Lark Trills for Language Drills: Text-to-speech technology for language learners

Elena Volodina
University of Gothenburg, Dpt of Swedish,
Swedish Language Bank (Språkbanken)
Box 200, 405 30, Gothenburg, Sweden
elena.volodina@svenska.gu.se

Dijana Pijetlovic
Nuance Communications Switzerland AG
Baslerstrasse 30
8048 Zürich, Switzerland
dijana.pijetlovic@nuance.com

Abstract

This paper reports on the development and the initial evaluation of a dictation&spelling prototype exercise for second language (L2) learners of Swedish based on text-to-speech (TTS) technology. Implemented on an already existing Intelligent Computer-Assisted Language Learning (ICALL) platform, the exercise has not only served as a test case for TTS in L2 environment, but has also shown a potential to train listening and orthographic skills, as well as has become a way of collecting learner-specific spelling errors into a database. Exercise generation re-uses well-annotated corpora, lexical resources, and text-to-speech technology with an accompanying talking head.

1 Introduction and background

ICALL – Intelligent Computer-Assisted Language Learning - is an intersection between Computer-Assisted Language Learning (CALL) and Natural Language Processing (NLP) where interests of the one side and technical possibilities of the other meet, e.g. automatic error detection and automatic essay scoring.

Multiple research projects worldwide explore the benefits of NLP in educational applications (Mitkov & Ha 2003; Monaghan & Bridgeman 2005; Heilman & Eskenazi, 2006; Antonsen 2012), some of them being exploited for real-life language teaching (Amaral and Meurers, 2011; Heift, 2003; Nagata, 2009), most of them though staying within academic research not reaching actual users (Nilsson & Borin, 2002; François & Fairon, 2012) or remaining limited by commercial usage (Attali & Burstein, 2006; Burstein et al., 2007).

In the past five decades the area of NLP has witnessed intensive development in Sweden. However, ICALL has remained rather on the periphery of NLP community interests. Among the directions in which ICALL research developed in Sweden, one can name supportive writing systems (Bigert et al., 2005; Östling et al., 2013); exercise generators (Bick 2001, 2005; Borin & Saxena, 2004; Volodina et al., 2014); tutoring systems (Wik 2004, 2011; Wik & Hjalmarsson, 2009).

As can be seen, the number of directions for Swedish ICALL projects is relatively small. Given the potential that NLP holds for CALL community, this fact is rather surprising, if not remarkable.

1.1 Pedagogical Framework

More than a decade ago Council of Europe has adopted a new framework for language learning, teaching and assessment, the *Common European Framework of Reference for Languages* (CEFR; COE, 2001). The CEFR guidelines describe language skills and competences at six proficiency levels (from beginner to proficient): A1, A2, B1, B2, C1, C2. Among those skills, *orthographic skills*, *listening comprehension*, *vocabulary range and control*, and *knowledge of lexical elements* are relevant in the context of the exercise described in the paper.

Orthographic control, as defined by the CEFR, is ranging from *"Can copy familiar words and short phrases ... used regularly"* at the beginner level (A1) to *"Writing is orthographically free of error"* at the mastery level (C2) (COE, 2001:118). The same applies to *listening comprehension* which ranges from *"I can recognise fami-*

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 107–117,
Denver, Colorado, June 4, 2015.

liar words and very basic phrases..." at A1 to "*I have no difficulty in understanding any kind of spoken language...*" at C1 (COE 2001:26-27). Criteria for *lexical competence* include *vocabulary range and control* and knowledge of *lexical elements* that stretch over the limits of one single word (2001:110-112).

The proposed *dictation&spelling* exercise is a possible way to improve the above-mentioned competences and skills. Learners first hear the item pronounced by a talking head, and afterwards spell it - item in this context being understood as either a single word, a phrase or a sentence. For teachers, it is rather time-consuming to engage in dictation in an attempt to help students improve their lexical, listening and orthographic skills. In this case, NLP can successfully replace a teacher in this drill-like exercise.

1.2 Use of TTS for L2 learning

TTS is being increasingly used in CALL systems for multiple tasks, such as for listening and dictation practice (Santiago-Oriola, 1999; Huang et al., 2005; Pellegrini et al., 2012; Coniam, 2013), for reading texts aloud (Lopes et al., 2010), and for pronunciation training (Wik, 2011; Wik & Hjalmarsson, 2009).

The Swedish TTS in CALL environment is represented by *Ville* and *Deal* (Wik, 2011; Wik & Hjalmarsson, 2009). *Ville* is a virtual language teacher that assists learners in training vocabulary and pronunciation. The system makes a selection of words that the student has to pronounce. The system analyses students' input and provides feedback on their pronunciation. The freestanding part of *Ville*, called *DEAL*, is a role-playing game for practicing conversational skills. While *Ville* provides exercises in the form of isolated speech segments, *DEAL* offers the possibility to practice them in conversations (Wik & Hjalmarsson, 2009).

Like *Ville*, the dictation&spelling exercise presented here uses TTS technology for training vocabulary. However, unlike *Ville*, the dictation&spelling exercise is (1) focused on spelling rather than pronunciation, and in this respect complements the functionality offered by Ville; (2) is web-based and does not need prior installation; and (3) is designed to address students at different CEFR proficiency levels.

1.3 Research questions

Two important research questions, raised in connection to this project, have influenced the design of the implemented exercise.

