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

Eckhard Bick

Institute of Language and Communication

University of Southern Denmark

Co-funded by the
Erasmus+ Programme
of the European Union

Introduction

- Automated writing evaluation(AWE) vs. human grading of L2 student production
- Spelling & grammar error detection (e.g. Lee et al 2011)
  - typically **sentence** level
  - many available tools and methods, both ML, rule-based and hybrid (e.g. Ng et al. 2014)
- Semantic assessment
  - typically **text** level (e.g. Yannakoudakis 2013: word-embedding similarity of adjacent sentences)
- AWE typically does not address the semantic viability of the individual sentence
  - its object are human-generated sentences, likely to have a coherent meaning (even students know what they **want** to say
  - any semantic oddities are usually local and a byproduct of lower-level errors or word pair confusion that are recognizable as such in context

Non-human sentences

- Topic of this paper:
  - automatic rather than human production (i.e. no controlling intellect)
  - (random) sentence generation rather than text generation
- Automatic generation of meaningful sentences is useful
  - language learning exercises
  - create training data for text-to-speech (TTS) or voice recognition systems
    - e.g. Lilley et al. 2012: HPSG, ingredients: noun ontology, lexicon (1100 lemmas), production rules (39)
- Special features of the approach presented here
  - the method handles both **sentence generation** and **evaluation of existing sentences**
  - For evaluation, **lexical frames** and **valency patterns** are used in conjunction with combinatorial **corpus statistics**
  - Vocabulary size and content are parameters controlled by the user, lesson or text book
  - Almost unabridged lexicon support for the target language - the **semantic ontology** has a high coverage even on unabridged text

The project

- Pedagogical framework
  - creation of meaningful sentences for CALL exercises with a pre-defined vocabulary
    - substitution table exercises
    - slot filler exercises
    - one-word substitution exercises
    - question-answering sentence pairs
- 2-year CALL project (co-funded by the **EU's Erasmus+ programme**)
  - first grades of primary school in several different countries in parallel
  - teach Esperanto as a propedeutic foreign language
    - transparent, modular and highly regular linguistic system
    - facilitates general linguistic awareness and subsequent learning of other languages
- Set-up
  - All course materials (lessons, songs, dialogues, exercises) were lexicographically analyzed (EspGram parser Bick 2009, 2016), determening which roots and morphemes were introduced when
  - Sentence generator: create exercise sentences for a given course stage
  - Sentence evaluator: filter possible sentences from teacher-defined substitution tables
  - Fun factor: identify/create wrong, but not absurd sentences by substitution with a semantically close word
    - YES: an animal for a human agent in an activity sentence
    - NO: a abstract feature or activity noun for a food object in an eating sentence

