Automatic Generation and Semantic Grading of Esperanto Sentences in a Teaching Context

Eckhard Bick
University of Southern Denmark
eckhard.bick@mail.dk

Abstract

This paper presents a method for the automatic generation and semantic evaluation of exercise sentences for Esperanto teaching. Our sentence grader exploits both corpus data and lexical resources (verb frames and noun/adjective ontologies) to either generate meaningful sentences from scratch, or to determine the acceptability of a given input sentence. Individual words receive scores for how well they match the semantic conditions projected onto their place in the syntactic tree. In a CALL context, the system works with a lesson-/level-constrained vocabulary and can be integrated into e.g. substitution table or slot filler exercises. While the method as such is language-independent, we also discuss how morphological clues (affixes) can be exploited for semantic purposes. When evaluated on out-of-corpus course materials and short stories, the system achieved a rejection precision, in terms of false positives, of 98-99% at the sentence level, and 93-97% at the word level.

1 Introduction

Automated writing evaluation (AWE) can be a cost-efficient and consistent, albeit controversial, alternative to human grading of L2 student production. One possible approach is to focus on learner error detection in terms of spelling and grammar at the sentence level (e.g. Lee et al. 2011), a task for which a wide range of tools and methods is available, covering both rule-based, machine learning (ML) and hybrid approaches (e.g. Ng et al. 2014). Semantic assessment is usually seen as having a wider scope than the individual sentence, and is mainly used to address properties of a text as a whole. In Yannakoudakis (2013), for instance, machine learning (ML) techniques based on word embeddings are employed to assess textual coherence through the semantic similarity of adjacent sentences. By contrast, the semantic correctness of the individual sentence is less important in AWE, since human-produced sentences generally do have a coherent meaning, and by and large sentence understanding is quite robust even in the face of multiple spelling and grammatical errors. Therefore, semantic oddities are usually not independent errors, but either a byproduct of lower-level errors or word pair confusion errors that are recognizable as such in context. Even beginner-level L2 students know what they *want* to say.

In the research presented here, however, we focus on the semantics of the individual sentence, and we are interested not only in human-generated sentences, but also in automatically generated random sentences. In the latter - unlike in AWE - meaning coherence at the sentence level is not governed by an underlying intellect, but has to be controlled and evaluated.

While sentence generatio in our own project is intended for use in language learning exercises with a controlled vocabulary and controlled syntactic complexity, adding a semantic component to random sentence generation is also useful for other tasks, such as creating training data for text-to-speech (TTS) or voice recognition systems. Thus, Lilley et al. (2012) describe an HPSG-based generator with a 20-category noun ontology, a lexicon of 2181 wordforms (1100 lemmas) and 39 production rules that achieved a human meaningfulness

rating of 3.09 on a 0-6 scale, as opposed to 0.66 for a semantics-free system and 4.52 for human text.

We here adopt a similar approach, matching head words with semantically and categorically constrained slot fillers for valency and modifier slots. However, we go beyond this framework on several important accounts:

- We apply the method to both sentence generation and the evaluation of existing sentences
- For evaluation, lexical frames and valency patterns are combined with combinatorial corpus statistics
- Vocabulary size and content are parameters controlled by the user, lesson or text book, and has no upper boundary - in principle, free input sentences can be evaluated, and the semantic ontology has a high coverage even on unabridged text

2 The project

Though it can be used for different purposes, the sentence generator/evaluator was developed primarily with a pedagogical framework in mind. Specifically, the idea was to allow the creation of CALL exercises, where a pre-defined vocabulary would yield a maximum number of meaningful sentences for substitution table exercises and their more constrained variants, such as slot filler exercises, one word substitution exercises, question-answering sentence pairs etc. The currently funded international project aims at teaching Esperanto to children in the first grades of primary school, as propedeutic foreign language with a transparent, modular and highly regular linguistic system supposed to facilitate general linguistic awareness and subsequent learning of other foreign languages. All course materials (lessons, songs, dialogues, exercises) were lexicographically analyzed with the EspGram parser (Bick 2009, 2016) in order to determine the introduced morpheme vocabulary of root words, affixes and inflection categories. Our tool is then used to generate sentences from the accumulated vocabulary at a given stage of the course, to evaluate and semantically filter combinatorial sentences from teacher-defined substitution tables or to suggest wrong, but not absurd, semantic alternatives for word substitution exercises, the idea being that substitution with a semantically close word (e.g. an animal for a human agent subject in an activity sentence) would create fun effects that substitution with something completely unrelated would not (e.g. an abstract feature or activity noun for a food object in an eating sentence).

