Measuring Encoding Efficiency in Swedish and English Language Learner Speech Production

Gintarė Grigonytė[1], *Gerold Schneider*[2]

[1]Department of Linguistics, Stockholm University, Sweden
[2]English Department, University of Zurich, Switzerland
gintare@ling.su.se, gschneid@es.uzh.ch

Abstract

We use n-gram language models to investigate how far language approximates an optimal code for human communication in terms of Information Theory [1], and what differences there are between Learner proficiency levels. Although the language of lower level learners is simpler, it is less optimal in terms of information theory, and as a consequence more difficult to process.

Index Terms: L2, spoken corpora, proficiency levels, surprisal

1. Introduction

Natural Language is subject to two opposing forces, both for the process of production and understanding: on the one hand the need for efficiency (economy), on the other hand the need for expressivity. Production and understanding is easiest (most economic) if a small vocabulary and highly routinised forms are used, which leads to highly redundant utterances in terms of Shannon's noisy channel model [1], but then expressivity is severely restricted. Too expressive language, on the other hand, leads to an increased processing burden, up to the breakdown of the communication channel due to the inevitable presence of noise, according to Shannon's model. [2] has shown that predicting patterns in situations is crucial for understanding.

Routinised forms, which manifest themselves as recurrent patterns, tend to dominate language use. [3] estimates that up to 80% of the words in a corpus are part of a recurring sequence. Language is full of partly prefabricated structures, phraseological expressions, and frequent word combinations. [4] sees frequent multi-word expressions as not very salient psychologically and to be used as the preferred building blocks in speech and writing. [5] shows that also in first language acquisition, formulaic language precedes the use of creative constructions.

If this were to hold equally for second language (L2) acquisition, we could also expect learner language to be more redundant than native language. But second language learners often resort to transfer from their native language, thus showing highly creative language. As a consequence, learner language is simpler, but at the same time more difficult to process for native speakers [6].

Some of the creative constructions used by language learners lead to subtle failures, they use routinized language, idioms and lexcial preferences in a less native-like fashion [7]. In the following examples from the ASU corpus [8], (1) shows a morphological error and a lexical preference error and (2) shows word order, verb construction and tense errors.

(1a) Ori: Nu ska jag beskrivar lite om svenskarna.
(1b) En: Now I will describe a little bit about the Swedes.
(1c) Cor: Nu ska jag berätta lite om svenskarna.
(2a) Ori: Kan jag vill stoppa striderna.
(2b) En: Can I want to stop the fights.
(2c) Cor: Kunde jag, skulle jag stoppa striderna.

From the perspective of analysing learner production, Shannon's tug-of-war between routinisation/economy and expressivity results in detectable patterns of language use. We aim to measure encoding efficiency with word level n-grams models. We employ two models: first, word-level surprisal and second, model fit to a POS tagger, in order to study encoding efficiency, which we see as a skill acquired through learning process and thus comparable across proficiency levels in spoken corpora. If [7] and [1] are right, then we should see measurable differences correlated to language speaker levels.

The overarching goal of our proposed n-gram measures is to provide a basis for grammatical error detection that goes beyond spellcheckers, to test features for automated learner assessment, and to facilitate new methods for investigating the idiom principle in learner language.

2. Related work

Related research falls into two major categories, namely formulaicity in written and spoken production of L2 learners, and the assessment of L2 language learning.

As for the former, numerous studies investigating language usage of L2 learners on the basis of substantial corpora have been published since the 1990s (e.g. [3] (for Swedish), [9]). In [10] and [11] quantitative corpus studies showed that learners tend to overuse and underuse adverbial connectors in terms of frequencies. In [9] the same pattern of misuse is observed for lexical phrases, collocations and active/passive verb constructions. Additionally, the interlanguage of L2 learners often include syntactically incorrect sentences, but equally often learners fail to use the subtle idiom principle factors as successfully as native speakers do, as [7] and [12] observed. These subtle factors involve idiomatic constraints and collocations, lexical preferences, choice of determiner, tense, avoidance of ambiguity, and frequency effects [13].