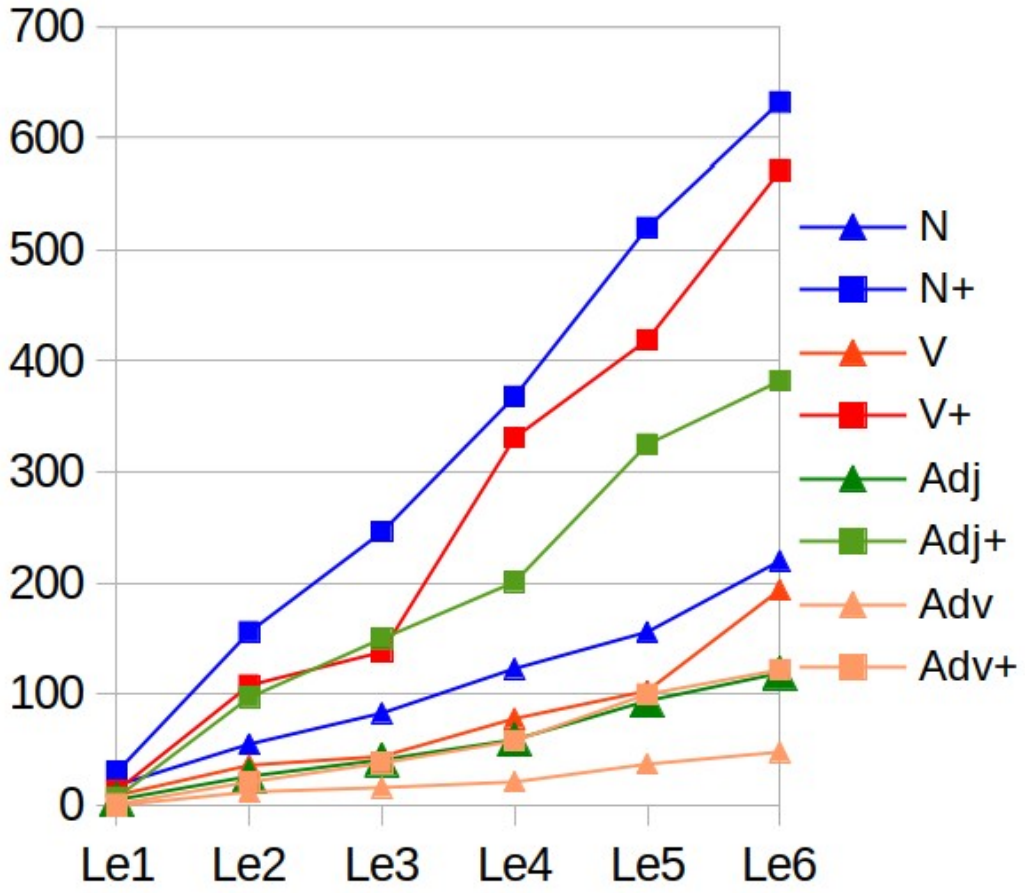
Sentence generation: Vocabulary base

- words, roots and affixes from 6 lessons with 5 blocks each
  - Esperanto marks word class systematically with vowel endings, and roots can change word class --> 13% more words than roots:
    - *amiko (friend), amika/amike (friendly), amiki (be friends)*
  - Esperanto has a productive morphology with compounding and affixes --> growth of root vocabulary is steeper than morpheme growth

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

Sentence generation: Vocabulary base

- To provide more lexical variation, the generator can optionally expand its lexicon with additional compounds and affix-derivations that can be formed from course material morphemes
- For this, a large Eo word list from the parser lexicon (58.000 words) was morphologically analysed. Words made up of only "allowed" morphemes, and not marked as rare or archaic, were then added to the generator lexicon
- The extended vocabulary is about **3 times the size** of the course material vocabulary, at all levels