3 Sentence generation

3.1 Vocabulary base

In preparation for sentence generation, a morpheme and word lexicon is established for the current teaching level, analyzing all words up to and including the current lesson block, in our case 6 lesson levels with 5 blocks each. Fig 1 plots vocabulary growth per lesson for the four inflection-marked content word classes (POS). Esperanto marks word class systematically with an endings vowel (nouns -o, verbs -i, adjectives -a, adverbs -e). Where semantically feasible, Esperanto word roots can change word class by changing this vowel, e.g. *amiko (friend), amika/amike (friendly), amiki (be friends)*. Thus, in our course material, the number of words was about 10-13% higher than the number of roots (table 1). The language also allows compounding (e.g *amletero 'love letter')* and uses a number of semantically transparent agglutinative affixes e.g. *-ej* for places (*vendejo 'shop')* and *mal-* for antonyms (*malvarma 'cold'*). Therefore, the number of N/V/A-roots exhibits a steeper per-lesson increase than the number of morphemes (table 1).

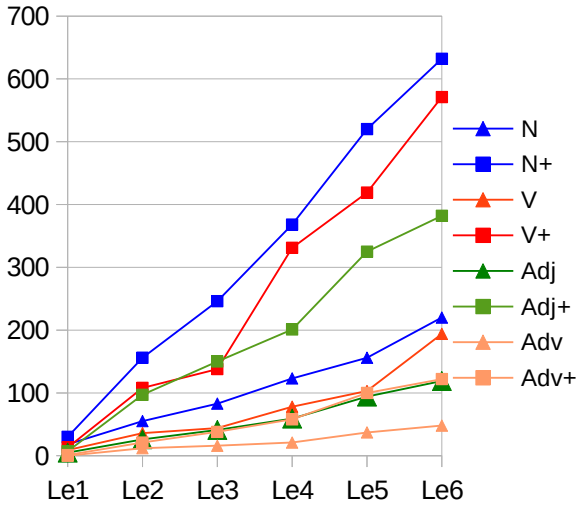

Fig. 1: Vocabulary size by POS for Lesson 1-6, as-is and extended (POS+)

In order to provide more lexical variation, the sentence generator can optionally expand its

lexicon with compounds and affix-derivations that do not themselves occur in the course material, but can be formed using known morphemes. For this, EspGram's parser lexicon (58,000 lemmas) is morphologically analyzed. All words that do not contain unknown morphemes, and are not marked as "rare" or "archaic", will then be added to the respective POS lexica in the generator. As can be seen in table 1, this extended lexicon is about three times the size of the original course material at all levels.

L	mor-phemes	N/V/A roots	N/V/A words	N/V/A extended	affixes
1	45	32	32	51	-in,-ist
2	139	118	129	382	-ebl,-ej, ge-,mal-
3	177	165	184	572	-ul
4	246	252	281	958	-iĝ,re-
5	335	341	390	1364	-et,-ind
6	392	516	581	1707	-eg,-er

Table 1: Overall vocabulary size for content POS, roots vs. words

3.2 Valency frames for verbs

As a sentence skeleton for generation we use valency frames based on framenet entries for verbs. The Esperanto frames follow the system described in (Bick 2012) and contain, besides the semantic category of the verb, a list of arguments with semantic and morphosyntactic slot filler information.

(a) manĝi <FN:eat/S§AG'H|A/O§PAT'food>

(b) instrui <FN:instrui/S§AG'H/O§BEN'H/P-pri§TP'all><FN:teach/S§AG'H/O§TP'domain|ling|fcl/P-al§BEN'H>

(c) diri <FN:say/S§SP'H/O§MES'sem-s|fcl/P-al§REC'H>

The word for 'eat' (a), for instance, has two frame arguments, an agent (§AG) and a patient (§PAT), the former as subject (S), the latter as object (O). Semantic types are added with an apostrophe - 'human' (H) and 'food', respectively. Sometimes, more than one frame construction is possible. Thus, the word for 'teach' (instrui), can have either the human beneficiary (§BEN) or the teaching topic (§TP) as object, with the other argument as a prepositional complement ('P-' plus the preposition, *pri 'about'* or *al 'to')*. The second construction, and example (c) can also vary in syntactic form, allowing a finite clause object (fcl) rather than a noun phrase. All in all, we created frame entries for 6235 verb lemmas, providing 100% coverage for the course material. In a 1.3 million word newspaper corpus[1] coverage for verb tokens was 96%. There are similar frames for 157 nouns and 50 adjectives, but they are not used by the sentence generator.