(1) Is TTS technology for Swedish mature enough for use in ICALL applications? To answer this question, we included evaluation and a follow-up questionnaire by the end of the project, where users could assess several parameters of the speech synthesizer and express an overall impression of the exercise (Section 3).

(2) What way should feedback on L2 misspellings be delivered? To have a better idea about what typical L2 spelling errors learners of Swedish make, we designed an error database that stores incorrect answers during the exercise. Based on the analysis of the inititally collected errors, we suggest a way to generate meaningful feedback to Swedish L2 learners (Section 4).

The rest of the paper is structured as follows: Section 2 describes the implementation details of the exercise and the database. Section 3 presents the results of the evaluation. Section 4 focuses on the first explorations of the SPEED (SPElling Error Database) and suggests a feedback generation flow. Section 5 concludes the paper and outlines future prospects.

2 Exercise design and implementation

2.1 Resources

A number of computational resources for Swedish have been used in the exercise, namely:

- Corpora available through *Korp*, Språkbanken's infrastrusture for maintaining and searching Swedish corpora (Borin et al., 2012b). All corpora in Korp are accessible via web services and contain linguistic annotation: lemmas, parts-of-speech, morphosyntactic information, dependency relations.
- Lexical resources available through *Karp*, Språkbanken's lexical infrastructure (Borin et al., 2012a*): Kelly word list*, a frequency-based word list of modern Swedish containing 8,500 most important words for language learners with associated CEFR proficiency levels (Volodina & Johansson Kokkinakis, 2012); and *Saldo morphology*, a morphology lexicon of Swedish containing all in-

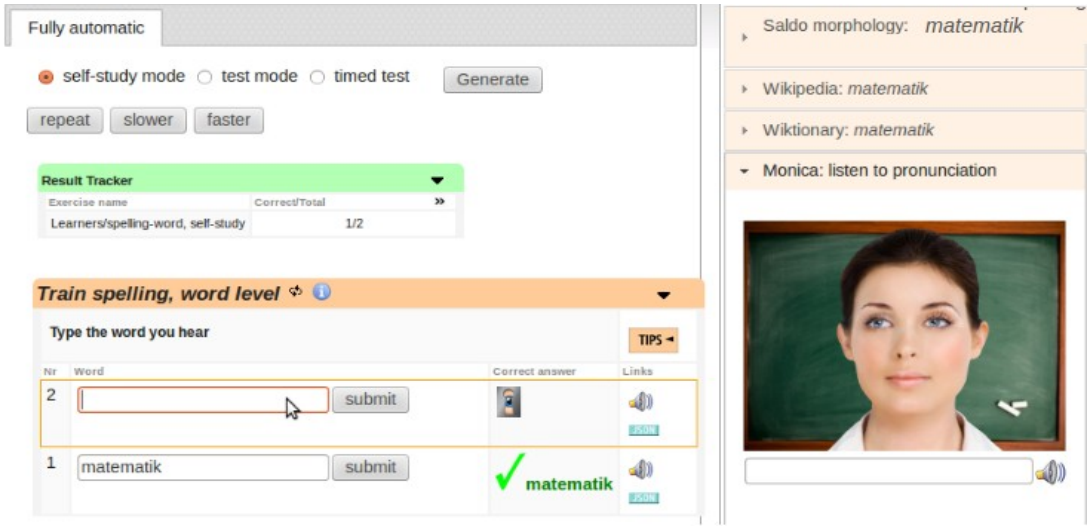

Figure 1. User interface for dictation&spelling exercise, version 2

flected forms for each lemgram (base form + part of speech pair) (Borin, Forsberg & Lönngren, 2013). Karp resources are also accessible through web services.

- *SitePal*'s TTS synthesizer module and a talking head, Monica, who is addressed that way in the paper
- *Lärka*, an ICALL platform for Swedish where the exercise is deployed (Volodina et al., 2014). *Lärka* is an ICALL platform for studying Swedish (in broad sense). It targets two major user groups – students of Linguistics, and L2 learners. The exercise repertoire comprises (1) exercises for training parts-of-speech, syntactic relations and semantic roles for students of Linguistics; and (2) exercises for training word knowledge and inflectional paradigms for L2 learners (Volodina et al., 2014). Features common to all exercises include corpora and lexical resources, training modes, access to reference materials (Figure 1).

2.2 Linguistic levels

According to Nation (2001), aspects of word knowledge include: (1) *Form*: spoken (recognition in speech, pronunciation); written (recognition in texts, spelling); word parts (inflection, derivation, word-building); (2) *Meaning*: form and meaning; concept and references; associations; (3) *Use*: grammatical functions; collocations; constraints on use (register/frequency/etc.)

While the two previously available exercises in Lärka – for training vocabulary knowledge and inflectional paradigms – focus on some aspects of meaning, use and form, the newly added dictation&spelling exercise has extended the spectrum of trained word knowledge aspects to cover other dimensions of form-aspect, namely spoken and written forms, and therefore the exercise has become a natural and welcome addition to the exercise arsenal offered by Lärka.

The exercise is offered at four linguistic levels, each targeting different aspects of word knowledge. The *word level* focuses on pronunciation and spelling of the base form of a word. A target word of an indicated CEFR level is randomly selected from the Kelly list or from a *user-defined list*, an option provided by Lärka where learners can type words they need to train. The target item is then sent to the TTS module to obtain its pronunciation. TTS pronounces the word, while the user needs to spell it (Figure 2).