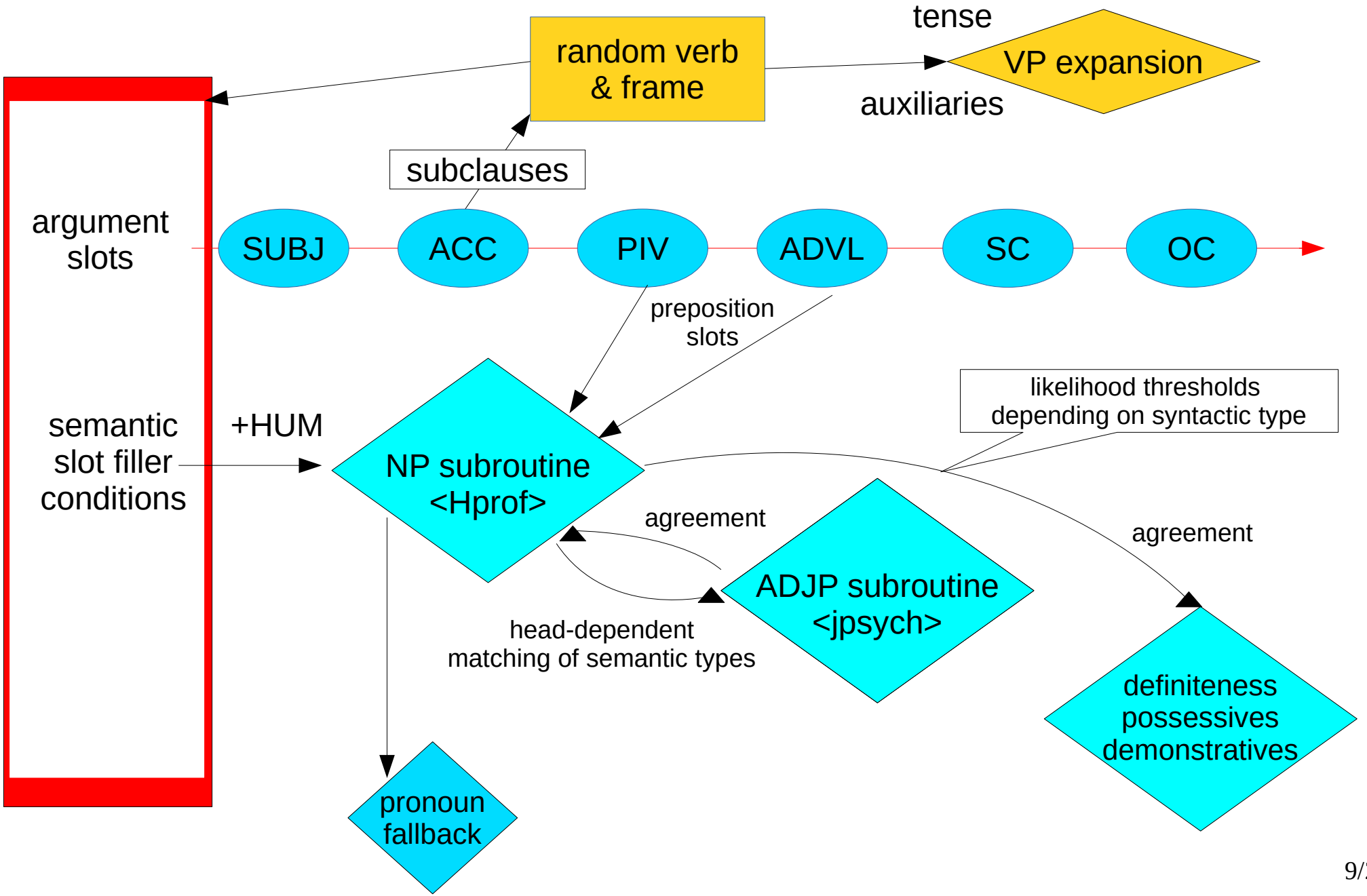
Valency frames for verbs

- As a syntactic skeleton for sentence generation, verb frames are used
- Each frame contains the semantic category of the verb and 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:**teach**/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>
- frame entries for 6235 verb lemmas
- 100% coverage for verbs in the course material
- 96% coverage in a 1.3 million news corpus
- frames were also created for 157 nouns and 50 adjectives, but they are not currently used by the generator

Semantic ontologies for nouns and adjectives

- Expanding frames into sentences
  - semantic verb type <--(match)--> type of slot filler nominals
- noun ontology from EspGram (~ 200 categories)
  - <Hprof> (profession), <Hideo> (follower), <Hfam> (family relation)
  - <tool-cut>, <tool-mus> (musical instruments), <tool-light> etc.
- Most Esperanto affixes allow a safe prediction of semantic class of out-of-lexicon words (99.3% coverage on news)
  - e.g. -ej for <L> 'place' or -uj for <con> 'container')
- adjective ontology adapted from Danish (~ 110 categories in a shallow hierarchy, Bick 2019)
  - <jcol> (colour) <jpsych> (psychological state)
  - categories are sem. prototypes, but also intended for distributional restriction (e.g. <jcol> & physical objects, <jpsych> & human or semiotic nouns)
  - for Esperanto, 4140 adjectives were tagged: 100% course material, 71% news coverage - but 2/3 of failures could be covered through derivation from other POS

Syntactic and morphological generation



Semantic sentence grading

- For certain exercises, student-produced, student-altered or student-completed sentences have to be judged
  - substitution tables, slot filler exercises, QA
  - here, sentence grading allows the use of free input
- semantic mismatches indicate that the student has not understood the context, or is unsure of his own (filler) words
  - here, our tool not only grades the sentence as a whole, but can also flag words or structures as "probably known" or "probably problematic", even in a monolingual L2 setting
- Two-thronged co-occurence approach based on corpus data
  - (a) bigrams and trigrams are checked against a frequency table
  - (b) *depgrams* (word pairs with a syntactic link, e.g. verb-object) are scored for lexical, syntactic and semantic acceptability, based on corpus data

Corpus statistics

50 mill. word corpus

morphosyntactically and semantically annotated with the EspGram parser

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**

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

Depgram types

func = syntactic function
semtype = semantic type
prp = preposition header

(POS is vowel-encoded and hence superfluous in Eo)

Depgram levels

- (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)