3.3 Semantic ontologies for nouns and adjectives

When the generator expands a verb frame into a sentence, it randomly picks slot-filler nouns from the lesson-constrained vocabulary, making sure the noun in question is semantically compatible with the head verb. For this we use Esp Gram's existing ontology of semantic prototypes for nouns. The ontology has about 200 categories, organized in a shallow hierarchy. For instance, the human <H> category has sub-types like <Hprof> (profession), <Hideo> (follower of an ideology), <Hfam> (family relation term) etc., and the <tool> class is subdivided into <tool-cut>, <tool-mus> (musical instruments), <tool-light> etc. Since the majority of Esperanto affixes allows a safe prediction of semantic class (e.g. *-ej* for <L> 'place' or *-uj* for <con> 'container'), it is possible to increase coverage through morphological analysis, creating semantic entries for the productive, unlisted part of the lexicon, too. Thus, 99.3% of the nouns in the test corpus received a semantic entry. For the course material, all entries were manually checked.

For adjectives, we used a scheme with about 110 categories, suggested for Danish in (Bick, 2019). The categories can be seen as semantic prototypes (e.g. *<jcol> colour* or *<jpsych> psychological state*), but are at the same time intended for distributional restriction, i.e. to purvey information on which (semantic) type of head noun they can combine with. Thus, *<jcol>* will combine with physical objects, and *<jpsych>* with human and semiotic nouns. The adjective ontology can be ordered into 14 primary and 25 secondary umbrella categories. For Esperanto, we tagged about 4140 adjective lemmas in the dictionary, amounting to a token coverage of 100% for the course material and 71% for the news corpus. Inspection indicates

[1] The corpus is based on *Monato*, a monthly news magazine published in Esperanto by *Flandra Esperanto-Ligo*.

that coverage could probably be increased considerably for the latter by systematic class transfer, because 2/3 of untagged cases were derivations from nouns or verbs, either by direct conversion or by affixation.

3.4 Syntactic and morphological generation

The sentence generator builds and joins phrases for a list of main syntactic constituents (subject, object, subject complement, object complement, adverbial arguments and adjuncts) around a governing verb frame. Therefore, as a first step, a random verb is chosen and assigned a random tense, expanding into a VP in the case of auxiliaries. Second, the list of arguments is culled depending on the frame's valency, and for each argument, an phrase-generating subroutine is called. In most cases, this will be an NP, but if the frame demands a subclause, step 1 will be iterated and a conjunction added.

In the NP subroutine, a random head noun is chosen, looping until the frame's semantic slot filler condition is matched. In most cases, frames only ask for a supercategory such as 'human' or 'food', and here we allow all subcategories to match on the noun side. The chosen noun is then inflected depending on syntactic function (*-n* for direct objects). Number (singular/plural) is in principle assigned randomly, but has to observe the fact that some verbs take only plural subjects or plural objects, and that the semantics of the noun may make a plural meaningless (e.g. domain and mass words). If no semantically matching noun is found in the lesson vocabulary, a matching pronoun is used instead, e.g. *li (he), io 'something'*) or *ĝi (it).*

With a likelihood of 0.25% for subjects and objects, and 0.20% for other constituents, NP's will be expanded with a definite article, possessive or demonstrative pronoun. The same likelihood threshold holds for expansion with an adjective phrase (ADJP), and both types of modifiers will be inflected in agreement with the head noun. Like for nouns, there are semantic restrictions on the choice of adjective, depending on the semantic category of the phrase head. In the absence of explicit selection information on the noun, we constructed a table of many-to-many matches for groups of semantic noun prototypes on the one side and groups of adjective prototypes on the other, roughly mirroring the granularity of the secondary umbrella terms in the adjective ontology, e.g. a list of concrete object types (containers, clothes, furniture, tools, machines, vehicles ...) matching a list of physical property adjectives (colour, size, shape, weight, state, texture ...). In some cases, there was some isomorphism in the ontologies, allowing one-on-one matches, e.g. <*cloH*> (clothing) - <*jclo*> (clothing adjective), <*f-q*> (quantifiable feature) - <*jchange*> (change adjective).