Patterns of formulaic language use proved not only to be helpful at discriminating between L1 and L2 production, but also be useful in estimating the level of proficiency ([14], [15], [16]). Several studies have found that high-scoring essays have significantly and consistently more formulaic language sequences than low-scoring ones. Applications of some of these patterns of language use have been reported in research fields that relate to assessment in L2: in CALL (Computer assisted language learning) systems ([17]); and in automatic essay grading methods, which currently are relying on text categorization techniques ([18], [19]). For the assessment of L2 production, [20] showed that collocation use is a testable phenomenon in discriminating among L2 proficiency levels. Their proposed collocation test was found to be reliably correlating to TOEFL scores and ESL teachers' proficiency rankings.

Automated assessment is a relatively new field of study that is highly cross-disciplinary and combines linguistics, cognitive psychology, and methods from natural language processing and machine learning. The recent work on automated assessment mainly covers English learners' written text and it aims at assigning grades based on textual features that try to balance performance errors and language competency. Most of the work in this area falls into a category of a supervised text classification ([17], [21], [22], [19]).

The above approaches are mainly based on a descriptive linguistic approach, and use a fixed set of collocations or formulaic utterances. We believe that a measure which expresses the full cline from completely fixed to fully creative expressions is a useful addition. We suggest to use surprisal [23] as such a gradient measure in model 1.

In model 2, we measure if L2 fits the language model less well. This approach follows the tradition of statistical anomaly and outlier detection [24].

3. Methods for encoding efficiency

We employ a one word level (unigram) measure of lexical richness, and two n-gram models, corresponding to different processing levels, namely surprisal [23], and a part-of-speech tagger [25].

3.1. Lexical Diversity: TTR

Lexical Diversity can be observed at the unigram level of learner production, and is traditionally assumed to be an important factor for assessing learner level. A frequently used measure is Type-Token Ratio (TTR). As TTR is size-dependent [26], documents of equal length need to be used. We have normalised all subparts to the length of the smallest available subpart, per language. While Lexical Diversity measures vocabulary richness, it does not measure formulaic language, as it does not take sequence information into account. According to TTR and to most readability measures, learner language is rated as very easily readable.

We test TTR in order to show if the corpora that we use, behave as expected (small lexical diversity, i.e. low TTR for low-level speakers) and because our model 1 measure, surprisal, has a correlation to TTR: if the vocabulary is small, the continuation of the sequence is also on average less surprising, simply as there are fewer words to choose from, as there is less entropy.

3.2. Model 1: Surprisal

We now turn to our word level models. First, we use surprisal to measure fixedness and the influence of the idiom principle [27]. Fixedness and entrenchment have earlier shown to be closely related [13, 28] and we want to measure relations between fixedness and language learner proficiency. For first language acquisition it has been shown that lexical-specific idiom-based language use precedes creativity [5], for second language acquisition the situation is less clear [29].

According to the Uniform Information Density Principle UID [23] the tug-of-war between routinisation/economy and expressivity prefers utterances which show a balance between high routinization (i.e. low surprisal) and dense expression (i.e. high surprisal) thus UID can be seen as minimising comprehension difficulty. Surprisal has been used to describe conditions in which zero-elements can appear as the context is redundant enough (ibid.), but it also holds as a general ordering principle: successful communication exhibits a trend towards normal distribution of surprisal, unless factors such as compression push it towards high surprisal (e.g scientific writing), or lower speaker competence leads to more sequences of high surprisal [30].

3.3. Model 2: POS tagger confidence

Second, we use a part-of-speech tagger as a model of surface ambiguity. By this model we expect that entrenched structures and collocations get higher scores, as they are expected and create no ambiguity at the part-of-speech level. Our hypothesis is that low learner proficiency leads to lower tagger confidence.

Our approach follows the tradition of statistical anomaly and outlier detection [24]. If a model is trained on L1 speakers's data, it can be expected that L2 speakers' data will fit less well, that they are more often outliers, because they produce more ungrammatical sentences, and because, according to [7], they master the subtleties of formulaic language less well. We could have measured surprisal at the POS level, but as really surprising sequences typically lead to tagging errors, and as potential ambiguity is a particularly important factor for processing load, we decided to use the POS tagger confidence as a model.

4. Data

We use the following corpora. For English, the British National Corpus (BNC) for obtaining bigram surprisal, and the NICT JLE learner corpus [31] for applying our methods. We use the Treetagger [25] as pretrained on the Penn Treebank in its distribution. English learner corpus contains 9 proficiency levels.