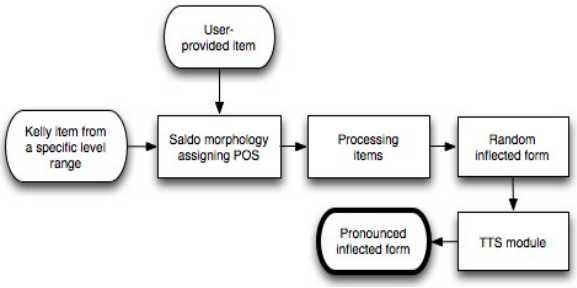

Figure 2. NLP pipeline for word levels. At the non-inflected word level *Saldo morphology* is excluded from the pipeline

The *inflected word level* (Figure 2) also focuses on a single word, however, the learner is made aware of its inflectional patterns, in addition to

pronunciation and spelling (learners have to spell the inflected form they hear). Analogous to the word level, the target word is randomly selected from the Kelly list or the user-defined list. Before the item is sent to the TTS module, its different inflected forms are checked in Saldo-morphology, whereas some of the forms, e.g. possessives, are excluded as inappropriate for training through dictation. One random form is used for training.

The *phrase level* offers the target word in some typical context, which alongside demonstrating the item's collocational and distributional patterns, also requires the user to identify (via listening) the number of separate words constituting the phrase. While the implementation for the word and the inflected word levels was straightforward, the implementation for the phrase level needed some workaround to achieve the best phrase accuracy. In this exercise version only noun and verb phrases have been taken into consideration.

For retrieval of the typical phrase patterns, word pictures associated with the target item are retrieved from Korp. A fragment of a word picture for the noun *ord* [word], is shown in Figure 3. The columns on top of Figure 4 provide the most distinguished collocation patterns (prepositions, premodifiers, post-modifiers), underneath followed by the actual lemmas alongside with the number of hits in the corpora. Most typical prepositions used with the noun *ord* are (in translation): *with*, *without*, *behind*, *against*, *beyond*. Most typical premodifiers are: *free*, *ugly*, *beautiful*, *hard*, *empty*.

Preposition		Pre-Modifier		ord	Post-Modifier	
1.	med	9191	1.	fri	2353	
2.	utan	494	2.	ful	314	
3.	bakom	235	3.	vacker	376	
4.	mot	484	4.	hård	395	
5.	bortom	50	5.	tom	208	

Figure 3. Word picture for the noun *ord* [word] in Korp

The number accompanying each of the collocates reflects the number of hits in the corpus. For example, *fri 2353* on top of the second column means that the phrase starting with a pre-modifier *fri* [free] has a pattern *fri + ord* and has been used 2353 times in the corpora where we performed our search. To extract the actual phrase containing *fri ord*, another request is forwarded to Korp where the actual corpus hits are returned (the 2353 of them). Then, any of the sentences can be used for extracting the actual phrase preserving inflections and words that come in-between, e.g. *fria tankar och ord* [free thoughts and words]. After some experiments, we have set the limit at max 6 tokens per phrase.

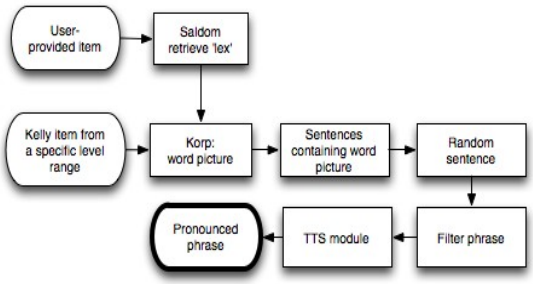

Figure 4. NLP pipeline for the phrase level

The final flow of the exercise generation at the phrase level is shown in Figure 4: A random item from the Kelly list is forwarded to the Korp's word picture web-service, one of the top frequent patterns is selected and the actual KWIC hits are consulted. After the phrase has been selected and adjusted, it is sent to the TTS module for pronunciation. In case of a user-defined word list, the randomly selected item is first sent to Saldo-morphology to check possible word classes associated with the item, one is selected and sent further to Korp for extracting a word picture.

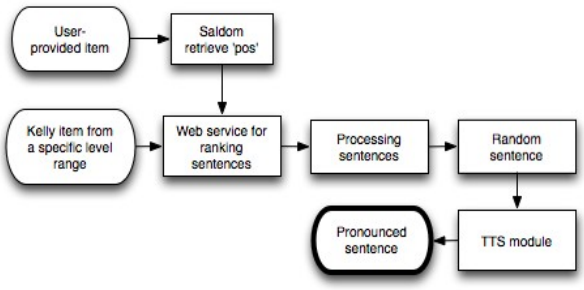

Figure 5. NLP pipeline for the sentence level.

The *sentence level* offers the target item in a sentence context, which sets further demands on listening comprehension and awareness of structures that the target word can be used in. The sentence level is the most challenging for the users, since sentences are usually long and it is difficult to remember all information. Programmatically, though, it is less challenging than the phrase level,

unless you want to ensure that learners understand the sentences they get for training. We have used algorithms developed by Pílan et al. (2014) for automatic retrieval of sentences understandable by learners at B1/B2 proficiency levels. Before the sentence is sent to the TTS module, some additional filtering is performed blacklisting sentences of inappropriate length or containing inappropriate tokens (e.g. dates with slashes *30/11/2013*), see Figure 5.