On average, the generator produces sentences with 4-5 words, with no clear correlation to the size of the accumulated lesson vocabulary. Word length increased slightly, from 6.8 letters/word in the first lesson to 7.4 for the last lesson. Using the extended vocabulary increases both word and sentence length for the first lesson. At later stages, there is a tendency towards longer but fewer words, probably due to a higher percentage of words that are morphologically complex, but at the same time rarer and simpler in terms of valency.

4 Semantic Sentence Grading

While vocabulary-constrained semantic sentence generation is useful for creating random training sentences, other course-related tasks call for semantic grading of student-produced sentences rather than the generation of completely new sentences. Such sentences can, for instance, be prompted online in a question-answering scenario, or they can be the result of substitution table or slot filler exercise in a graphical user interface (GUI). Here, a backend program has to decide, if a certain combination of words from the substitution table, or a suggested slot filler word creates a semantically acceptable sentence or not. Impossible or odd sentences indicate that the student has not understood the sentence context and is unsure of one or more words. An automatic tutoring program can use this information to flag words or structures as "probably known" or "problematic", even in a monolingual L2 setting.

Our sentence grader implements a two-thronged co-occurrence approach based on annotated corpus data. First bigrams and trigrams (combinations of two or three consecutive words, respectively)[2] are checked against a corpus-

[2] BOS (beginning-of-sentence) and EOS (end-of-sentence) dummies are used to create ngrams for the first and last words in a sentence.

derived frequency table of such co-occurrences. Second, we apply a similar check to what we call *depgrams* (word pairs with a syntactic dependency link such as verb-object or noun-attribute), in a fashion similar to the method used in (Sidorov et al. 2013) to detect and correct grammatical errors. While the former (ngrams) is a surface fluency check of what is "normal", the latter (depgrams) evaluates deeper, word order-independent, syntactic relations.

4.1 Corpus-based statistics

For the ngram and depgram frequency data, mixed corpora were used, amounting to a total of 50 million tokens (words and punctuation):

Corpus	**Size** (million tokens)
Classical literature[3]	9.76
Eventoj[4] (news magazine)	1.80
Monato[5] (news magazine)	1.44
Wikipedia	16.48
Internet (mixed)[6]	19.78
All	**49.26**

Table 2: Corpus sources

After morphosyntactic analysis with EspGram, we counted word bigrams and trigrams, as well as depgrams. For the former, only word form was used, for the latter we also stored - where applicable - syntactic function (func), semantic type (semtype) and a possible preposition header (prp), in the following combinations of dependent (left) and head (right), cf. table 3:

head (prp) dependent (func)	word_2	semtype_2
lemma_1	word-word	word-sem
semtype_1	sem-word	sem-sem

Table 3: Depgram types

Part-of-speech (POS) is a (vowel-coded) morpheme category in Esperanto, and hence need not be stored separately. For PP's (b2, d2), we stored both the preposition and its - semantically more important - head. The syntactic function field (also called edge label) was added for clause level dependents (subjects, objects, adverbials etc., cf. a2, c, d1-2), but not for phrases (a1, a3, b1), since function here is already almost unambiguously implied by POS and head type. Subject and object complement relations were treated like in-NP attributes (a4).

(a1) *tute -> same* ('completely equal')
(a2) *PHUM@SU -> organizis* ('Peter organized')
(a3) *NUM -> mm/m2* (e.g. 37 mm/m2)
(a4) *NUM -> aĝo* (e.g. 'her age was 23')

(b1) *aŭtomata -> <act>* ('automatic action')
(b2) *al/EWORD -> <FN:send>*
(e.g. 'send [email] to xxx@gmail.com')

(c) *<f-right>@AC -> havis* ('have the right')

(d1) *<Hprof>@SU -> <FN:create>*
(e.g. 'the carpenter built ...')
(d2) *per/<Vground>@AD -> <FN:run>*
(e.g. 'he went by train)

As semantic types we used semantic prototypes for nouns (e.g. <act>, <f-right>) and adjectives (e.g. <jshape>), and framenet categories for verbs (<FN:....>). In order to avoid sparse data problems and to keep the database manageable, certain highly productive word types were replaced with a letter dummy: PHUM (human proper noun), PTIT (work-of-art title), PORG (organization name) POTH (other proper nouns), PTWIT (twitter name), EWORD (emails, URLs), YEAR, DATE and NUM (non-letter cardinal numbers).