For Swedish, we use the largest dataset available - Swedish Wikipedia articles together with Swedish spontaneuos speech - Spontal - corpus[1] for obtaining bigram surprisal, and the ASU learner corpus oral part for applying our methods. The ASU corpus comprises 100 audio recordings (50 hours), 10 with each of the 10 informants. The oral trascripts measure ca. 269,000 word tokens in total, out of which ca. 147,000 constitute the learners utterances [8]. We have trained the Treetagger on the Swedish SUC corpus [33] ourselves. We have mapped 3 proficiency levels on the ASU speech transcipts on the basis of university term annotations (3 corresponding semesters as described in [8]). Note that proficiency levels for both datasets are not directly comparable.

5. Results

5.1. Lexical Diversity

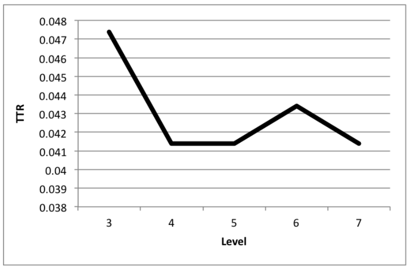

Figure 1: *Token per type ratio (TTR) for English learner productions.*

Since TTR is a size-dependent measure, one needs to compare texts of the same length. As the NICT levels 1, 2, 8, and 9 only have very little data, we have excluded them, and cut

[1] Spontal [32] covers 60 hours of recordings

the aggregated texts for each level to the shortest text, containing 105544 tokens. The type per token ratio (TTR) for NICT is given in Figure 1. We can see that there is no clear trend. Looking more closely at the data reveals, however, that the Zipfian curve tails off faster at the lower levels. At position 30, for example, the type at level 3 has 452 occurrences, while the type at level 7 has 588 occurrences. The relatively high number of types for low-level speakers is largely due to L1 (Japanese) words and proper names.

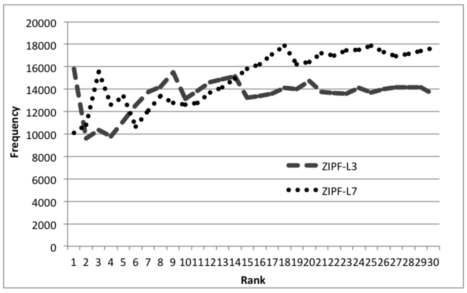

Figure 2: *Zipfian constant for English learner productions.*

Figure 2 plots the Zipf constant (rank * frequency) of level 3 against level 7. After some initial fluctuations, the constant stays higher for the higher level learners for about the top 100 types. This indicates a richer vocabulary of function words and routinised expressions. We will further investigate this hypothesis in section 5.2.

Also the Swedish data shows a similar behaviour: there is no clear trend in TTR, but the Zipfian distribution tails off faster for low-level speakers. Closer inspection of the data reveals that there are many L1/L3 (English) language transfer words and many false starts (transcribed hesitations) in the low-level speaker Swedish data.

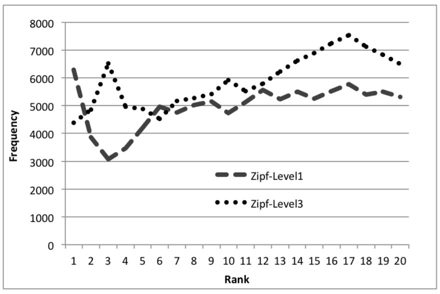

Figure 3: *Zipfian constant for Swedish learner production.*

TTR is affected by direct borrowings and code switching to the native language of the learners. In order to assess the impact, we also measure out-of-vocabulary words (Table 2). In ASU, we see that beginners make more false starts and hesitations and use direct borrowings. We have learnt from this that the naive application of monogram measures is not without complications. On the one hand, TTR does not separate errors from advanced vocabulary, a point which we currently address by looking at out-of-vocabulary words.

On the other hand, it often seems to be the sequence of words rather than individual words which are characteristic of learner language. We address this point with surprisal (5.2) and tagging (5.3) models.