Finally the *performance-based variant* of the exercise offers a path from the word to the sentence level, allowing the user to go over from one level to another according to his/her performance. If 10 items have been spelled correctly, a new level is offered.

2.3 Error database

All user answers are logged in SPEED, the SPElling Error Database, which has been deployed on Karp's backend. SPEED keeps track of:
(1) the session which consists of the date and time when the user has started the exercise. All errors made by that particular user have the same session ID. This way we have a chance to identify some user-specific behaviour and error patterns.
(2) the correct item, its parts-of-speech, the misspelling and the time when the misspelling is added. If an entry for the correct item has already been created, a new misspelling is added to the list of misspellings. Otherwise, a new entry is created.

Since no login information is required to use Lärka (which is a choice made at the departmental level), we cannot log information about learners' first language (L1) background.

3 User evaluation

We have used an off-the-shelf TTS solution offered by SitePal (www.sitepal.com), which offers an optimal combination of voice quality, availability of talking heads, user-friendliness and a reasonable subscription price.

A critical question for this project has been whether the TTS quality of the SitePal's synthesizer is mature enough for use in an ICALL application. A quality of a TTS synthesizer is generally judged by its *naturalness* (i.e. similarity to the human voice), *understandability* (comprehensibility of the message and intelligibility of individual sounds), and *accuracy* (Handley & Hamel, 2005). This is especially significant when applied to L2 context where TTS is used both for setting an example of correct pronunciation and for testing listening comprehension. Besides the three criteria above, the criteria of *language learning potential* and *opportunity to focus on linguistic form* are critical in CALL environment (Chapelle, 2001a, 2001b). If the technology doesn't live up to the demands, this type of exercise should be excluded in want of better technological solutions.

A few studies have evaluated TTS in CALL applications. A study by Pellegrini et al. (2012) compared TTS-produced versus human pre-recorded speech in L2 dictation exercises (sentence level). They found that L2 learners make more mistakes when human voice is heard, thus establishing that (at A2 level) TTS speech is more understandable by L2 learners of Portuguese, most probably due to the speed difference, TTS version being 15% slower. Handley (2009) evaluated TTS modules in four CALL applications using criteria of comprehensibility, acceptability, and appropriateness, and found TTS technology mature enough for use in L2 applications, emphasizing that expressiveness was insufficient. Handley & Hamel (2005) discuss a benchmark for evaluation of speech synthesis for CALL applications. Evaluation focus should differ depending on uses of TTS, since different features play roles for various learning scenarios. They explored appropriateness, acceptability and comprehensibility as potential criteria for the three TTS tasks: reading texts, pronunciation training and dialogue partner, and found that the same TTS module has been evaluated differently depending upon the task it was used for.

3.1 Participants and setting

The evaluation of the exercise was carried out with 10 participants who represented three user groups: beginner levels A1/A2, intermediate levels B1/B2 and advanced levels C1/C2 with 3 participants in each. A native speaker is categorized separately as his/her language knowledge exceeds the CEFR-defined proficiency levels.

The participants have been asked to fill an evaluation form following the experience of working with the exercise. During the exercise, each of the participants spelled at least 40 items: 10 at each of the four linguistic levels. They were also encoura-

ged to test performance-based level. All along the misspellings have been saved to the error database.

3.2 Questionnaire

The purpose of the evaluation has been primarily to evaluate the text-to-speech module and to assess the usefulness of the exercise, based on L2 learner preferences. We have used criteria suggested by Handley and Hamel (2005) and Chapelle (2001a, 2001b) as the basis for our evaluation adding some more questions.

The questionnaire contained 15 questions, of which five were focused on the TTS quality (questions #3-7, Table 1), six - on the exercise and its effectiveness (#8-14), one explicitly asking for the type of feedback learners expect from the program (#15), and the rest were devoted to the user-friendliness of the GUI (#1-2)[1].

All questions (except #15) were evaluated according to a 5-grade scale, where 1 corresponded to *very good* and 5 to *very poor*. Additionally, the evaluators had the possibility to add comments for every question and at the end of the questionnaire.

	A1/A2	B1/B2	C1/C2	Native	all levels
Q1 - Instructions	1	2	2.67	1	**1.8**
Q2 - GUI	1.33	2	2	1	**1.7**
Q3 - Comprehensibility	2	3	2.67	2	**2.5**
Q4 - Intelligibility	2	2.33	3.33	1	**2.4**
Q5 - Avatar	4	3.67	3.67	4	**3.8**
Q6 - Naturalness of speech	1	2.33	2.67	2	**2**
Q7 - Pronunciation	1.67	1.33	2	2	**1.7**
Q8 - Difficulty	2	1.67	3.67	1	**2.3**
Q9 - Word level	1.33	1.67	2	1	**1.8**
Q10 - Phrase level	2.67	2.33	2.67	1	**2.4**
Q11 - Sentence level	3.67	2.67	4	1	**3.2**
Q12 - Sentence length	4	2.33	4	2	**3.3**
Q13 - Speed	2.33	2.33	1.67	2	**2.1**
Q14 - Effectiveness	1.67	2	1.67	2	**1.8**
overall results	2.19	2.26	2.76	1.64	**2.34**

Table 1. Results by question & proficiency level, on the scale 1=very good … 5=very poor

3.3 Evaluation results and discussion

According to the evaluation results (Table 1), the talking head (#5) appears to be the least effective element in the spelling exercises. The unsatisfying results for the speaking head are based hypothetically on the missing facial expression and on its location within the spelling game. Compared to the virtual language teacher Ville, which was developed specifically for educational purposes, the SitePal's talking head seems to have a rather entertaining function. The expressive lip movement that is characteristic of Ville, is clearly missing from Monica.