In the database, ngram counts are stored as *relative* frequencies (i.e. divided by unigram/lower order frequencies), so they can be used to predict the likelihood of a given ngram given its left part. For depgrams, mutual information is used:

- bigrams: f(ab) / f(a)
- trigrams: f(abd) / f(ab)
- depgrams: f(a->b) * n / f(a)*f(b)

where a, b and c are adjacent words (or, for depgrams, lemmas), and n is the number of tokens in the corpus.

[3] A mixed corpus of internet-available Esperanto books

[4] An Esperanto biweekly published 1992-2002

[5] http://www.monato.be/

[6] A 2012 crawl downloaded from: http://wortschatz.uni-leipzig.de/en/download/

4.2 Sentence grading

We compute the acceptability score of a sentence as the sum of its bigram, trigram and depgram scores, which again are the sums of the corresponding scores for the individual words in context, divided by the number of bigrams, trigrams and depgrams that contributed to the score.

$$\frac{\sum_{i=1}^{n} bg_i}{n-1} + \frac{\sum_{i=1}^{n} tg_i}{n-2} + \frac{\sum_{j=1}^{d} dg_j}{d}$$

Since bigrams (bg) and trigrams (tg) are used to measure cohesion at the surface level, they only are computed for words, and their number therefore depends directly on the number of words in the sentence (n).

The depgram score, on the other hand, is designed to measure the semantic acceptability of word relations, and is computed in a more complex and more semantic fashion. First, inflected forms are lemmatized, function labels added for clause level dependents and relations copied from preposition dependents to their arguments, regarding the preposition as a kind of case marker. Second, for all content words, corpus depgram frequencies are checked for all 4 combinations of lemma and semantic type (cf. table 3).

By using different weights for the different types[7] of ngrams and depgrams, their respective impact on the sum count can be controlled. For instance, since trigrams contain more information (are more constrained) than bigrams, we use higher weights for the former. Depgrams are scored with the square root of their likelihood, and in addition assigned the following weight factors:

depgram type	**weighting**
clause level dependent (with function label)	* 3
sem -> sem relation	log 2
coarse/simplified sem categories	* 0.2

Table 4: Weights

[7] Multiple occurrence of the same ngram in the sentence is weighted down by using the square root.

When no depgram corpus match is found for any of the 4 word/sem combinations[8], a second round of look-ups is performed, where a more coarse-grained semantic ontology is used for nouns, collapsing all subcategories for the types of 'human', 'animal', 'plant', 'things', 'place', 'vehicle', 'tool', 'food', 'domain', 'semantical' and 'substance', and by allowing internal cross-matching for all action/event categories and for all unit categories. A few ambiguous categories are tested twice, for different umbrella categories. For instance, <sem-r> ('readable', e.g. *book*) can be used as either a 'thing' (combining with verbs like *throw, put* and *borrow*) or 'semantical', being *read, written* or *translated.*

In addition, zero frequencies are punished by negative weights, i.e. when there is no example of a given depgram relation in the corpus:

depgram type	**clause level**	**other**
sem -> sem	-12	-8
word -> sem sem -> word	-6	-3
word -> word	-2	-0.1

Table 5: Punishments

Finally, there is a grammatically motivated punishment (-5) for NP agreement mismatches, and a frame-based punishment (-20), if neither the ordinary or coarse-grained semantic type of a verb argument matches the semantic condition of the corresponding argument slot in the verb frame.

4.3 Word grading

In addtion to grading whole sentences, it is pedagogically useful to be able to identify "outlier" words, that do not fit the rest of the sentence. Possible applications are slot filler exercises, but also as a kind of "fuzzy" proofing tool, that goes beyond simple spell- and grammar checking and is able to flag odd lexical choices in written L2 production.