Table 1: *Sentence length, out-of-vocabulary words (per 100 words), and TTR for English and Swedish learner production.*

Sentence Length				
Level	ASU	NICT 1-4	Level	NICT 5-9
1	6.203	3.075	5	8.465
2	7.766	4.328	6	9.420
3	8.864	5.948	7	9.741
4	-	7.354	8	9.663
			9	10.460
Out-of-vocabulary Words				
Level	ASU	NICT 1-4	Level	NICT 5-9
1	11.820	10.534	5	1.234
2	6.705	4.879	6	0.988
3	4.344	2.773	7	0.893
4	-	1.729	8	0.852
			9	0.716
Type-Token Ratio				
Level	ASU			for NICT
1	0.086			see
2	0.093			fig. 1
3	0.080			

Another simple complexity measure is sentence length (Table 3), which steadily increases across learner levels. Longer sentences and richer vocabulary (which advanced learners use) lead to more complex language, but also lexical errors, unusual word order or collocations (which beginners may use) increase processing complexity. In the following subsection, we now turn to surprisal to see if we can tease apart these two opposing forces.

5.2. Surprisal

We compare surprisal between the learner levels in Figure 5 for the NICT corpus, and in Table 2 for the NICT and ASU corpora. Figure 4 shows that the corrected utterances have lower surprisal, as expected.

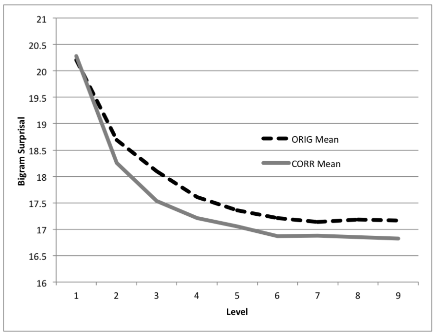

Figure 4: *Bigram Surprisal for English learners, by level and comparing original versus corrected utterances*

Figure 5 shows that surprisal indeed forms a normal distribution, as expected due to UID [23], except for sparse data issues: sparse data on the one hand leads to a considerable amount of fluctuations, and on the other hand unseen bigrams. We have given the highest value to unseen bigrams, which explains the peak at the right of the histograms in Figures 5. In addition, we can observe that the peak of level 3 is lower, while the right tail is much larger, and the left tail hardly smaller, which illustrates that the standard deviation is higher for low level learners, another sign that UID is observed less well. The exact stan-

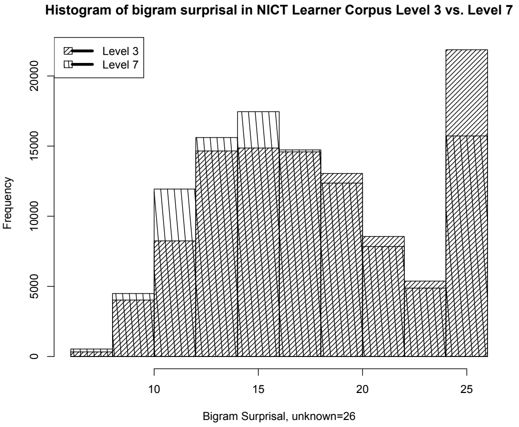

Figure 5: *Bigram Surprisal for English learners, comparing the low level 3 to the high level 7*

dard deviations are given in Table 2. Except for the very sparse Level 1, the English data follows the expectations. The smaller and noisier Swedish data (in which false starts and repetitions are kept, and words have often been transcribed to mimic pronunciation rather than orthographic conventions, no trend can emerge, unfortunately.

Table 2: *Mean and standard deviation of bigram surprisal across proficiency levels.*

	Mean		Standard Deviation	
Level	ASU	NICT (orig)	ASU	NICT (orig)
1	23.7	20.202	6.245	(5.089)
2	22.9	18.692	6.432	5.291
3	23.6	18.098	5.757	5.182
4	–	17.612	–	5.061
5		17.362		4.999
6		17.208		4.998
7		17.138		4.996
8		17.181		4.986
9		17.162		4.949

5.3. POS Tagger Confidence

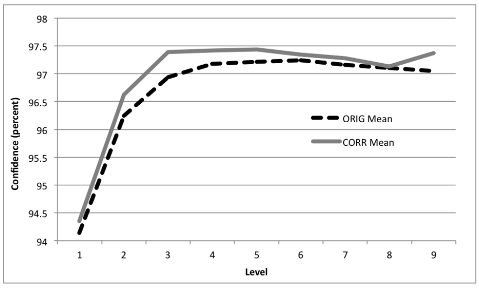

Figure 6: *Tagger confidence for English learners, by level and comparing original versus corrected utterances*

The mean confidence of the Treetagger for the English data is given in Figure 6. As a trend, the confidence of the tagger generally increases, which means that with increasing learner level, the tagger model fits better, despite the fact that the learners' vocabulary grows richer and the language more complex.