The *pronunciation* generated by the TTS module (#7), however, is regarded as good despite comments on some smaller pronunciation errors. The user interface (#2) and the quality of pronunciation (#7) are the most satisfactory features. The *naturalness* of speech (#6) is perceived differently among the participants. While the beginner group finds the TTS-produced speech natural and human-like, the advanced group perceives it as least natural. This result is not very surprising as the beginner group is not familiar with the language and therefore is not able to critically judge the naturalness of speech. The native speaker is in general very positive towards the TTS system.

Table 1 shows clearly that the word/inflected word levels (#9) are the most appropriate units for training spelling followed by the phrase level (#10). Phrases need to be adapted to the respective proficiency level in order to achieve the best learning effect. The sentence level (#11) is assessed as the least appropriate one, as the length and the speed rate have been perceived unsuitable for training spelling and listening. The results demonstrate that the *learning potential* at the word and phrase levels is higher than at the sentence level, as perceived by L2 learners.

The results by proficiency level (Table 1) show that there is an obvious tendency to become more critical as the level grows. The proficiency group C1/C2 is the least satisfied one, while the native speaker is the most positive. The reason for that might be that language learners from higher proficiency levels are more critical as their knowledge of the language is better and therefore TTS mistakes are more noticeable, while TTS mistakes might not be that obvious to the learners with lower levels of proficiency. The vocabulary chosen for training spelling and listening at lower levels may also be easier for the TTS system to pronounce. The native speaker shows in general a very positive attitude towards the spelling game as (s)he might be more aware of the difficulty of the language and is therefore more 'forgiving'. Another reason might be that the native speaker does not assess the spelling exercise from the learner's point of view and might therefore be less critical.

When it comes to the word level (#9), with the increase of learners' proficiency dissatisfaction also increases (Figure 6). The reason for that might be

[1]Full questionnaire form can be downloaded from http://spraakbanken.gu.se/eng/larka/tts

that the words in the Kelly-list are too advanced for the intermediate level. Some of the advanced participants find the word level not challenging enough as the target words are displayed quickly before they are pronounced. This kind of spelling tip needs to be adapted to the proficiency level.

As for the appropriateness of phrases (#10), the intermediate group is more positive to them than the beginner and advanced groups. The reason for that may lie in the implementation approach. Since words within a phrase do not all belong to the same difficulty level, phrases extracted for the beginner level might be too advanced.

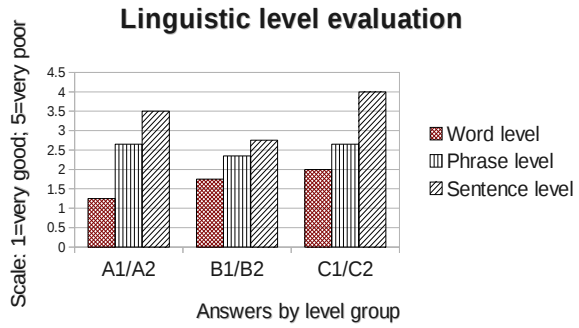

Figure 6. Results by proficiency levels & linguistic levels, on the scale 1=very good … 5=very poor

The sentence level (#11) is in general the most challenging linguistic level for training spelling and listening (Figure 6). The C1/C2 group finds the sentence level in general inappropriate for spelling exercises. The obtained sentences were difficult to follow not only for beginners but even for advanced Swedish L2 learners.

Especially interesting are the comments provided for the question on feedback (#15). The feedback that the participants would like to see in this exercise is grouped into several suggestions:

- A hint on the word form for the inflected forms
- Tips regarding grapheme-phoneme mappings
- English translation of the spelled items
- Possibility to see the correct answer by choice
- Possibility to notify the pronunciation mistakes made by the TTS module
- Detailed feedback on the wrong answers
- Run-time marking of spelling errors

4 Feedback on L2 misspellings

In the pedagogical and psychological studies on feedback one can encounter an extensive amount of different terms, e.g. achievement feedback, assessment feedback (Higgins et al., 2002), formative and summative feedback, feedback on performance (Hyland, 2001), etc. The common ground for all types of feedback is that the student performance (actual level) is compared with the expected performance (reference level) and some information is provided to the learners that should help them develop the target skills further in order to alter the gap between the actual and the reference levels (Ramaprasad, 1983).

Obviously, just stating the presence of the gap ("incorrect") is not sufficient. Feedback becomes useful when ways to improve or change the situation are outlined. To do that, we need to understand the nature of a spelling mistake, and to point learners to the specific aspects of the target language orthography, the phoneme-grapheme mappings in L2; or even to the relation between L1 and the target L2 spelling and pronunciation systems. A lot of studies argue that it is vital to know a learner's L1 for successful error analysis (Tingbjörn & Andersson, 1981; Abrahamsson, 2004; Koppel et al., 2005; Nicolai et al., 2013). Unfortunately, the ICALL platform that is used as a basis for the exercise does not offer any login facility, which makes it impossible to log learners' L1, at least at present. Given that constraint we had to make the best out of the situation. We started looking for a taxonomy of most typical L2 spelling errors which students should be addressed to, independent of their L1.

