexikon version 2). Göteborg: Språkbanken, Göteborgs universitet.

Diderichsen, Paul (1946). Elementær Dansk Grammatik. Köpenhamn: Gyldendahl.

Ejerhed, Eva, Källgren, Gunnel, & Brodda, Benny (2006). Stockholm-Umeå corpus version 2.0. Institutionen för Lingvistik, Stockholms universitet, Institutionen för Lingvistik, Umeå universitet.

Fraurud, Kari (1988), Pronoun Resolution in Unrestricted Text, *Nordic Journal of Linguistics* 11.

Hassel, Martin (2000), Pronominal Resolution in Automatic Text Summarisation. Master thesis, Department of Computer and Systems Sciences (DSV), Stockholm University

Heilman, Michael, and Smith, Noah A. (2009) Ranking Automatically Generated Questions as a Shared Task. Proceedings of the AIED Workshop on Question Generation. Brighton, UK, 2009. 30-37.

Kann, Viggo, and Magnus Rosell (2005) Free Construction of a Free Swedish Dictionary of Synonyms. Proceedings of 15th Nordic Conference on Computational Linguistics – (NODALIDA 05). Joensuu, 2005.

Lefevre Marie, Jean-Daubias, Stéphanie and Guin, Nathalie (2009) Generation of Exercises within the PERLEA project, Proceedings of the AIED Workshop on Question Generation. Brighton, UK, 2009.

Lexin – Svenska ord. Norstedts Akademiska Förlag, (1998)

Mitkov, Ruslan (1998). Robust pronoun resolution with limited knowledge. In In Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, pages 869–875, Montreal, Canada.

Nilsson, Kristina (2010) Hybrid Methods for Coreference Resolution in Swedish. Ph.D thesis. Department of Linguistics, Stockholm University

Nivre, Joakim. Inductive Dependency Parsing (2006) Dordrecht: Springer, Text, speech, and language technology series, Volume 34.

Rus, Vasile, and Graesser, Arthur. The Question Generation Shared Task and Evaluation Challenge (2009) Workshop Report, Memphis, USA: The University of Memphis, 2009.

Viberg, Åke, Kerstin Lindmark, Ann Lindvall, and Ingmarie Mellenius (2002)" The Swedish WordNet Project." Proceedings of Euralex 2002. Köpenhamn, 407-412.

Wikipedia: www.wikipedia.org; Swedish version. http://sv.wikipedia.org.

Wilhelmsson, Kenneth (2010). Heuristisk analys med Diderichsens satsschema - Tillämpningar för svensk text (doktorsavhandling). Göteborgs universitet: Institutionen för filosofi, lingvistik och vetenskapsteori

Wilhelmsson, Kenneth (2011). Automatic Question Generation from Swedish Documents as a Tool for Information Extraction. Proceedings of the 18th Nordic Conference of Computational Linguistics. Riga.

Wilhelmsson, Kenneth (2012). Adverbialkarakteristik för praktisk informationsextraktion i svensk text - (project report) GU-ISS, Forskningsrapporter från Institutionen för svenska språket, University of Gothenburg, ISSN 1401-5919; nr 2012-03

Wyse, Brendan and Piwek, Paul (2009) Generating Questions from OpenLearn study units. Proceedings of the AIED Workshop on Question Generation. Brighton, UK.