Every time, an ngram or depgram combination is evaluated, the sum score of the participating words is adjusted correspondingly. Implementing

[8] Obviously, words can be semantically ambiguous and carry more than one semantic tag, with the parser being unable to choose. In these cases, the matching check is performed twice, before progressing to the fallback option of more coarse-grained umbrella tags.

a left-to-right prediction approach, the affected word is the last one in bigrams and trigrams. For depgrams it is intuitively more likely that the head is seen as the primary part, and the dependent as either matching or offending, assuming that the brain thinks *idea* first, and then is more likely to expand the concept into a *good idea* than a *blue idea*, rather than coming up with a colour and then finding a head noun for it. Therefore, when inheriting depgram scores, dependents are weighted twice as heavily as heads, and heads cannot inherit negative (punishing) scores. Still, heads should get some depgram scoring, too, because learners may make lexical errors and wrong synonym choices for heads, too, and in these cases it is useful to know if a given head is supported by a matching choice of attributes (for nouns) or arguments (for verbs).

After a long row of experiments and improvements, weights and punishments were ultimately chosen in a mostly empirical fashion, designed to yield positive scores for acceptable word choices and negative scores for conflicting words, balancing the positive values from corpus data matches on the one hand with negative values for lexical punishments on the other.

However, while both sentence and word scores represent a cline of acceptability without discrete breaks, we also introduced three types of unary flags for specific dependent mismatches:

'?' frame mismatch at the clause level

'*' missing corpus evidence for a sem/sem match or a clause-level word/sem match[9]

'%' agreement mismatch in noun phrases

These markers can either be used for specific error flagging, or as a secondary filter (3) after eliminating sentences with (1) negative overall scores or (2) one or more negative word scores.

(a) 12.4 *patrino* (33.5 *bakis* (62.2) *bongustan* (18.3) *kukon* (98) . 'mother baked a delicious cake'

(b) 5.25 *viro* (8.0) *vendis* (25) *bluan* (4.67) auxton (41.41) . 'a man sold a blue car'

(c) 2.04 *patrino* (11.9) *vendis* (11) ***bluan*** (-4.29) *kukon** (5.15) . 'mother sold a blue cake'

(d) 0.24 *patrino* (11.9) *vendis* (11) ***bluajn***% (-12.7) *kukon** (5.91) . 'mother sold blue cakes'

(e) -0.09 *viro* (2.8) *mangxis* (3.35) *bluan* (4.67) ***auxton***?* (-21.7) . 'a man ate a blue car'

(f) -3.51 ***floro****?* (-40) *bakis* (1.1) *bluan* (4.67) ***auxton***?* (-24). 'a flower baked a blue car'

(g) -4.05 ***floro****? *(-27) *bakis* (0) ***bluan*** (-5.36) ***songon***?* (-32.14). 'a flower baked a blue dream'

In the graded example sentences, (a-d) have positive scores and would be semantically acceptable, but (d) could be filtered out for grammatical reasons - it exhibits an agreement error, because the object (-n) NP *'blue cakes'* has the adjective in the plural (-j), and its head in the singular. Furthermore, both (c) and (d) contain a negative word score, due to the lack of corpus evidence for blue cakes. At the clause level, the selling of cakes has lexical frame support, but an asterisk is added, because the corpus does not have an example of a selling verb with 'cake' as object. At the top end, a *mother baking* and a *cake* being *delicious* (a) are both more typical (i.e. have higher mutual information in statistical terms) than a *man selling* and a *car* being *blue* (b). The negatively scored sentences all contain one or more words that not only fail the corpus test, but also violate generalized frame conditions ('?'-mark). Thus, cars can't be eaten (e), plants can't bake (f). *Dream* (g) mismatches both at the clause level (with *bake*) and at the phrase level (with *blue*).

5 Language-specific adaptations

In principle, the sentence-grading system presented here is largely language independent, as long as a corpus of sufficient size, a dependency parser of sufficient quality and - not least - a framenet and semantic ontology are available for the language in question. However, two areas of language-specific adaptation should be born in mind, both concerning morphology.

First, we used words rather than lemmas, and treated prepositions as a kind of noun-prefix. While this is an easy way to capture surface clues for co-occurrence, and would work for e.g. Germanic languages in general, it is problematic for morphologically rich languages, where inflectional variation would create a sparse data problem. For such languages, lemmas should be

[9] There can be more than one asterisk, because they are assigned for each level separately. The worst case (***) means a mismatch even at the coarse-grained semantic level.

used instead of words. In the absence of prepositions, case categories should be used instead.