Table 3: *Treetagger confidence for Swedish learner production.*

Level	ASU	NICT 1-4	Level	NICT 5-9
1	0.96486	0.941	5	0.974
2	0.96549	0.963	6	0.974
3	0.96901	0.969	7	0.973
4	-	0.972	8	0.973
			9	0.971

Also this data supports our hypothesis that low level learner language, although simpler, is often more difficult to process. With higher levels, the curve flattens out and decreases again from about level 6 onwards, indicating that the increasingly complex language is slightly more challenging for the tagger.

The mean confidence of the Treetagger for the Swedish ASU corpus is given in Table 3. The confidence of the tagger generally increases, which means that with increasing learner proficiency level, the tagger model fits better.

6. Conclusions

We used n-gram language models of surprisal and POS tagger confidence to investigate Swedish and English learner language, and differences across proficiency levels. Our starting hypothesis that, although the language of lower level learners is simpler, it is also less optimal in terms of information theory, and as a consequence more difficult to process, also has been confirmed.

At the level of the independent word, i.e. at the unigram level, we found correlations in terms of TTR, out-of-vocabulary words and sentence length across proficiency levels for Swedish and English, as expected. We also saw that sentence length and out-of-vocabulary ratio are better predictors than the traditional measure of TTR.

At the bigram level, we have used a surprisal model, and POS tagger confidence. For surprisal, we have seen that it forms a normal distribution as UID predicts, and we have also shown that dispersion of surprisal decreases with higher proficiency levels, indicating increased abidance to UID. These findings hold less well for the particular situation of the Swedish corpus, in which false starts and repetitions are kept, and words have often been transcribed to mimic pronounciation rather than orthographic conventions.

Particularly concerning the mean of surprisal, further investigations are needed, as higher surprisal can on the one hand originate from language borrowing, hesitations, false starts and nonnative use of collocations, hence indicating low language proficiency. On the other hand, it can also be due higher compression, denser language, hence indicating higher proficiency level, particularly at very advanced levels. These two opposing factors can also lead to a situation of no clear trend in the mean of the surprisal value across L2 levels, as we observe it in the Swedish data.

Although the variation in the POS tagger confidence appears to be minimal, it must be noted that the POS tagger is trained on a large data set and has a fairly stable performance. Therefore even a slight increase of the tagger confidence indicates improvements in learners' language use.

7. Acknowledgements

The authors would like to thank Björn Hammarberg for the courtesy of accessing the ASU corpus and Johan Sjons for the preprocessing the Swedish Spontal corpus.

8. References

[1] C. E. Shannon, "Prediction and entropy of printed english," *The Bell System Technical Journal*, vol. 30, pp. 50–64, 1951.

[2] J. R. Nattinger, "A lexical phrase-grammar for ESL," *TESOL quarterly*, vol. 14, no. 3, pp. 337–344, 1980.

[3] B. Altenberg, "On the phraseology of spoken english: The evidence of recurrent word combinations," in *Phraseology: Theory, analysis, and applications*, A. P. Cowie, Ed. Oxford: Oxford University Press, 1998.

[4] S. De Cock, "Repetitive phrasal chunkiness and advanced efl speech and writing," in *Corpus Linguistics and Linguistic Theory*, C. Mair and M. Hundt, Eds. Amsterdam: Rodopi, 2000.

[5] M. Tomasello, "The item based nature of children's early syntactic development." *Trends in Cognitive Sciences*, vol. 4, pp. 156–163, 2000.

[6] N. Millar, "The processing of malformed learner collocations," *Applied Linguistics*, vol. 32, no. 2, pp. 129–148, 2011.

[7] A. Pawley and F. H. Syder, "Two puzzles for linguistic theory: Native-like selection and native-like fluency," in *Language and Communication*, J. C. Richards and R. W. Schmidt, Eds. London: Longman, 1983, pp. 191–226.