- semtype is used for N,V and ADJ, from ontologies and framenet
- To prevent sparce data problems, certain highly productive word types were replaced with a dummy:
  - PHUM - human proper noun
  - PORG - organization
  - PTIT - work-of-art title
  - EWORD - emails, URLs
  - NUM - non-letter cardinals
  - YEAR, DATE

-- Ngram counts are stored as *relative* frequencies
-- For depgrams, mutual information is used

[a,b,c=words, n = corpus size]

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

Sentence grading

- Sentence acceptability = sum of ngram and depgram scores
  - bigrams and trigrams measure surface cohesion (words only)
  - depgrams for semantic acceptability, 4 combinations, 3 layers
  - weighting system, e.g. trigrams > bigrams, clause level++, coarse--
  - punishments for lookup failures for depgrams
  - fallback lookup of depgrams with simplified ontology

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

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

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

Word grading

- in addition to sentence grading, it can be useful to identify the "outlier" words that do not fit the rest of the sentence
  - why: fuzzy proofing tool, flagging odd lexical L2 choices
- method: each time an ngram or depgram is evaluated, the sum score of the participating words is adjusted accordingly
  - dependents are weighted heavier (2x), because heads are supposed to be mentally primary (*'idea' -> +'good' -'blue',* but the latter are much less restricted in terms of head choice)
  - in addition to incremental sums, a few unary flags for specific mismatches are used:
    - '?' - frame mismatch at the clause level
    - '*' - missing corpus evidence for a sem/sem match or a clause-level word/sem match
    - '%' agreement mismatch in noun phrases
- word grading can supplement sentence grading when used as a quality filter
  - 1. overall low sentence score
  - 2. one ore more negative word score
  - 3. one or more unary punishment flags

Grading examples

(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) ***songĝon***?* (-32.14). 'a flower baked a blue dream'

Language-specific adaptations

- The method is in general language-independent, given a large corpus, a good parser reasonable ontology coverage
- Language-group specific adaptations:
  - (1) We use words rather than lemmas
  - (2) We use prepositions as a kind of noun-prefix
  - Both make it easier to capture surface clues and would work for all Germanic languages, but would be problematic in languages with a rich morphology, where case and lemma should be used
- Esperanto-specific coverage boosters: Stripping semantically transparent affixes off their roots:
  - verb affixes: *ek-* (inchoative), *-ad* (durative), *re-* (iterative), *fin-* (resultative)
  - general affixes: *-eg* (big, intensely), *-et* (small, moderately)
  - *-ig* (make [object] do) and -igx (get x'ed), when added to verbs, change their transitivity, but the original verb frame can be reused with argument switches (objetct/subject)
  - *-ebl* (x-able) and *-ind* (worth x-ing) adjectivize verbs, but the verb's object slot can be used as a head-condition for the resulting adjective: *manĝ-ebla planto* (eat-able/edible plant)

Evaluation 1

- sentence generation performs better than sentence grading, because the former can simply stick to corpus- and ontology-sanctioned combinations, while the latter - on open text - may suffer from low corpus coverage, and missing or - in particular - too-constrained frames.
- (1) Very good performance with only course-based material, and the course as part of the corpus
- (2) Precision evaluation (false positives) by runinng the evaluator on ordinary (not-learner) text:
  - (a) on all course-sentences (7,363 words) with a course-free corpus:
  - (b) on Lernu.net short stories (61,676 words)

Evaluation

	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

- The short stories had more complex language and far more neg. word scores (false positives)
- However, there were fewer sentences with neg. scores
  - Possible reasons: milder individual neg. scores, lower incidence of frame failures, better corpus n-gram support
- For real-world use: Minimize false positives by combining scores and flags

Conclusion & outlook

- Syntactic-analytic and semantic-lexical resources can be brought together to build a **robust system of sentence generation/grading**
- Esperanto exhibits specific morphological traits facilitating semantic classification (**affix classes**)
- The method was developed for a specific teaching project (with 100% coverage), but will also work in other scenarios and on unrestricted input
- Error inspection suggests that future work should focus not just on the number of frames, but rather on the **semantic slot filler information** for frame arguments.
- Negatively scored words from a corpus run could be used as a point of departure for improvements.

Co-funded by the
Erasmus+ Programme
of the European Union

parses:

http://visl.sdu.dk --> esperanto

sentence grading:

https://visl.sdu.dk/frazilo-eo.html

eckhard.bick@mail.dk