The second language-specific adaptation was the exploitation of the regular affixation system of Esperanto for semantic purposes. We already discussed the use of semantic class-conveying affixes in section 3.3. A different method is to strip semantically transparent affixes off roots in order to classify otherwise unlisted, productive forms:

verb affixes: *ek-* (inchoative), *-ad* (durative), *re-* (iterative), *fin-* (resultative)

general affixes: *-eg* (big, intensely), *-et* (small, moderately)

The suffixes *-ig* (turn into) and *-iĝ* (become) are not transparent when added to a noun root. With verbs, however, they are used to increase *(-ig)* or decrease *(-iĝ)* transitivity. Thus, *-ig* means 'make (object) do', and the frame of the root verb can be used for semantic matching, with the original object becoming the subject of the recovered root verb. Conversely, *-iĝ* has a passivization effect on a transitive verb, and the original subject can be matched against the object slot of the root verb.

Finally, two verb-adjectivizing affixes, *-ebl* (x-able) and *-ind* (worth x-ing) can be used to match an NP head against the object slot of the adjectivized verb. For instance, in *manĝ-ebla planto* (eat-able/edible plant) the NP head 'plant' can be matched against the object slot of 'eat'. Similarly, in *manĝ-inda kuko*, 'cake' becomes the object worth eating.

6 Evaluation

Obviously, sentence generation (section 3) is more robust than sentence grading (section 4), because the former will only produce what is sanctioned by the input vocabulary and dictionary-based frames. Sentence grading, however, has to work on unknown sentences, and its performance may suffer from lexical coverage problems, sparse data in the corpus database and, not least, missing or too-constrained frames. Especially the latter is a problem in the face of free, creative input from ordinary language users. Thus, in the controlled environment of course-based substitution table sentences, and with the course itself as part of its database, the tool performed very well, and accepted very few sentences that should have been discarded. In order to test for false positives (falsely discarded) sentences in a less constrained environment, we removed the course texts from the database, and submitted the entire teaching material[10] (7,363 words) to the sentence grader. In this run, only 1.4% of the (supposedly correct) sentences received negative scores. At the word level, there were 3.3% negative scores. Flags for missing frame support (?-flag) and missing dependency corpus support (*-flag) appeared in 5.3% and 2.4% of cases. However, neither negative score nor flags are a safe marker for unacceptability. For instance, negative scores can be caused by bad ngram scores even in the face of a frame match, and conversely, corpus "proof" can compensate for a missing frame, preventing a negative score. Therefore, to increase precision and limit the number of false positives, flags and scores should be combined for automatic use. Thus, at the clause-level, a combined frame failure (?) and corpus evidence failure (*) only occurred in 0.1% of words, and with a double condition for negative score AND a '?' or '*' flag, false positives were down to 0.64% and 1.6%, respectively.

Without a corpus of L2 learner texts, and in the absence of funding for extensive manual evaluation, it is difficult to evaluate the risk of false negatives (i.e. accepted sentences that should have been discarded) in external texts, but it is still possible to use the above method and estimate the prevalence of false positives, even on a larger scale, by simply running the evaluator on text that is not produced by learners, and is supposed to be correct. For our experiments we used a collection of short stories[11] (61,676 words), that are part of the advanced-learner material on the Esperanto teaching site lernu.net.

As expected, the short stories, with almost 4 times as many morphemes and lemmas, and sentences that were on average 75% longer, were more difficult for the evaluator program, with considerably more false positives, i.e. words with negative scores in supposedly normal sentences.

[10] after removing non-sentence parts such as word lists.

[11] 46 texts, written by Claude Piron, accessed at https://lernu.net/eo/biblioteko/106 (June 2019). Together, the texts contain approximately 5000 non-name lemmas, built from about 1200 morphemes.

	Course materials		Lernu.net short stories	
sentences	*1,233*		*5,921*	
with neg. score	1.4 %		0.86 %	
words	*7,363*		*61,676*	
per sentence	*5.97*		*10.4*	
with neg. score	2.6%		7.3 %	
	%	*neg.*	*%*	*neg.*
frame failure (?)	5.3	0.64	3.5	1.5
corpus failure (*)	2.4	1.6	5.9	3.7
both ? and *	0.1	0.1	0.5	0.5

Table 6: Evaluation - False positives

However, at the sentence level there were fewer false positives, indicating that the individual negative scores were milder, or that there was more corpus support from ngrams. One reason for the former could be the lower incidence of frame failures[12], which are punished harder than corpus failures and therefore contribute more to a negative word score than corpus failures. In any case the unclear balance between ? and * flags is another reason for a 'in dubio pro reo' approach, where words are regarded as problematic only, if they fail on all accounts (both frames and corpus). With this condition, the rate of false positives is very low even for the more complex short story corpus (0.5%).