[8] B. Hammarberg, "Introduction to the asu corpus: A longitudinal oral and written text corpus of adult learner swedish with a corresponding part from native swedes. white paper. version 2010-11-16." 2010.

[9] S. Granger, "Prefabricated patterns in advanced efl writing: Collocations and formulae (oup, 1998)," in *Phraseology: Theory, analysis, and applications*, A. P. Cowie, Ed. Tokyo: Kurosio Publishers, 2009, pp. 185–204.

[10] S. Granger and S. Tyson, "Connector usage in the english essay writing of native and non-native efl speakers of english," *World Englishes*, vol. 15, no. 1, pp. 17–27, 1996.

[11] B. Altenberg and M. Tapper, "The use of adverbial connectors in advanced swedish learners' writtten english," 1998.

[12] B. Erman, "Formulaic language from a learner perspective: What the learner needs to know," in *Formulaic Language*, R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley, Eds., 2009, vol. II: Acquisition, loss, psychological reality, and functional explanations, pp. 323–346.

[13] J. Bybee, *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press, 2007.

[14] J. Read and P. Nation, "An investigation of the lexical dimension of the IELTS speaking test," in *IELTS Research Reports*, P. McGovern and S. Walsh, Eds. IELTS Australia and British Council, 2006, vol. 6.

[15] C. Kennedy and D. Thorp, "A corpus investigation of linguistic responses to an ielts academic writing task," in *IELTS collected paper: research in speaking and writing assessment*, L. Taylor and P. Falvey, Eds. Cambridge: Cambridge University Press, 2007, pp. 316–378.

[16] A. Ohlrogge, "Formulaic expressions in intermediate EFL writing assessment," in *Formulaic Language*, R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley, Eds. Amsterdam: Benjamins, 2009, vol. vol. II: Acquisition, loss, psychological reality, and functional explanations, pp. 375–386.

[17] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® v. 2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, 2006.

[18] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 90–95.

[19] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading esol texts," in *Human Language Technologies-Volume 1*. ACL, 2011, pp. 180–189.

[20] W. J. Bonk, *Testing ESL learners' knowledge of collocations*. Illinois: Clearinghouse, 2000.

[21] T. K. Landauer, "Automatic essay assessment," *Assessment in education: Principles, policy & practice*, vol. 10, no. 3, pp. 295–308, 2003.

[22] L. M. Rudner and T. Liang, "Automated essay scoring using bayes' theorem," *The Journal of Technology, Learning and Assessment*, vol. 1, no. 2, 2002.

[23] R. Levy and T. F. Jaeger, "Speakers optimize information density through syntactic reduction," in *20th Annual Conference on Neural Information Processing Systems*, 2007.

[24] C. C. Aggarval, *Outlier Analysis*. Boston/Dordrecht/London: Kluwer, 2013.

[25] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[26] D. D. Malvern, B. J. Richards, N. Chipere, and P. Durán, *Lexical Diversity and Language Development*. Houndmills, UK: Palgrave MacMillan, 2004.

[27] J. Sinclair, *Corpus, Concordance, Collocation*. Oxford: OUP, 1991.

[28] A. Blumenthal-Dramé, *Entrenchment in Usage-Based Theories. What Corpus Data Do and Do Not Reveal About The Mind*. Berlin, Boston: De Gruyter, 2012.

[29] N. C. Ellis, "Formulaic language and second language acquisition: Zipf and the phrasal teddy bear," *Annual Review of Applied Linguistics*, vol. 32, pp. 17–44, 2012.

[30] G. Schneider and G. Grigonyte, "Studies in language companion series," in *New Approaches to English Linguistics: Building bridges*, A. H. Olga Timofeeva, Anne-Christine Gardner and S. Chevalier, Eds. John Benjamins, 2016, vol. Volume 177, pp. 281–320.

[31] E. Izumi, K. Uchimoto, and H. Isahara, "The nict jle corpus: Exploiting the language learners speech database for research and education," *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

[32] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture." in *LREC*, 2010, pp. 2992–2995.

[33] G. Källgren, "Documentation of the Stockholm-Umeå corpus. in: Manual of the Stockholm Umeå Corpus version 2.0." *Department of Linguistics, Stockholm University*, 2006.