While there are several available error corpora for other languages (Granger 2003; Tenfjord, Meurer & Hofland 2006), we are aware of only one error database for Swedish, an Error Corpora Database (ECD), which is a collection of different types of errors, among others spelling ones. They have been collected from Swedish newspapers, and analyzed to create an error typology used for developing proof-reading tools for professional writers of Swedish (Wedbjer Rambell, 1999a; Wedbjer Rambell, 1999b). Being a good source for comparison, ECD, however, cannot be applied as it is to the context of Swedish L2 learning. Antonsen (2012) points out that L2 errors differ in nature and type from L1 errors. Rimrott & Heift (2005) found that generic spell-checkers fail to identify L2 errors and therefore special care should be taken to study specific L2 errors. We faced therefore the necessity of collecting a special database of Swedish L2 errors as the first step on the way to useful feedback.

Collecting errors into a database from corpora is a time-consuming process which we could not afford. We have opted for another alternative, inspired by Rodrigues & Rytting (2012), where errors are collected into a database while learners do exercises. Advantages of collecting a corpus by applying this method are numerous: participants are quickly attracted, while cost, time and effort of collecting a corpus are reduced.

While the feedback has not been implemented at this stage, the database has been populated with misspellings and has given us the first insights into the nature of typical L2 errors and prompted some ideas on useful feedback.

4.1 Error log analysis

The initial analysis of the error logs focused on word-level errors, which have been categorized into several error types. The same spelling errors could often be classified into more than one category; e.g. a real word error can be at the same time a performance- or a competence-based error.

There are two major groups of errors, competence-based (55%) and performance-based (17%) ones, that are described here. The rest of the errors (28%) are connected to a group of errors occurring in sentences or phrases where e.g. wrong segmentation or total absence of one or several words are the cause of the error. These errors have been left out of the present analysis.

While performance-based errors are accidental and are easily corrected with a hint to the learner, competence-based errors depend on the lack of or insecure knowledge and need to be explained. Learners need to be made aware of the mappings between orthography and pronunciation in the target language. L1 speakers usually make performance-based errors while in L2 learners' writing competence-based errors dominate (Rimrott & Heift, 2005).

Competence-based errors (55%) occur as a result of not knowing a word's spelling or confusing words. L2 spelling errors are mostly competence-based. This type of errors mainly occurs when the orthographic rules of L2 differ from the ones of L1 or when a language contains special characters or sounds that do not exist in L1. The competence-based errors from the evaluation fall into the four categories described below.

Spelling errors based on *consonant doubling* (28%) belong to one of the most common errors, where either a single consonant is written instead of a double (e.g. *sto**p**a* instead of *sto**pp**a* [thrust]) or a double consonant instead of a single (e.g. *ri**mm**ligen* instead of *ri**m**ligen* [reasonably]).

Spelling words that contain *characters with accents/diacritics* (ä, å, ö) present challenge for Swedish L2 learners, due to the difficulty to distinguish between special characters and the orthographically or phonetically similar vowels (23%). For example, the sound of the letter *å* was frequently mistaken for the vowel *o*.

Phonetic errors (25%) appear when parts of words are spelled as they are heard. The most frequent phonetic error in our logs is caused by confusing voiced and voiceless consonants.

Another cause of a typical Swedish L2 misspelling are *consonant clusters* that follow special rules for grapheme-to-phoneme mapping (20%). The letter combination *rl*, for example, is pronounced [l]. The drop of "r"-sound applies also for the combinations *rs*, *rd* and *rt*. Some other problematic clusters are *tj*, *ch*, *hj*, *sk*.

Performance-based errors (17%), the so called 'typos', are caused by addition, deletion, insertion or replacement of one or several letters in a word, often a result of hitting a wrong key or two keys at the same time on the keyboard. Performance-based errors are not always obvious, for example, the misspelling *s**j**ön* (corr. *s**k**ön* [beautiful]) could have been created by confusing the keys *j*' and *k* on the keyboard but could also be categorized as a competence-based phonetic error. The spelling error *förb**'**ttra* (corr. *forb**ä**ttra* [improve]) clearly belongs to the performance-based category.

Spelling mistakes can also result in *real words* (14%) either by chance or because a word is misheard and therefore mistaken for another word. For example, the word *liknan**d**e* [similar] could either be mistaken for *liknade* [resembled] or the letter *n* was omitted accidentally, while the word *livs**stil*** [life style] is more likely to be misheard as *livs**tid*** [life time]. Overall, the results show that non-word errors (86%) are significantly more likely to occur than real-word errors (14%).

The first analysis of the error logs inspired us to propose a feedback generation tree (Figure 7). The analysis of a larger database might lead to a more specific decision tree. The tree is build up from the easiest spelling errors to identify to the

more difficult ones. All along the error analysis, relevant feedback is provided. If multiple changes are necessary, they are advised step-wise. In case the spelling error cannot be classified, the correct item is shortly exposed.