7 Conclusions and outlook

We have shown how a combination of syntactic-analytic (dependency parsing) and semantic-lexical resources (verb frames and noun/adjective ontologies) on the one hand, and corpus data on the other can be used to build a semantically constrained sentence generator and grader for Esperanto. While intended for teaching use with a restricted vocabulary, the method is also applicable to unrestricted input. However, performance is dependent on corpus size and variety, as well as on lexical coverage for frames. Inspection of scoring errors suggested that future work should focus not just on the number of frames, but - maybe more importantly - on the the coverage of the semantic slot filler information for frame arguments. Still, with a limited vocabulary, this is a tractable task and can be addressed for a given course or textbook individually (as was done for our own course materials), not least by simply making the annotated materials part of the corpus[13]. In practical terms, words with a negative score in our evaluation (i.e. false positives) could be used as a point of departure for this work.

Apart from improving linguistic resources for a CALL-prioritized lexicon, future work should include evaluation with human annotators grading both automatic test sentences and learner sentences for (semantic) acceptability. This would make it possible to evaluate recall (false negatives) rather than just precision (false positives), and also to examine if machine scores can emulate a human scale of *degree* of sentence acceptability from least to most acceptable.

[12] With correct input, frame failures are basically lexical coverage errors. Since frequent verbs usually have the most complex grammar, the larger lemma spread in the short stories possibly leads to a higher percentage of less frequent, but easier verbs, while the teaching corpus exploits its relative few verbs to the full in combinatorial terms.

[13] It should be born in mind that for our evaluation on course materials, these were intentionally removed from the corpus first.

Acknowledgments

This research was carried out at the University of Southern Denmark, in the framework of an international 2-year CALL project co-funded by the European Union's Erasmus+ programme.

References

Bick, Eckhard. 2009. A Dependency Constraint Grammar for Esperanto. Constraint Grammar Workshop at NODALIDA 2009, Odense. NEALT Proceedings Series, Vol 8, pp. 8-12. Tartu: Tartu University Library. ISSN 1736-6305

Bick, Eckhard. 2012. Towards a Semantic Annotation of English Television News - Building and Evaluating a Constraint Grammar FrameNet, In: Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (Bali, 7-10 November, 2012). pp. 60-69. Faculty of Computer Science, Universitas Indonesia. ISBN 78-979-1421-17-1

Bick, Eckhard. 2016. A Morphological Lexicon of Esperanto with Morpheme Frequencies, In: Calzolari, Nicoletta et al. (eds.), Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC2016 (Portorož, May 23-28, 2016). pp. 1075-1078. ISBN 978-2-9517408-9-1

Bick, Eckhard. 2019. A Semantic Anthology of Danish Adjectives. In: Simon Dobnik, Stergios Chatzikyriakidis, Vera Demberg (editors):

Proceedings of IWCS 2019 - 13th International Conference on Computational Semantics (Gothenburg, 23-27 May 2019). ACL Anthology W19-04. pp. 71-78. URL: http://aclweb.org/anthology/W19-04

Lee, K.J.; Y. Choi and J.E.Kim. 2011. Building an Automated English Sentence Evaluation System for Students Learning English as a Second Language. Computer Speech and Language, 25. pp. 246-60

Lilley, Jason; Amanda Stent; Ilija Zeljkovic. 2012. A Random, Semantically Appropriate Sentence Generator for Speaker Verification. In: Proceedings of ICSLP (INTERSPEECH-2012). pp. 739-742.

Ng, Hwee Tou; Siew Mei Wu; Ted Briscoe; Christian Hadiwinoto; Raymond Hendy Susanto; Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning (2014, Baltimore): Shared Task, pp. 1–14. ACL

Sidorov, Grigori; Anubhav Gupta; Martin Tozer; Dolors Catala; Angels Catena; Sandrine Fuentes. 2013. Rule-based System for Automatic Grammar Correction Using Syntactic N-grams for English Language Learning (L2). In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL2013, Sofia): Shared Task, pp. 96-101

Yannakoudakis, Helen. 2013. Automated Assessment of English-learner Writing. Technical Report Number 842. University of Cambridge Computer Laboratory. UCAM-CL-TR-842. ISSN 1476-2986