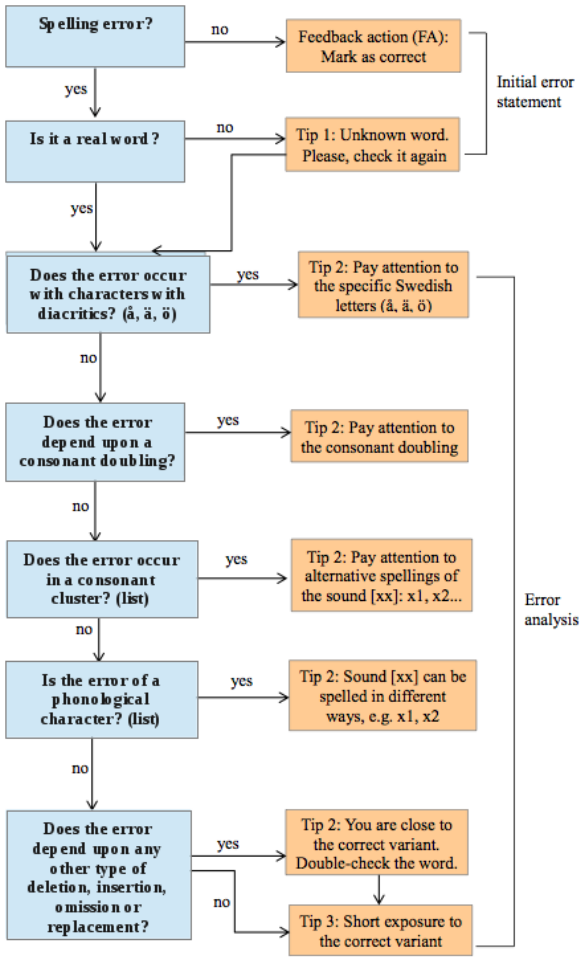

Figure 7. Feedback generation tree

The proposed feedback generation flow would allow to offer the kind of information that can help learners to fill the gap between the reference and the actual level of the assessed spelling error.

5 Concluding remarks

The goals of this project have been, firstly, the implementation of a Swedish dictation&spelling exercise that can provide L2 learners with a tool for training spelling and listening at different linguistic levels; and secondly, the evaluation of the newly implemented module regarding its effectiveness and usefulness. The main focus of the evaluation, in its turn, was to find out whether the TTS technology is mature enough for the use in L2 context and to suggest a way to provide useful feedback on L2 specific misspellings.

The state of TTS development looks very promising for integration of the current TTS synthesizer for Swedish L2 learning. Some improvements might be in place on the Lärka side, especially regarding the placement of the talking head on the screen and adjustment of the pronunciation speed to the level of the learner. However, the naturalness and understandability of the SitePal's TTS module hold a very good level.

The issue of homophones should be solved at word levels, either by counting alternative spellings as correct ones (e.g. flour vs flower) or by offering learners an additional possibility to hear the item in a context of a phrase or a sentence. The latter should help distinguish errors that arise due to learners' inability to recognize the word pronounced out of context versus their not knowing how to spell the word.

Besides, a broader spectrum of lexical resources and detailed feedback are necessary. The taxonomy of spelling errors shows that generating feedback for easily identifiable spelling errors is straightforward while more work is necessary to understand the nature of other types of errors. More detailed evaluations with larger number of participants, and repeated analysis of more extensive error logs are necessary to refine the feedback generation tree. Other suggestions on feedback proposed by evaluation participants will be considered for implementation.

The vocabulary for the word level needs to be expanded with larger lexical resources and domain specific vocabulary lists. The generation pace of phrases has to be accelerated, and the phrase level needs to be adapted to the proficiency level. Since the sentence level is regarded as the least effective one, most improvements are due on this level. The sentence length as well as the speech rate need to be adapted to the proficiency level.

In order to assess the spelling exercises from the pedagogical point of view, an in-class evaluation with teachers needs to be carried out once a new version is in place.

References

Niclas Abrahamsson. 2004. Fonologiska aspekter på andraspråksinlärning och svenska som andraspråk. In: *Hyltenstam Kenneth & Lindberg Inger (eds.) Svenska som andraspråk: i forskning, undervisning och samhälle*. Lund: Studentlitteratur.

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

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.

Yigal Attali & Jill Burstein. 2006. Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment 4 (3)*.

Eckhard Bick. 2001. The VISL System: Research and Applicative Aspects of IT-based learning. *Proceedings of NoDaLiDa*. Uppsala, Sweden.

Eckhard Bick. 2005. Grammar for Fun: IT-based Grammar Learning with VISL. In: *Henriksen, Peter Juel (ed.), CALL for the Nordic Languages*. p.49-64. Copenhagen: Samfundslitteratur (Copenhagen Studies in Language).

Johnny Bigert, Viggo Kann, Ola Knutsson & Jonas Sjöbergh. 2005. Grammar Checking for Swedish Second Language Learners. *Chapter in CALL for the Nordic Languages p. 33-47*. Copenhagen Studies in Language 30, Copenhagen Business School. Samfundslitteratur.

Lars Borin, Markus Forsberg & Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 1-21.

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

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

Lars Borin & Anju Saxena. 2004. Grammar, Incorporated. In *Peter Juel Henrichsen (ed.), CALL for the Nordic Languages*. Copenhagen Studies in Language 30. p.125-145. Copenhagen: Samfundslitteratur.

Jill Burstein,Jane Shore, John Sabatini, Y Lee, Matthew Ventura. 2007. Developing a reading support tool for English language learners. *Demo Proceedings of NAACL-HLT.*

Carol Chapelle. 2001a. Innovative language learning: Achieving the vision. *ReCALL, 23*(10), 3-14

Carol Chapelle. 2001b. *Computer applications in second language acquisition: Foundations for teaching testing and research.* Cambridge, England: Cambridge University Press.

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

David Coniam. 2013. Computerized dictation for assessing listening proficiency. *CALICO Journal*, Vol.13., Nr 2&3.

Thomas François & Cedrik Fairon. 2012. An "AI readability" formula for French as a foreign language In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, Jeju, 466-477.

Sylviane Granger. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20(3), p-p 465-480.

Zöe Handley. 2009. Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication* , vol. 51, no. 10, pp. 906–919, 2009.

Zöe Handley & Marie-Josée Hamel. 2005. Establishing a methodology for benchmarking speech synthesis for Computer-Assisted Language Learning (CALL). *Language Learning & Technology,* Vol. 9, No. 3, pp. 99-119

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

Shang-Ming Huang, Chao-Lin Liu, and Zhao-Ming Gao. 2005. Computer-assisted item generation for listening cloze tests and dictation practice in English. in *Proc. of ICWL*, pp. 197–208.

Michael Heilman, & Maxine Eskenazi. (2006). Language Learning: Challenges for Intelligent Tutoring Systems. *Workshop on Ill-defined Domains in Intelligent Tutoring*, Taiwan

Richard Higgins, Peter Hartley & Alan Skelton. 2002. "The Conscientious Consumer: reconsidering the role of assessment feedback in student learning". *Studies in Higher Education*, Vol.27, No.1, p.53-64

Fiona Hyland. 2001. Providing Effective Support: investigating feedback to distance language learners. *Open Learning*, Vol.16. No.3, p.233-247.

Moshe Koppel, Jonathan Schler, & Kfir Zigdon. 2005. Determining an authors' native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD international conference*, 624–628.

José Lopes, Isabel Trancoso, Rui Correia, Thomas Pellegrini, Hugo Meinedo, Nuno Mamede, & Maxine Eskenazi. 2010. Multimedia Learning Materials. In *Proc. IEEE Workshop on Spoken Language Technology SLT, Berkeley*, pp. 109–114.

Ruslan Mitkov & Le An Ha. 2003. Computer-Aided Generation of Multiple-Choice Tests. *Proceedings of the HLT-NAACL 2003* Workshop on Building Educational Applications Using Natural Language Processing, 17-22.

William Monaghan & Brent Bridgeman. 2005. E-Rater as a Quality Control on Human Scores. *ETS R&D Connections*: Princeton, NJ: ETS.

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

Paul Nation. 2001. *Learning Vocabulary in Another Language*, Cambridge University Press, p.477

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, LeiYao, Grzegorz Kondrak. 2013. Cognate and Misspelling Features for Natural Language Identification. In *Proceedings of NAACL-BEA8*.

Kristina Nilsson & 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*, Third International Conference on Language Resources and Evaluation p.411-418. Las Palmas: ELRA.

Thomas Pellegrini, Ângela Costa, Isabel Trancoso. 2012. Less errors with TTS? A dictation experiment with foreign language learners. *Thirteenth Annual Conference of the International Speech Communication Association*

Ildikó Pilán, Elena Volodina and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the 9th workshop on Building Educational Applications Using NLP*, ACL 2014.

Arkalgud Ramaprasad. 1983. On the definition of feedback. *Behavioural Science*, Vol.28, p.4-13.

Anne Rimrott & Trude Heift. 2005. Language Learners and Generic Spell Checkers in CALL. *CALICO journal*, Vol.23, No.1.

Paul Rodrigues & C. Anton Rytting. 2012. Typing Race Games as a Method to Create Spelling Error Corpora. *LREC 2012*.

C. Santiago-Oriola. 1999. Vocal synthesis in a computerized dictation exercise. In *Proc. of Eurospeech, 1999*.

Gunnar Tingbjörn, Anders-Börje Andersson. 1981. *Invandrarbarnen och tvåspråkigheten : rapport från ett forskningsprojekt om hur invandrarbarn med olika förstaspråk lär sig svenska.* Liber: Skolöverstyrelsen.

Kari Tenfjord, Paul Meurer & Knut Hofland. 2006. The ASK corpus - A Language Learner Corpus of Norwegian as a Second Language. *Proceedings of LREC 2006.*

Elena Volodina & Sofie Johansson Kokkinakis. 2012. Introducing Swedish Kelly-list, a new lexical e-resource for Swedish. *LREC 2012*, Turkey.

Elena Volodina, Ildikó Pilán, Lars Borin & Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. *LREC 2014*, Iceland.

Olga Wedbjer Rambell. 1999a. Error Typology for Automatic Proof-reading. *Reports from the SCARRIE project*, Ed. Anna Sågvall Hein.

Olga Wedbjer Rambell. 1999b. An Error Database of Swedish. *Reports from the SCARRIE project*, Ed. Anna Sågvall Hein.

Preben Wik. 2004. Designing a Virtual Language Tutor. In *Proc. of The XVIIth Swedish Phonetics Conference, Fonetik 2004*. p. 136-139. Stockholm University, Sweden.

Preben Wik. 2011. *The Virtual Language Teacher: Models and applications for language learning using embodied conversational agents.* PhD Thesis. KTH Royal Institute of Technology

Preben Wik & Anna Hjalmarsson. 2009. Embodied conversational agents in computer assisted language learning. *Speech communication* 51 (10), 1024-1037

Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich & Erik Höglin. 2013. Automated Essay Scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia, June 13 2013. Association for Computational Linguistics.